Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Hejie Cui Date

AI-Assisted Healthcare with Multimodal Structured Knowledge Extraction and Augmented Inference

By

Hejie Cui
Doctor of Philosophy

Computer Science and Informatics

Carl Yang, Ph.D.
Advisor

Joyce C Ho, Ph.D.
Committee Member

Yana Bromberg, Ph.D.
Committee Member

Fei Wang, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

AI-Assisted Healthcare with Multimodal Structured Knowledge Extraction and Augmented Inference

By

Hejie Cui
B.Eng., Tongji University, Shanghai, 2019
M.Sc., Emory University, GA, 2022

Advisor: Carl Yang, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2024

Abstract

AI-Assisted Healthcare with Multimodal Structured Knowledge Extraction and Augmented Inference
By Hejie Cui

The rapid advancement of artificial intelligence (AI) has unlocked new opportunities for enhancing healthcare. However, the heterogeneity and complexity of healthcare data, spanning scientific literature, clinical texts, medical images, and electronic health records, pose significant challenges in extracting useful knowledge and leveraging AI models effectively for clinical decision-making. This thesis addresses these challenges through two key themes: (i) multimodal structured knowledge extraction, focusing on integrating knowledge from diverse data sources and pre-trained models to enable comprehensive data understanding, and (ii) augmented inference, developing techniques to improve the domain-specific reasoning capabilities and reliability of AI models by incorporating the extracted or external knowledge resources. The proposed methods enhance the breadth of multimodal data understanding and the depth of AI models' capabilities in specialized applications. The effectiveness of the proposed core ideas is demonstrated in various domains, including brain disorder analysis, scientific literature understanding, disease prediction, and biomedical reasoning, paving the way for more personalized, precise, and reliable AI-assisted care delivery.

AI-Assisted Healthcare with Multimodal Structured Knowledge Extraction and Augmented Inference

By

Hejie Cui
B.Eng., Tongji University, Shanghai, 2019
M.Sc., Emory University, GA, 2022

Advisor: Carl Yang, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Computer Science and Informatics
2024

Acknowledgments

My Ph.D. journey over the past five years has been an incredible and transformative experience. Starting with a broad exploration of the graph mining domain, I gradually narrowed my focus to the specialized field of healthcare. This journey has been driven by my inner passion and a strong desire to make a positive impact on the world. I have been fortunate enough to meet many people who have supported me along my journey in various ways. I am deeply grateful for their presence and the inspiration they have provided as I continue to grow with courage.

I want to express my heartfelt thanks to my advisor, Prof. Carl Yang. His constant support, invaluable guidance, and the flexibility he gave me have shaped my growth during the Ph.D. journey. He taught me to delve into the meaningful problems that drive our work. His patient guidance and insightful advice helped me navigate research challenges and grow enormously. Under his guidance, I have discovered true fulfillment in research and have learned to be a brave leader when confronted with obstacles. The experiences I have gained working with him have transformed me into a more skilled and independent researcher in my field. I will be forever grateful for the time, effort, and care he has invested in my academic and personal growth.

I am thankful to Prof. Joyce C. Ho for being on my thesis committee and for her pivotal role in shaping my vision for the future of ML and data mining in healthcare. Her expertise and guidance have been a constant source of inspiration, motivating me to expand my research horizons and explore new frontiers. Prof. Ho has encouraged and influenced me to communicate effectively and actively engage with the research community. As a female role model in my field, her support and mentorship have been instrumental in my development as a researcher and as an individual.

I am also deeply grateful to Prof. Yana Bromberg for being on my thesis committee and for her generosity in sharing her extensive knowledge and expertise in the field of biomedical informatics. Prof. Bromberg's encouragement and support in

my pursuit of academia, along with her willingness to share her personal story and experiences, have been a guiding light in my journey. I am incredibly thankful for her encouragement on my academic path.

It is also my great honor to have Prof. Fei Wang on my thesis committee, whose profound knowledge and insights into AI health have been a learning resource for me. His guidance on identifying and tackling significant research problems has been crucial in shaping my research direction and has helped me solidify the foundation for my future endeavors. I am deeply appreciative of the time and effort he has dedicated to mentoring me and for the positive influence he has had on my academic growth.

I want to express my heartfelt gratitude to my mom for all the care, love, and sacrifice she has given to me. I have been grateful to have her as my mom so that I learn to grow up into a caring and kind-hearted person. Thank you for being my closest friend and for being strong for me. I love you from the bottom of my heart. I also want to thank my father, who has always been a responsible and hardworking role model for me. I cherish the memories of chatting with him back in the day, the wisdom he shared, and the encouragement he gave me when I faced challenges. Thank you for being the strong man in our family and for all your dedication to us.

The multidisciplinary nature of my research has made collaboration a central part of my Ph.D. studies. I am fortunate to have worked with and learned from numerous researchers, including Xiaoxiao Li, Pan Li, Lifang He, Lichao Sun, Liang Zhan, Ying Guo, Joshua Lukemire, Jingbo Shang, Manling Li, Yangqiu Song, Xin Liu, Kun Zhang, Liang Zhao, Wei Jin, Lianhui Qin, Hui Shao, Xiang Li, QuanZheng Li, Jon Kleinberg, etc. I also extend my gratitude to the mentors from my industry internships: Rongmei Lin, Nasser Zalmout, Chenwei Zhang, Xian Li, Tobias Schnabel, Jennifer Neville, Mengting Wan, Longqi Yang, Stojan Trajanovski, Lu Cao, and Masrour Zoghi. I am sincerely thankful to the researchers who have provided valuable career suggestions, including Sheng Wang, Ruogu Fang, Liyue Shen, Tianfan Fu,

Jiliang Tang, Wei Wang, etc. Their kindness and willingness to offer their experience and guidance have been instrumental to me, and I am very grateful for their help.

I want to thank the Emory Graph Mining Group, especially Xuan Kan, Jiaying Lu, Ran Xu, Han Xie, Ziyang Zhang, Keqi Han, Yuzhang Xie, and the students I worked with, including Wei Dai, Owen Yang, Tong Gu, Xinyu Fang, etc. I cherish the happy memories we shared, from hiking to playing board games and making Tang-Yuan. Thank you for your presence in my life. I want to express my gratitude to Prof. Vaidy Sunderam, the chair of the CS department, for the warm smile and caring he showed when we met in the department hallways. I thank my friends Edgar Leon and Sergio Gramacho in the IT department of Emory CS, who have always been very responsive and helpful when I encountered problems with computational resources.

I feel fortunate to connect with several great friends who bring warmth to me in different ways, including Yanqiao Zhu, Jieyu Zhang, Yanbang Wang, Yuhao Zhang, Yuanqi Du, Yue Yu, Haitao Mao, Zijie Huang, Chulin Xie, Shuowei Jin, Yupeng Hou, Yujia Zheng, etc. I am thankful for the enjoyable intersections with them.

Finally, I can't wait to thank those who have brightened my life during my challenging times. I am especially grateful to Xuan, whose inspiration, encouragement, support, and humor have made this journey possible and easier along the way. I also want to thank my bestie, Yuquan Zhou, for our years of mutual understanding and thank Yukai Jiang for holding together my dear close friends from undergraduates. I feel so fortunate to have all the love and encouragement from them. I cannot forget to thank S, the other half of my inner soul, for the beautiful memories and experiences I have while wandering in time. She has been encouraging me to reflect on the true meaning of life and struggled to make me understand myself, others, and the nature of life's gains and losses. Her presence has enriched my soul in many ways and inspired me to pursue my deepest dreams as I continue to move forward.

Thank you all for accompanying me on this once-in-a-lifetime adventure.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

The rapid advancement of artificial intelligence (AI) has unlocked unprecedented opportunities for revolutionizing healthcare. AI-assisted healthcare holds the promise of improving patient outcomes, streamlining clinical workflows, and enabling personalized medicine. By harnessing the vast amounts of multimodal data available in healthcare, such as electronic health records, medical images, and scientific literature, AI can potentially extract extensive, valuable insights and support clinical decision-making. However, efficiently analyzing these diverse data sources and effectively leveraging state-of-the-art AI models for healthcare applications poses significant challenges that need to be addressed to realize the potential of AI in healthcare.

1.2 Challenges

1.2.1 Healthcare Data Can be Complex and Heterogeneous

One of the primary challenges in AI-assisted healthcare is the heterogeneity of medical data. As demonstrated in Figure 1.1, healthcare data comes from various sources

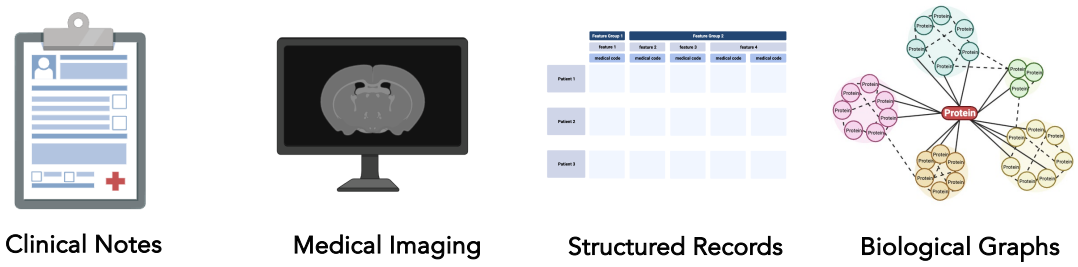


Figure 1.1: The complex and heterogeneous data in healthcare.

and modalities, each with its own unique characteristics and formats. For example, electronic health records (EHRs) contain structured and unstructured data, including patient demographics, clinical notes, and laboratory results. Medical imaging data, such as X-rays, CT scans, and MRIs, provide visual details of patient conditions. Additionally, scientific literature and clinical guidelines offer valuable knowledge for clinical decision-making. Extracting knowledge from these diverse data sources and integrating them into a unified framework is a complex task. Moreover, the complexity of medical knowledge, which often involves intricate relationships between various factors, adds to the challenge of developing AI systems that can effectively reason over healthcare data.

1.2.2 AI-assisted Healthcare Suffers from Limited Labeled Data

Another significant challenge in AI-assisted healthcare is the limited availability of labeled data for adapting AI models to specific medical domains. While general-purpose AI models have been trained on large-scale datasets, such as ImageNet for computer vision and Wikipedia for natural language processing, domain-specific labeled data in healthcare is often scarce. This scarcity can be attributed to several factors, including the high cost of expert annotation, privacy concerns, and the rare occurrence of certain medical conditions. Consequently, adapting AI models to healthcare applications often requires training on small, domain-specific datasets, which can limit the

models' performance and generalizability. Developing techniques that can effectively leverage limited labeled data or utilize unsupervised learning approaches is crucial for the successful application of AI in healthcare.

1.2.3 Pre-trained Models Have Wide Knowledge Base but May Not be Adequately Reliable for Healthcare

Large pre-trained models, such as language models and multimodal models, have demonstrated remarkable capabilities in various AI tasks. These models are trained on vast amounts of diverse data, allowing them to capture a wide range of knowledge. However, when it comes to healthcare applications, the breadth of knowledge in these models is often accompanied by a lack of reliability and domain-specific accuracy. While large pre-trained models can generate plausible outputs, they may not always produce factually correct or clinically relevant information. This is particularly concerning in healthcare, where the consequences of incorrectness can be severe. To effectively harness the breadth of knowledge in large pre-trained models for healthcare applications, it is crucial to address the reliability challenges. This involves developing techniques to align the models' outputs with domain-specific knowledge and incorporating external knowledge sources.

1.3 Research Contributions

To address the aforementioned challenges, this thesis focuses on two main themes: **Structured Knowledge Extraction** and **Augmented Inference**, for multimodal data from broad resources. The overall research roadmap is shown in Figure 1.2. The contributions aim to extract useful structured knowledge from diverse data sources and augment the extracted or external knowledge to improve the performance and reliability of AI models in healthcare scenarios.

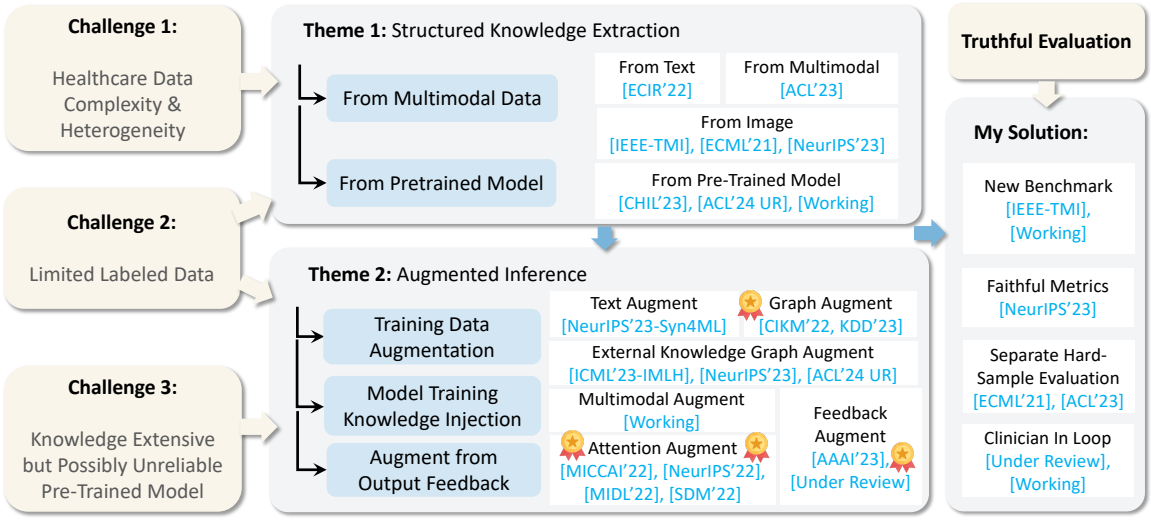


Figure 1.2: An overall research contribution on AI-assisted healthcare with multimodal structured knowledge extraction and augment inference.

1.3.1 Multimodal Structured Knowledge Extraction

A significant portion of valuable knowledge often lies hidden within *heterogeneous data sources* and *pre-trained models*. Extracting and leveraging this knowledge is crucial for effectively integrating multimodal resources and supporting data-driven decision-making.

Extracting information *from heterogeneous data* presents both opportunities and challenges. Effective multimodal fusion, seamless integration, and efficient conversion between different modalities are key to unlocking the potential of vast amounts of data. We develop techniques to understand and distill essential information from heterogeneous sources, facilitating the application of AI in real-world scenarios.

Pre-trained models, such as language models and multimodal models, have been trained on extensive datasets and possess a wealth of encoded knowledge. Extracting information *from pre-trained models* can provide valuable insights for unseen tasks. However, when it comes to healthcare applications, the breadth of knowledge in these models may not always translate to factual correctness or clinical relevance. Therefore, it is crucial to address potential unreliability by carefully designing the process or curating knowledge extraction under well-designed supervision.

Overall, the contributions in multimodal structured knowledge extraction aim to address the challenges of data heterogeneity through automated knowledge conversion and integration. By developing techniques to effectively extract and integrate knowledge from diverse data sources and pre-trained models, we aim to create a more comprehensive and accurate foundation that can support various applications, ultimately leading to more precise models for data-driven decision-making.

1.3.2 Augmented Inference

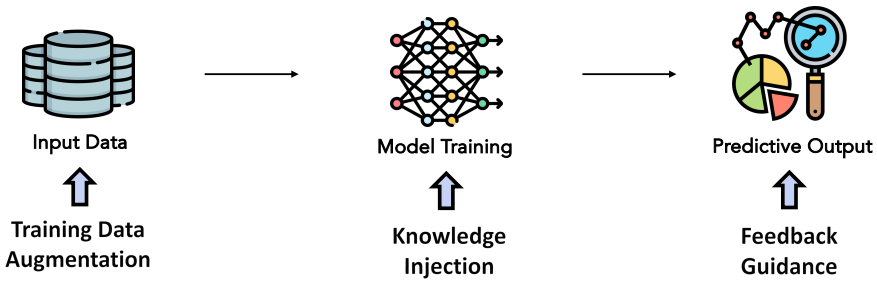


Figure 1.3: This thesis explores augmented inference techniques that are applied at various stages of AI model development.

Augmented inference focuses on enhancing the ability and reliability of AI models by leveraging extracted or external knowledge. This thesis explores techniques that can be applied at different stages of the AI model development process as shown in Figure 1.3, including *(1) augmentation with the model input*, *(2) knowledge injection during model training*, and *(3) augmentation through output feedback*.

Augmentation through Model Input Data augmentation during the input phase is a direct approach to enhance AI models. As AI models often require large amounts of training data, the extracted or external knowledge can serve as additional training data for model learning. Another way is to augment the model input with related contextual information, such as visual knowledge from extracted knowledge corpora or structured information from external knowledge graphs. By providing this contextual information, the models can generate more accurate and relevant responses in various downstream tasks.

Augmentation through Model Training Process Incorporating additional knowledge during the model training phase can explicitly influence the model's capabilities. We introduce novel learned attention masks that guide the model to focus on task-relevant regions or relations while filtering out potentially biased or erroneous information. Furthermore, supplementary training targets and optimization regularizations are proposed and incorporated in the training process, which can directly improve the model's inference performance and generalization ability.

Augmentation through Output Feedback Leveraging output feedback as an augmentation strategy can significantly boost the model's inference performance. We develop techniques to measure the discrepancy between the model's predictions and the ground truth, using these observations to augment the model. By identifying and addressing problematic areas, our approach guides the model to focus on the most relevant factors, leading to more reliable and accurate predictions.

Overall, our contributions in augmented inference aim to improve the reliability and accuracy of AI models for reasoning and inference tasks. We achieve this by incorporating extracted structured knowledge and external knowledge resources into the models' learning processes. By enhancing the models' ability to reason over complex problems and make more accurate and dependable predictions, we push the boundaries of AI-assisted decision-making in healthcare and beyond.

1.4 Dissertation Outline

The thesis is organized into two main chapters, each exploring the core ideas of multimodal structured knowledge extraction and augmented inference in different contexts. We first illustrate the core idea of this thesis using a specific scenario of brain analysis, as brain imaging is inherently multimodal and contains complex knowledge that can aid interpretation. We then generalize the discussion to broader data types, including

texts, images, and multimodal input, demonstrating how specialized and foundational models can be optimized and adapted for multimodal knowledge extraction and augmented inference in various application scenarios.

- Chapter 2 focuses on brain analysis as a specific case study to explore the core ideas of multimodal structured knowledge extraction and inference. Brain imaging is inherently multimodal, with data from various modalities such as fMRI and DTI, where connectomes representing different kinds of knowledge can be constructed and analyzed. We first focus on connectome extraction and introduce our established benchmark, BrainGB, in Section 2.2. Section 2.3 covers the design space of graph neural network baselines for brain network inference. Moving forward, Section 2.4 proposes an advanced interpretable model for brain network inference and presents several interesting visualizations of important edges and ROI interpretations generated by the framework.

- Chapter 3 generalizes the core ideas of multimodal structured knowledge extraction and augmented reasoning to broader data types, including texts, images, and multimodal data. Section 3.1 introduces an effective framework for concept map extraction from scientific literature and its application in long document retrieval. Section 3.2 presents a novel open visual knowledge extraction framework from images, where the proposed designs significantly increase the diversity of the generated structured knowledge to reflect real-world richness and complement textual knowledge documented in the literature. Section 3.3 focuses on structured knowledge extraction from multimodal data, proposing techniques to better align multimodalities and mitigate the inherent modality biases in multimodal fusion and cross-modality reasoning. Moving on to knowledge-augmented inference, Section 3.4 delves into the exciting world of Electronic Health Records (EHRs) and explores how we can leverage the power of language foundation models with augmented inference for disease prediction. In the final project, Section 3.5 investigates

adapting generic multimodal foundation models to the specific healthcare domain with knowledge augmentation from clinician expertise and encoded knowledge from GPT models.

- Chapter 4 concludes the thesis by providing a holistic summary and discussing future research directions for advancing AI-assisted healthcare based on the proposed methods. The chapter also outlines several future directions to extend the proposed ideas in this dissertation for future advancements in the field.

Chapter 2

Multimodal Neurobiological Data: Brain Connectome Extraction and Inference

Brain Imaging is internally multimodal and contains complex knowledge that can aid in interpretation. In this chapter, we focus on illustrating the core ideas of this thesis using this specific scenario of brain analysis. Mapping the connectome of the human brain using structural or functional connectivity has become one of the most pervasive paradigms for neuroimaging analysis. Recently, Graph Neural Networks (GNNs) motivated from geometric deep learning have attracted broad interest due to their established power for modeling complex networked data. Despite their superior performance in many fields, there has not yet been a systematic study of how to design effective GNNs for brain network analysis. To bridge this gap, we present BrainGB, a benchmark for brain network analysis with GNNs. BrainGB standardizes the process by (1) summarizing brain network construction pipelines for both functional and structural neuroimaging modalities and (2) modularizing the implementation of GNN designs. We conduct extensive experiments on datasets across cohorts and modalities

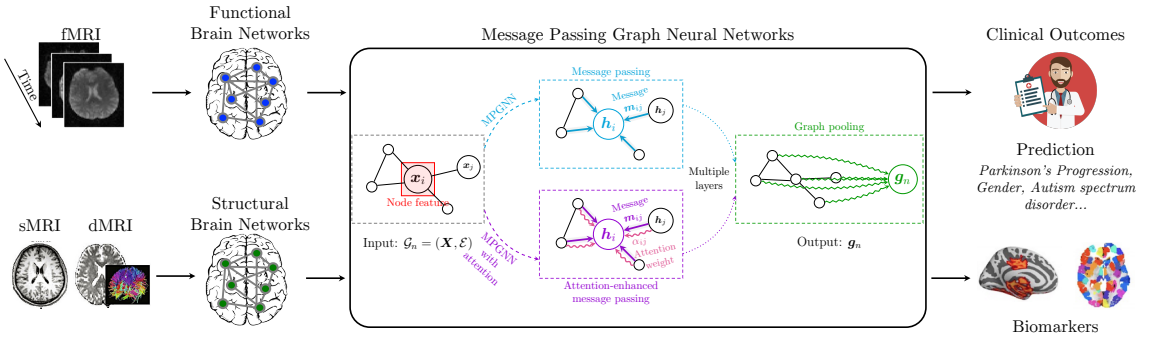


Figure 2.1: An overview of our BrainGB framework for brain network analysis with graph neural networks.

and recommend a set of general recipes for effective GNN designs on brain networks. To support open and reproducible research on GNN-based brain network analysis, we host the BrainGB website at `https://braingb.us` with models, tutorials, examples, as well as an out-of-box Python package. Furthermore, we introduce another advanced interpretable model design for brain network analysis. We hope that these work will provide useful empirical evidence and offer insights for future research in this novel and promising direction.

2.1 Introduction

Human brains are at the center of complex neurobiological systems in which neurons, circuits, and subsystems interact to orchestrate behavior and cognition. Understanding the structures, functions, and mechanisms of human brains has been an intriguing pursuit for researchers with various goals, including neural system simulation, mental disorder therapy, as well as general artificial intelligence. Recent studies in neuroscience and brain imaging have reached the consensus that interactions between brain regions are key driving factors for neural development and disorder analysis [146, 69]. Inspired by graph theory, brain networks composed of nodes and edges are developed to describe the interactions among brain regions.

The human brain can be scanned through various medical imaging techniques, in-

cluding Magnetic-Resonance Imaging (MRI), Electrogastrography (EGG), Positron Emission Tomography (PET), and so on. Among all these acquisitions, MRI data are the most widely used for brain analysis research. There are also different modalities of MRI data such as functional MRI (fMRI) and Diffusion Tensor Imaging (DTI), from which functional and structural brain networks can be constructed respectively. Specifically, the connectivity in functional brain networks describes correlations between time-series signals of brain regions, while the connectivity in structural brain networks models the physical connectivity between gray matter regions [188]. Both functional and structural connections are widely acknowledged as valuable resources of information for brain investigation [170, 20].

Previous work on brain network analysis has studied shallow models based on graph theory [20, 227] and tensor factorization [154, 313] extensively, which focuses on proposing neurobiologically insightful graph measures and approaches from the node, motif, and graph level to detect network communities or modules and identify central network elements. Methodological developments in graph research enable us to quantify more topological characteristics of complex systems, many of which have already been assessed in brain networks, such as modularity, hierarchy, centrality, and the distribution of network hubs. However, shallow modeling techniques can be inadequate for the sophisticated connectome structures of brain networks [71]. On the other hand, deep learning models have become extraordinarily popular in machine learning, achieving impressive performance on images [61, 199], videos [6], and speech processing tasks [85]. These regular data are represented in 1D/2D/3D Euclidean spaces and can be suitably handled by traditional Recurrent (RNNs) or Convolutional Neural Networks (CNNs). In contrast, the irregular structural and functional brain connectivity networks constructed from neuroimaging data are more complex due to their non-Euclidean characteristics. In recent years, Graph Neural Networks (GNNs) have attracted broad interest due to their established power for analyzing graph-

structured data [126, 280, 249]. Several pioneering deep models have been devised to predict brain diseases by learning graph structures of brain networks. For instance, Li et al. [146] propose BrainGNN to analyze fMRI data, where ROI-aware graph convolutional layers and ROI-selection pooling layers are designed for neurological biomarker prediction. Kawahara et al. [120] design a CNN framework BrainNetCNN composed of edge-to-edge, edge-to-node, and node-to-graph convolutional filters that leverage the topological locality of structural brain networks. However, they mainly experiment with their proposed models on specific private datasets. Due to the ethical issue of human-related research, the datasets used are usually not publicly available and the details of imaging preprocessing are not disclosed, rendering the experiments irreproducible for other researchers.

To address the aforementioned limitations, there is an urgent need for a public benchmark platform to evaluate deep graph models for brain network analysis. However, it is non-trivial to integrate different components within a unified benchmarking platform. Current brain network analyses are typically composed of two steps. The first step is to construct brain networks from neuroimaging data. Then, in the second stage, the resulting brain connectivity between all node pairs is used to classify individuals or predict clinical outcomes. The difficulties in the initial stage are mostly due to restricted data accessibility and sophisticated brain imaging preprocessing and network construction pipelines that differ across cohorts and modalities. The difficulty of the second stage is to establish a standard evaluation pipeline based on fair experimental settings, metrics, and modular-designed baselines that can be easily validated and extended for future research.

2.2 Brain Connectome Extraction and Benchmark

2.2.1 Background: Diverse Modalities of Brain Imaging

Models of the human brain as a complex network have attracted increasing attention due to their potential for helping understand human cognition and neurological disorders. In practice, human brain data can be acquired through various scanning techniques [209], such as Magnetic-Resonance Imaging (MRI), Electroencephalography (EEG) and Magnetoencephalography (MEG), Positron Emission Tomography (PET), Single-Photon Emission Computed Tomography (SPECT), and X-ray Computed Tomography (CT). Among them, MRI is one of the most widely used techniques in brain research and clinical practice, due to its large range of available tissue contrast, detailed anatomical visualization, and high sensitivity to abnormalities [15].

MRI Data

In this paper, we focus on MRI-derived brain networks. Specifically, for different modalities of MRI data, we can reconstruct different types of brain networks. Functional MRI (fMRI) is one of the most popular modalities for investigating brain function and organization [78, 149, 223] by detecting changes in blood oxygenation and blood flow that occur in response to neural activity. Diffusion-weighted MRI (dMRI), on the other hand, can enable inference about the underlying connection structure in the brain's white matter by recording the diffusion trajectory of molecules (usually water). fMRI focuses on functional activity, while dMRI presents brain structural information from different perspectives. Specifically, two types of brain networks, functional and structural, can be constructed from the aforementioned modalities by following different connectivity generation paradigms [132].

Challenges in MRI Preprocessings

The raw MRI data collected from scanners is not directly usable for brain network construction or imaging analysis. A complicated preprocessing pipeline is necessary to remove unwanted artifacts, transform the data into a standard format, and perform structure discovery. Although there are several widely-used neuroimaging data preprocessing tools, such as SPM[1], AFNI[2] and FSL[3], each of them still needs considerable training and learning efforts. Moreover, the functionality of these software varies, and for dMRI, no one software contains all the necessary preprocessing capabilities. In addition, many neuroimaging datasets cannot be made public due to privacy or ethical concerns. Due to the variety of preprocessing approaches and issues with making data publically available, there are difficulties in reproducibility in neuroimaging studies. Additionally, the preprocessing steps are distinctive across modalities. All these challenges make it difficult for deep learning researchers with little knowledge in medical imaging processing to get into the field.

2.2.2 Brain Network Extraction from Multimodal Imaging

In this section, we provide a general overview of the standard preprocessing pipelines for the construction of brain networks of different modalities. Due to the regulation restrictions for direct sharing of the brain network data, we provide two complete pipelines, one for functional brain networks (ABCD[4] specifically) and one for structural brain networks (PPMI[5] specifically), with step-by-step commands and parameter settings on our hosted website for public access[6].

[1] `https://www.fil.ion.ucl.ac.uk/spm/software/spm12`
[2] `https://afni.nimh.nih.gov`
[3] `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL`
[4] `https://nda.nih.gov/abcd`
[5] `https://www.ppmi-info.org`
[6] `https://braingb.us/preprocessing`

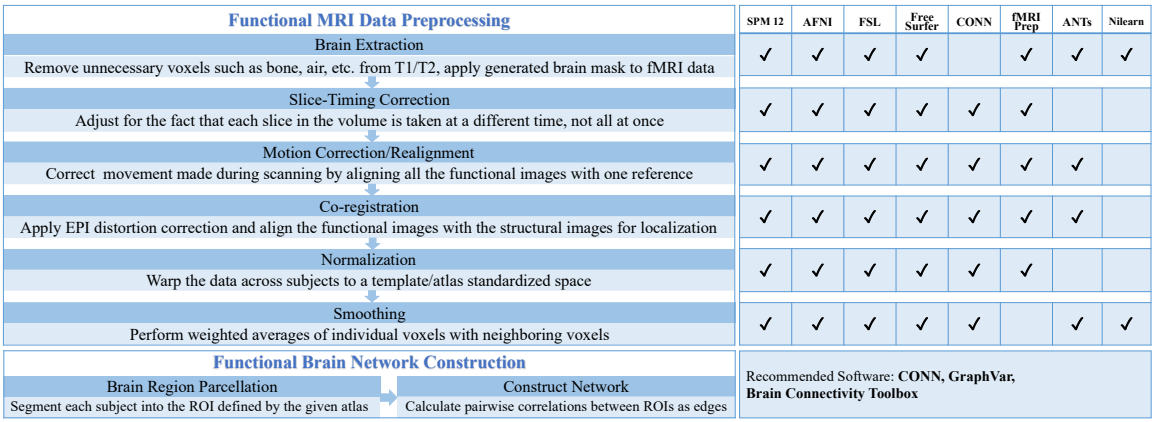


Figure 2.2: The framework of fMRI data preprocessing and functional brain network construction procedures, with recommended tools for each step shown on the right. The more commonly-used tools for the functional modality are placed at the front.

Functional Brain Network Extraction

The left side of Fig. 2.2 shows a standard preprocessing procedure for functional brain imaging, with the corresponding commonly-used toolboxes (i.e., SPM12[1], AFNI[2], FSL[3], FreeSurfer[7], CONN[8], fMRI Prep[9], ANTs[10], Nilearn[11]) shown on the right side. Note that each step in the preprocessing and network construction pipeline needs quality control by the experts, and the specific order of preprocessing steps may change slightly based on the acquisition conditions of the dataset. Some representative functional neuroimaging datasets in literature to facilitate scientific research include ADHD 200 [14], ADNI (fMRI part) [194], HCP 900 [246], ABIDE [57], etc.

To measure functional connectivity, some preprocessing of the fMRI time series is often performed including detrending, demeaning, and whitening fMRI BOLD time series at each voxel [261]. To construct the brain networks, a brain atlas or a set of Regions of Interest (ROI) are selected to define the nodes. Then, the representative fMRI BOLD series from each node are obtained by either averaging or performing Singular Value Decomposition (SVD) on the time series from all the voxels within

[7] https://surfer.nmr.mgh.harvard.edu
[8] https://web.conn-toolbox.org/home
[9] https://fmriprep.org/en/stable/index.html
[10] http://stnava.github.io/ANTs
[11] https://nilearn.github.io/stable/index.html

the node. Various measures have been proposed for assessing brain connectivity between pairs of nodes. One of the simplest and most frequently used methods in the neuroimaging community is via pairwise correlations between BOLD time courses from two ROIs. Other methods include partial correlations [261], mutual information, coherence, Granger causality [225]. After selecting the Functional Connectivity (FC) measure, one can evaluate the strength of connectivity between each pair of ROIs. Often, some transformation, such as the Fisher's transformation, is performed to transform the original FC measures to improve their distribution properties. The transformed FC measures can then be utilized for the subsequent analysis of functional brain networks.

To facilitate public testing, we take Adolescent Brain Cognitive Development Study (ABCD) as an example and provide a step-by-step instruction for functional brain network construction on our hosted BrainGB website[6]. The ABCD-HCP BIDS[12] pipeline is used to preprocess the data. In brief, anatomical preprocessing included normalization, co-registration, segmentation, and brain extraction. Functional data preprocessing included slice-time correction, motion correction, distortion correction, co-registration, normalization, and spatial smoothing. Brain parcellation schemes were then applied to the functional data to obtain time courses for each ROI, and Pearson correlation was used to construct brain networks representing the connectivity between ROIs.

Structural Brain Network Extraction

Structural brain networks provide a systematic perspective for studying the anatomical and physiological organization of human brains and help to understand how brain structure influences function. Some representative neuroimaging studies include diffusion MRI data are PPMI [174], ADNI [194], HCP [246], AIBL [65], OASIS [65], etc.

[12] `https://github.com/DCAN-Labs/abcd-hcp-pipeline`

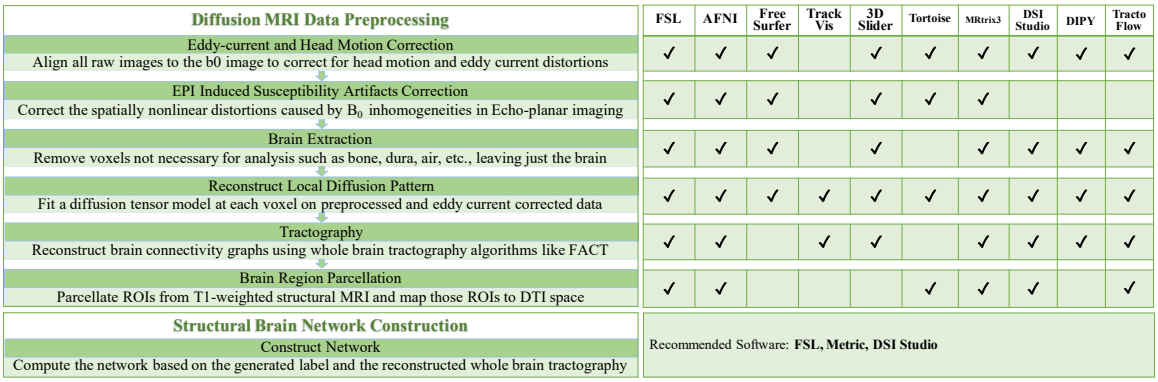


Figure 2.3: The framework of dMRI data preprocessing and structural brain network construction procedures, with recommended tools for each step shown on the right. The more commonly-used tools for the structural modality are placed at the front.

The commonly-used toolboxes for dMRI include FSL[3], AFNI[2], FreeSurfer[7], TrackVis[13], 3D Slicer[14], Tortoise[15], MRtrix3[16], DSI Studio[17].

The left side of Fig. 2.3 summarizes the pipeline for reconstructing the structural brain network. Preprocessing steps for the dMRI data include removal of eddy current-induced distortions, brain extraction, and co-registration between diffusion and structural images. Next, some modeling strategies are applied to reconstruct the local diffusion patterns. Commonly adopted models include the DTI modeling, which fits a tensor model or muti-tensor model [136] to capture the local diffusion patterns, and the Ball and Sticks model [11]. After reconstructing the local diffusion patterns, a tractography algorithm is performed to computationally reconstruct fiber tract connections between brain regions. Commonly-used algorithms include the deterministic tractography [9] and the probabilistic tractography [12]. The deterministic tractography connects neighboring voxels from seed regions based on the major direction of the DTI tensor. The probabilistic tractography involves first estimating fiber orientation and its uncertainty at each voxel and building a diffusion path probability map based on the estimated orientation and uncertainty. While deterministic tractography is

[13] http://trackvis.org
[14] https://www.slicer.org
[15] https://tortoise.nibib.nih.gov
[16] https://www.mrtrix.org
[17] https://dsi-studio.labsolver.org

a more computationally efficient approach to reconstruct major fiber bundles in the brain, probabilistic tractography has become more popular because it is more robust to noise and allows tractography to progress beyond uncertain regions by taking into account uncertainty in fiber orientations at each voxel [314]. To construct the structural network, the structure connectivity for each node pair is calculated based on the empirical probability of fiber tracts connecting the two regions. Note that each step of network construction ideally needs quality control from experts.

Similarly to functional brain network construction, we take PPMI as an example and provide an instruction pipeline for structural brain network construction on our hosted BrainGB website[6]. Specifically, the Diffusion Toolkit from TrackVis is used to reconstruct local diffusion patterns and tractography. The brain region parcellation is completed with both FSL and Freesurfer. Then local diffusion pattern reconstruction and the network computation are further performed by calculating the number of fibers within each ROI after removing the false positive ones.

In addition to the mainstream methods of constructing connections in brain networks discussed above, there are also other ways to construct different types of edges. For example, directional connectivity that characterizes effective interactions for fMRI [53]; hybrid functional brain networks where different orders of relationships can be sensitive to different levels of signal changes [327]; and dynamic functional brain networks which include derivatives of windowed functional network connectivity in the identification of reoccurring states of connectivity [67, 53]. Apart from fMRI and DTI, the most commonly used modalities to construct functional and structural brain networks, other neuroimaging modalities have also been explored in literature, such as metabolic brain network constructed from PET imaging [130], functional brain network constructed from EEG signals [113], etc. Recent studies have shown that the combination of both functional and structural neuroimaging modalities can be more effective than using only a single one, which can exploit complementary information

across different modalities [170, 23].

2.2.3 Open Source Benchmark Platform

To foster future research, we provide an out-of-box package that can be directly installed through pip, with installation and tutorials on our hosted BrainGB website `https://braingb.us` for brain connectome construction and modeling. The BrainGB package is also open-sourced at `https://github.com/HennyJie/BrainGB`. We provide examples of GNN-based brain network analysis, trained models, and instructions on imaging preprocessing and functional and structural brain networks construction from raw fMRI and dMRI respectively.

2.3 Graph Neural Network Baselines for Brain Network Inference

In this work, we propose Brain Graph Neural Network Benchmark (BrainGB)—a novel attempt to benchmark brain network analysis with GNNs to the best of our knowledge. The overview of BrainGB is demonstrated in Fig. 2.1 and the main contributions are four-fold:

- A *unified*, *modular*, *scalable*, and *reproducible* framework is established for brain network analysis with GNNs to facilitate reproducibility. It is designed to enable fair evaluation with accessible datasets, standard settings, and baselines to foster a collaborative environment within computational neuroscience and other related communities.
- We summarize the preprocessing and construction pipelines for both functional and structural brain networks to bridge the gap between the neuroimaging and deep learning community.

- We decompose the design space of interest for GNN-based brain network analysis into four modules: (1) node features, (b) message passing mechanisms, (c) attention mechanisms, and (d) pooling strategies. Different combinations based on these four dimensions are provided as baselines, and the framework can be easily extended to new variants.

- We conduct a variety of empirical studies and suggest a set of general recipes for effective GNN designs on brain networks, which could be a starting point for further studies.

To foster future research, we release the source code of BrainGB at `https://github.com/HennyJie/BrainGB` and provide an out-of-box package that can be installed directly, with detailed tutorials available on our hosted website at `https://braingb.us`. Preprocessing instructions and models are provided for standardized model evaluations. We enable the community to collaboratively contribute by submitting their own custom models, and we will maintain a leaderboard to ensure such efforts will be recorded.

Specifically, the process of applying GNNs to brain networks starts from initialization of the ROI features, followed by the forward pass which includes two phases, message passing, and pooling. The learned graph-level representation then can be utilized for brain disease analysis. In the machine learning domain, the rapid evolution of GNNs has led to a growing number of new architectures. Specifically for GNNs on brain network analysis, we decompose the design space of interest for basic message passing GNNs into four modules: node feature construction, message passing, attention enhanced message passing, and pooling strategies. An illustration of these modules is shown in the middle of Fig. 2.1.

2.3.1 Node Feature Construction

In neuroscience analysis, researchers mostly focus on brain connectivity represented by a featureless graph. To apply GNNs on non-attributed brain networks, researchers in the graph machine learning domain have studied several practical methods to initialize node features [43, 62]. In this paper, we focus on the following node features that can be categorized as positional or structural:

- *Identity*: A unique one-hot feature vector is initialized for each node [66, 298]. By giving each ROI in the brain network a unique high-dimensional vector, this identity node feature allows the GNN model to learn the relative positions of the nodes by memorizing their k-hop neighbors. They are essentially the same as random initialization considering the parameters in the first linear layer of the GNN are randomly initialized.

- *Eigen*: Eigen decomposition is performed on the weighted matrix describing the connection strengths between ROIs and then the top k eigenvectors are used to generate a k-dimensional feature vector for each node [103, 27, 319]. The optimal value of k is decided by grid search. This feature is essentially dimension reduction and targets at grouping brain regions with respect to their positions, with global graph information condensed into a low-dimensional representation.

- *Degree*: The degree value of each node is obtained as a one-dimensional vector as the node feature. This feature captures structural information of brain regions, meaning that neighborhood structural similarity of two regions will be partially recorded in the initialized node features.

- *Degree profile*: This method takes advantages of existing local statistical measures on degree profiles [22], where each feature $\boldsymbol{x}_i$ of node v_i on graph $\mathcal{G}_n$ is computed as

$$\begin{aligned} \boldsymbol{x}_i = [\deg(v_i) \parallel \min(\mathcal{D}_i) \parallel \max(\mathcal{D}_i) \\ \parallel \text{mean}(\mathcal{D}_i) \parallel \text{std}(\mathcal{D}_i)], \end{aligned} \tag{2.1}$$

where $\mathcal{D}_i = \{\deg(v_i) \mid (i,j) \in \mathcal{E}_n\}$ describes the degree values of node v_i's one-hop neighborhood and $\|$ denotes concatenation.

- *Connection profile*: The corresponding row for each node in the edge weight matrix is utilized as the initial node feature, which contains connections with respect to all other nodes in the brain network. This feature aligns with the common practice of using pairwise connections to perform brain parcellation. Also, it reflects the whole picture of connection information in the brain network.

2.3.2 Message Passing Mechanisms

The power of most GNNs to learn structures lies in their message passing schemes, where the node representation is updated iteratively by aggregating neighbor features through local connections. In each layer l, the node representation $\boldsymbol{h}_i^l$ is updated through two steps, namely message passing and update respectively. In the message passing step (Eq. 2.2), each node v_i receives messages from all its neighbors, and then all the messages are aggregated with a sum function:

$$\boldsymbol{m}_i^l = \sum_{j \in \mathcal{N}_i} \boldsymbol{m}_{ij} = \sum_{j \in \mathcal{N}_i} M_l \left(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, w_{ij} \right), \tag{2.2}$$

where $\mathcal{N}_i$ denotes the neighbors of node v_i in graph $\mathcal{G}$, w_{ij} represents the edge weights between node v_i and v_j, M_l is the message function. In the update step (Eq. 2.3), the embedding of each node is updated based on the aggregated messages from Eq. 2.2 and optionally the previous embedding of node v_i, where the update function can be arbitrary differentiable functions (e.g., concat the aggregated message with the previous node embedding and then pass them into a learnable linear layer).

$$\boldsymbol{h}_i^{l+1} = U_l \left(\boldsymbol{h}_i^l, \boldsymbol{m}_i^l \right), \tag{2.3}$$

where U_l stands for the update function and the number of running steps L is defined by the number of GNN layers. The message passing mechanism can leverage both permutation equivariance and inductive bias towards learning local structures and

achieve good generalization on new graphs. For brain networks, whether incorporating connections into the message function is beneficial for graph-level prediction tasks remains to be investigated. In this paper, we discuss the influence of different message function M_l designs including:

- *Edge weighted*: The message $\boldsymbol{m}_{ij}$ passed from node v_j to node v_i is calculated as the representation of node v_j weighted by the corresponding edge weight w_{ij}, that is

$$\boldsymbol{m}_{ij} = \boldsymbol{h}_j \cdot w_{ij}. \tag{2.4}$$

  This is the standard message passing implementation in Graph Convolutional Network (GCN) [126] when $w_{ij} = 1/N_i$. With this message vector design, the update of each brain region representation is influenced by its neighbor regions weighted by the connection strength between them.

- *Bin concat*: In this scheme, we map the edge w_{ij} into one of the equally split T buckets based on its weight value. Each bucket corresponds to a learnable representation $\boldsymbol{b}_t$, $t = \{1 \dots T\}$. The total bucket number encompassing the entire value range of edge weights is determined by grid search and the representation dimension of each bin is set to the same as node features. Specifically, given the number of buckets is T, we first rank all the edge weights and then divide them into the equally divided T buckets from the lowest to the highest. All edges in the same bucket will be mapped to the same learnable vector $\boldsymbol{b}_t$, so region connections with similar strength are binned together. In our experiment, we simply select from [5, 10, 15, 20] as the possible number of buckets for grid search, which is a common practice in machine learning for hyperparameter tuning. The message $\boldsymbol{m}_j$ passed from node v_j to node v_i is calculated as the concatenation of the representation of node v_j and its corresponding bucket

representation $\boldsymbol{b}_t$ followed by an MLP,

$$\boldsymbol{m}_{ij} = \text{MLP}(\boldsymbol{h}_j \parallel \boldsymbol{b}_t). \tag{2.5}$$

The usage of bins helps to clusters region connections with similar strengths. By concatenating with the unique neighbor node representation, this message captures both common and peculiar characteristics of each neighbor.

- *Edge weight concat*: The message $\boldsymbol{m}_{ij}$ passed from node v_j to node v_i is represented as the concatenation of the representation of node v_j and the scaled edge weight $d \cdot w_{ij}$, followed by a MLP,

$$\boldsymbol{m}_{ij} = \text{MLP}(\boldsymbol{h}_j \parallel d \cdot w_{ij}), \tag{2.6}$$

where d is a constant equal to the dimension number of node features. The motivation behind edge weight scaling is to increase the influence of edge features to the same scale as node features. Compared with bin concat where edges with weight values in the same bin interval share the same initial edge representation, directly concatenating the scaled edge weights as the edge representations can retain the original edge information, therefore reserving more uniqueness on the pairwise connection when performing the aggregation from neighboring brain regions.

- *Node edge concat*: To investigate the influence of preserving the brain region representation from the last time step while iterative updating the new representation, we design a message $\boldsymbol{m}_j$ as the concatenation of both embeddings of node v_i, v_i and the edge weight w_{ij} between them, followed by an MLP, that is

$$\boldsymbol{m}_{ij} = \text{MLP}(\boldsymbol{h}_i \parallel \boldsymbol{h}_j \parallel w_{ij}). \tag{2.7}$$

In this paradigm, every message passed from the local neighbors of each central node is reinforced with its representation from the last time step. This design may alleviate the over-smoothing problem of GNNs, where the feature distance between all nodes becomes too close and not distinguishable after layers of

convolutions.

- *Node concat*: Since the effect of involving connection weights into message passing is still unknown, we also include another message $\boldsymbol{m}_{ij}$ similar to *node edge concat* but without the concatenation of edge weights, where

$$\boldsymbol{m}_{ij} = \text{MLP}(\boldsymbol{h}_i \parallel \boldsymbol{h}_j). \tag{2.8}$$

2.3.3 Attention-Enhanced Message Passing

Attention is arguably one of the most important mechanisms in modern deep learning [248, 185]. It is inspired by human cognitive systems that tend to selectively concentrate on the important parts as needed when processing large amounts of information. Various fields in deep learning communities such as natural language processing [56] and computer vision [87] have widely benefited from attention mechanisms in terms of model efficiency and accuracy. The attention mechanism can also be used to enhance the message passing scheme of GNNs, while also providing interpretations over the edge importance.

Specifically in brain network analysis, by utilizing the attention-enhanced version of message passing, the model updates the brain region representation in a data-driven way, where adjustable attention weights from each local neighbor perform as an additional influence factor besides the neural signals represented by edge weights. It is worth noting that the traditional designs of graph attention mechanisms on general graphs usually do not take the edge attributes (i.e., connection weights in the brain network scenario) into consideration. However, for brain networks, the correlation between two regions contains meaningful biomedical information and might be helpful for graph-level tasks. In this paper, we design several attention-enhanced message passing mechanisms including:

- *Attention weighted*: This is the original GAT [249] on general graphs without

involving edge attributes. The message from node v_j to v_i is weighted by the corresponding attention score α_{ij} as

$$\boldsymbol{m}_{ij} = \boldsymbol{h}_j \cdot \alpha_{ij}. \tag{2.9}$$

The α_{ij} is calculated from a single-layer feed-forward neural network parameterized by a weight vector $\mathbf{a}$, followed by the LeakyReLU nonlinearity σ,

$$\alpha_{ij} = \frac{\exp\left(\sigma\left(\boldsymbol{a}^\top \left[\boldsymbol{\Theta}\boldsymbol{x}_i \parallel \boldsymbol{\Theta}\boldsymbol{x}_j\right]\right)\right)}{\sum_{k\in\mathcal{N}(i)\cup\{i\}} \exp\left(\sigma\left(\boldsymbol{a}^\top \left[\boldsymbol{\Theta}\boldsymbol{x}_i \parallel \boldsymbol{\Theta}\boldsymbol{x}_k\right]\right)\right)}, \tag{2.10}$$

where $\boldsymbol{\Theta}$ is a learnable linear transformation matrix.

- *Edge weighted w/ attn*: This is the attention-enhanced version of *edge weighted* message passing in Eq. 2.4. The message from v_j to v_i is obtained as the multiplication of node v_j's representation $\boldsymbol{h}_j$, the edge weight w_{ij} and the attention score α_{ij} in Eq. 2.10,

$$\boldsymbol{m}_{ij} = \boldsymbol{h}_j \cdot \alpha_{ij} \cdot w_{ij}. \tag{2.11}$$

- *Attention edge sum*: This is another version of attention-enhanced *edge weighted* (Eq. 2.4) message passing. The edge weight w_{ij} and the attention score α_{ij} are first summed, then used as the impact factor on the node embedding $\boldsymbol{h}_j$,

$$\boldsymbol{m}_{ij} = \boldsymbol{h}_j \cdot (\alpha_{ij} + w_{ij}). \tag{2.12}$$

- *Node edge concat w/ attn*: This is the attention-enhanced version of *node edge concat* (Eq. 2.7) message passing, where the attention score α_{ij} (Eq. 2.10) between node v_i and v_j is multiplied on the node representation $\boldsymbol{h}_j$ before concatenation, followed by a MLP,

$$\boldsymbol{m}_{ij} = \mathrm{MLP}(\boldsymbol{h}_i \parallel (\boldsymbol{h}_j \cdot \alpha_{ij}) \parallel w_{ij}). \tag{2.13}$$

- *Node concat w/ attn*: This design corresponds to the attention-enhanced version of *node concat* (Eq. 2.8) message passing, where the attention score α_{ij} (Eq. 2.10) between node v_i and node v_j is multiplied on the node representation $\boldsymbol{h}_j$

before concatenation, followed by a MLP,

$$\boldsymbol{m}_{ij} = \text{MLP}(\boldsymbol{h}_i \parallel (\boldsymbol{h}_j \cdot \alpha_{ij})). \tag{2.14}$$

2.3.4 Pooling Strategies

In the second phase of GNNs, a feature vector for the whole graph $\boldsymbol{g}_n$ is computed using the pooling strategy R, where

$$\boldsymbol{g}_n = R\left(\{\boldsymbol{h}_k \mid v_k \in \mathcal{G}_n\}\right). \tag{2.15}$$

The pooling function R operates on the set of node vectors and is invariant to permutations of the node vectors. In this paper, we cover three basic global pooling operators [84, 180]:

- *Mean pooling*: The graph-level representation is obtained by averaging node features. For each single graph $\mathcal{G}_n$, the graph-level representation is computed as

$$\boldsymbol{g}_n = \frac{1}{M}\sum_{k=1}^{M} \boldsymbol{h}_k. \tag{2.16}$$

- *Sum pooling*: The graph-level representation is obtained by summing up all node features. For each single graph $\mathcal{G}_n$, the graph-level representation is computed as

$$\boldsymbol{g}_n = \sum_{k=1}^{M} \boldsymbol{h}_k. \tag{2.17}$$

- *Concat pooling*: The graph-level representation is obtained by concatenating node features of all nodes contained in the graph. For each single graph $\mathcal{G}_n$, the graph-level representation is computed as

$$\boldsymbol{g}_n = \|_{k=1}^{M} \boldsymbol{h}_k = \boldsymbol{h}_1 \parallel \boldsymbol{h}_2 \parallel \ \ldots \ \parallel \boldsymbol{h}_M. \tag{2.18}$$

Note that there are also other complex pooling strategies such as hierarchical pooling [296], learnable pooling [83] and clustering readout [118], which are usually viewed as independent GNN architecture designs that are not defined based on combinative

Table 2.1: Dataset summarization.

Dataset	Modality	# Samples	Atlas	Size	Response	# Classes
HIV	fMRI	70	AAL 116	90×90	Disease	2
PNC	fMRI	503	Power 264	232×232	Gender	2
PPMI	DTI	754	Desikan-Killiany	84×84	Disease	2
ABCD	fMRI	7,901	HCP 360	360×360	Gender	2

modules. Here we include the representative method of DiffPool [296] to provide a view of the comparison between basic and more complex pooling methods.

2.3.5 Experimental Analysis and Insights

In this section, we show experimental results on brain networks generated from real-world neuroimaging studies with different GNN modular designs. Varying each design dimension under each module results in a total of 375 different architectures. Note that here we do not aim to cover all combinations, but to quickly find a relatively good one. Furthermore, we emphasize that the design space can be expanded as new design dimensions emerge.

Experimental Settings

Datasets To establish a benchmark for generic brain network analysis models, we include four datasets processed and constructed from different neuroimaging modalities, specifically fMRI (HIV [155], PNC[18], ABCD[4]) and dMRI (PPMI[5]), based on different brain atlas. For the HIV and PPMI datasets, the task is to classify patients from healthy control (Patient, Normal Control); while for the PNC and ABCD datasets, the task is gender prediction (Male, Female). We intentionally cover such a diverse set of datasets from different modalities (and preprocessing procedures/parcellations/tasks), because our purpose is to establish a benchmark for generic brain network analysis models. Thus observations on a diverse set of datasets can be more instructive for methodology focused studies. All the datasets we used have been visually checked by imaging experts in our team for quality control. Among these four datasets, PNC,

[18] `https://www.nitrc.org/projects/pnc`

PPMI, and ABCD are restrictively publicly available ones that can be requested and downloaded from their official website. The dataset information is summarized in TABLE A.1. Since the datasets can be acquired from multiple sites, multisite issues need to be addressed when performing the analysis on the constructed networks. Over the past few years, ComBat techniques [29, 75] from the microarray literature have started to be used more frequently to deal with multi-site batch effects. Since our benchmark focuses more on a comprehensive overview of brain network construction and effective GNN designs for brain networks, advanced methods for handling multi-site issues are out of the scope of this work. Interested readers can refer to [30, 13, 196, 285, 195] for more advanced multisite data handling methods.

- *Human Immunodeficiency Virus Infection (HIV):* This dataset is collected from the Chicago Early HIV Infection Study at Northwestern University [200]. The clinical cohort includes fMRI imaging of 70 subjects, 35 of which are early HIV patients and the other 35 are seronegative controls. The preprocessing includes realignment to the first volume, followed by slice timing correction, normalization, and spatial smoothness, band-pass filtering, and linear trend removal of the time series. We focus on the 116 anatomical ROIs [244] and extract a sequence of time courses from them. Finally, brain networks with 90 cerebral regions are constructed, with links representing the correlations between ROIs.

- *Philadelphia Neuroimaging Cohort (PNC):* This rs-fMRI dataset is from the Brain Behavior Laboratory at the University of Pennsylvania and the Children's Hospital of Philadelphia. 289 (57.46%) of the 503 included subjects are female, indicating this dataset is balanced across genders. The regions are parcellated based on the 264-node atlas defined by Power et al. [197]. The preprocessing includes slice timing correction, motion correction, registration, normalization, removal of linear trends, bandpass filtering, and spatial smoothing. In the re-

sulting data, each sample contains 264 nodes with time-series data collected through 120 time steps. We focus on the 232 nodes in the Power's atlas associated with major resting-state functional modules [224].

- *Parkinson's Progression Markers Initiative (PPMI):* This dataset is from a collaborative study for Parkinson's Research to improve PD therapeutics. We consider the DTI acquisition of 754 subjects, with 596 Parkinson's disease patients and 158 healthy controls. The raw data are first aligned to correct for head motion and eddy current distortions. Then the non-brain tissue is removed and the skull-stripped images are linearly aligned and registered. 84 ROIs are parcellated from T1-weighted structural MRI based on the Desikan-Killiany' cortical atlas [54] and the brain network is reconstructed using the deterministic 2nd-order Runge-Kutta (RK2) whole-brain tractography algorithm [314].

- *Adolescent Brain Cognitive Development Study (ABCD)*: This study recruits children aged 9-10 years across 21 sites in the U.S. Each child is followed into early adulthood, with repeated imaging scans, as well as extensive psychological and cognitive tests [24]. After selection, 7,901 children are included in the analysis, with 3,961 (50.1%) female. We use rs-fMRI scans for the baseline visit processed with the standard and open-source ABCD-HCP BIDS fMRI Pipeline[12]. After processing, each sample contains a connectivity matrix whose size is 360×360 and BOLD time-series for each node. The region definition is based on the HCP 360 ROI atlas [81].

Structural connectivity and functional connectivity are different in their strength and sparsity, thus need to be handled differently. For structural connectivity, we normalize the edge weights by dividing each value by the maximum value in a sample. The processed edge weights are thus ranged from 0 to 1. For functional connectivity, we follow common practice to remove the negative values for GNNs that cannot handle

negative values (like GCN), and keep them for GNNs that can handle negative values (like GAT).

Baselines For comprehensiveness, we compare our modular design with competitors of both shallow and deep models. The shallow methods we consider include M2E [154], MPCA [162], and MK-SVM [63], where the output graph-level embeddings are evaluated using logistic regression classifiers. Specifically, M2E is a partially-symmetric tensor factorization based method for brain network analysis, and it has been empirically compared with spectral embedding clustering methods such as SEC [183] or spectral learning frameworks such as AMGL [184]; MPCA is proposed for the feature extraction and analysis of tensor objects such as neuroimaging; multiple kernel SVM (MK-SVM) is essentially an extension of the conventional SVM algorithm and has been applied for the analysis of functional and structural connectivity in Alzheimer's disease. We also include two state-of-the-art deep models specifically designed for brain networks: BrainGNN [146] and BrainNetCNN [120]. The message passing in BrainGNN is Edge weighted and it further leverages additional regional information (such as coordinates or ROI ordering based one-hot embeddings) to assign a separate GCN kernel for each ROI where ROIs in the same community are embedded by the similar kernel and those in different communities are embedded in different ways, but this will introduce a lot of additional model parameters and make the model hard to train. On the other hand, BrainNetCNN models the adjacency matrix of a brain network as a 2D image and does not follow the message passing mechanism as we discussed in Section 2.3.2. Note that the purpose of our paper, and of most benchmark papers, is not to establish superior performance of a certain method, but rather to provide an effective and fair ground for comparing different methods.

Implementation Details The proposed model is implemented using PyTorch 1.10.2 [192] and PyTorch Geometric 2.0.3 [72]. A Quadro RTX 8000 GPU with 48GB of memory is used for model training. The optimizer we used is Adam. We train all of our models through 20 epochs, and the learning rate is 1e-3. We use a weight decay of 1e-4 as a means of regularization. The loss function is cross entropy. Hyper-parameters are selected automatically with an open-source AutoML toolkit NNI[19]. Please refer to our repository for comprehensive parameter configurations. When tuning the hyperparameters, we first split the dataset into a train set and a test set with the ratio of 8:2. The k-fold validation is performed on the train set, where we further divide the train set into 10 parts and take one in each run to use as the validation set. The selection of the best hyperparameter is based on the average performance of the model on the validation sets. The reported metrics in Table II, on the other hand, is the average performance on the test set, with each run trained on different train sets. The competing methods are also tuned in the same way. For BrainGNN, we used the author's open-source code[20]. For BrainNetCNN, we implemented it by ourselves with PyTorch, which is publicly available in our BrainGB package[21]. For the hyper-parameter tuning, we selected several important hyper-parameters and performed the grid search on them based on the provided best setting as claimed in their paper. To be specific, for BrainGNN, we searched for different learning rates in {0.01, 0.005, 0.001} with different feature dimensions in {100, 200} and the number of GNN layers in {2, 3}. For BrainNetCNN, we searched for different dropout rates in {0.3, 0.5, 0.7} with learning rates in {0.001, 0.0005, 0.0001} and the number of layers in MLP in {1, 2, 3}. The reported results of these two baselines in Table II are from the best performing groups, where for BrainGNN, the learning rate is 0.01, the feature dimension is 200 and the number of GNN layers is 2, and for BrainNetCNN, the

[19] `https://github.com/microsoft/nni`
[20] `https://github.com/xxlya/BrainGNN_Pytorch`
[21] `https://github.com/HennyJie/BrainGB`

Table 2.2: Performance report (%) of different message passing GNNs in the four-modular design space with other two representative baselines on four datasets. We highlight the best performed one in each module based on AUC, since it is not sensitive to the changes in the class distribution, providing a fair evaluation on unbalanced datasets like PPMI.

Module	Method	HIV			PNC			PPMI			ABCD		
		Accuracy	F1	AUC	Accuracy	F1	AUC	Accuracy	F1	AUC	Accuracy	F1	AUC
Node Features	*Identity*	$50.00_{\pm 0.00}$	$33.33_{\pm 0.00}$	$46.73_{\pm 10.57}$	$57.34_{\pm 0.17}$	$36.44_{\pm 0.17}$	$52.58_{\pm 4.80}$	$79.25_{\pm 0.24}$	$44.21_{\pm 0.08}$	$59.65_{\pm 6.80}$	$49.97_{\pm 0.13}$	$33.32_{\pm 0.06}$	$50.00_{\pm 0.20}$
	Eigen	$65.71_{\pm 2.86}$	$65.45_{\pm 2.69}$	$65.31_{\pm 2.89}$	$51.40_{\pm 3.92}$	$48.63_{\pm 5.42}$	$50.18_{\pm 7.57}$	$74.09_{\pm 2.77}$	$47.36_{\pm 4.26}$	$49.21_{\pm 1.58}$	$50.79_{\pm 0.82}$	$50.79_{\pm 0.83}$	$51.18_{\pm 1.16}$
	Degree	$44.29_{\pm 5.35}$	$35.50_{\pm 6.10}$	$42.04_{\pm 4.00}$	$63.89_{\pm 2.27}$	$59.69_{\pm 3.85}$	$70.25_{\pm 4.38}$	$79.52_{\pm 2.31}$	$49.40_{\pm 5.17}$	$59.73_{\pm 4.31}$	$63.46_{\pm 1.29}$	$63.45_{\pm 1.28}$	$68.16_{\pm 1.41}$
	Degree profile	$50.00_{\pm 0.00}$	$33.33_{\pm 0.00}$	$50.00_{\pm 0.00}$	$51.40_{\pm 7.21}$	$33.80_{\pm 3.21}$	$50.00_{\pm 0.00}$	$77.02_{\pm 1.97}$	$49.45_{\pm 3.51}$	$58.65_{\pm 2.44}$	$49.92_{\pm 0.11}$	$33.30_{\pm 0.05}$	$50.00_{\pm 0.00}$
	Connection profile	$65.71_{\pm 13.85}$	$64.11_{\pm 13.99}$	**$75.10_{\pm 16.95}$**	$69.83_{\pm 4.15}$	$66.20_{\pm 4.74}$	**$76.69_{\pm 8.04}$**	$77.99_{\pm 2.78}$	$52.96_{\pm 4.52}$	**$65.77_{\pm 4.09}$**	$82.42_{\pm 1.93}$	$82.30_{\pm 2.08}$	**$91.33_{\pm 0.77}$**
Message Passing	*Edge weighted*	$50.00_{\pm 0.00}$	$33.33_{\pm 0.00}$	$49.80_{\pm 4.20}$	$64.87_{\pm 5.44}$	$59.70_{\pm 7.04}$	$69.98_{\pm 4.19}$	$79.25_{\pm 0.24}$	$44.21_{\pm 0.08}$	$62.26_{\pm 2.80}$	$74.47_{\pm 1.17}$	$74.36_{\pm 1.23}$	$82.37_{\pm 1.46}$
	Bin concat	$50.00_{\pm 0.00}$	$33.33_{\pm 0.00}$	$49.39_{\pm 9.25}$	$54.74_{\pm 5.88}$	$36.42_{\pm 3.97}$	$61.68_{\pm 3.91}$	$79.25_{\pm 0.24}$	$44.21_{\pm 0.08}$	$52.67_{\pm 7.16}$	$53.72_{\pm 4.97}$	$43.26_{\pm 12.43}$	$61.86_{\pm 5.79}$
	Edge weight concat	$51.43_{\pm 2.86}$	$44.36_{\pm 6.88}$	$48.16_{\pm 10.13}$	$63.68_{\pm 3.31}$	$60.27_{\pm 5.97}$	$67.34_{\pm 3.02}$	$79.25_{\pm 0.24}$	$44.21_{\pm 0.08}$	$59.72_{\pm 4.65}$	$64.59_{\pm 1.30}$	$64.30_{\pm 1.43}$	$70.63_{\pm 1.02}$
	Node edge concat	$65.71_{\pm 13.85}$	$64.11_{\pm 13.99}$	$75.10_{\pm 16.95}$	$69.83_{\pm 4.15}$	$66.20_{\pm 4.74}$	$76.69_{\pm 5.04}$	$77.99_{\pm 2.78}$	$52.96_{\pm 4.52}$	$65.77_{\pm 4.09}$	$82.42_{\pm 1.93}$	$82.30_{\pm 2.08}$	$91.33_{\pm 0.77}$
	Node concat	$70.00_{\pm 15.91}$	$68.83_{\pm 17.57}$	**$77.96_{\pm 8.20}$**	$70.63_{\pm 2.35}$	$67.12_{\pm 1.81}$	**$78.32_{\pm 1.42}$**	$78.41_{\pm 1.62}$	$54.46_{\pm 3.08}$	**$68.34_{\pm 1.89}$**	$80.50_{\pm 2.27}$	$80.10_{\pm 2.47}$	**$91.36_{\pm 0.92}$**
Message Passing w/ Attention	*Attention weighted*	$50.00_{\pm 0.00}$	$33.33_{\pm 0.00}$	$49.80_{\pm 8.52}$	$65.09_{\pm 2.21}$	$60.74_{\pm 4.89}$	$69.79_{\pm 4.24}$	$79.25_{\pm 0.24}$	$44.21_{\pm 0.08}$	$63.24_{\pm 3.77}$	$77.74_{\pm 0.97}$	$77.70_{\pm 1.01}$	$85.10_{\pm 1.10}$
	Edge weighted w/ attn	$50.00_{\pm 0.00}$	$33.33_{\pm 0.00}$	$42.04_{\pm 15.63}$	$62.90_{\pm 1.22}$	$61.14_{\pm 0.57}$	$69.74_{\pm 2.37}$	$79.25_{\pm 0.24}$	$44.21_{\pm 0.08}$	$54.92_{\pm 4.80}$	$78.04_{\pm 1.96}$	$77.81_{\pm 2.33}$	$86.86_{\pm 0.63}$
	Attention edge sum	$51.43_{\pm 7.00}$	$49.13_{\pm 5.65}$	$54.49_{\pm 15.67}$	$61.51_{\pm 2.86}$	$55.36_{\pm 4.76}$	$69.38_{\pm 3.50}$	$79.11_{\pm 0.40}$	$44.17_{\pm 0.12}$	$60.47_{\pm 6.26}$	$75.71_{\pm 1.52}$	$75.59_{\pm 1.68}$	$83.78_{\pm 0.82}$
	Node edge concat w/ attn	$72.86_{\pm 11.43}$	$72.52_{\pm 11.72}$	$78.37_{\pm 10.85}$	$67.66_{\pm 5.07}$	$64.69_{\pm 5.36}$	$74.52_{\pm 1.20}$	$77.30_{\pm 1.52}$	$50.96_{\pm 4.20}$	$63.93_{\pm 4.89}$	$83.10_{\pm 0.47}$	$83.03_{\pm 0.52}$	**$91.85_{\pm 0.29}$**
	Node concat w/ attn	$71.43_{\pm 9.04}$	$70.47_{\pm 9.26}$	**$82.04_{\pm 11.21}$**	$68.85_{\pm 6.42}$	$64.29_{\pm 10.15}$	**$75.36_{\pm 6.09}$**	$78.41_{\pm 1.43}$	$49.98_{\pm 1.87}$	**$68.14_{\pm 5.01}$**	$83.19_{\pm 0.93}$	$83.12_{\pm 0.96}$	$91.55_{\pm 0.59}$
Pooling Strategies	*Mean pooling*	$47.14_{\pm 15.39}$	$41.71_{\pm 17.36}$	$58.78_{\pm 18.63}$	$66.86_{\pm 2.33}$	$61.39_{\pm 4.88}$	$74.20_{\pm 3.39}$	$79.25_{\pm 0.24}$	$44.21_{\pm 0.08}$	$59.64_{\pm 5.47}$	$81.13_{\pm 0.35}$	$81.06_{\pm 0.34}$	$88.49_{\pm 1.12}$
	Sum pooling	$57.14_{\pm 9.04}$	$52.23_{\pm 12.65}$	$57.96_{\pm 11.15}$	$60.13_{\pm 2.87}$	$53.96_{\pm 7.61}$	$66.11_{\pm 4.22}$	$79.39_{\pm 0.52}$	$47.68_{\pm 3.12}$	$61.29_{\pm 2.11}$	$77.48_{\pm 3.75}$	$76.96_{\pm 4.58}$	$87.90_{\pm 0.65}$
	Concat pooling	$65.71_{\pm 13.85}$	$64.11_{\pm 13.99}$	$75.10_{\pm 16.95}$	$69.83_{\pm 4.15}$	$66.20_{\pm 4.74}$	**$76.69_{\pm 8.04}$**	$77.99_{\pm 2.78}$	$52.96_{\pm 4.52}$	**$65.77_{\pm 4.09}$**	$82.42_{\pm 1.93}$	$82.30_{\pm 2.08}$	**$91.33_{\pm 0.77}$**
	DiffPool	$72.86_{\pm 21.19}$	$70.22_{\pm 23.91}$	**$76.57_{\pm 17.16}$**	$62.72_{\pm 12.40}$	$75.95_{\pm 4.28}$	$64.08_{\pm 16.71}$	$78.42_{\pm 3.53}$	$56.55_{\pm 6.48}$	$63.07_{\pm 7.77}$	$76.45_{\pm 1.44}$	$76.35_{\pm 1.52}$	$83.92_{\pm 1.25}$
Shallow Baselines	M2E	$57.14_{\pm 19.17}$	$53.71_{\pm 19.80}$	$57.50_{\pm 18.71}$	$53.76_{\pm 4.94}$	$46.10_{\pm 6.94}$	$49.70_{\pm 5.18}$	$78.69_{\pm 1.78}$	$45.81_{\pm 4.17}$	$50.39_{\pm 2.59}$	$50.10_{\pm 1.90}$	$49.95_{\pm 1.88}$	$50.10_{\pm 1.90}$
	MPCA	$67.14_{\pm 20.25}$	$64.28_{\pm 23.47}$	$69.17_{\pm 20.17}$	$76.76_{\pm 4.30}$	$75.95_{\pm 4.28}$	$76.05_{\pm 4.34}$	$79.15_{\pm 0.57}$	$44.18_{\pm 0.18}$	$50.00_{\pm 0.00}$	$88.94_{\pm 1.64}$	$88.94_{\pm 1.64}$	$88.94_{\pm 1.64}$
	MK-SVM	$65.71_{\pm 7.00}$	$62.08_{\pm 7.49}$	$65.83_{\pm 7.41}$	$78.38_{\pm 5.09}$	$77.55_{\pm 5.83}$	$77.57_{\pm 5.65}$	$79.15_{\pm 0.57}$	$44.18_{\pm 0.18}$	$50.00_{\pm 0.00}$	$89.42_{\pm 0.97}$	$89.42_{\pm 0.97}$	$89.42_{\pm 0.97}$
Deep Baselines	BrainNetCNN	$60.21_{\pm 17.16}$	$60.12_{\pm 13.56}$	$70.93_{\pm 4.01}$	$71.93_{\pm 4.90}$	$69.94_{\pm 5.42}$	$78.50_{\pm 3.28}$	$77.24_{\pm 2.09}$	$50.24_{\pm 3.09}$	$58.76_{\pm 8.95}$	$85.1_{\pm 0.92}$	$85.7_{\pm 0.83}$	$93.5_{\pm 0.34}$
	BrainGNN	$62.98_{\pm 11.15}$	$60.45_{\pm 8.96}$	$68.03_{\pm 9.16}$	$70.62_{\pm 4.85}$	$68.93_{\pm 4.01}$	$77.53_{\pm 3.23}$	$79.17_{\pm 1.22}$	$44.19_{\pm 3.11}$	$45.26_{\pm 3.65}$	OOM	OOM	OOM

dropout rate is 0.3, the learning rate is 0.0001 and the number of layers in MLP is 3. The metrics used to evaluate performance are Accuracy, F1 score, and Area Under the ROC Curve (AUC), which are widely used for disease identification. To indicate the robustness of each model, all the reported results are the average performance of ten-fold cross-validation conducted on different train/test splits.

Performance Report

Node Feature On comparing node features, we set the other modules as the well-performed settings in individual tests. Specifically, we use *node edge concat* in Eq. 2.7 as the message passing scheme, and *concat pooling* in Eq. 2.18 as the pooling strategy. Our experimental results demonstrate that the ***connection profile*** which uses the corresponding row in the adjacency matrix as the node features achieves the best performance across all datasets, with up to 33.99% improvements over the second-best, *degree*, on ABCD. We believe this is because the *connection profile* captures the whole picture of structural information in the brain network, and preserves rich information on pairwise connections that can be used to perform brain parcellation.

In general, the structure node features (e.g., *degree*, *connection profile*) perform better than the positional ones (e.g., *identity*, *eigen*), indicating that the overall structural information of graph and the structural role of each node are important in the task of brain network analysis. This conclusion is consistent with previous findings in the literature that structural artificial node features work well for graph-level tasks on general graphs [43].

Message Passing To study the effectiveness of different message passing schemes, we initialize the node features with *connection profile* and apply the *concat pooling* to produce graph-level representations, which both perform best when examined separately in each module. Our results reveal that ***node concat*** (Eq. 2.8) message passing has the highest AUC performance across four datasets, followed by *node edge concat* (Eq. 2.7), which achieves a similar AUC performance with sometimes slightly better accuracy and F1 scores (ABCD). The performance superiority of the last two methods may arise from their advantage of reinforcing self-representation of the central node during each step of message passing. This helps to retain the original information from the last step and avoid over-fitting towards a biased direction in the optimization process. Surprisingly, the edge involved *node edge concat* performs slightly worse than the pure *node concat*, though the gap gets closer on larger datasets. This indicates that encoding edge weights as a single value may not be useful when the global structure has already been used as the initial node features.

Attention Enhanced Message Passing When evaluating the effectiveness of different attention-enhanced message passing schemes, we set the node features as *connection profile* and apply the *concat pooling* strategy, just as for the evaluation of message passing without attention mechanisms. It is shown that ***node concat w/ attn*** (Eq. 2.14) and ***node edge concat w/ attn*** (Eq. 2.13) yield very close results across four datasets and they alternately perform the best. Furthermore, the

attention-enhanced version achieves better outcomes most of the time (up to 5.23% relative improvements) vs. the corresponding message passing architecture without attention. This demonstrates the effectiveness of utilizing learnable attention weights in the GNN aggregation and update process in addition to the fixed edge weights. Also, *node edge concat w/ attn* surpasses *node concat w/ attn* on larger datasets (e.g., ABCD), which may imply potential advantages of involving edge weights into message design when there are enough training samples.

Pooling Strategies For studying pooling strategies, we employ the *node edge concat* (Eq. 2.7) as the message passing scheme and *connection profile* as the initial node features. Our findings reveal that the ***concat pooling*** strategy (Eq. 2.18) consistently outperforms the other two methods across all four datasets. This is likely because when *concat* is used, the final node representations of all the brain regions are kept in the graph-level representation for classifiers. The other two paradigms, on the other hand, obtain a graph-level embedding with the same dimension of node features. Thus they lose some information that could be helpful for graph-level prediction tasks. Though *concat* does not ensure permutation invariance, it is actually not needed for brain network analysis since the node order given a parcellation is fixed. The compared hierarchical pooling method *DiffPool* demonstrates some advantages on the small HIV dataset but fails to surpass the simple *concat* pooling on three other larger datasets.

Other Baselines In general, we expect deep models like GNNs to perform better on larger datasets. For example, the performance of GNN models on the ABCD dataset clearly surpasses all shallow models by about 2 percent. However, this trend should not prohibit one from experimenting with GNN models on smaller datasets. GNNs do perform well on some small datasets, such as the HIV dataset. Despite running on a small dataset, GNN models in BrainGB have an over 5% advantage over all shallow

models. As for the deep baselines, BrainGNN can be out-of-memory (OOM) on large datasets. The best combination based on our modular design outperforms BrainGNN on all four datasets (HIV, PNC, PPMI and ABCD) and achieves comparable results with BrainNetCNN in most cases especially on smaller datasets. These findings prove the need to carefully experiment with our modular designs of GNNs before further developing more complicated architectures, which might overfit certain datasets.

Insights on Density Levels Functional connectivity and structural connectivity have distinctive differences in sparsity levels. Functional networks like ABCD are fully connected. Structural networks like PPMI contain approximately 22.64% edges on average. Through our experiments, we found sparsity levels do have an impact on the choices of hyperparameters. For example, GNNs on the sparser structural networks of PPMI reach the maximum performance with a hidden dimension of 64, whereas on the functional network of ABCD, they have an optimal hidden dimension of 256, which indicates that GNN models should more complicated with more learnable parameters when the input networks are denser. This observation can be instructive for designing GNN architectures on brain networks constructed from different modalities.

2.3.6 Discussion and Extensions

In this paper, we first present BrainGB, a *unified*, *modular*, *scalable*, and *reproducible* framework for brain network analysis with GNNs. While the dataset generation, baselines, and evaluations we provide in BrainGB are thorough, we consider several limitations in the current paradigm:

- The aggregation mechanism in GNN is known to be effective for node-level tasks with the effect of node feature smoothing, and for graph-level tasks due to its capability in structure differentiation. However, for brain networks, what kinds of graph structures (e.g., communities, subgraphs) are effective beyond

the pairwise connections are still unknown.

- The small size of neuroimaging datasets may limit the effectiveness and generalization ability of complex deep learning models.

Towards these two limitations, we envision several future directions that can be potentially helpful to fully unleash the power of GNNs for brain network analysis:

- Neurology-driven GNN designs: to design the GNN architectures based on neurological understandings of predictive brain signals, especially disease-specific ones.

- Pre-training and transfer learning of GNNs: to design techniques that can train complex GNN models across studies and cohorts [292]. Besides, information sharing across different diseases could lead to a better understanding of cross-disorder commonalities.

2.4 Interpretable Brain Network Inference

Human brains lie at the core of complex neurobiological systems, where the neurons, circuits, and subsystems interact in enigmatic ways. Understanding the structural and functional mechanisms of the brain has long been an intriguing pursuit for neuroscience research and clinical disorder therapy. Mapping the connections of the human brain as a network is one of the most pervasive paradigms in neuroscience. Graph Neural Networks (GNNs) have recently emerged as a potential method for modeling complex network data. Deep models, on the other hand, have low interpretability, which prevents their usage in decision-critical contexts like healthcare. To bridge this gap, we propose an interpretable framework to analyze disorder-specific Regions of Interest (ROIs) and prominent connections. The proposed framework consists of two modules: a brain-network-oriented backbone model for disease prediction and

a globally shared explanation generator that highlights disorder-specific biomarkers including salient ROIs and important connections. We conduct experiments on three real-world datasets of brain disorders. The results verify that our framework can obtain outstanding performance and also identify meaningful biomarkers. All code for this work is available at `https://github.com/HennyJie/IBGNN.git`.

2.4.1 Introduction

Brain networks (a.k.a the connectome) are complex graphs with anatomic regions represented as nodes and connectivities between the regions as links. Interpretable models on brain networks for disorder analysis are vital for understanding the biological functions of neural systems, which can facilitate early diagnosis of neurological disorders and neuroscience research [177]. Previous work on brain networks has studied models from shallow to deep, such as graph kernels [109], tensor factorizations [154], and convolutional neural networks [121, 146].

Recently, Graph Neural Networks (GNNs) attract broad interest due to their established power for analyzing graph-structured data [126, 251]. Compared with shallow models, GNNs are suitable for brain network analysis with universal expressiveness to capture the sophisticated connectome structures [175, 40, 329, 292]. However, GNNs as a family of deep models are prone to overfitting and lack transparency in predictions, preventing their usage in decision-critical areas like disorder analysis. Although several methods have been proposed for GNN explanation [297, 168, 252], most of them focus on node-level prediction tasks and will produce a unique explanation for each subject when applied to graph-level tasks. However, for graph-level connectome-based disorder analysis, it is recognized that subjects having the same disorder share similar brain network patterns [117], which means disorder-specific explanations across instances are preferable. Moreover, brain networks have unique properties such that directly applying vanilla GNN models will obtain suboptimal

performance.

In this work, we propose an interpretable GNN framework to investigate disease-specific patterns that are common across the group and robust to individual image quality. Meanwhile, the group-level interpretation can be combined with subject-specific brain networks for different levels of interpretation. As shown in Fig. 3.16, it is composed of two modules: a backbone model IBGNN which adapts a message passing GNN designed for connectome-based disease prediction and an explanation generator that learns a globally shared mask to highlight disorder-specific biomarkers including salient Regions of Interest (ROIs) and important connections. Furthermore, we combine the two modules by enhancing the original brain networks with the learned explanation mask and further tune the backbone model. The resulting model, which we term IBGNN+ for brevity, produces predictions and interpretations simultaneously.

Through experiments on three real-world brain disorder datasets (i.e. HIV, BP, and PPMI), we show our backbone model performs well across brain networks constructed from different neuroimaging modalities. Also, it is demonstrated that the explanation generator can reveal disorder-specific biomarkers coinciding with neuroscience findings. Last, we show that the combination of explanation generator and backbone model can further boost disorder prediction performance.

2.4.2 Preliminaries

Brain Network Analysis Brain networks are complex graphs with anatomic Regions of Interest (ROIs) represented as nodes and connectivities between the ROIs as links [182]. In recent years, the analysis of brain networks has become increasingly important in neuroimaging studies to understand human brain organization across different groups of individuals [228, 210, 51, 260, 300]. Abundant findings in neuroscience research suggest that neural circuits are highly related to brain func-

tions, with aberrations in these neural circuits being identified in diseased individuals [105, 267, 144].

Formally, in the task of brain network analysis, the input is a brain network dataset $\mathcal{D} = \{\mathcal{G}_n, y_n\}_{n=1}^N$ consisting of N subjects, where $\mathcal{G}_n = \{\mathcal{V}_n, \mathcal{E}_n\}$ represents the brain network of subject n and y_n is the subject's label of the prediction, such as neural diseases. In $\mathcal{D}$, the brain network $\mathcal{G}_n$ of every subject n involves the same set of M nodes defined by the ROIs on a specific brain parcellation, i.e., $\forall n, \mathcal{V}_n = \mathcal{V} = \{v_i\}_{i=1}^M$. The difference across subjects lies in the edge connections $\mathcal{E}_n$ among M brain regions, which are often represented by a weighted adjacency matrix $\boldsymbol{W}_n \in \mathbb{R}^{M \times M}$ describing the connection strengths between ROIs. The edge weights in $\boldsymbol{W}$ are real-valued and the edges are potentially dense and noisy. The model outputs a prediction $\hat{y}_n$ for each subject n, which can be further analyzed in terms of features and biomarkers.

Given brain networks constructed from different modalities such as Diffusion Tensor Imaging (DTI) and functional Magnetic Resonance Imaging (fMRI) [20, 331, 102], effective analysis of the neural connectivities of different label groups (e.g., disease, gender) plays a pivotal role in understanding the biological structures and functions of the complex neural system, which can be helpful in the early diagnosis of neurological disorders and facilitate neuroscience research [177, 284, 149, 223, 215, 46, 98]. Previous models on brain networks are mostly shallow, such as graph kernels [109] and tensor factorization [92, 155], which are unable to model the complex graph structures of the brain networks [71].

Graph Neural Network Graph Neural Networks (GNNs) have revolutionized the field of graph modeling and analysis for real-world networked data such as social networks [126], knowledge graphs [212], protein or gene interaction networks [280], and recommendation systems [272]. The advantage of GNNs is that they can combine node features and graph structures in an end-to-end fashion as needed for specific

prediction tasks. A generic framework of GNN could be represented in two phases. In the first phase, it computes the representation $\boldsymbol{h}_i$ of each node $v_i \in \mathcal{V}_n$ by recursively aggregating messages from v_i's multi-hop neighborhood, where $\boldsymbol{h}_i^0$ is initialized with node features. After getting the last-layer node representation $\boldsymbol{h}^{(L)}$, an extra pooling strategy is adopted to obtain the graph representation. Thereafter, a Multi-Layer Perceptron (MLP) can be applied to make predictions on the downstream tasks.

It is worth noting that brain networks are different from other real-world graphs such as social networks or knowledge graphs, due to (1) the lack of useful initial node (ROI) features on brain networks represented by featureless graphs, (2) the real-valued connection weights that can be both positive or negative, and (3) the ROI identities and their orders are fixed across individual graph samples within the same dataset. The design of GNN models should be customized to fit the unique nature of brain network data. Recently, there have been emerging efforts on GNN-based brain network analysis [145, 146, 120, 16, 41, 329, 116, 234, 233]. However, these models are only tested on specific local datasets, mainly due to the convention in neuroscience that researchers are more used to developing methods that are applicable to their specific datasets and the regulatory restrictions that most brain imaging datasets are usually restrictively public, meaning that qualified researchers need to request access to the raw imaging data and preprocess them to obtain brain network data, but they are not allowed to release the preprocessed data afterwards. These challenges largely prohibit the methodology development in computational neuroscience research.

2.4.3 Method

Problem definition. The input to the proposed framework is a set of N weighted brain networks. For each network $G = (V, E, \boldsymbol{W})$, $V = \{v_i\}_{i=1}^{M}$ is the node set of size M defined by the Regions Of Interest (ROIs) on a specific brain parcellation [73, 217], with each v_i initialized with the node feature $\boldsymbol{x}_i$, $E = V \times V$ is the edge set of

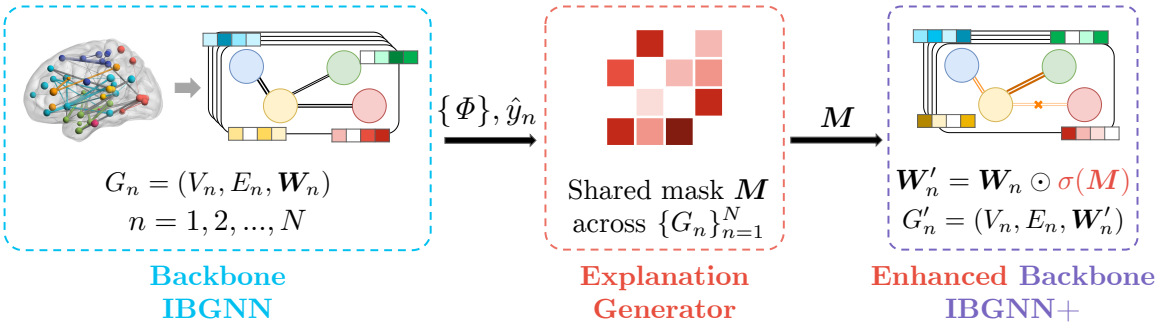


Figure 2.4: An overview of our proposed framework. The backbone model is firstly trained on the original data. Then, the explanation generator learns a globally shared mask across subjects. Finally, we enhance the backbone by applying the learned explanation mask and fine-tune the whole model.

brain connectome, and $\boldsymbol{W} \in \mathbb{R}^{M \times M}$ is the weighted adjacency matrix describing the connection strengths between ROIs. The model outputs a brain disorder prediction $\hat{y}_n$ for each subject n and learns a disorder-specific interpretation matrix $\boldsymbol{M} \in \mathbb{R}^{M \times M}$ that is shared across all subjects to highlight the disorder-specific biomarkers.

The backbone model IBGNN. Edge weights in brain networks are often determined by the signal correlation between brain areas, which may have both positive and negative values, and thus cannot be handled correctly by conventional GNNs. To avoid this issue and better utilize edge weights in the GNN model, we design an edge-weight-aware message passing mechanism specifically for brain networks. Specifically, we first construct a message vector $\boldsymbol{m}_{ij} \in \mathbb{R}^D$ by concatenating embeddings of a node v_i and its neighbor v_j, and the edge weight w_{ij}:

$$\boldsymbol{m}_{ij}^{(l)} = \text{MLP}_1\left(\left[\boldsymbol{h}_i^{(l)}; \boldsymbol{h}_j^{(l)}; w_{ij}\right]\right), \tag{2.19}$$

where l is the index of the GNN layer. Then, for each node v_i, we aggregate messages from all its neighbors $\mathcal{N}_i$ with the following propagation rule:

$$\boldsymbol{h}_i^{(l)} = \xi\left(\sum\nolimits_{v_j \in \mathcal{N}_i \cup \{v_i\}} \boldsymbol{m}_{ij}^{(l-1)}\right), \tag{2.20}$$

where ξ is a non-linear activation function such as ReLU, and $\boldsymbol{h}_i^{(0)}$ is initialized with node feature $\boldsymbol{x}_i$ reflecting the connectivity information in brain networks [43]. After stacking L layers, a readout function summarizing all node embeddings is employed

to obtain a graph-level embedding $\boldsymbol{g}$. Formally, we instantiate this function with another Multi-Layer Perceptron (MLP) and residual connections:

$$\boldsymbol{z} = \sum\nolimits_{i \in V} \boldsymbol{h}_i^{(L)}, \qquad \boldsymbol{g} = \mathrm{MLP}_2(\boldsymbol{z}) + \boldsymbol{z}. \tag{2.21}$$

This backbone model IBGNN can be trained with the conventional supervised cross-entropy objective towards ground-truth disorder prediction, defined as

$$\mathcal{L}_{\mathrm{CLF}} = -\frac{1}{N} \sum\nolimits_{n=1}^{N} (y_n \log(\hat{y}_n) + (1 - y_n) \log(1 - \hat{y}_n)). \tag{2.22}$$

The globally shared explanation generator. A general paradigm to generate explanations for GNNs is to find an explanation graph G' that has the maximum agreement with the label distribution on the original graph $G = (V, E, \boldsymbol{W})$, where G' can be a subgraph [297] or other variations of G [168, 305]. However, these explanation methods for GNNs mostly work on node-level prediction tasks and will produce a unique explanation graph for each subject when applied to graph-level tasks. On the other hand, directly using attention weights in some attention-based GNN models [251, 307] as explanations is known to be problematic [106, 8]. Note that brain networks have some unique properties. For example, the node number and order are fixed under a given atlas. Also, brain networks assume that subjects with the same brain disorder have similar brain connection patterns. Therefore, a globally shared explanation graph G' capture common patterns for specific disorders at the group level is preferable.

In this work, we propose to learn a globally shared edge mask $\boldsymbol{M} \in \mathbb{R}^{M \times M}$ that is applied to all brain network subjects in a dataset. Specifically, we maximize the agreement between the predictions $\hat{y}$ on the original graph G and $\hat{y}'$ on an explanation graph $G' = (V, E, \boldsymbol{W}')$ induced by a masking matrix $\boldsymbol{M}$, where $\boldsymbol{W}' = \boldsymbol{W} \odot \sigma(\boldsymbol{M})$, $\odot$ denotes element-wise multiplication, and σ denotes the sigmoid function. Formally this objective is implemented as a cross-entropy loss:

$$\mathcal{L}_{\mathrm{MASK}} = -\frac{1}{N} \sum\nolimits_{n=1}^{N} \sum\nolimits_{c=1}^{C} \mathbb{1}[\hat{y}_n = c] \log P_{\Phi}\left(\hat{y}'_n = \hat{y}_n \mid G'_n\right), \tag{2.23}$$

where $\sum_{n=1}^{N} P_{\Phi}\left(\hat{y}'_n = \hat{y}_n \mid G'_n\right)$ represents the conditional probability that the backbone model Φ's prediction $\hat{y}'_n$ on the masked graph G'_n is consistent with the prediction $\hat{y}_n$ on the original graph G_n, C is the number of possible prediction labels. Besides, following the practice in GNNExplainer [297], we further apply two regularization terms $\mathcal{L}_{\text{SPS}}$ and $\mathcal{L}_{\text{ENT}}$ to encourage the compactness of the explanation and the discreteness of the mask values, respectively:

$$\mathcal{L}_{\text{SPS}} = \sum\nolimits_{i,j} \boldsymbol{M}_{i,j}, \qquad \mathcal{L}_{\text{ENT}} = -(\boldsymbol{M}\log(\boldsymbol{M}) + (1-\boldsymbol{M})\log(1-\boldsymbol{M})). \tag{2.24}$$

The final training objective is given as:

$$\mathcal{L} = \mathcal{L}_{\text{CLF}} + \alpha\mathcal{L}_{\text{MASK}} + \beta\mathcal{L}_{\text{SPS}} + \gamma\mathcal{L}_{\text{ENT}}, \tag{2.25}$$

where α, β and γ scale the numerical value of each loss item to the same order of magnitude to balance their influence. Our explanation generator will generate a globally shared edge mask that can be used for all testing graphs to investigate neurological biomarkers and highlight disorder-specific salient connections.

Enhancing the backbone with the learned explanations. The learned explanation mask can further improve the disorder prediction considering that raw brain networked data inevitably contain random noise. Specifically, we enhance the original backbone by applying essential disorder-specific signals. We note that this strategy is compatible with any backbone model, not limited to our proposed IBGNN. We combined the aforementioned two modules so that predictions and interpretations are produced in a closed-loop for brain disorder analysis. We term the enhanced model by IBGNN+ hereafter.

The whole training pipeline is summarized in Fig. 3.16. The original brain networks are firstly input to train the backbone model. Then, a globally shared explanation mask is learned based on the backbone model Φ and prediction $\hat{y}_n$. Finally, we enhance the backbone model by highlighting salient ROIs and important connections on the raw data and tune the backbone model again.

2.4.4 Experiments

Dataset acquisition and preprocessing. We evaluate our framework using three real-world neuroimaging datasets of different modalities. Specifically, groups in each dataset have balanced age and gender portions and are collected with the same image acquisition procedure.

- *Human Immunodeficiency Virus Infection (HIV):* This dataset is collected from Early HIV Infection Study at Northwestern University. It includes fMRI imaging of 70 subjects, 35 of which are early HIV patients, and the others are seronegative controls. We perform image preprocessing using the DPARSF[22] toolbox. The images are realigned to the first volume, followed by slice timing correction, normalization, spatial smoothness using an 8-mm Gaussian kernel, band-pass filtering (0.01-0.08 Hz), and linear trend removing of the time series. We focus on the 116 anatomical regions of interest (ROI), and extract a sequence of responses from them. Finally, brain networks with 90 cerebral regions are constructed, where each node represents a brain region and links are created based on correlations between different brain regions.
- *Bipolar Disorder (BP):* This DTI imaging dataset is collected from 52 bipolar I subjects and 45 healthy controls. We use the FSL toolbox[23] for preprocessing which includes distortion correction, noise filtering, and repetitive sampling from the distributions of principal diffusion directions for each voxel. Each subject is parcellated into 82 regions based on FreeSurfer-generated cortical/subcortical gray matter regions.
- *Parkinson's Progression Markers Initiative (PPMI):* This large-scale, publicly available dataset[24] is from a collaborative study[25] to improve PD therapeutics.

[22] `http://rfmri.org/DPARSF/`
[23] `https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/`
[24] `https://www.ppmi-info.org/`
[25] `https://www.michaeljfox.org/`

We consider brain imaging in the DTI modality of 754 subjects, 596 of whom are Parkinson's disorder patients, and the rest 158 are healthy controls. The raw data are aligned using the FSL eddy-correct tool to correct head motion and eddy current distortions. The brain extraction tool (BET) from FSL is used to remove non-brain tissue. The skull-stripped images are linearly aligned and registered using Advanced Normalization Tools (ANTs[26]). 84 ROIs are parcellated from T1-weighted structural MRI using FreeSurfer[27] and the brain network connectivity is reconstructed using the deterministic 2nd-order Runge-Kutta (RK2) whole-brain tractography algorithm [314].

Experimental settings. The proposed model is implemented using PyTorch 1.10.2 [192] and PyTorch Geometric 2.0.3 [72]. A Quadro RTX 8000 GPU with 48GB of memory is used for our model training. Hyper-parameters are selected automatically with the open source AutoML toolkit NNI[28]. We refer readers of interest to supplementary materials for implementation details. All reported results are averaged of ten-fold cross validation.

Baselines. We compare our proposed models, i.e., the backbone model IBGNN and the explanation enhanced IBGNN+, with competitors of both shallow and deep models. Shallow methods include M2E [154], MIC [214], MPCA [162], and MK-SVM [63], where the output graph-level embeddings are evaluated using logistic regression classifiers. We also include three representative deep graph models: GAT [250], GCN [126], PNA [38] and two state-of-the-art deep models specifically design for brain networks: BrainNetCNN [121] and BrainGNN [146].

[26] `http://stnava.github.io/ANTs/`
[27] `https://surfer.nmr.mgh.harvard.edu/`
[28] `https://github.com/microsoft/nni`

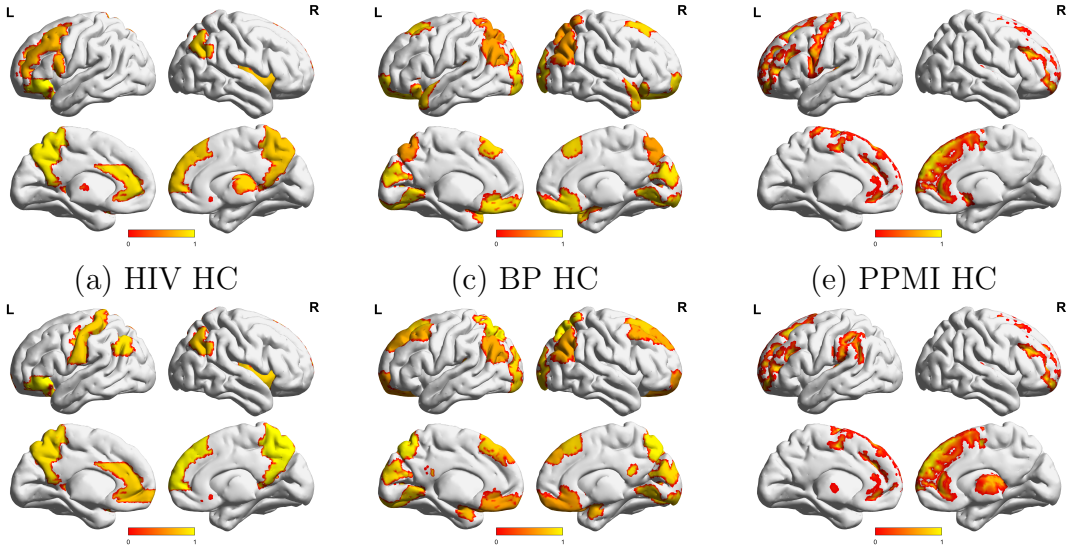
Table 2.3: Experimental results (%) on three datasets, where * denotes a significant improvement according to paired t-test with $p = 0.05$ compared with baselines. The best performances are in bold and the second runners are underlined.

Method	HIV			BP			PPMI		
	Accuracy	F1	AUC	Accuracy	F1	AUC	Accuracy	F1	AUC
M2E	$57.14_{\pm 19.17}$	$53.71_{\pm 19.80}$	$57.50_{\pm 18.71}$	$52.56_{\pm 13.86}$	$51.65_{\pm 13.38}$	$52.42_{\pm 13.83}$	$78.69_{\pm 1.78}$	$45.81_{\pm 4.17}$	$50.39_{\pm 2.59}$
MIC	$54.29_{\pm 18.95}$	$53.63_{\pm 19.44}$	$55.42_{\pm 19.10}$	$62.67_{\pm 20.92}$	$63.00_{\pm 21.61}$	$61.79_{\pm 21.74}$	$79.11_{\pm 2.16}$	$49.65_{\pm 5.10}$	$52.39_{\pm 2.94}$
MPCA	$67.14_{\pm 20.25}$	$64.28_{\pm 23.47}$	$69.17_{\pm 20.17}$	$52.56_{\pm 13.12}$	$50.43_{\pm 14.99}$	$52.42_{\pm 13.69}$	$79.15_{\pm 0.57}$	$44.18_{\pm 0.18}$	$50.00_{\pm 0.00}$
MK-SVM	$65.71_{\pm 7.00}$	$62.08_{\pm 7.49}$	$65.83_{\pm 7.41}$	$57.00_{\pm 8.89}$	$41.08_{\pm 13.44}$	$53.75_{\pm 8.00}$	$79.15_{\pm 0.57}$	$44.18_{\pm 0.18}$	$50.00_{\pm 0.00}$
GCN	$70.00_{\pm 12.51}$	$68.35_{\pm 13.28}$	$73.58_{\pm 9.49}$	$55.56_{\pm 13.86}$	$50.71_{\pm 11.75}$	$61.55_{\pm 28.77}$	$78.55_{\pm 1.58}$	$47.87_{\pm 4.40}$	$59.43_{\pm 8.64}$
GAT	$71.43_{\pm 11.66}$	$69.79_{\pm 10.83}$	$77.17_{\pm 9.42}$	$63.34_{\pm 9.15}$	$60.42_{\pm 7.56}$	$67.07_{\pm 5.98}$	$79.02_{\pm 1.25}$	$45.85_{\pm 3.16}$	$64.40_{\pm 6.87}$
PNA	$57.14_{\pm 12.78}$	$45.09_{\pm 19.62}$	$57.14_{\pm 12.78}$	$63.71_{\pm 11.34}$	$55.54_{\pm 14.06}$	$60.30_{\pm 11.89}$	$79.36_{\pm 1.84}$	$51.76_{\pm 10.32}$	$54.71_{\pm 6.77}$
BrainNetCNN	$69.24_{\pm 19.04}$	$67.08_{\pm 11.11}$	$72.09_{\pm 19.01}$	$65.83_{\pm 20.64}$	$64.74_{\pm 17.42}$	$64.32_{\pm 13.72}$	$55.20_{\pm 12.63}$	$\underline{55.45}_{\pm 9.15}$	$52.54_{\pm 10.21}$
BrainGNN	$74.29_{\pm 12.10}$	$73.49_{\pm 10.75}$	$75.00_{\pm 10.56}$	$68.00_{\pm 12.45}$	$62.33_{\pm 13.01}$	$74.20_{\pm 12.93}$	$69.17_{\pm 0.00}$	$44.19_{\pm 0.00}$	$45.26_{\pm 3.65}$
IBGNN	$\underline{82.14}_{\pm 10.81}^{*}$	$\underline{82.02}_{\pm 10.86}^{*}$	$\underline{86.86}_{\pm 11.65}^{*}$	$\underline{73.19}_{\pm 12.20}$	$\underline{72.87}_{\pm 12.09}^{*}$	$\underline{83.64}_{\pm 9.61}^{*}$	$\mathbf{79.82}_{\pm 1.47}$	$51.58_{\pm 4.66}$	$\underline{70.65}_{\pm 6.55}^{*}$
IBGNN+	$\mathbf{84.29}_{\pm 12.94}^{*}$	$\mathbf{83.86}_{\pm 13.42}^{*}$	$\mathbf{88.57}_{\pm 10.89}^{*}$	$\mathbf{76.33}_{\pm 13.00}^{*}$	$\mathbf{76.13}_{\pm 13.01}^{*}$	$\mathbf{84.61}_{\pm 9.08}^{*}$	$\underline{79.55}_{\pm 1.67}$	$\mathbf{56.58}_{\pm 7.43}$	$\mathbf{72.76}_{\pm 6.73}^{*}$

Prediction performance. The overall results are presented in Table 3.5. Both our proposed models yield impressive improvements over SOTA shallow and deep baselines. Compared with shallow models such as MK-SVM, our backbone model IBGNN outperforms them by large margins, with up to 11% absolute improvements on BP. Besides, the effectiveness of our brain network-oriented design is supported by its superiority compared with other SOTA deep models. Moreover, the performance of the explanation enhanced model IBGNN+ can further increase the backbone by about 9.7% relative improvements, which demonstrates that IBGNN+ effectively highlights the disorder-specific signals while also achieving the benefit of restraining random noises in individual graphs.

2.4.5 Interpretation Analysis

Neural system mapping. The ROIs on brain networks can be partitioned into neural systems based on their structural and functional roles under a specific parcellation atlas, which facilitates the understanding of generated explanations from a neuroscience perspective. In this paper, we map the ROI nodes as defined on each dataset into eight commonly used neural systems, including Visual Network (VN), Auditory Network (AN), Bilateral Limbic Network (BLN), Default Mode Network (DMN), Somato-Motor Network (SMN), Subcortical Network (SN), Memory Net-



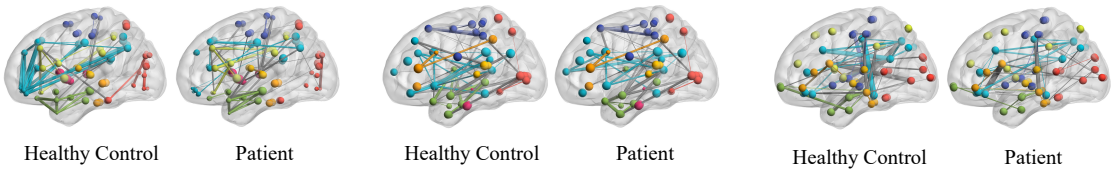
(a) HIV HC (c) BP HC (e) PPMI HC

(b) HIV Patient (d) BP Patient (f) PPMI Patient

Figure 2.5: Visualization of salient ROIs on the explanation enhanced brain connection networks for Health Control (HC) and Patient. The color of regions represents ROI's average importance in the given group. The bright-yellow color indicates a high score, while dark-red indicates a low score.

work (MN), and Cognitive Control Network (CCN).

Salient ROIs. We provide both group-level and individual-level interpretations to understand which ROIs contribute most to the prediction of a specific disorder. On the group level, we rank the most salient ROIs on the learned explanation mask by calculating the sum of the edge weights connected to each node. Then on the individual level, we use the BrainNet Viewer [273] to plot the salient ROIs on the average brain connectivity graph enhanced by the learned explanation mask. For the HIV disease, anterior cingulate, paracingulate gyri, and inferior frontal gyrus are selected as salient ROIs. This complies with scientific findings that the regional homogeneity value of the anterior cingulate and paracingulate gyri are decreased [169] and lower gray matter volumes are found in inferior frontal gyrus in HIV patients [147]. The individual-level visualizations in Fig. 2.5(a)(b) show the difference between Health Control (HC) and HIV patients in those salient ROIs. For the BP disease, secondary visual cortex and medial to superior temporal gyrus are selected as salient ROIs. This



(a) HIV (b) BP (c) PPMI

Figure 2.6: Visualization of important connections on the explanation enhanced brain connection network. Edges connecting nodes within the same neural system (VN, AN, BLN, DMN, SMN, SN, MN, CCN) are colored accordingly, while edges across different systems are colored gray. Edge width indicates its weight in the explanation graph.

observation is in line with existing studies that visual processing abnormalities have been characterized in bipolar disorder patients [202, 187], which is also confirmed in Fig. 2.5(c)(d). For the PPMI disease, rostral middle frontal gyrus and superior frontal gyrus are selected as salient ROIs and Fig. 2.5(e)(f) display the difference. This is in accordance with MRI analysis revealing a significant decrease in PD patients in the rostral medial frontal gyrus and superior, middle, and inferior frontal gyri [122]. All these observed salient ROIs can be potential biomarkers to identify brain disorders from each cohort.

Important connections. The globally shared explanation mask $\boldsymbol{M}$ provides interpretations of important connections. We obtain an explanation subgraph G'_s by taking the top 100 weighted edges from the masked G' with all other edges removed. The connection comparisons are shown in Fig. 2.6, which helps identify connections related to specific disorders. For the HIV dataset, the explanation subgraph of patients excludes rich interactions within the DMN (colored blue) system. Also, interactions within the VN (colored red) system of patients are significantly less than those of HCs. These patterns are consistent with the findings in earlier studies [97, 74] that connectivity alterations within- and between-network DMN and VN may relate to known visual processing difficulties for HIV patients. For the BP dataset, compared with tight interactions within the BLN (colored green) system of the healthy control, the connections within BLN system of the patient subject are much sparser, which

may signal pathological changes in this neural system. This observation is in line with previous studies [49], which finds that the parietal lobe, one of the major lobes in the brain roughly located at the upper back area in the skull and is in charge of processing sensory information received from the outside world, is mainly related to Bipolar disorder attack. Since parietal lobe ROIs are contained in BLN under our parcellation, the connections missing within the BLN system in our visualization are consistent with existing clinical understanding. For the PPMI dataset, the connectivity in the patient group decreases in the SMN (colored purple) system, which integrates primary sensorimotor, premotor, and supplementary motor areas to facilitate voluntary movements. This observation confirms existing neuroimaging studies that have repeatedly shown disorder-related alteration in sensorimotor areas of Parkinson's patients [25]. Furthermore, individuals with PD have lower connectivity within the DMN (colored blue) system compared with healthy controls, which is consistent with the cognition recession study on Parkinson's patients [245, 238].

2.4.6 Conclusion

In this work, we propose a novel interpretable GNN framework for connectome-based brain disorder analysis, which consists of a brain network-oriented GNN predictor and a globally shared explanation generator. Experiments on real-world neuroimaging datasets show the superior prediction performance of both our backbone and the explanation enhanced models and validate the disorder-specific interpretations from the generated explanation mask. The limitation of the proposed framework might arise from the small size of neuroimaging datasets, which restraints the effectiveness and generalization ability of deep learning models. A direct future direction based on this work is to utilize pre-training and transfer learning techniques to learn across datasets. This allows for the sharing of information and explanations across different cohorts, which could lead to a better understanding of cross-disorder commonalities.

Chapter 3

Broader Types of Multimodal Data: Structured Knowledge Extraction and Augmented Inference

3.1 Specialized Models for Structured Knowledge Extraction from Textual Data

3.1.1 Introduction

Concept map, which models texts as a graph with words/phrases as vertices and relations between them as edges, has been studied to improve information retrieval tasks previously [321, 70, 114]. Recently, graph neural networks (GNNs) attract tremendous attention due to their superior power established both in theory and through experiments [126, 88, 249, 157, 39]. Empowered by the structured document representation of concept maps, it is intriguing to apply powerful GNNs for tasks

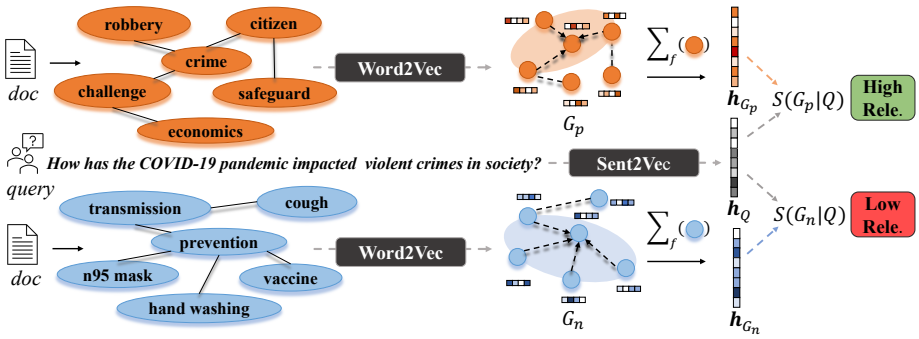


Figure 3.1: An overview of GNN-based document retrieval.

like document classification [289] and retrieval [320]. Take Fig. 3.1 as an example. Towards the query about "violent crimes in society", a proper GNN might be able to highlight query-relevant concept of "crime" and its connection to "robbery" and "citizen", thus ranking the document as highly relevant. On the other hand, for another document about precaution, the GNN can capture concepts like "n95 mask" and "vaccine", together with their connections to "prevention", thus ranking it as not so relevant.

Present work. In this work, we explore how GNNs can help document retrieval with generated concept maps. The core contributions are three-fold:

- We use constituency parsing to construct semantically rich concept maps from documents and design quality evaluations for them towards document retrieval.
- We investigate two types of graph models for document retrieval: the structure-oriented complex GNNs and our proposed semantics-oriented graph functions.
- By comparing the retrieval results from different graph models, we provide insights towards GNN model design for textual retrieval, with the hope to prompt more discussions on the emerging areas such as IR with GNNs.

3.1.2 Concept Map based Document Retrieval

Overview

In this section, we describe the process of GNN-based document retrieval. As is shown in Fig. 3.1, concept maps $G = \{V, E\}$ are first constructed for documents. Each node $v_i \in V$ is a concept (usually a word or phrase) in the document, associated with a frequency f_i and an initial feature vector $\boldsymbol{a}_i$ from the pretrained model. The edges in E denote the interactions between concepts. GNNs are then applied to each individual concept map, where node representation $\boldsymbol{h}_i \in \mathbb{R}^d$ is updated through neighborhood transformation and aggregation. The graph-level embedding $\boldsymbol{h}_G \in \mathbb{R}^d$ is summarized over all nodes with a read-out function.

For the training of GNN models, the widely-used triplet loss in retrieval tasks [171, 295, 288] is adopted. Given a triplet (Q, G_p, G_n) composed by a relevant document G_p (denoted as positive) and an irrelevant document G_n (denoted as negative) to the query Q, the loss function is defined as:

$$L(Q, G_p, G_n) = \max\left\{S(G_n \mid Q) - S(G_p \mid Q) + \mathit{margin}, 0\right\}. \tag{3.1}$$

The relevance score $S\left(G \mid Q\right)$ is calculated as $\frac{\boldsymbol{h}_G \cdot \boldsymbol{h}_Q}{\|\boldsymbol{h}_G\| \|\boldsymbol{h}_Q\|}$, where $\boldsymbol{h}_G$ is the learned graph representation from GNN models and $\boldsymbol{h}_Q$ is the query representation from a pretrained model. In the training process, the embeddings of relevant documents are pulled towards the query representation, whereas those of the irrelevant ones are pushed away. For retrieval in the testing phrase, documents are ranked according to the learned relevance score $S(G \mid Q)$.

Concept Maps and Their Generation

Concept map generation, which aims to distill structured information hidden under unstructured text and represent it with a graph, has been studied extensively in literature [31, 287, 290, 320]. Since entities and events often convey rich semantics, they

are widely used to represent core information of documents [36, 143, 164]. However, according to our pilot trials, existing concept map construction methods based on name entity recognition (NER) or relation extraction (RE) often suffer from limited nodes and sparse edges. Moreover, these techniques rely on significant amounts of training data and predefined entities and relation types, which restricts the semantic richness of the generated concept maps [259].

To increase node/edge coverage, we propose to identify entities and events by POS-tagging and constituency parsing [173]. Compared to concept maps derived from NER or RE, our graphs can identify more sufficient phrases as nodes and connect them with denser edges, since pos-tagging and parsing are robust to domain shift [178, 299]. The identified phrases are filtered via articles removing and lemmas replacing, and then merged by the same mentions. To capture the interactions (edges in graphs) among extracted nodes, we follow the common practice in phrase graph construction [181, 206, 128] that uses the sliding window technique to capture node co-occurrence. The window size is selected through grid search. Our proposed constituency parsing approach for concept map generation alleviates the limited vocabulary problem of existing NER-based methods, thus bolstering the semantic richness of the concept maps for retrieval.

GNN-based Concept Map Representation Learning

Structure-oriented complex GNNs Various GNNs have been proposed for graph representation learning [126, 88, 279, 249]. The discriminative power of complex GNNs mainly stems from the 1-WL test for graph isomorphism, which exhaustively capture possible graph structures so as to differentiate non-isomorphic graphs [279]. To investigate the effectiveness of structured-oriented GNNs towards document retrieval, we adopt two state-of-the-art ones, Graph isomorphism network (GIN) [279] and Graph attention network (GAT) [249], as representatives.

Semantics-oriented permutation-invariant graph functions The advantage of complex GNNs in modelling interactions may become insignificant for semantically important task. In contrast, we propose the following series of graph functions oriented from semantics perspectives.

- **N-Pool**: independently process each single node v_i in the concept map by multi-layer perceptions and then apply a read-out function to aggregate all node embeddings $\boldsymbol{a}_i$ into the graph embedding $\boldsymbol{h}_G$, i.e.,

$$\boldsymbol{h}_G = \text{READOUT}\Big(\{\text{MLP}(\boldsymbol{a}_i) \mid v_i \in V\}\Big). \tag{3.2}$$

- **E-Pool**: for each edge $e_{ij} = (v_i, v_j)$ in the concept map, the edge embedding is obtained by concatenating the projected node embedding $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ on its two ends to encode first-order interactions, i.e.,

$$\boldsymbol{h}_G = \text{READOUT}\Big(\{cat\left(\text{MLP}(\boldsymbol{a}_i), \text{MLP}(\boldsymbol{a}_j)\right) \mid e_{ij} \in E\}\Big). \tag{3.3}$$

- **RW-Pool**: for each sampled random walk $p_i = (v_1, v_2, \ldots, v_m)$ that encode higher-order interactions among concepts ($m = 2, 3, 4$ in our experiments), the embedding is computed by the sum of all node embeddings on it, i.e.,

$$\boldsymbol{h}_G = \text{READOUT}\Big(\{sum\left(\text{MLP}(\boldsymbol{a}_1), \text{MLP}(\boldsymbol{a}_2), \ldots, \text{MLP}(\boldsymbol{a}_m)\right) \mid p_i \in P\}\Big). \tag{3.4}$$

All of the three proposed graph functions are easier to train and generalize. They preserve the *message passing* mechanism of complex GNNs [80], which is essentially *permutation invariant* [176, 175, 123], meaning that the results of GNNs are not influenced by the orders of nodes or edges in the graph; while focusing on the basic semantic units and different level of interactions between them.

3.1.3 Experiments

Experimental Setup

Dataset We adopt a large scale multi-discipline dataset from the TREC-COVID[1]

[1]https://ir.nist.gov/covidSubmit/

Table 3.1: The similarity of different concept map pairs.

Pair Type	# Pairs	**NCR** (%)	**NCR+** (%)	**ECR** (%)	**ECR+** (%)
Pos-Pos	762,084	4.96	19.19	0.60	0.78
Pos-Neg	1,518,617	4.12	11.75	0.39	0.52
(*t-score*)	-	(*187.041*)	(*487.078*)	(*83.569*)	(*105.034*)
Pos-BM	140,640	3.80	14.98	0.37	0.43
(*t-score*)	-	(*126.977*)	(*108.808*)	(*35.870*)	(*56.981*)

challenge [204] based on the CORD-19[2] collection [255]. The raw data includes a corpus of 192,509 documents from broad research areas, 50 queries about the pandemic that interest people, and 46,167 query-document relevance labels.

Experimental settings and metrics We follow the common two-step practice for the large-scale document retrieval task [153, 48, 186]. The initial retrieval is performed on the whole corpus with full texts through BM25 [205], a traditional yet widely-used baseline. In the second stage, we further conduct re-ranking on the top 100 candidates using different graph models. The node features and query embeddings are initialized with pretrained models from [318, 32]. NDCG@20 is adopted as the main evaluation metric for retrieval, which is used for the competition leader board. Besides NDCG@K, we also provide Precision@K and Recall@K (K=10, 20 for all metrics).

Evaluation of Concept Maps

We empirically evaluate the quality of concept maps generated from Section 3.1.2. The purpose is to validate that information in concept maps can indicate query-document relevance, and provide additional discriminative signals based on the initial candidates. Three types of pairs are constructed: a Pos-Pos pair consists of two documents both relevant to a query; a Pos-Neg pair consists of a relevant and an irrelevant one; and a Pos-BM pair consists of a relevant one and a top-20 one from BM25. Given a graph pair G_i and G_j, their similarity is calculated via four measures: the node coincidence rate (NCR) defined as $\frac{|V_i \cap V_j|}{|V_i \cup V_j|}$; NCR+ defined as NCR weighted

[2]https://github.com/allenai/cord19

Table 3.2: The retrieval performance results of different models.

Type	Methods	**Precision** (%)		**Recall** (%)		**NDCG** (%)	
		k=10	k=20	k=10	k=20	k=10	k=20
Traditional	BM25	55.20	49.00	1.36	2.39	51.37	45.91
	Anserini	54.00	49.60	1.22	2.25	47.09	43.82
Structure-Oriented	GIN	35.24	34.36	0.77	1.50	30.59	29.91
	GAT	46.48	43.26	1.08	2.00	42.24	39.49
Semantics-Oriented	N-Pool	58.24	52.20	1.38	2.41	53.38	48.80
	E-Pool	59.60	53.88	1.40	2.49	56.11	51.16
	RW-Pool	**59.84**	**53.92**	**1.42**	**2.53**	**56.19**	**51.41**

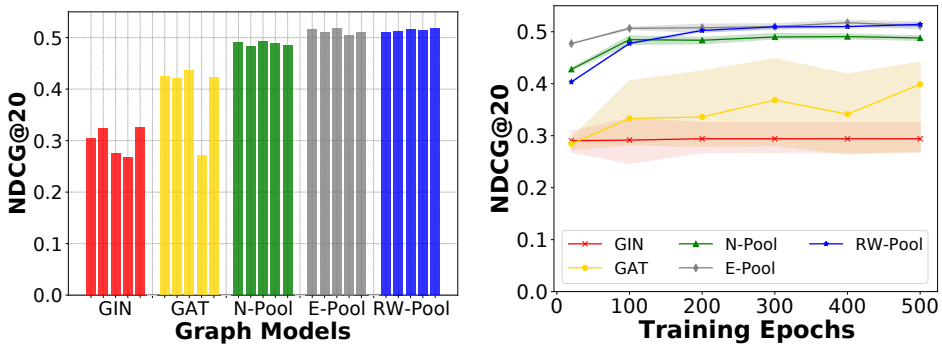
by the tf-idf score [7] of each node; the edge coincidence rate (ECR) where an edge is coincident when its two ends are contained in both graphs; and ECR+ defined as ECR weighted by the tf-idf scores of both ends.

It is shown in Table 3.1 that Pos-Neg pairs are less similar than Pos-Pos under all measures, indicating that concept maps can effectively reflect document semantics. Moreover, Pos-BM pairs are not close to Pos-Pos and even further away than Pos-Neg. This is because the labeled "irrelevant" documents are actually hard negative ones difficult to distinguish. Such results indicate the potential for improving sketchy candidates with concept maps. Besides, student's t-Test[99] is performed, where standard critical values of (Pos-Pos, Pos-Neg) and (Pos-Pos, Pos-BM) under 95% confidence are 1.6440 and 1.6450, respectively. The calculated *t-scores* shown in Table 3.1 strongly support the significance of differences.

Retrieval Performance Results

In this study, we focus on the performance improvement of GNN models based on sketchy candidates. Therefore, two widely-used and simple models, the forementioned BM25 and Anserini[3], are adopted as baselines, instead of the heavier language models such as BERT-based [55, 294, 52] and learning to rank (LTR)-based [21, 270] ones. The retrieval performance are shown in Table 3.2. All the values are reported as the averaged results of five runs under the best settings.

[3]https://git.uwaterloo.ca/jimmylin/covidex-trec-covid-runs/-/tree/master/round5, which is recognized by the competition organizers as a baseline result.



(a) Stability comparison (b) Efficiency comparison

Figure 3.2: Stability and efficiency comparison of different graph models.

For the structure-oriented GIN and GAT, different read-out functions including mean, sum, max and a novel proposed tf-idf (i.e., weight the nodes using the tf-idf scores) are experimented, and tf-idf achieves the best performance. It is shown that GIN constantly fails to distinguish relevant documents while GAT is relatively better. However, they both fail to improve the baselines. This performance deviation may arise from the major inductive bias on complex structures, which makes limited contribution to document retrieval and is easily misled by noises. In contrast, our three proposed semantics-oriented graph functions yield significant and consistent improvements over both baselines and structure-oriented GNNs. Notably, E-Pool and RW-Pool improve the document retrieval from the initial candidates of BM25 by 11.4% and 12.0% on NDCG@20, respectively. Such results demonstrate the potential of designing semantics-oriented GNNs for textual reasoning tasks such as classification, retrieval, etc.

Stability and Efficiency

We further examine the stability and efficiency of different models across runs. As is shown in Fig. 3.2(a), GIN and GAT are less consistent, indicating the difficulty in training over-complex models. The training efficiency in Fig. 3.2(b) shows that GIN can hardly improve during training, while GAT fluctuates a lot and suffers from overfitting. In contrast, our proposed semantics-oriented functions perform more stable

in Fig. 3.2(a), and improve efficiently during training in Fig. 3.2(b), demonstrating their abilities to model the concepts and interactions important for the retrieval task. Among the three graph functions, E-Pool and RW-Pool are consistently better than N-Pool, revealing the utility of simple graph structures. Moreover, RW-Pool converges slower but achieves better and more stable results in the end, indicating the potential advantage of higher-order interactions.

3.1.4 Conclusion

In this paper, we investigate how can GNNs help document retrieval through a case study. Concept maps with rich semantics are generated from unstructured texts with constituency parsing. Two types of GNNs, structure-oriented complex models and our proposed semantics-oriented graph functions are experimented and the latter achieves consistently better and stable results, demonstrating the importance of semantic units as well as their simple interactions in GNN design for textual reasoning tasks like retrieval. In the future, more textual datasets such as news, journalism and downstream tasks can be included for validation. Other types of semantics-oriented graph functions can also be designed based on our permutation-invariant schema, such as graphlet based-pooling.

3.2 Specialized Models for Structured Knowledge Extraction from Visual Data

Images contain rich relational knowledge that can help machines understand the world. Existing methods of visual knowledge extraction often rely on the pre-defined format (e.g., sub-verb-obj tuples) or vocabulary (e.g., relation types), restricting the expressiveness of the extracted knowledge. In this work, we take a first exploration of a new paradigm of open visual knowledge extraction. To achieve this, we present

`OpenVik`, which consists of an open relational region detector to detect regions potentially containing relational knowledge and a visual knowledge generator that generates format-free knowledge by prompting the large multimodality model with the detected region of interest. We also explore two data enhancement techniques for diversifying the generated format-free visual knowledge. Extensive knowledge quality evaluations highlight the correctness and uniqueness of the extracted open visual knowledge by `OpenVik`. Moreover, integrating our extracted knowledge across various visual reasoning applications shows consistent improvements, indicating the real-world applicability of `OpenVik`.

3.2.1 Introduction

Knowledge extraction has been widely studied on texts [42, 3, 68, 45] for enhancing logical reasoning [237, 77, 33] and explainable AI [100, 301, 26, 281], and recent studies have explored *open* knowledge extraction through categorizing seed relations [315, 208] and eliciting from language models [253]. Visual knowledge extraction, on the other hand, captures intricate details like tools, sizes, and positional relationships, which are often difficult to express exhaustively in texts [207, 142, 256, 44]. Yet existing approaches of visual knowledge extraction are either restricted by a fixed knowledge format [277, 311, 112, 124] or the predefined sets of objects/relations [277, 311, 115]. While efficient at capturing interactions between objects, the produced visual knowledge is often limited in richness and confined to a single format, falling short of representing the diverse real-world information that can be complemented by visual data.

In this endeavor, we propose to further explore a new paradigm of open visual knowledge extraction (`OpenVik`). Specifically, we propose to generate relation-oriented, but format-free knowledge that includes a wider variety of elements, such as descriptions, insertions, and attributes, among others. Drawing inspiration from

the wealth of knowledge encapsulated in large models [264, 310, 242], we propose to leverage pre-trained large multimodality models by eliciting open visual knowledge through relation-oriented visual prompting. This approach allows for a more nuanced understanding of visual data, mirroring how humans naturally emphasize certain aspects of visual scenes when perceiving and describing visual information, leading to more flexible visual knowledge extraction.

Our proposed `OpenVik` framework consists of two modules, an open relational region detector and a format-free visual knowledge generator. It is a unique challenge to detect the regions potentially containing relational knowledge since traditional region detectors primarily focus on learning predefined object classes. To learn the regression of relational regions, we propose to use free-form knowledge descriptions as supervision and leverage knowledge generation as a training objective. With the detected regions, the remaining question is how to interpret these regions into free-form knowledge. We propose a visual knowledge generator by harnessing the power of language variety enhancement in large pre-trained multimodality models. Specifically, we prompt them to generate knowledge descriptions of any formats and condition the generation on the detected relational regions.

However, establishing a new paradigm of open visual knowledge extraction is challenging due to the absence of comprehensive and diverse training data. Existing datasets sources such as scene graphs [276, 129], dense captions [112], and dense relational subsets [124] often exhibit a long-tail distribution biased to more prevalent relations and entities [235]. Brute-force merging of these datasets could exacerbate the distribution bias inherent in the data. To alleviate the bias, we propose two diversity-driven data enhancement strategies based on an adapted TF-IDF+ score, involving random dropping and data augmentation with external knowledge resources. These strategies optimize data distributions and richness, thus fostering diverse open visual knowledge extraction.

We implement extensive evaluations to assess the quality and utility of the open visual knowledge extracted by `OpenVik`, encompassing: 1) directly evaluating the performance of knowledge generation; 2) engaging human evaluators for a multi-faceted assessment of in-depth knowledge quality; and 3) comparing the open visual knowledge extracted with `OpenVik` with existing knowledge sources, such as non-parametric knowledge from the ConceptNet knowledge graph, and parametric knowledge from the GPT-3.5 large language model. Furthermore, the utility of the extracted open visual knowledge is validated through its integration with several common applications that require visual understanding, including text-to-image retrieval, grounded situation recognition, and visual commonsense reasoning. These applications demonstrate consistent improvements, affirming the practical utility of `OpenVik`.

3.2.2 Related Work

Visual knowledge extraction. Recent advancements in knowledge extraction have extended from being purely text-driven to incorporating images [58, 156]. VisKE [207] is designed to verify relations between pairs of entities, e.g., *eat*(*horse*, *hay*). Scene graphs, which locate objects in the image and identify visual predicates between subjects and objects in a triple format, e.g., (*man, on, chair*), are extensively studied for vision understanding [277, 311, 309]. A recent work OpenSGG [93] extends SGG to open-vocabulary objects, enabling the relation prediction for unseen objects. Other studies have explored caption-like formats, like dense captioning [112] with a set of object-centric descriptions across regions, and relational captioning [124] focusing on relational information between objects. Despite these advancements, existing methods either adhere to a pre-defined format and vocabulary or are constrained by the biased distribution of training sets. This highlights the pressing need for a format-free approach in visual knowledge extraction with knowledge diversity.

Large model prompting. Recently, large language and multimodality models have

exhibited remarkable successes in capturing commonsense knowledge across various tasks, especially facilitating few-shot [79, 282, 133, 303] and zero-shot learning [127, 324, 304]. The potential of prompt-based learning for pre-trained vision-language models [4, 199, 229] has been explored for handling diverse data types across multiple modalities, such as images and texts, with improved performance in tasks including image classification [179, 325], segmentation [167] and visual question answering [86]. Leveraging the substantial information encapsulated within these pre-trained multimodality models to extract explicit knowledge can enrich existing resources, potentially laying the groundwork for advances in interpretability research and mitigating the hallucination issue associated with large models [107, 47].

3.2.3 Method

In this section, we introduce our new paradigm and two key model design novelty featuring `OpenVik`, relation-oriented multimodality model prompting and diversity-

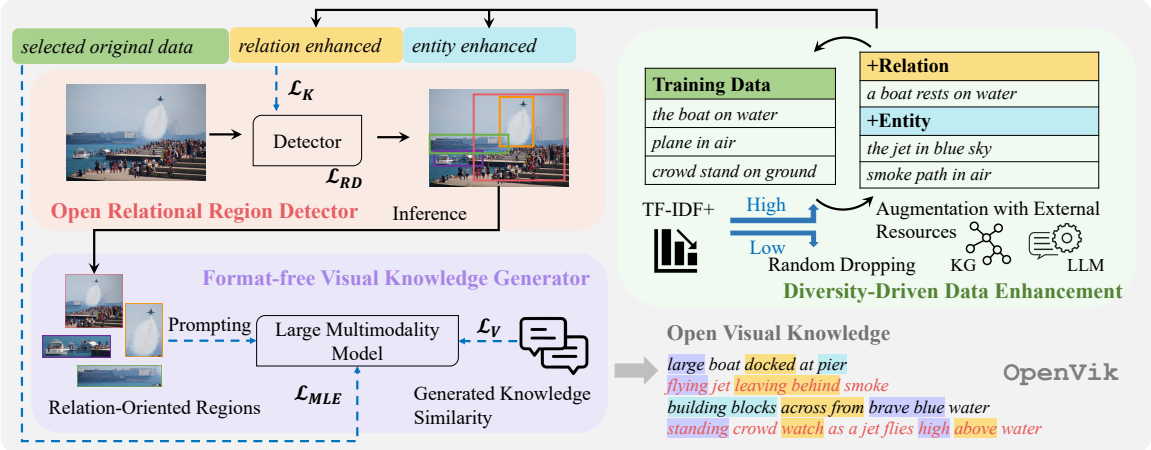


Figure 3.3: The overview of `OpenVik`. The left orange and purple panels illustrate key components of relation-oriented multimodality model prompting: open relational region detector and format-free visual knowledge generator. The right green one depicts diversity-driven data enhancement strategy. `OpenVik` is designed to extract relation-oriented format-free open visual knowledge with novel entities, diverse relations, and nuanced descriptive details.

Open Visual Knowledge Extraction

Given a dataset $\mathcal{D} = \{(\mathcal{I}_i, \mathbf{T}_i, \mathbf{U}_i)\}_{i=1}^{M}$ consisting of M samples, $\mathcal{I}_i$ is the i-th image (such as the input image in Figure 3.3), $\mathbf{T}_i = \{\mathcal{T}_j\}_{j=1}^{n_i}$ is a set of n_i region descriptions (such as "*the boat on water*" in Figure 3.3), $\mathbf{U}_i = \{\mathcal{U}_j\}_{j=1}^{n_i}$ is the set of n_i relation-oriented visual regions, where each $\mathcal{T}_j$ corresponds to a visual region $\mathcal{U}_j \in \mathbf{U}_i$ in image $\mathcal{I}_i$. The goal of our open visual knowledge discovery is to train a model $\mathcal{M}$ capable of producing a set of format-free knowledge descriptions (such as "*large boat docked at pier*" in Figure 3.3) given any image $\mathcal{I}_k$ during the inference stage.

Relation-Oriented Multimodality Model Prompting

The overall architecture of `OpenVik` is shown in Figure 3.3. It comprises two modules: an open relational region detector $\mathcal{M}_v$ and a format-free visual knowledge generator $\mathcal{M}_t$. The two modules are learned separately during training with our diversity-enhanced data (Section 3.2.3) and combined to produce format-free visual knowledge at inference. Specifically, the relational region detector $\mathcal{M}_v$ takes an image $\mathcal{I}_i$ as the input and learns to select a flexible number of relational regions $\mathbf{U}_i = \{(\mathcal{U}_j)\}_{j=1}^{n_i}$ that captures object interactions, each corresponding to a description $\mathcal{T}_j$ in $\mathbf{T}_i$; the visual knowledge generator $\mathcal{M}_t$ generates format-free knowledge descriptions by prompting and fine-tuning the multimodality model with the guidance of detected visual region $\mathcal{U}_j$. All notations for the region detector and knowledge generator are detailed in Table 3.3 and Table 3.4, respectively.

Table 3.3: Notations for open region detector.

Notation	Meaning
$\mathcal{I}_i$	input image of the relational region detector
$\mathcal{U}_j$	relation-centric box label
$\mathbf{U}_i$	set of relation-centric boxes of an image
$\mathcal{T}_j$	region description of a box
$\mathbf{T}_i$	set of region descriptions of the an image
$\mathcal{L}_{\text{RD}}$	region regression loss supervised by union regional boxes
$\mathcal{L}_{\text{K}}$	knowledge generation loss supervised by GT relational knowledge
$\mathcal{L}_v$	the overall objective of the relational region detector

Table 3.4: Notations for knowledge generator.

Notation	Meaning
$\mathcal{I}_i$	the input image of the knowledge generator
T_a, T_b	two regional knowledge descriptions of one same image
N_i	the number of generated knowledge descriptions of an image
ϕ	hyper-parameter controlling the penalty slightly different sequences
$\mathcal{L}_{\text{MLE}}$	the language modeling loss of the generation decoder
$\mathcal{L}_{\text{V}}$	inter-sequence information variety regularizer
α	weight hyper-parameter balancing generation accuracy and variety
$\mathcal{L}_l$	the overall objective of the knowledge generator

◇ **Open relational region detector.** Although existing object detection algorithms have been widely recognized for their efficiency in object detection, they are usually restricted to object-centric visual regions in a predefined set, and thus cannot directly capture open relational information with a single box. Detecting regions containing relational knowledge remains to be a challenge. We make two adaptions on the object detection FasterRCNN [203] to train the open relational region detector:

- *Region Regression*: we change the original object-centric region labels to our newly created relation-centric box labels, denoted as $\mathbf{U}_j$. The foreground of each relation-centric region label $\mathcal{U}_j$ is created by taking the union of the object-level bounding boxes of the entities, i.e., *boat, water*, contained in a ground truth region knowledge description $\mathcal{T}_j$. This forms the region regression loss $\mathcal{L}_{\text{RD}}$.
- *Knowledge Supervision*: To assist with the refinement of the bounding box, we replaced the object-centric label classification in traditional object detectors with knowledge supervision. A pre-trained generator is finetuned to create the regional description grounded to the given region. This is supervised by the cross-entropy loss $\mathcal{L}_{\text{K}}$ with region description $\mathbf{T}_j$.

The training objective $\mathcal{L}_l$ of the relational region detector is formulated as below, where $\mathcal{L}_{\text{RD}}$ is the regional regression loss and $\mathcal{L}_{\text{K}}$ is the knowledge supervision loss,

$$\mathcal{L}_v = \mathcal{L}_{\text{RD}} + \mathcal{L}_{\text{K}}. \tag{3.5}$$

◇ **Format-free visual knowledge generator.** OpenVik provides better knowledge grounding by conditioning the generator on the detected relational region, leading to a reasoning-driven generation. Specifically, the detected bounding box (such as the box containing "*boat*" and "*pier*" on the far left) is utilized as a visual prompt when fine-tuning the visual knowledge generator. The model architecture of the knowledge generator is built upon a combined large multimodality model, which composes a pre-trained vision transformer ViT-B [61] and the image-grounded text decoder of BLIP [139]. The two modules are jointly trained on a generic image-text paired

dataset comprising over 14 million entries and fine-tuned on the image captioning task, which delivered state-of-the-art performance.

In our visual knowledge generator, the decoder takes the ViT visual representation of the entire image as input and leverages the detected regional mask as a binary visual prompt. This prompt aids in filtering out the background and directing attention toward the relational foreground. The generation of format-free knowledge from the decoder is supervised by the language modeling loss $\mathcal{L}_{\text{MLE}}$, which further refines visual attention during the knowledge generation process. As a result, our approach facilitates the production of format-free outcomes that extend beyond the conventional sub-verb-obj form. Besides, to improve information variety, we introduce an amplifying penalty factor for highly similar knowledge generation. For any two generated sequences T_a and T_b describing image $\mathcal{I}_i$,

$$\mathcal{L}_{\text{V}} = \frac{1}{N_i} \sum_{N_i} \text{ReLU}\left(-\log\left(1 - \left(s\left(T_a, T_b\right) - \phi\right)\right)\right), \tag{3.6}$$

where N_i is the number of generated knowledge of image $\mathcal{I}_i$, $s\left(T_a, T_b\right)$ indicates the semantic cosine similarity, and ϕ is a hyper-parameter set as 0.01 controlling the penalty on sequences with only slight difference (e.g. "*dog chasing the man*" and "*dog licking the man*") to be relatively small.

The training objective $\mathcal{L}_l$ of the format-free visual knowledge generator is formulated as

$$\mathcal{L}_l = \alpha \times \mathcal{L}_{\text{MLE}} + (1 - \alpha) \times \mathcal{L}_{\text{V}}, \tag{3.7}$$

where α is a weighting hyper-parameter we set as 0.7. The trained relational region detector and visual knowledge generator are combined during inference. Given any image $\mathcal{I}$, the open relational region detector first detects a flexible number of open relations regions of interest, then each detected region $\mathcal{R}$ is passed to the format-free visual knowledge generator, where a relation-oriented format-free knowledge phrase (such as "*flying jet leaving behind smoke*" in Figure 3.3) is generated to describe the given

visual focus subarea $\mathcal{R}$ of the image. To further encourage within-sequence language variety during inference, we leverage the contrastive decoding strategy from [230], which improves over nucleus sampling and beam search.

Diversity-driven Data Enhancement

The training data for relational knowledge extraction usually exhibits a long-tail distribution, where more prevalent but simple relations such as *in*, *on*, and *wear* dominate the training set [235]. Consequently, the model trained with such a biased dataset may render limited and repetitive knowledge. As a remedy, we propose two data enhancement techniques to optimize the data distribution. As the foundational measure for given relation r's importance, we design a grid TF-IDF+ score $\mathcal{S}_r$ [190, 283]:

$$\mathcal{S}_r = (\log(\frac{N}{1 + f_r * \alpha_1}))^{\alpha_2}, \tag{3.8}$$

where N is the total number of knowledge phrases in the datasets, f_r is the number of occurrences of the relation r, α_1 and α_2 are the grid scales whose values are selected based on f_r.

◇ **Random dropping on low-quality data.** We first remove repeated knowledge descriptions in the same image and then randomly drop descriptions that contain frequently occurring yet meaningless relations with a low $\mathcal{S}_r$ (e.g., "*people on ground*") from the original dataset. Specifically, if the $\mathcal{S}_r$ of the relation in a description is relatively low, i.e., 0.4, we remove it at a random dropping rate of 0.5. This process repeats for all descriptions in an image until the remaining set is 0.6 times the size of the original training set. Consequently, the training data bias is mitigated by removing low-quality data.

◇ **Data augmentation with external knowledge resources.** For the relations with high TF-IDF+ scores, we leverage external knowledge resources from both non-parametric (i.e., ConceptNet [226]) and parametric (i.e., COMET [18]) knowledge re-

sources to promote diverse knowledge generation [293]. ✔ *Enhance Relation Recognition*: For each training description with a high TF-IDF+ score, we perform semantic parsing to get all the objects and complement additional relations (e.g., "*rest*" in Figure 3.3) between each pair of them by mapping the nodes and retrieving edges from the ConceptNet. Each retrieved knowledge triplet is converted to a knowledge phrase and added to the training set for generator training. With this introduced external knowledge, the knowledge generator ultimately yields a more robust and detailed representation of the underlying visual information of objects. This, in turn, bolsters the relation recognition of the visual knowledge generator. ✔ *Boost Entity Perception*: For the description with the highest-scored TF-IDF+ relation given each image, we also leverage ConceptNet to enrich similar objects (e.g., "*jet*") to the original object (e.g., "*plane*"). Additionally, we further introduce new entities (e.g., "*smoke*" in Figure 3.3) and attribute descriptions (e.g., "*blue*") by prompting the pre-trained attribute commonsense branch of the COMET model (Refer to Appendix A.1 for more details). The entity-based enrichment potentially helps in boosting entity understanding and at the same time enhances the occurrence of important but rare relations in the training set.

Implementation Details

Our training data are built based on Visual Genome [129] and its relation-enhanced version Dense Relational Captioning [124]. Each sample includes an image identified by a unique ID and a set of relational descriptors describing interactions among objects in the image. Specifically, each relational descriptor includes the full description text, the subject and object names contained in the description text, the relation between them, as well as the bounding box coordinates of the subject and object. The dataset statistic information is summarized in Table A.1 in the Appendix A.2.

Our model is implemented in PyTorch [191] and trained on two Quadro RTX 8000

GPUs. The open relational region detector is initialized from the ResNet50-FPN backbone, then finetuned for another 20 epochs with the relational bounding box. The model detects a maximum of 30 bounding boxes for each image with the highest confidence to avoid misleading noises. The format-free visual knowledge generator is initialized from $\text{BLIP}_{\text{base}}$ with the basic ViT-B/16 and finetuned for 20 epochs. Full details on learning parameters can be referred to in Appendix A.3.

3.2.4 Evaluation

In this section, we directly evaluate the extracted open visual knowledge from `OpenVik` from two perspectives: (1) knowledge generation performance with traditional generative metrics and in-depth knowledge quality assessment; (2) comparison with existing knowledge sources. Besides, ablation studies are conducted to study the influence of diversity design on the generated knowledge and data.

Evaluation on Generated Knowledge

Table 3.5: Knowledge comparison of `OpenVik` and baselines on performance and in-depth quality (%).

Method	Generation Performance			In-Depth Knowledge Quality			
	BLEU↑	ROUGE-L↑	METEOR↑	Validity↑	Conformity↑	Freshness↑	Diversity↑
Closed/Open Scene Graph Generation							
IMP [277]	0.075	0.123	0.118	0.800	0.823	0.676	0.316
Neural Motifs [311]	0.229	<u>0.283</u>	**0.273**	0.822	0.767	0.667	0.349
UnbiasSGG [235]	0.217	0.258	0.194	0.739	0.733	0.666	0.357
Ov-SGG [93]	0.167	0.210	0.183	0.712	0.633	0.693	0.413
Dense Relational Captioning							
MTTSNet+REM [124]	0.240	0.226	0.228	<u>0.897</u>	0.852	0.754	0.375
Region Captioning							
DenseCap [112]	0.248	0.245	0.196	0.883	0.843	0.790	0.543
Sub-GC [323]	0.272	0.263	0.221	0.892	<u>0.871</u>	<u>0.795</u>	<u>0.547</u>
BLIP [139]	0.264	0.266	0.252	0.886	0.855	0.760	0.531
BLIP2 [140]	<u>0.275</u>	**0.285**	<u>0.257</u>	0.892	<u>0.871</u>	0.766	0.535
Open Visual Knowledge Extraction							
`OpenVik`	**0.280**	<u>0.283</u>	0.250	**0.907**	**0.883**	**0.809**	**0.619**

◇ **Generation performance.** To directly evaluate the visual knowledge generator, we compare the knowledge generated by `OpenVik` with a variety of baselines, including

scene graph generation [277, 311, 235, 93] (of which Ov-SGG employs an open vocabulary), dense relational captioning [124], and region captioning [112, 323, 139, 140]. Evaluation metrics are traditional language generation measures such as BLEU, ROUGE-L, and METEOR. The results, displayed in the left side of Table 3.5, reveal that `OpenVik` outperforms captioning-based approaches and yields results on par with the best scene graph generation baseline. These findings underscore the effectiveness of the format-free visual knowledge generator through relation-oriented prompting of the large multimodality model.

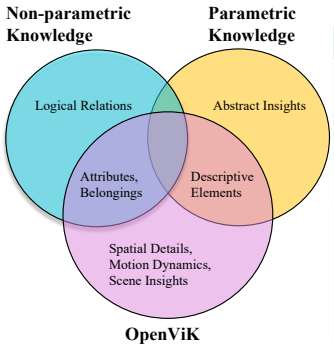
◇ **In-depth knowledge quality.** To more thoroughly evaluate the quality and richness of the format-free visual knowledge extraction, beyond simply evaluating it as a language generation model with the limitation of training data, we incorporate four additional metrics [165], which delve into an in-depth quality evaluation of the extracted visual knowledge from four distinct perspectives:

- *Validity* (↑): whether the generated visual knowledge is valid to human.
- *Conformity* (↑): whether the generated knowledge faithfully depicts the scenarios in the images.
- *Freshness* (↑): the novelty of the knowledge, i.e., the proportion not present in the training set.
- *Diversity* (↑): the language variance between a randomly sampled pair of knowledge pieces.

Among the four metrics, both the validity and conformity metrics involve human annotators. We randomly selected 100 images as the evaluative subset. Details regarding the scoring guidance and the interface provided to the annotators can be found in Appendix A.4. The remaining metrics, i.e., freshness and diversity, are calculated automatically. The in-depth knowledge quality evaluation results are displayed in the right part of Table 3.5, where the average pairwise Cohen's κ on human evaluation results is 0.76 (good agreement). The findings demonstrate that trained

with the diversity-enhanced datasets, the format-free visual knowledge extracted by `OpenVik` significantly outperforms other types of baselines in terms of all four metrics. The improvement of diversity, in particular, reaches 14% relatively compared with the inference results from the second runner DenseCap, indicating the advantage of `OpenVik` in generating rich and comprehensive visual knowledge.

Comparison with Existing Knowledge Sources



Knowledge Source	Examples
Non-parametric Knowledge	dog IsA animal (ConceptNet); conceptnet IsA knowledge graph (ConceptNet);
	dog HasProperty black (ConceptNet); dog and brown fur covering black (OpenVik)
	computer HasA keyboard (ConceptNet); keyboard with computer (OpenVik)
Parametric Knowledge	people using light bulbs to illuminate the room; (LLM)
	yellow sign in corner (both); black seat attached to bike (both);
Open Visual Knowledge (OpenViK)	three layer cake on table; blue trash can full of garbage next to brown dresser; blue box sitting beside a sneaky garage; (OpenVik)
	people wearing fashionable black hats are skiing; baby elephants walking around adventurous wood; (OpenVik)
	the light shining from bright black background; hanging fan are above tall shelf; brown chair in the background of the room; (OpenVik)

Figure 3.4: The Venn diagram of knowledge comparison between the open visual knowledge from `OpenVik` with the non-parametric knowledge from existing knowledge graph (i.e., ConceptNet) and parametric knowledge from large language model (i.e., COMET).

We compare the extracted visual knowledge with the non-parametric knowledge in the existing knowledge graph (KG) and the parametric knowledge from the large language model (LLM). The comparison insights from the three knowledge resources are shown in the Venn Diagram in Figure 3.4.

◇ **Compare with non-parametric knowledge.** We take ConceptNet [226] as the representative in the comparison with non-parametric knowledge. To map the knowledge generated by `OpenVik` to ConceptNet, we parse the knowledge into triplets and associate the endpoints of these triplets with nodes in ConceptNet. Then we calculate the similarity of embeddings[4] between the parsed relation and all the edge relations among the mapped nodes in ConceptNet. If the similarity score exceeds a

[4]Embeddings are produced by ConceptNet API: https://github.com/commonsense/conceptnet-numberbatch.

predetermined threshold, i.e., 0.75, we consider the mapping successful. As illustrated in Figure 3.4, we observe that compared with the non-parametric knowledge in KG, the extracted visual knowledge captures richer and more meaningful spatial details, e.g., "*three layer cake on table*", and motion dynamics, e.g., "*baby elephants walking around adventurous wood*".

◇ **Compare with parametric knowledge.** We compare with parametric knowledge contained in LLM by prompting the gpt-3.5-turbo[5] model with the object information in the image. The prompt template used is detailed in Appendix A.5. The mapping process follows the approach mentioned earlier. It is found that compared with the parametric knowledge in LLM, the extracted visual knowledge exhibits unique fine-grained visual details, e.g., "*red sticker on fence*", and provides precise scene information, e.g., "*the light shining from bright black background*".

Ablation Study

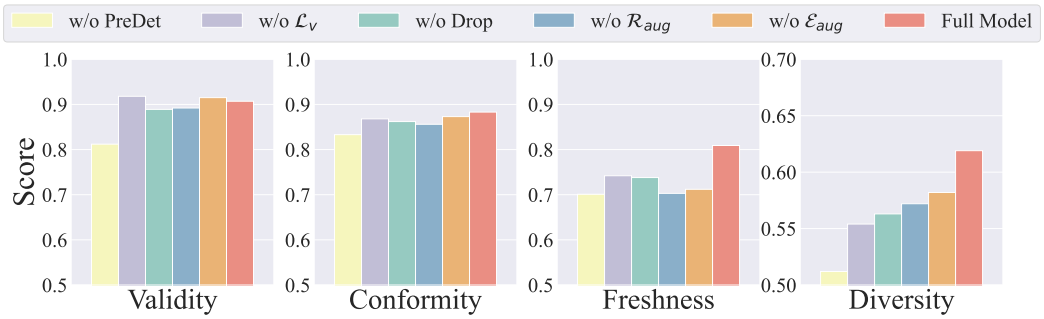


Figure 3.5: The influence of information variety regularization and diversity-driven data enhancement strategies.

◇ **The influence on knowledge quality with information variety regularization and data strategies.** We conducted ablation studies to evaluate the effectiveness of the information variety regularizer, $\mathcal{L}_{\mathrm{V}}$, and our diversity-driven data enhancement strategies. This involves an in-depth assessment of knowledge quality on the same evaluation subset. The results are presented in Figure 3.5. It is evident from the results that our proposed information variety design primarily impacts

[5]https://platform.openai.com/docs/models/gpt-3-5

freshness and diversity, without compromising validity and conformity. For the freshness, the omission of data augmentation for entities and relations results in the most significant performance degradation. This implies the crucial role these strategies play in infusing novel knowledge into the generation process. As for diversity, the most notable changes in metrics are observed when the $\mathcal{L}_{\mathrm{V}}$ and random dropping are removed. The strategy for augmenting entities and relations also plays a valuable role in enriching diversity.

◇ **Ablation of the pre-training for the open relational region detector.** We conducted a comparison of the outcomes when loading a pre-trained detector backbone versus training the detector from scratch, as shown by the yellow bar in Figure 3.5. Results demonstrate a noticeable decrease in both knowledge diversity and freshness, which indicates the importance of loading the pre-trained model for region detection. This may be because omitting the pre-training step of the FasterRCNN model tends to result in the detection of more overlapping regions, which in turn causes the drop.

◇ **The influence on dataset diversity with data strategies.** We conduct a direct analysis of the knowledge diversity of the existing datasets and our diversity-enhanced one, compared with the visual knowledge generated from `OpenVik`. The findings, presented in Table 3.6, show that the diversity-driven data enhancement strategies significantly boost knowledge diversity. Trained with this enhanced data, `OpenVik` can extract visual knowledge that exhibits greater diversity than that found in the *Visual Genome* and *Relational Caps*, indicating the advantage of `OpenVik` to format-free visual knowledge generation and its ability to yield richer knowledge diversity.

Table 3.6: Diversity of existing and enhanced datasets and generated knowledge from `OpenVik`.

Metrics	Training Dataset			Generate Knowledge
	Visual Genome [129]	*Relational Caps [124]*	*Diversity Enhanced (Ours)*	`OpenVik` *(Ours)*
Diversity	0.589	0.604	0.632	0.619

Case Study

We present two case studies in Figure 3.6 (See Appendix A.6 for more) to showcase the format-free visual knowledge generated by `OpenVik`, in comparison to Visual Genome (Scene Graph and Region Description) and Relational Caps. Contrary to the rigidity of scene graphs, which strictly adhere to a predefined format, `OpenVik` can generate knowledge with a flexible semantic structure, not strictly bound to the sub-verb-obj format (e.g., "*blue post attached to wall with white letter*"). Examples of this adaptability are highlighted in red. When compared to dense region descriptions, the relational knowledge extracted by `OpenVik` offers a deeper understanding of the multiple entity interactions within an image. In comparison to Relational Caps, which mainly focus on interactions between two objects, `OpenVik` significantly broadens the diversity of relation with vivid verbs (e.g., "*attached to*", "*adorning*"). Moreover, it introduces novel entities (e.g., "*post*", "*mane*") and enriches the knowledge representation with nuanced details (e.g., "*full of*", "*striped*") that are missed by Relational Caps.

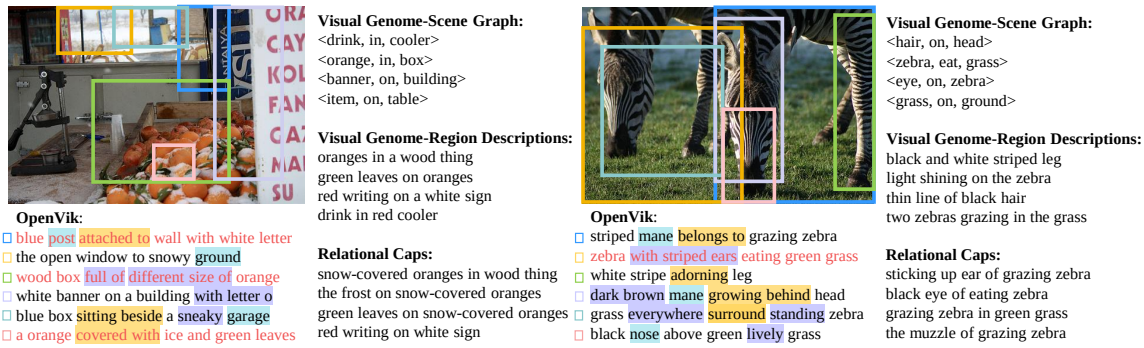


Figure 3.6: Case study on the extracted open visual knowledge from `OpenVik`. Examples of format-free knowledge are highlighted in red. Compared with VG and Relational Caps, `OpenVik` performs better at capturing novel entities, broadening object interactions with diverse relations, and enriching the knowledge representation with nuanced descriptive details.

Note that we observe the unbalanced and noisy distributions within the training data can lead to errors in the knowledge produced. Viewing hallucinations as erroneous inferences based on input, the inaccuracies observed in OpenVik and similar

baselines often stem from detection errors. These errors are typically caused by data biases that incorrectly associate features with a specific class or label. We further two illustrative failure cases in Figure 3.7. For example, a "*black speaker by flat tv*" is generated, although the speaker is not present in the image—possibly reflecting common co-occurrences within the dataset. Similarly, a ladder in the right figure has been misidentified as a towel, leading to the erroneous description of a "*blue towel hanging from dry shower*". The key to mitigating such incorrect inference is identifying the cofounder feature of class labeling.

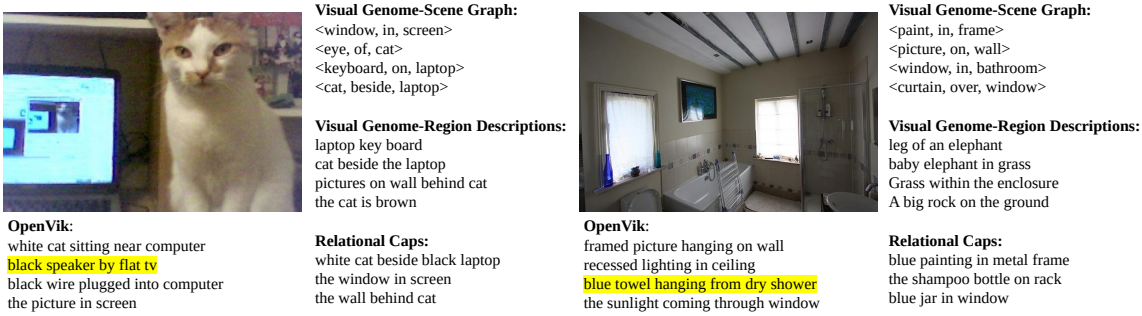


Figure 3.7: Examples of incorrectly knowledge resulting from distribution bias are highlighted.

3.2.5 Application

This section explores whether the extracted open visual knowledge from `OpenVik` can bolster reasoning and inference capabilities in multimodality downstream tasks by augmenting a baseline in the challenging zero-shot setting.

Text-to-Image Retrieval

◇ **Task setting.** In the text-to-image retrieval task, a given caption is matched to a large set of candidate images, with the most relevant image returned as the result. Adopting the challenging zero-shot setting, we generate the visual representation $\boldsymbol{v}$ and textual representation $\boldsymbol{t}$ of the given image $\mathcal{I}$ and caption $\mathcal{T}$ using a pre-trained clip-retrieval model [10]. The baseline involves the image and text em-

bedding similarly based on zero-shot CLIP Retrieval [10] and the fine-tuned model from BLIP [140].

To explore the potential of the extracted visual knowledge from `OpenVik`, we enrich the given caption $\mathcal{T}$ with related contexts derived from the extracted visual knowledge. Specifically, for each query caption, we parse the caption to extract all subject-object pairs (s, o) with the NLTK parser. Then s and o are mapped to the open visual knowledge, where knowledge phrases that contain relations occurring more than 30% of the time between s and o are enriched to the original caption $\mathcal{T}$.

Original text: A little room and dining room area with furniture. A living room with a big table next to a book shelf. A living room decorated with a modern theme. A living room with wooden floors and furniture. The large room has a wooden table with chairs and a couch.

Enriched text: big table in room. a decorated living room with wooden furniture. brown couch in room. book on table. wooden table with shelf. shelf next to couch. wooden bookshelf with books next to table.

Figure 3.8: An example of `OpenVik` enrichment on text-to-image retrieval (See Appendix A.7.1 for more).

Method	Recall@1	Recall@5	Recall@10	*Avg*
ZS-CLIP	36.16	65.47	78.66	60.10
`OpenVik` + ZS-CLIP	40.55	73.29	84.53	66.12
BLIP	63.11	86.30	91.10	80.17
`OpenVik` + BLIP	65.23	87.71	91.90	81.61

Table 3.7: Text-to-image retrieval results (%) of `OpenVik` enrichment compared with zero-shot baselines.

◇ **Qualitative examples.** Figure 3.8 presents an example of `OpenVik`-based visual knowledge enrichment on captions. By incorporating related contexts from the generated open visual knowledge, the enriched captions convey more precise visual details, which enhances the alignment for text-image alignment.

◇ **Quantitative results.** We curated a subset of 680 images from the testing set of the MS-COCO dataset containing parsed knowledge with at least eight nouns. This ensures an adequate degree of enrichment is achieved through the use of `OpenVik`. Standard image retrieval metrics, i.e., *Recall@1/5/10/* and *Avg*, are employed to evaluate the performance. The results are presented in Table 3.7. It is evident that relational context enrichment leads to the average correction of more than 6.0% of the initial zero-shot, highlighting the practical benefits of extracted visual knowledge in visual reasoning tasks.

Grounded Situation Recognition

◇ **Task setting.** The event type prediction for the grounded situation recognition task is to predict the best match from predefined 504 event types [198] based on the image. We convert each candidate event verb into a description $\mathcal{T}$: "*An image of <verb>*" for image description matching. Similarly to text-to-image retrieval, we include zero-shot CLIP and the fine-tuned model from BLIP as baselines.

To enrich with contextual knowledge from `OpenVik`, for each given verb v, we find its nearest synonym in the extracted open visual knowledge and enrich the text description with the most common knowledge phrase containing it, regularized by the objects present in the image. Instead of directly concatenating the retrieved knowledge triplets to the original textual description, we employ an additive decomposition strategy: the similarity $s(\mathcal{I}, v)$ of the candidate verb v with respect to the given image $\mathcal{I}$ is calculated as $s(\mathcal{I}, v) = \frac{1}{|D(v)|} \sum_{d \in D(v)} \phi(\mathcal{I}, v)$, where $D(v)$ is the set of descriptors, including the original description and the enriched ones, and ϕ represents the single log probability that descriptor d pertains to the image $\mathcal{I}$.

Verb: talking
Original text: This is an image of talking.
Enriched text: man talking on a small white telephone. adult male with white shirt talking on chatty cellphone. the man waving arms is talking on phone.

Figure 3.9: An example of `OpenVik` context enrichment on task GSR (See Appendix A.7.2 for more).

Method	Accuracy	Precision	Recall	F_1
ZS-CLIP	53.14	42.54	45.19	43.82
`OpenVik` + ZS-CLIP	75.16	61.63	62.75	62.18
BLIP	70.42	65.32	69.25	67.23
`OpenVik` + BLIP	80.25	72.55	70.61	71.57

Table 3.8: Grounded situation recognition results (%) of `OpenVik` enrichment compared with zero-shot baselines.

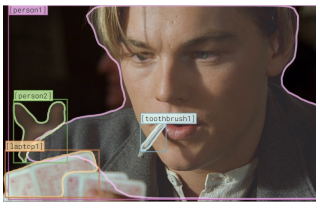
◇ **Qualitative examples.** Figure 3.9 presents a qualitative example of `OpenVik`-based context enrichment in the grounded situation recognition task. We observed that verbs like "shopping" and "talking" were appropriately enriched with their frequently occurring contexts from the open visual knowledge, leading to a reduced embedding distance between the description and its matching image.

◇ **Quantitative results.** We assembled a test set of 900 samples from the testing set of GSR that included verbs such as "talking", "filming", and "picking", among

others, from a list of 256 words that can be accurately mapped to extracted visual knowledge, as well as 138 verbs that have a fuzzy match through ConceptNet embedding comparison. The full lists of the exact and fuzzy-matched verbs are detailed in Appendix A.8. The evaluated metrics include Accuracy, Precision, Recall, and F_1. The results are presented in Table 3.8. It can be observed that knowledge enrichment significantly outperforms the zero-shot and BLIP baselines. This suggests that the verb-related contexts introduced by `OpenVik`-generated knowledge are intuitive and greatly assist in understanding the semantics of event verbs, bolstered by related visual information.

Visual Commonsense Reasoning

◇ **Task setting.** The goal of visual commonsense reasoning is to predict an answer from four given option candidates for a given image and question. For the baseline approach, we compare the backbone model R2C from the VCR paper [312] and BLIP [139]. In the visual knowledge-enhanced `OpenVik` Enriched approach, we perform two-level context augmentation, incorporating both entities and relations: (1) we parse the question and options to obtain all `(S, O)` pairs and, for each entity pair, apply the same relation augmentation as in the image retrieval task; (2) for the `V` in each option, we enrich the visual context using the same method as illustrated in grounded situation recognition.



Question : Is Person1 winning the game? the person engaged in a game.
A Yes, he is about to go for a run. the person engaged in a game. the person walking near runway.
B No, he is losing. the person engaged in a game. frustrated person lose game.
C No, he' s not really enjoying it. the person engaged in a game. person enjoy at celebration.
D Yes, he looks like he has a good hand. the person engaged in a game. the person is watching left hand.
Answer: D Yes, he looks like he has a good hand.

Figure 3.10: An example of `OpenVik` context enrichment on the VCR task (See Appendix A.7.3 for more).

Method	Accuracy	Precision	Recall	F_1
R2C	56.66	56.73	56.72	56.72
`OpenVik` + R2C	59.96	60.01	60.03	60.02
BLIP	62.50	62.50	62.45	62.47
`OpenVik` + BLIP	67.40	67.54	67.43	67.48

Table 3.9: Visual commonsense reasoning results (%) of `OpenVik` context enrichment compared with zero-shot baselines.

◇ **Qualitative examples.** Figure 3.10 presents an example before and after applying the two-level visual knowledge-based enrichment for visual commonsense reasoning.

The results indicate that visual knowledge enhances the correspondence between the correct answer and the image itself.

◇ **Quantitative results.** We assembled a test set of 939 samples from the validation set of the VCR dataset [312]. Each sample in this test set contains questions and answers with a minimum of five nouns and two relations, guaranteeing an adequate level of information complexity for meaningful engagement with open visual knowledge. The results can be found in Table 3.9. We observe that the enriched visual knowledge helps especially when solving reasoning questions on humans and their interactions with visually impressive entities, such as "*game*" in Figure 3.10. This enhancement results in a performance improvement above 3.0% over the zero-shot baseline.

3.2.6 Conclusion, Limitations, and Future Work

This work is the first exploration of a new paradigm of open visual knowledge extraction, which combines an open relational region detector to flexibly pinpoint relational regions and a format-free visual knowledge generator that generates visual knowledge by prompting a multimodality model conditioned on the region of interest. To further enhance the diversity of the generated knowledge, we explore two distinct data enhancement techniques. Extensive knowledge evaluations underscore the correctness and uniqueness of our extracted open visual knowledge, and the consistent improvements observed across various visual reasoning tasks highlight the real-world applicability of `OpenVik`.

While our approach has been shown effective in various scenarios, its performance at larger scales or on more diverse datasets remains to be studied. Future work could investigate its effectiveness across a broader range of tasks and contexts. Also, the current model requires fine-tuning for the visual knowledge extractor. Developing a model that can generalize well with prompt tuning or demonstration augmentation

could be another interesting direction for future work.

3.3 Specialized Models for Structured Knowledge Extraction from Multimodal Data

Information extraction, e.g., attribute value extraction, has been extensively studied and formulated based only on text. However, many attributes can benefit from image-based extraction, like color, shape, pattern, among others. The visual modality has long been underutilized, mainly due to multimodal annotation difficulty. In this paper, we aim to patch the visual modality to the textual-established attribute information extractor. The cross-modality integration faces several unique challenges: (C1) images and textual descriptions are loosely paired intra-sample and inter-samples; (C2) images usually contain rich backgrounds that can mislead the prediction; (C3) weakly supervised labels from textual-established extractors are biased for multimodal training. We present PV2TEA, an encoder-decoder architecture equipped with three bias reduction schemes: (S1) Augmented label-smoothed contrast to improve the cross-modality alignment for loosely-paired image and text; (S2) Attention-pruning that adaptively distinguishes the visual foreground; (S3) Two-level neighborhood regularization that mitigates the label textual bias via reliability estimation. Empirical results on real-world e-Commerce datasets demonstrate up to 11.74% absolute (20.97% relatively) F_1 increase over unimodal baselines.

3.3.1 Introduction

Information extraction, e.g., attribute value extraction, aims to extract structured knowledge triples, i.e., (*sample_id, attribute, value*), from the unstructured information. As shown in Figure 3.11, the inputs include text descriptions and images (optional) along with the queried attribute, and the output is the extracted value. In

Image

Textual Descriptions: "Best Price Mattress 12 Inch Memory Foam Mattress, Calming ***Green Tea***-Infused Foam, Pressure Relieving, Bed-in-a-Box, Queen"
Question: What is the ***color*** of the mattress?
Weakly Supervised Label: green **True Value:** white



...

Challenge Explanations:

C1 Loosely-aligned image and textual descriptions:
- intra-sample: weakly related across modalities and difficult to ground
- inter-samples: images of other samples can also pair with this text

C2 Visual bias: noisy contextual backgrounds, e.g., pillow, bed frame, etc.
C3 Textual bias: the training label is misled/biased by 'green tea' in text

Figure 3.11: Illustration of multimodal attribute extraction and the challenges in cross-modality integration.

practice, textual description has played as the main or only input in mainstream approaches for automatic attribute value extraction [322, 278, 257, 119, 286, 59]. Such models perform well when the prediction targets are inferrable from the text.

As the datasets evolve, interest in incorporating visual modality naturally arises, especially for image-driven attributes, e.g., *Color*, *Pattern*, *Item Shape*. Such extraction tasks rely heavily on visual information to obtain the correct attribute values. The complementary information contained in the images can improve recall in cases where the target values are not mentioned in the texts. In the meantime, the cross-modality information can help with ambiguous cases and improve precision.

However, extending a single-modality task to multi-modality can be very challenging, especially due to the lack of annotations in the new modality. Performing accurate labeling based on multiple modalities requires the annotator to refer to multiple information resources, leading to a high cost of human labor. Although there are some initial explorations on multimodal attribute value extraction [328, 148, 50], all of them are fully supervised and overlook the resource-constrained setting of building a multimodal attribute extraction framework based on the previous textual-established models. In this paper, we aim to patch the visual modality to attribute value extraction by leveraging textual-based models for weak supervision, thus reducing the

manual labeling effort.

Challenges. Several unique challenges exist in visual modality patching: **C1.** Images and their textual descriptions are usually *loosely aligned* in two aspects: From the intra-sample aspect, they are usually weakly related considering the rich characteristics, making it difficult to ground the language fragments to the corresponding image regions; From the inter-samples aspect, it is commonly observed that the text description of one sample may also partially match the image of another. As illustrated in Figure 3.11, the textual description of the *mattress* product is fragmented and can also correspond to other images in the training data. Therefore, traditional training objectives for multimodal learning such as binary matching [125] or contrastive loss [199] that only treat the text and image of the same sample as positive pairs may not be appropriate. **C2.** Bias can be brought by the *visual input* from the *noisy contextual background*. The images usually not only contain the interested object itself but also demonstrate a complex background scene. Although the backgrounds are helpful for scene understanding, they may also introduce spurious correlation in a fine-grained task such as attribute value extraction, which leads to imprecise prediction [275, 115]. **C3.** Bias also exists in *language perspective* regarding the *biased weak labels* from textual-based models. As illustrated in Figure 3.11, the color label of *mattress* is misled by '*green tea infused*' in the text. These noisy labels can be more catastrophic for a multimodal model due to their incorrect grounding in images. Directly training the model with these biased labels can lead to gaps between the stronger language modality and the weaker vision modality [302].

Solutions. We propose PV2TEA, a sequence-to-sequence backbone composed of three modules: visual encoding, cross-modality fusion and grounding, and attribute value generation, each with a bias-reduction scheme dedicated to the above challenges: **S1.** To better integrate the *loosely-aligned texts and images*, we design an augmented label-smoothed contrast schema for cross-modality fusion and ground-

ing, which considers both the intra-sample weak correlation and the inter-sample potential alignment, encouraging knowledge transfer from the strong textual modality to the weak visual one. **S2.** During the visual encoding, we equip PV2TEA with an attention-pruning mechanism that adaptively distinguishes the distracting background and *attends to the most relevant regions* given the entire input image, aiming to improve precision in the fine-grained task of attribute extraction. **S3.** To mitigate the bias from *textual-biased weak labels*, a two-level neighborhood regularization based on visual features and previous predictions, is designed to emphasize trustworthy training samples while mitigating the influence of textual-biased labels. In this way, the model learns to generate more balanced results rather than being dominated by one modality of information. In summary, the main contributions of PV2TEA are three-fold:

- We propose PV2TEA, an encoder-decoder framework effectively patching up visual modality to textual-established attribute value extraction.
- We identify three unique challenges in patching visual modality for information extraction, with solutions for *intra-sample and inter-samples loose alignment* and bias from *complex visual background* and *textual-biased labels*.
- We release three human-annotated datasets with modality source labels of the gold values to facilitate fine-grained evaluation. Extensive results validate the effectiveness of PV2TEA.

3.3.2 Preliminaries

Problem Definition

We consider the task of automatic attribute extraction from multimodal input, i.e., textual descriptions and images. Formally, the input is a query attribute $\mathcal{R}$ and a text-image pairs dataset $\mathcal{D} = \{\mathcal{X}_n\}_{n=1}^{N} = \{(\mathcal{I}_n, \mathcal{T}_n, c_n)\}_{n=1}^{N}$ consisting of N samples (e.g., products), where $\mathcal{I}_n$ represents the profile image of $\mathcal{X}_n$, $\mathcal{T}_n$ represents the textual

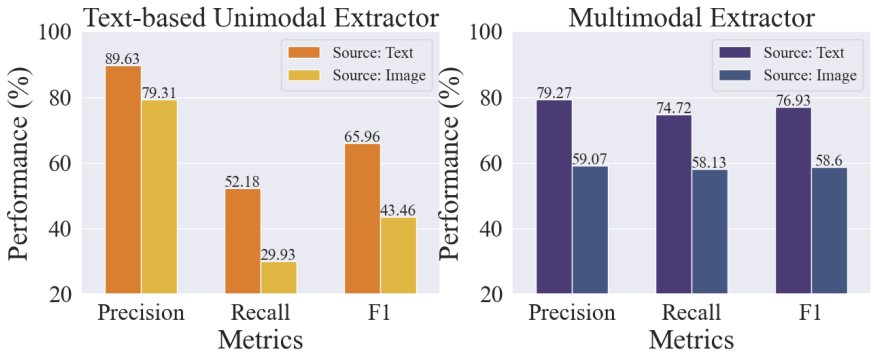


Figure 3.12: Source-aware evaluation of existing unimodal and multimodal models on the textual-biased issue.

description and c_n is the sample category (e.g., product type). The model is expected to infer attribute value y_n of the query attribute $\mathcal{R}$ for sample $\mathcal{X}_n$. We consider the challenging setting with open-vocabulary attributes, where the number of candidate values is extensive and y_n can contain either single or multiple values.

Motivating Analysis on the Modality Bias towards Text

Existing textual-based models or multimodal models directly trained with weak labels suffer from a strong bias toward the texts. As illustrated in Figure 3.11, the training label for the *color* attribute of the *mattress* is misled by '*green tea infused*' from the textual profile. Models trained with such textual-shifted labels will result in a learning ability gap between modalities, where the model learns better from the textual than the visual modality. To quantitatively study the learning bias, we conduct fine-grained source-aware evaluations on a real-world e-Commerce dataset with representative unimodal and multimodal methods, namely OpenTag [322] with the classification setup and PAM [148]. Specifically, for each sample in the test set, we collect the source of the gold value (i.e., text or image). Experiment results are shown in Figure 3.12, where label *Source: Text* indicates the gold value is present in the text, while label *Source: Image* indicates the gold value is absent from the text and must be inferred from the image. It is shown that both the text-based unimodal extractor and multimodal extractor achieve impressive results when the gold value is contained

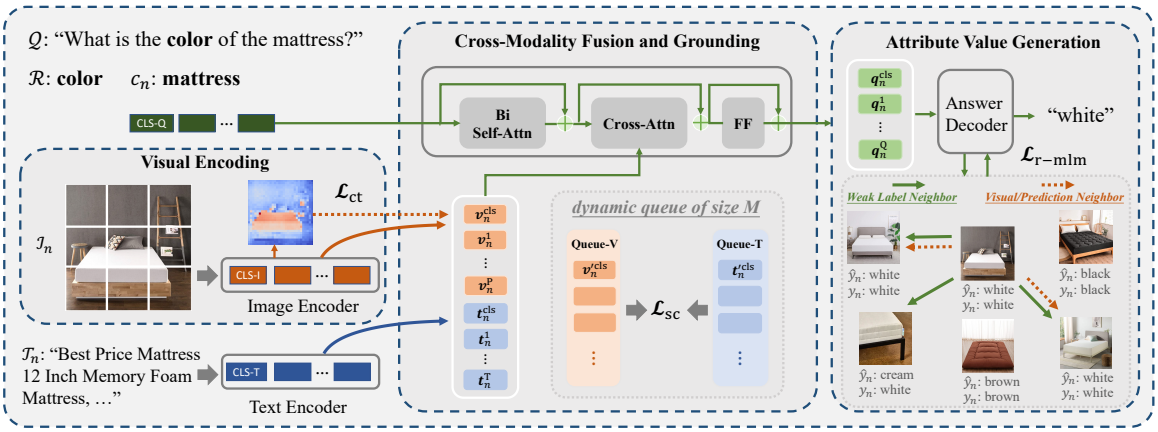


Figure 3.13: The overview of PV2TEA model architecture with three modules, where each of them is equipped with a bias reduction scheme corresponding to the discussed challenges in Figure 3.11.

in the text. However, when the gold value is not contained in the text and must be derived from visual input, the performance of all three metrics drops dramatically, indicating a strong textual bias and dependence of existing models.

3.3.3 Patching Visual Modality to Textual-Established Multimodal Information Extraction

We present the backbone architecture and three bias reduction designs of PV2TEA, shown in Figure 3.16. The backbone is formulated based on visual question answering (VQA) composed of three modules:

◇ *Visual Encoding* We adopt the Vision Transformer (ViT) [61] as the visual encoder. The given product image $\mathcal{I}_n$ is divided into patches and featured as a sequence of tokens, with a special token `[CLS-I]` appended at the head of the sequence, whose representation $\boldsymbol{v}_n^{\text{cls}}$ stands for the whole input image $\mathcal{I}_n$.

◇ *Cross-Modality Fusion and Grounding* Following the VQA paradigm, we define the question prompt as "*What is the* $\mathcal{R}$ *of the* c_n?", with a special token `[CLS-Q]` appended at the beginning. A unimodal BERT [55] encoder is adopted to produce token-wise textual representation from product profiles (title, bullets, and descriptions). The visual representations of P image patches $\boldsymbol{v}_n = [\boldsymbol{v}_n^{\text{cls}}, \boldsymbol{v}_n^1, \ldots, \boldsymbol{v}_n^{\text{P}}]$ are

concatenated with the textual representation of T tokens $\boldsymbol{t}_n = [\boldsymbol{t}_n^{\text{cls}}, \boldsymbol{t}_n^1, \ldots, \boldsymbol{t}_n^{\text{T}}]$, which is further used to perform cross-modality fusion and grounding with the question prompt through cross-attention. The output $\boldsymbol{q}_n = [\boldsymbol{q}_n^{\text{cls}}, \boldsymbol{q}_n^1, \ldots, \boldsymbol{q}_n^{\text{Q}}]$ is then used as the grounded representation for the answer decoder.

◇ *Attribute Value Generation* We follow the design from [139], where each block of the decoder is composed of a causal self-attention layer, a cross-attention layer, and a feed-forward network. The decoder takes the grounded multimodal representation as input and predicts the attribute value $\hat{y}_n$ in a generative manner.

◇ *Training Objectives* The overall training objective of PV2TEA is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{sc}} + \mathcal{L}_{\text{pt}} + \mathcal{L}_{\text{r-mlm}}, \tag{3.9}$$

where the three loss terms, namely augmented label-smoothed contrastive loss $\mathcal{L}_{\text{sc}}$ (Section 3.3.3), product type aware ViT loss $\mathcal{L}_{\text{pt}}$ (Section 3.3.3), and neighborhood-regularized mask language modeling loss $\mathcal{L}_{\text{r-mlm}}$ (Section 3.3.3) correspond to each of the three prementioned modules respectively.

Augmented Label-Smoothed Contrast Loss

Contrastive objectives have been proven effective in multimodal pre-training [199] by minimizing the representation distance between different modalities of the same data point while keeping those of different samples away. However, for product attribute extraction, the image and textual descriptions of products are typically *loosely aligned* from two perspectives: (1) *Intra-sample weak alignment*: The product description usually does not form a coherent and complete sentence, but a set of semantic fragments describing multiple facets of the product. Thus, grounding the language to corresponding visual regions is difficult. (2) *Potential inter-samples alignment*: Due to the commonality of products, the textual description of one product may also correspond to the image of another. Thus, traditional binary matching and contrastive objectives become suboptimal for these loosely-aligned texts and images.

To handle the looseness of product images and texts, we augment the contrast to include sample comparison outside the batch with two queues storing the most recent M ($M \gg$ batch size B) visual and textual representations, inspired by the momentum contrast in MoCo [91] and ALBEF [138]. For the *intra-sample weak alignment* of each given sample $\mathcal{X}_n$, instead of using the one-hot pairing label $\mathbf{p}_n^{i2t}$, we smooth the pairing target with the pseudo-similarity $\mathbf{q}_n^{i2t}$,

$$\widetilde{\mathbf{p}}_n^{i2t} = (1-\alpha)\mathbf{p}_n^{i2t} + \alpha\mathbf{q}_n^{i2t}, \tag{3.10}$$

where α is a hyper-parameter and $\mathbf{q}_n^{i2t}$ is calculated by softmax over the representation multiplication of the `[CLS]` tokens, $\boldsymbol{v}_n^{'\text{cls}}$ and $\boldsymbol{t}_n^{'\text{cls}}$, from momentum unimodal encoders ${\mathcal{F}'}_v$ and ${\mathcal{F}'}_t$,

$$\mathbf{q}_n^{i2t} = \sigma\left({\mathcal{F}'}_v\left(\mathcal{I}_n\right)^\top {\mathcal{F}'}_t\left(\mathcal{T}_n\right)\right) = \sigma\left({\boldsymbol{v}_n^{'\text{cls}}}^\top \boldsymbol{t}_n^{'\text{cls}}\right). \tag{3.11}$$

For *potential inter-samples product pairing relations*, the visual representation $\boldsymbol{v}_n^{'\text{cls}}$ is compared with all textual representations $\boldsymbol{T}'$ in the queue to augment contrastive loss. Formally, the predicted image-to-text matching probability of $\mathcal{X}_n$ is

$$\mathbf{d}_n^{i2t} = \frac{\exp\left({\boldsymbol{v}_n^{'\text{cls}}}^\top \boldsymbol{T}'_m/\tau\right)}{\sum_{m=1}^{M}\exp\left({\boldsymbol{v}_n^{'\text{cls}}}^\top \boldsymbol{T}'_m/\tau\right)}. \tag{3.12}$$

With the smoothed targets from Equation (3.10), the *image-to-text* contrastive loss L_{i2t} is calculated as the cross-entropy between the smoothed targets $\widetilde{\mathbf{p}}_n^{i2t}$ and contrast-augmented predictions $\mathbf{d}_n^{i2t}$,

$$L_{i2t} = -\frac{1}{N}\left(\sum_{n=1}^{N}\widetilde{\mathbf{p}}_n^{i2t}\cdot\log\left(\mathbf{d}_n^{i2t}\right)\right), \tag{3.13}$$

and vise versa for the *text-to-image* contrastive loss L_{t2i}. Finally, the augmented label-smoothed contrastive loss L_{sc} is the average of these two terms,

$$L_{\text{sc}} = \left(L_{i2t} + L_{t2i}\right)/2. \tag{3.14}$$

Visual Attention Pruning

Product images on e-Commerce services usually contain not only the product itself but also rich background contexts. Although previous studies indicate context can

serve as an effective cue for visual understanding [60, 316, 275], it has been found that the output of ViT is often based on supportive signals in the background rather than the actual object [28]. Especially in a fine-grained task such as product attribute value extraction, the associated backgrounds could distract the visual model and harm the prediction precision. For example, when predicting the color of *birthday balloons*, commonly co-occurring contexts such as *flowers* could mislead the model and result in wrongly predicted values.

To encourage the ViT encoder $\mathcal{F}$ focus on task-relevant foregrounds given the input image $\mathcal{I}_n$, we add a product type aware attention pruning schema, supervised with product type classification,

$$L_{\text{pt}} = -\frac{1}{N}\left(\sum_{n=1}^{N} c_n \cdot \log\left(\mathcal{F}(\mathcal{I}_n)\right)\right). \tag{3.15}$$

The learned attention mask $\boldsymbol{M}$ in ViT can gradually resemble the product boundary and distinguishes the most important task-related regions from backgrounds by assigning different attention weights to the image patches [213]. The learned $\boldsymbol{M}$ is then applied on the visual representation sequences $\boldsymbol{v}_n$ of the whole image,

$$\boldsymbol{v}_n^{pt} = \boldsymbol{v}_n \odot \sigma(\boldsymbol{M}), \tag{3.16}$$

to screen out noisy background and task-irrelevant patches before concatenating with the textual representation $\boldsymbol{t}_n$ for further cross-modal grounding.

Neighborhood-regularized Sample Weight Adjustment

Weak labels from established models can be noisy and biased toward the textual input. Directly training the models with these labels leads to a learning gap across modalities. Prior work on self-training shows that embedding similarity can help to mitigate the label errors issue [330, 134]. Inspired by this line of work, we design a two-level neighborhood-regularized sample weight adjustment. In each iteration, sample weight $s\left(\mathcal{X}_n\right)$ is updated based on its label reliability, which is then applied

to the training objective of attribute value generation in the next iteration,

$$\mathcal{L}_{\text{r-mlm}} = -\frac{1}{N}\left(\sum_{n=1}^{N} s\left(\mathcal{X}_n\right) \cdot g\left(y_n, \hat{y}_n\right)\right), \tag{3.17}$$

where g measures the element-wise cross entropy between the training label y_n and the prediction $\hat{y}_n$. As illustrated by the right example in Figure 3.16[6], where green arrows point to samples with the same training label as y_n, and red arrows point to either visual or prediction neighbors, a higher consistency between the two sets indicates a higher reliability of y_n, formally explained as below:

◇ *Visual Neighbor Regularization* The first level of regularization is based on the consistency between the sample set with the same training label y_n and visual feature neighbors of $\mathcal{X}_n$. For each sample $\mathcal{X}_n$ with visual representation $\boldsymbol{v}_n$, we adopt the K-nearest neighbors (KNN) algorithm to find its neighbor samples in the visual feature space:

$$\mathcal{N}_n = \left\{\mathcal{X}_n \cup \mathcal{X}_k \in \text{KNN}\left(\boldsymbol{v}_n, \mathcal{D}, K\right)\right\}, \tag{3.18}$$

where $\text{KNN}\left(\boldsymbol{v}_n, \mathcal{D}, K\right)$ demotes K samples in $\mathcal{D}$ with visual representation nearest to $\boldsymbol{v}_n$. Simultaneously, we obtain the set of samples in $\mathcal{D}$ with the same training label y_j as that of the sample $\mathcal{X}_n$,

$$\mathcal{Y}_n = \left\{\mathcal{X}_n \cup \mathcal{X}_j \in \mathcal{D}_{y_j = y_n}\right\}. \tag{3.19}$$

The reliability of sample $\mathcal{X}_n$ based on the visual neighborhood regularization is

$$s_v(\mathcal{X}_n) = \left|\mathcal{N}_n \cap \mathcal{Y}_n\right| / K. \tag{3.20}$$

◇ *Prediction Neighbor Regularization* The second level of regularization is based on the consistency between the sample set with the same training label and the prediction neighbors from the previous iteration, which represents the learned multimodal representation. Prediction regularization is further added after E epochs when the model can give relatively confident predictions, ensuring the predicted values are qualified for correcting potential noise. Formally, we obtain the set of samples in $\mathcal{D}$ whose

[6]See Appendix B.7 for additional demo examples.

Attr	# PT	Value Type	# Valid	# Train & Val	# Test
Item Form	14	Single	142	42,911	4,165
Color	255	Multiple	24	106,176	3,777
Pattern	31	Single	30	119,622	2,093

Table 3.10: Statistics of the attribute extraction datasets.

predicted attribute value p_j from the last iteration is the same as that of the sample $\mathcal{X}_n$,

$$\hat{\mathcal{Y}}_n = \left\{ \mathcal{X}_n \cup \mathcal{X}_j \in \mathcal{D}_{\hat{y}_j = \hat{y}_n} \right\}. \tag{3.21}$$

With the truth-value consensus set $\mathcal{Y}_n$ from Equation (3.19), the reliability based on previous prediction neighbor regularization of the sample $\mathcal{X}_n$ is

$$s_p\left(\mathcal{X}_n\right) = \left| \hat{\mathcal{Y}}_n \cap \mathcal{Y}_n \right| / \left| \hat{\mathcal{Y}}_n \cup \mathcal{Y}_n \right|. \tag{3.22}$$

Overall, $s(\mathcal{X}_n)$ is initially regularized with visual neighbors and jointly with prediction neighbors after E epochs when the model predicts credibly,

$$s\left(\mathcal{X}_n\right) = \begin{cases} s_v\left(\mathcal{X}_n\right) & e < E, \\ \mathrm{AVG}\left(s_v\left(\mathcal{X}_n\right), s_p\left(\mathcal{X}_n\right)\right) & e \geq E. \end{cases} \tag{3.23}$$

3.3.4 Experimental Setup

Dataset and Implementation Details

We build three multimodal attribute value extraction datasets by collecting profiles (title, bullets, and descriptions) and images from the public `amazon.com` web pages, where each dataset corresponds to one attribute $\mathcal{R}$. The dataset information is summarized in Table A.1, where **Attr** is the attribute $\mathcal{R}$, **# PT** represents the number of unique categories (i.e., product types), **Value Type** indicates whether y_n contain single or multiple values, and **# Valid** represents the number of valid values. To better reflect real-world scenarios, we use the attribute-value pairs from the product information section on web pages as weak training labels instead of highly processed data. We follow the same filtering strategy from prior text established work [308] to denoise training data. For the testing, we manually annotate gold labels on the

benchmark dataset to ensure preciseness. Besides, the label sources are marked down, indicating whether the attribute value is present or absent in the text, to facilitate fine-grained source-aware evaluation. The human-annotated benchmark datasets will be released to encourage the future development of modality-balanced multimodal extraction models.

Evaluation Protocol

We use Precision, Recall, and F1 score based on synonym normalized exact string matching. For single value type, an extracted value $\hat{y}_n$ is considered correct when it exactly matches the gold value string y_n. For multiple value type where the gold values for the query attribute $\mathcal{R}$ can contain multiple answers $y_n \in \{y_n^1, \ldots, y_n^m\}$, the extraction is considered correct when all the gold values are matched in the prediction. Macro-aggregation is performed across attribute values to avoid the influence of class imbalance. All reported results are the average of three runs under the best settings.

Baselines

We compare our proposed model with a series of baselines, spanning unimodal-based methods and multimodal-based ones. For unimodal baselines, OpenTag [322] is considered a strong text-based model for attribute extraction. $\text{OpenTag}_{\text{seq}}$ formulates the task as sequence tagging and uses the BiLSTM-CRF architecture with self-attention. $\text{OpenTag}_{\text{cls}}$ replaces the BiLSTM encoder with a transformer encoder and tackles the task as classification. TEA is another text-only unimodal generative model with the same architecture as PV2TEA but without the image patching, which is included to demonstrate the influence of the generation setting. For multimodal baselines, we consider discriminative encoder models, including ViLBERT [163], LXMERT [232] with dual encoders, and UNITER [34] with a joint encoder. We also add generative encoder-decoder models for comparisons. BLIP [139] adopts dual encoders and an

Type	Method	Dataset: Item Form			Dataset: Color			Dataset: Pattern		
		Precision	Recall	F_1	Precision	Recall	F_1	Precision	Recall	F_1
Unimodal	$\text{OpenTag}_{\text{seq}}$	91.37	44.97	60.27	83.94	24.73	38.20	79.65	19.83	31.75
	$\text{OpenTag}_{\text{cls}}$	89.40	51.67	65.49	81.13	28.61	42.30	78.10	24.41	37.19
	TEA	82.71	60.98	70.20	67.58	47.80	55.99	60.87	37.40	46.33
Multimodal	ViLBERT	75.97	65.67	70.45	60.22	51.12	55.30	60.10	40.52	48.40
	LXMERT	75.79	68.72	72.08	60.20	54.26	57.08	60.33	42.28	49.72
	UNITER	76.75	69.10	72.72	61.30	54.69	57.81	62.45	43.38	51.20
	BLIP	78.21	69.25	73.46	62.70	58.23	60.38	58.74	44.01	50.32
	PAM	78.83	74.35	<u>76.52</u>	63.34	60.43	<u>61.85</u>	61.80	44.29	<u>51.60</u>
Ours	PV2TEA w/o S1	80.03	72.49	76.07	71.00	58.41	64.09	60.03	45.59	51.82
	PV2TEA w/o S2	80.48	75.32	77.81	73.77	59.37	65.79	59.01	46.74	52.16
	PV2TEA w/o S3	80.87	72.71	76.57	74.29	59.04	65.79	59.92	44.92	51.35
	PV2TEA	82.46	75.40	**78.77**	77.44	60.19	**67.73**	62.10	46.84	**53.40**

Table 3.11: Performance comparison with different baselines (%). The performance gains over the baselines have passed the t-test with a p-value<0.05. The best performance is in bold, and the second runner baseline is underlined.

image-grounded text decoder. PAM [148] uses a shared encoder and decoder separated by a prefix causal mask.

3.3.5 Experimental Results

Overall Comparison

Table 3.11 shows the performance comparison of different types of extraction methods. It is shown that PV2TEA achieves the best F_1 performance, especially compared to unimodal baselines, demonstrating the advantages of patching visual modality to this text-established task. Comparing the unimodal methods with multimodal ones, textual-only models achieve impressive results on precision while greatly suffering from low recall, which indicates potential information loss when the gold value is not contained in the input text. With the generative setting, TEA sort of mitigates the information loss and improves recall over OpenTag under the tagging and classification settings. Besides, adding visual information can further improve recall, especially for the multi-value attribute *Color*, where multimodal models can even double that of text-only ones. However, the lower precision performance of the multimodal models implies the challenges beneath cross-modality integration. With the three proposed bias-reduction schemes, PV2TEA improves on all three metrics over

Method	Gold Value Source	Precision	Recall	F_1
$\text{OpenTag}_{\text{cls}}$	Text ✓	89.78	52.13	65.96
	Text ✗ Image ✓	78.95	31.25	44.78
	GAP ↓	10.83	20.88	21.18
PAM	Text ✓	79.16	74.53	76.78
	Text ✗ Image ✓	66.67	58.33	62.22
	GAP ↓	12.50	16.20	14.56
PV2TEA	Text ✓	82.64	75.71	79.02
	Text ✗ Image ✓	75.00	62.50	68.18
	GAP ↓	7.64	13.21	**10.84**

Table 3.12: Fine-grained source-aware evaluation of different methods. The *gold value source* indicates whether the gold value is contained in the text, or is not contained in the text and must be inferred from the image.

multimodal baselines and balances precision and recall to a great extent compared with unimodal models. Besides the full PV2TEA, we also include three variants that remove one proposed schema at a time. It shows that the visual attention pruning module mainly helps with precision while the other two benefit both precision and recall, leading to the best F_1 performance when all three schemes are equipped.

Source-Aware Evaluation

To investigate how the modality learning bias is addressed, we conduct fine-grained source-aware evaluation similarly to Section 3.3.2, as shown in Table 3.12 The performance gap between when the gold value is present or absent in the text is significantly reduced by PV2TEA when compared to both unimodal and multimodal representative methods, which suggests a more balanced and generalized capacity of PV2TEA to learn from different modalities. When the gold value is absent in the text, our method outperforms $\text{OpenTag}_{\text{cls}}$ by more than twice as much on recall, and also outperforms on precision under various scenarios compared to the multimodal PAM.

Ablation Studies

◇ *Augmented Label-Smoothed Contrast* We look into the impact of label-smoothed

Method	Single Value Dataset			Multiple Value Dataset		
	P	R	F_1	P	R	F_1
w/o L_{sc}	80.03	72.49	76.07	71.00	58.41	64.09
w/o Smooth	81.42	74.41	77.76	75.06	59.99	66.68
PV2TEA	82.46	75.40	78.77	77.44	60.19	67.73

Table 3.13: Ablation study on the augmented label-smoothed contrast for cross-modality alignment (%).

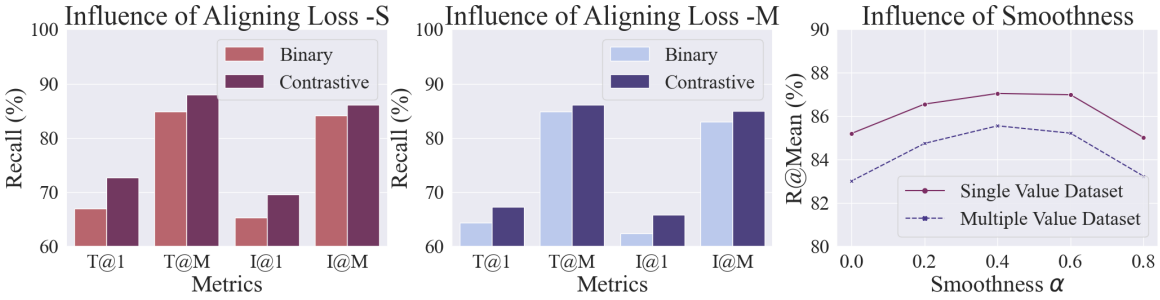


Figure 3.14: The influence study of alignment objectives, i.e., binary matching v.s. contrastive loss, and the influence of softness α via the task of image-to-text and text-to-image retrieval. The metric T/I@1 is the recall of text/image retrieval at rank 1, T/I@M means the rank average, and R@Mean further averages T@M and I@M.

contrast on both single- and multiple-value type datasets [7]. Table 3.13 shows that removing the contrastive objective leads to a drop in both precision and recall. For the multiple-value dataset, adding the contrastive objective significantly benefits precision, suggesting it encourages cross-modal validation when there are multiple valid answers in the visual input. With label smoothing, the recall can be further improved. This indicates that the augmented and smoothed contrast can effectively leverage the cross-modality alignment inter-samples, hence improving the coverage rate when making predictions.

In addition, we conduct cross-modality retrieval to study the efficacy of aligning objectives, i.e., binary matching and contrastive loss, for cross-modality alignment and the influence of the softness α, as shown in Figure 3.14. Across different datasets and metrics, the contrastive loss consistently outperforms the binary matching loss. This consolidates our choice of contrasting objectives and highlights the potential benefits of label-smoothing and contrast augmentation, given that both are neglected

[7]For ablation analysis, we select Item Form as the representative for single-value and Color for multiple-value type dataset. More ablation results can be referred in Appendix B.3.

Method	Single Value Dataset			Multiple Value Dataset		
	P	R	F_1	P	R	F_1
w/o L_{ct}	80.48	75.32	77.81	73.77	59.37	65.79
w/o Attn Prun	80.61	75.49	77.97	74.60	59.42	66.15
PV2TEA	82.46	75.40	78.77	77.44	60.19	67.73

Table 3.14: Ablation study on the category supervised visual attention pruning (%).

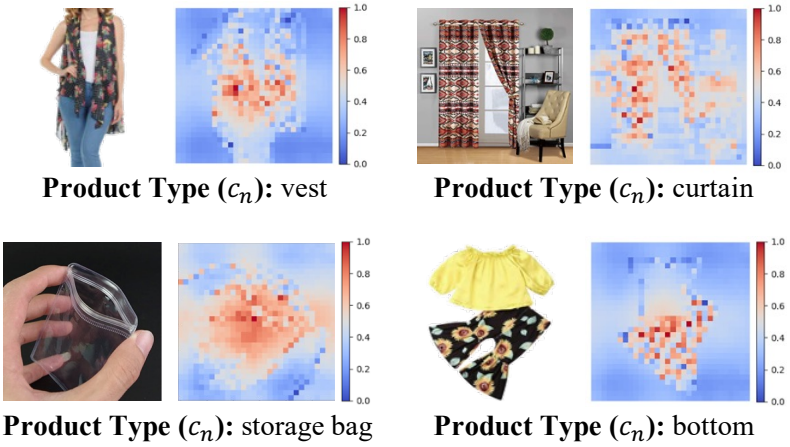


Figure 3.15: Visualization of learned attention mask with category (e.g., product type) aware ViT classification.

in a binary matching objective. Retrieval performance under different smoothness values shows a trend of first rising and then falling. We simply take 0.4 for α in our experiments.

◇ *Category Aware Attention Pruning* We study the influence of the category aware attention pruning, as shown in Table 3.14. The results imply that adding the category classification helps to improve precision performance without harming recall, and the learned attention mask can effectively highlight the foreground regions of the queried sample. Figure 3.15 presents several visualizations of the learned attention mask.

◇ *Neighborhood Regularization*

We consider the influence of the two-level neighborhood regularization by removing the visual neighborhood regularization (Vis-NR), prediction neighborhood regularization (Pred-NR), or both (NR) from the full model. Results in Table 3.15 show all the metrics decrease when both regularizations are removed, indicating the validity of the proposed neighborhood regularized sample weight adjustment in mitigating

Method	Single Value Dataset			Multiple Value Dataset		
	P	R	F_1	P	R	F_1
w/o NR	80.87	72.71	76.57	74.29	59.04	65.79
w/o Vis-NR	81.87	73.54	77.48	77.07	59.99	67.47
w/o Pred-NR	81.81	73.18	77.25	76.71	59.44	66.98
PV2TEA	82.46	75.40	78.77	77.44	60.19	67.73

Table 3.15: Ablation study on the two-level neighborhood-regularized sample weight adjustment (%).

Setting	$\mathcal{D}$: Item Form			$\mathcal{D}$: Color			$\mathcal{D}$: Pattern		
	P	R	F_1	P	R	F_1	P	R	F_1
Classification	79.93	70.47	74.90	72.21	50.18	59.21	59.08	42.16	49.21
Generation	82.46	75.40	78.77	77.44	60.19	67.73	62.10	46.84	53.40

Table 3.16: Attribute extraction performance comparison between the settings of classification and generation.

the influence of hard, noisy samples. Besides, since the second-level prediction-based neighbor regularization is independent of the multimodal extraction framework, it can be incorporated flexibly into other frameworks as well for future usage.

◇ *Classification vs. Generation* To determine which architecture is better for multimodal attribute value extraction, we compare the generation and classification settings for the module of the attribute information extractor. The results are demonstrated in Table 3.16. It is shown that the setting of generation achieves significant advantages over classification. Especially on the recall performance for multi-value type attribute Color, where the gold value can be multiple, the improvement of recall can be up to 20% relatively. This indicates that the generation setting can extract more complete results from the multimodal input, leading to a higher coverage rate. Therefore, we choose the generation setting in the attribute value extraction module in the final architecture design of PV2TEA.

3.3.6 Related Work

Attribute Information Extraction. Attribute extraction has been extensively studied in the literature primarily based on textual input. OpenTag [322] formalizes it as a sequence tagging task and proposes a combined model leveraging bi-

LSTM-CRF, and attention to perform end-to-end tagging. Xu et al. [278] scales the sequence-tagging-based model with a global set of BIO tags. AVEQA [257] develops a question-answering model by treating each attribute as a question and extracting the best answer span from the text. TXtract [119] uses a hierarchical taxonomy of categories and improves value extraction through multi-task learning. AdaTag [286] exploits an adaptive CRF-based decoder to handle multi-attribute value extractions. Additionally, there have been a few attempts at multimodal attribute value extraction. M-JAVE [328] introduces a gated attention layer to combine information from the image and text. PAM [148] proposes a transformer-based sequence-to-sequence generation model for multimodal attribute value extraction. Although the latter two use both visual and textual input, they fail to account for possible modality bias and are fully supervised.

Multi-modality Alignment and Fusion. The goal of multimodal learning is to process and relate information from diverse modalities. CLIP [199] makes a gigantic leap forward in bridging embedding spaces of image and text with contrastive language-image pretraining. ALBEF [138] applies a contrastive loss to align the image and text representation before merging with cross-modal attention, which fits loosely-aligned sample image and text. Using noisy picture alt-text data, ALIGN [108] jointly learns representations applicable to either vision-only or vision-language tasks. The novel Vision-Language Pre-training (VLP) framework established by BLIP [139] is flexibly applied to both vision-language understanding and generation tasks. GLIP [141] offers a grounded language-image paradigm for learning semantically rich visual representations. FLAVA [218] creates a foundational alignment that simultaneously addresses vision, language, and their interconnected multimodality. Flamingo [4] equips the model with in-context few-shot learning capabilities. SimVLM [263] is trained end-to-end with a single prefix language modeling and investigates large-scale weak supervision. Multi-way Transformers are introduced in BEIT-3 [258] for generic mod-

eling and modality-specific encoding.

3.3.7 Conclusion

In this work, we propose PV2TEA, a bias-mitigated visual modality patching-up model for multimodal information extraction. Specifically, we take attribution value extraction as an example for illustration. Results on our released source-aware benchmarks demonstrate remarkable improvements: the augmented label-smoothed contrast promotes a more accurate and complete alignment for loosely related images and texts; the visual attention pruning improves precision by masking out task-irrelevant regions; and the neighborhood-regularized sample weight adjustment reduces textual bias by lowering the influence of noisy samples. We anticipate the investigated challenges and proposed solutions will inspire future scenarios where the task is first established on the text and then expanded to multiple modalities.

3.4 Language Foundation Models with Augmented Inference for EHR-based Disease Prediction

Electronic health records (EHRs) contain valuable patient data for health-related prediction tasks, such as disease prediction. Traditional approaches rely on supervised learning methods that require large labeled datasets, which can be expensive and challenging to obtain. In this study, we investigate the feasibility of applying Large Language Models (LLMs) to convert structured patient visit data (e.g., diagnoses, labs, prescriptions) into natural language narratives. We evaluate the zero-shot and few-shot performance of LLMs using various EHR-prediction-oriented prompting strategies. Furthermore, we propose a novel approach that utilizes LLM agents with different roles: a predictor agent that makes predictions and generates reasoning processes and a critic agent that analyzes incorrect predictions and provides guidance for

improving the reasoning of the predictor agent. Our results demonstrate that with the proposed approach, LLMs can achieve decent few-shot performance compared to traditional supervised learning methods in EHR-based disease predictions, suggesting its potential for health-oriented applications.

3.4.1 Introduction

Large Language Models (LLMs) have emerged as a powerful tool in various domains, including healthcare. These models, such as GPT family [1] and PaLM [5], are trained on vast amounts of text data, allowing them to encode extensive knowledge across multiple fields. In the medical domain, the ability of LLMs to leverage their encoded medical knowledge has been showcased in recent studies [219, 96], with impressive performance on tasks such as medical question answering [220], clinical text summarization [247], and clinical decision support [95]. Certain very large language models demonstrate an emerging ability for few-shot learning, where the model can draw upon their existing understanding to quickly adapt to new tasks with limited examples [19, 211]. This raises the question of whether LLMs can be directly applied to perform few-shot disease predictions using Electronic Health Record (EHR) data.

EHRs contain a wealth of patient data for predictive modeling tasks such as disease prediction, readmission risk assessment, and mortality prediction [216]. Existing approaches to EHR-based prediction primarily rely on supervised learning methods, including traditional machine learning models, representation learning [201, 131, 76], and graph-based models [35]. While effective, these supervised approaches require training on large labeled datasets, which can be computationally expensive and challenging to obtain due to the high cost and difficulty of acquiring high-quality labeled EHR data [274]. In contrast, the capacity for few-shot learning enables LLMs to adapt to new tasks with minimal data, without any finetuning [19]. This adaptability raises the possibility of employing LLMs for few-shot disease prediction using EHR, a step

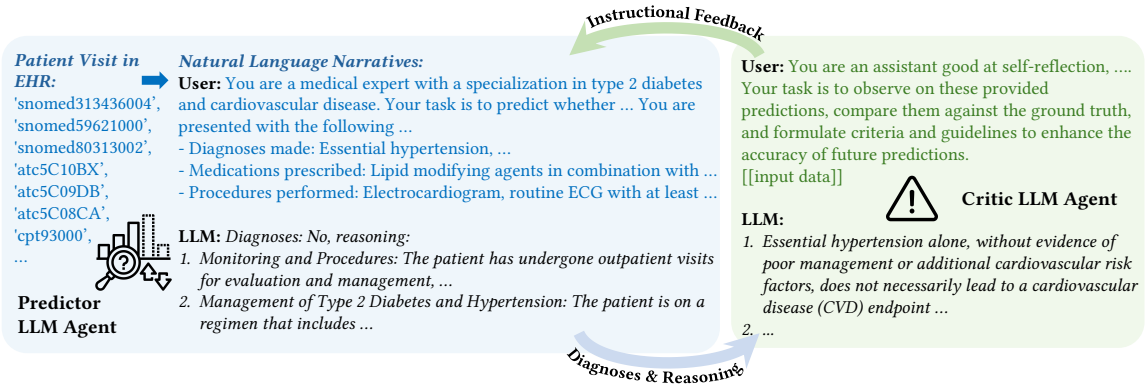


Figure 3.16: The framework of EHR-CoAgent employs two LLM agents: a predictor agent that makes predictions and generates reasoning processes and a critic agent that analyzes incorrect predictions and provides guidance for improvement. The critic agent's feedback is used to update the prompts given to the predictor agent, enabling the system to learn from its mistakes and adapt to the specific challenges of the EHR-based disease prediction task.

forward in making healthcare more precise and efficient [268].

In this study, we investigate the efficacy of LLMs-based few-shot disease prediction using the EHRs generated from clinical encounters that include three types of medical codes: disease, medications, and procedures. We convert the structured patient visit records into unstructured language narratives by mapping the ICD codes to their names and connecting them with proper conjunctives. This conversion process allows LLMs to better understand clinical records and retrieve related internal knowledge. We assess the zero-shot and few-shot diagnostic performance of LLMs using various prompting strategies, such as considering factor interactions and providing prevalence statistics and exemplars. The results of this evaluation provide insights into the potential of LLMs as a tool for EHR-based disease prediction and highlight the influence of prompting strategies on their performance.

Building upon the findings of our initial evaluation, we propose an innovative approach to further improve the few-shot diagnostic performance of LLMs on EHR data. Studies have shown the promise of specialized LLM agents working collaboratively [271, 231, 110], leveraging their diverse functionalities through few-shot learning. Our approach combines the strengths of predictive agent reasoning and critical

agent instruction to create a more robust and accurate prediction system. The overall framework is shown in Figure 3.16. Specifically, we employ two LLM agents with different roles: a *predictor agent* and a *critic agent*. The *predictor agent* makes few-shot predictions given the unstructured narratives, which are converted from structured records, and generates a reasoning process to support its predictions. The *critic agent* then takes the predictor's output alongside the ground-truth disease labels as input and identifies issues or biases in the predictor agent's reasoning process. Based on the analysis, the critic agent generates a set of instructions that draw the predictor agent's attention to potentially overlooked factors and offer specific recommendations for refining its reasoning process. These instructions are subsequently appended to the prompts used for the predictor agent, serving as additional context to inform its predictions. Our results show that by refining the prompts based on the critic agent's feedback, the overall diagnostic accuracy of the LLM-based few-shot prediction system improves significantly. This approach leverages the complementary strengths of predictive reasoning and critical analysis, enabling the system to learn from its mistakes and adapt to the specific challenges of EHR-based disease prediction. In summary, our main contributions are:

- We investigate the application of LLMs to EHR-based disease prediction tasks by converting structured data into natural language narratives and evaluating zero-shot and few-shot performance using various prompting strategies.
- We propose a novel approach combining two LLM agents with different roles: a predictor agent that makes predictions and provides reasoning processes, and a critic agent that analyzes incorrect predictions and provides feedback for improvement. The critic agent's feedback is used to update the predictor agent's prompts, enabling the system to learn from its mistakes and adapt to EHR-based disease prediction challenges.
- We summarize a set of insights into the performance of LLMs under various settings

and share practical guidance on leveraging LLMs for diagnostic tasks with limited labeled data. We hope this can contribute to developing efficient and effective clinical decision support systems in the era of LLMs.

3.4.2 Related Work

Large Language Models for Healthcare

LLMs have demonstrated remarkable capabilities in various application scenarios. Recently, there has been a growing interest in applying LLMs to the medical domain [239, 90, 193], particularly for tasks such as clinical note analysis [2, 172], medical question answering [158, 89], disease prediction [254], clinical trial matching [306], medical report generation [64]. For example, Yang et al. [291] introduced GatorTron, an LLM specifically designed for EHRs. They demonstrated the effectiveness of GatorTron in various clinical natural language processing (NLP) tasks, such as named entity recognition and relation extraction, showcasing the potential of LLMs to extract valuable information from unstructured EHR data. Peng et al. [193] investigated the use of generative LLMs for medical research and healthcare. They explored the capabilities of LLMs in tasks such as medical question answering, disease prediction, and clinical trial matching, highlighting their potential to support clinical decision-making and assist research.

However, applying LLMs to EHR-based disease prediction tasks remains underexplored. While some studies have investigated the use of LLMs for clinical NLP tasks on EHR [291], there is still a lack of research on leveraging the reasoning and instruction-following capabilities of LLMs for few-shot EHR-based prediction. Our research addresses this gap by exploring the use of LLMs for EHR-based disease prediction and proposes new methods to enable accurate prediction with minimal training data.

3.4.3 Method

In this study, we expand our investigations on two levels: (1) evaluating the zero-shot and few-shot performance of LLMs on EHR-based disease prediction tasks, and (2) proposing a novel approach that leverages collaborative LLM agents to enhance the predictive performance.

LLM Performance on Disease Prediction with EHR

The structured patient visit data are typically stored in tabular formats, where each row represents an individual patient visit record generated from clinical encounters, and columns correspond to different medical codes. In this study, we utilize EHR data that includes three types of medical codes $\mathcal{C}$: (1) diseases $\mathcal{C}_D$, (2) medications $\mathcal{C}_M$, and (3) procedures $\mathcal{C}_P$. Each patient visit sample v_i in the record $\mathcal{V}$ is represented by a set of medical codes $\{c_1, c_2, \ldots, c_n\}$, where $c_j \in \mathcal{C}$. We convert the structured EHR records into unstructured language narratives, denoted as $\mathcal{H}$, by mapping the medical codes to their names to enable the application of LLMs.

$\diamond$ Zero-Shot: Leveraging Pre-existing Knowledge Prompt engineering has emerged as a powerful technique for guiding the behavior of LLMs and improving their performance on various healthcare-related tasks, such as clinical named entity recognition [222, 101] and clinical text classification [166, 221]. We develop a set of prompting strategies tailored to EHR-based prediction tasks to provide additional context and guide the reasoning process of LLMs, including:

- Chain-of-thought (CoT) reasoning [265]: prompt the LLMs to generate step-by-step explanations;
- Incorporation of factor interactions: encourage LLMs to consider the interactions and dependencies among different medical factors (e.g., diseases, medications, and procedures);
- Prevalence information: integrate information about the prevalence statistics to

provide additional context.

◇ Few-Shot: Enhancing Performance with Limited Examples We randomly select a small number of positive and negative samples (e.g., 3 positive and 3 negative) from the training data to serve as exemplars for each prediction category. These exemplars are incorporated into the prompts to provide the LLMs with a limited set of task-specific examples to learn from. This leverages the LLMs' vast pre-existing knowledge while allowing them to adapt quickly to the specific characteristics of the EHR prediction task. By this, we aim to guide LLMs' attention toward the most relevant patterns associated with each prediction category.

EHR-CoAgent: Collaborative LLM Agents for Enhanced Prediction

Recently, the potential of LLMs has extended beyond single-agent applications. By leveraging the power of multiple LLMs with different roles working together in a collaborative framework, new possibilities have been unlocked for tackling complex problems and enhancing the performance of language models [271]. In this study, we propose a novel approach called EHR-CoAgent (as demonstrated in Figure 3.16), which harnesses the potential of collaborative LLM agents for enhanced prediction of EHR. Our framework consists of two components: a predictor agent $\mathcal{P}_{\text{LLM}}$ and a critic agent $\mathcal{K}_{\text{LLM}}$. The predictor agent focuses on generating predictions and providing explanatory reasoning, while the critic agent observes the predictor's outputs and provides instructional feedback to refine the prediction process. By integrating the feedback from the critic agent into the prompts used by the predictor agent, we aim to create an in-context learning process with feedback to continuously enhance disease prediction accuracy.

◇ Predictor Agent: Generating Predictions and Reasoning The predictor agent $\mathcal{P}_{\text{LLM}}$ is an LLM that performs few-shot disease predictions and provides explanatory reasoning based on the input EHR data. Given a patient's medical history $\mathcal{H}_i$, the

predictor LLM analyzes the relevant information and generates the most likely prediction $\hat{\mathcal{D}}_i$ and provides a step-by-step explanation of its reasoning process $\mathcal{R}_i$. Such explanatory reasoning is crucial for enhancing the interpretability of the generated predictions. By highlighting the key factors and evidence influencing the LLM agent's decision-making process, the reasoning serves as a transparent and informative basis for further analysis and validation. The detailed prompt we used for the predictor agent in EHR-CoAgent is shown in Figure C.1.

$\diamond$ Critic Agent: Providing Instructional Feedback The critic agent $\mathcal{K}_{\text{agent}}$ is another LLM that plays a different role in the EHR-CoAgent framework by observing a set of sampled wrong predictions from the predictor agent. Each set, denoted as $\mathcal{B}_j = \{(\hat{\mathcal{D}}_{ji}, \mathcal{R}_{ji})\}_{i=1}^{b}$, contains generated prediction $\hat{\mathcal{D}}_{ji}$ and their accompanying explanatory reasoning $\mathcal{R}_{ji}$ for b instances. The critic agent analyzes the inconsistency of the generated prediction to their corresponding ground truth label $\mathcal{D}_{ji}$ for each batch $\mathcal{B}_j$, identifying error patterns for improvement. Based on this analysis, we let the critic agent generate a set of instructional feedback $\{\mathcal{F}_j\}$ for batch $\mathcal{B}_j$ and repeat this process for m times. The detailed prompt we used for the critic agent in EHR-CoAgent is shown in Figure C.2.

To provide concise and coherent guidance, we employ GPT-4 to process the set of instructional feedback $\{\mathcal{F}_j\}_{j=1}^{m}$. GPT-4 analyzes the feedback across multiple batches and generates a consolidated set of instructions $\mathcal{F}_{\text{consolidated}}$ that captures the most important and recurring insights. This consolidated feedback highlights common biases or errors in the reasoning process, offers suggestions for considering additional factors, and provides insights into the relationships between different medical concepts.

$\diamond$ Instruction-Enhanced Prompting: Integrating Feedback for Refinement To effectively incorporate the feedback generated by the critic LLM, we introduce an instruction-enhanced prompting mechanism. This mechanism integrates the critic LLM's instructional feedback $\mathcal{F}_{\text{consolidated}}$ directly into the prompts $\mathcal{P}$ used by the predictor

LLM. By augmenting the prompts with specific instructions and guidance, we aim to steer the predictor LLM's attention toward the most relevant aspects of the input data and encourage it to consider the insights provided by the critic LLM. This iterative process of making predictions, receiving feedback, and refining the prompts allows the predictor LLM to continuously improve its performance and adapt to the specific challenges of EHR-based disease prediction.

3.4.4 Experimental Settings

Datasets

We conducted experiments on two datasets: the publicly accessible MIMIC-III dataset and the privately-owned CRADLE dataset. **MIMIC-III** [111] is a large, publicly accessible dataset comprising de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. Our task is to predict whether acute care conditions will be present during a patient's next visit, given their current ICU stay records. We focus on a specific chronic phenotype, Disorders of Lipid Metabolism, which is identified using Clinical Classifications Software (CCS) from the Healthcare Cost and Utilization Project (HCUP)[8]. During preprocessing, we extract patients with more than one hospital visit and create pairs of adjacent visits for each patient. For each pair, the former visit serves as the input, and the phenotypes in the latter visit are used as labels. This process yields 12,353 records with labels. For budget consideration, we randomly sample 1,000 records based on the data distribution of the prediction target as our testing set.

Project **CRADLE** (Emory Clinical Research Analytics Data Lake Environment) is a privately-owned database that contains de-identified electronic health records at Emory Healthcare from 2013 to 2017. In this study, we focus on the patients with

[8] `https://www.hcup-us.ahrq.gov/toolssoftware/ccs/AppendixASingleDX.txt`

type 2 diabetes and predict whether those patients will experience **cardiovascular disease** (CVD) endpoints within a year after the initial diabetes diagnosis. The CVD endpoints include coronary heart disease (CHD), congestive heart failure (CHF), myocardial infarction (MI), or stroke, which are identified by their ICD-9 and ICD-10 clinical codes. For patients who developed CVD complications within a year (positive cases), we select the earliest recorded encounter within a year of the CVD endpoint presence as the input. For patients without CVD complications (negative cases), we randomly select one encounter as the input from all encounters that occurred at least one year before the last recorded encounter. Patients are excluded if they (1) have less than two encounters at Emory Healthcare, (2) the time interval between their first and last encounter is less than one year, or (3) have a history of CVD conditions. After applying these exclusion criteria, 35,404 patients remain in the dataset. Similar to MIMIC-III, we randomly sample 1,000 records based on the data distribution of the prediction target

Evaluation Metrics

Both the MIMIC-III and CRADLE datasets exhibit class imbalance, with the prevalence of Disorders of Lipid Metabolism in MIMIC-III being 27.6% and the prevalence of cardiovascular disease (CVD) endpoints in CRADLE being 21.4%. To account for the imbalanced data distributions, we employ accuracy, sensitivity, specificity, and F1 score as evaluation metrics [35]. When evaluating LLM methods, we identify the presence of "Yes" or "No" tokens in the LLM responses and extract the top 5 probabilities associated with the predicting token. These probabilities are then normalized over both answers. We observed that GPT family models tend to provide highly confident answers (a confirmed prediction of either "Yes" or "No", with almost 0.0 probability for the other choice), often resulting in a majority probability of either 0.0 or 1.0.

Baselines

We compare the performance of EHR-CoAgent with traditional machine learning (ML), including Decision Trees, Logistic Regression, and Random Forests, which are widely used in EHR-based prediction tasks [269, 82], and single-agent LLM approaches using GPT-4 (`gpt-4-0125-preview`) and GPT-3.5 (`gpt-35-turbo-16k-0613`). The ML models are trained in both fully supervised and few-shot settings, while the LLM approaches are evaluated in pure zero-shot, zero-shot with additional prompt information as mentioned in section 3.4.3, and few-shot learning settings. By comparing EHR-CoAgent with these baselines, we aim to evaluate the effectiveness of diverse LLM agent frameworks in EHR-based disease prediction tasks.

Implementation Details

We implemented the empirical study methods in Python. The baseline machine learning models were trained and evaluated using the popular sklearn package, which provides a comprehensive set of tools for machine learning tasks. To access the various GPT models securely, we utilized the Azure OpenAI Service, a trusted and compliant cloud platform. Azure OpenAI offers a secure API interface that allows seamless integration of the GPT capabilities into our research pipeline while maintaining strict privacy and security controls. By leveraging Azure OpenAI, we ensured that the sensitive patient dataset was processed in a protected environment, adhering to necessary regulations and standards, such as HIPAA and GDPR.

3.4.5 Experimental Results

Table 3.17 presents the experimental results on the two datasets. The findings highlight several key observations:

- Traditional machine learning (ML) models achieve respectable performance when fully trained on large datasets (11,353 samples for MIMIC-III and 34,404 samples

Table 3.17: Performance (%) of different models under the zero-shot, few-shot, and fully-supervised settings on MIMIC-III and CRADLE datasets. The proposed method is colored in green. The reference results under the supervised training setting (trained on 11,353 samples for MIMIC-III and 34,404 samples for CRADLE) are colored in gray.

Type	Model	MIMIC-III (Pos : Neg = 27.6% : 72.4%)				CRADLE (Pos : Neg = 21.4% : 78.6%)			
		ACC	Sensitivity	Specificity	F1	ACC	Sensitivity	Specificity	F1
Fully-Supervised	Decision Tree	81.30	76.97	84.31	76.20	80.30	53.87	88.27	52.15
	Logistic Regression	79.70	70.48	83.56	73.18	80.90	58.34	86.15	59.74
	Random Forest	78.60	66.12	83.16	70.58	80.20	56.49	86.14	57.34
Few-Shot (N=6)	Decision Tree	71.10	53.14	77.62	51.16	31.90	54.81	25.99	31.71
	Logistic Regression	58.70	73.40	53.44	56.78	53.30	53.95	53.13	48.16
	Random Forest	69.70	62.88	72.18	63.61	65.00	51.50	68.43	51.04
GPT-4	Zero-Shot	51.90	76.15	42.56	51.89	24.10	51.81	16.82	22.33
	Zero-Shot+	62.90	59.30	64.29	58.58	30.00	53.25	23.76	29.67
	Few-Shot (N=6)	65.70	79.35	59.89	64.72	41.20	59.05	36.33	40.88
	EHR-CoAgent	79.10	73.11	81.43	73.88	70.00	62.88	71.72	60.21
GPT-3.5	Zero-Shot	78.00	66.87	82.37	68.56	56.50	59.88	55.45	52.29
	Zero-Shot+	72.40	50.00	80.37	42.00	62.60	57.62	63.96	54.40
	Few-Shot (N=6)	76.30	63.73	80.93	63.84	40.80	54.56	36.96	40.32
	EHR-CoAgent	79.30	74.49	80.98	71.59	66.60	58.31	68.83	55.83

for CRADLE). However, the performance of simpler models, such as Decision Trees and Logistic Regression, substantially deteriorates in the few-shot learning setting, emphasizing their limitations when labeled data is scarce.

- When comparing the performance of zero-shot or few-shot LLMs with ML methods under few-shot settings, we observe that LLMs exhibit higher sensitivity but lower specificity. This finding suggests that LLMs excel at correctly identifying positive cases (i.e., patients with the condition of interest) but at the cost of a higher false positive rate. In other words, LLMs are more prone to classifying a patient as having the condition, even when they do not. This tendency implies that LLMs, particularly GPT-4, adopt a more conservative mindset, possibly due to their alignment to err on the side of caution to mitigate the risk of potentially missing true positive cases.
- Zero-shot with additional prompting strategies (Zero-Shot+) can improve based on pure zero-shot, with occasionally produced errors. This observation underscores the importance of carefully crafting prompts to optimize the performance of LLMs in EHR-based disease prediction tasks.

- Most of the time, adding few-shot demonstrations enhance prediction performance compared to their respective Zero-Shot+ counterparts. This finding emphasizes providing even a limited number of labeled examples can potentially steer language models toward more precise predictions. By leveraging a small set of representative samples, LLMs can quickly adapt to the specific characteristics of the EHR-based disease prediction task.
- Our proposed approach EHR-CoAgent demonstrates remarkable performance, surpassing other methods and even fully supervised ML models in certain scenarios, with GPT-4 generally outperforming GPT-3.5. On the CRADLE dataset, EHR-CoAgent achieves an F1 score of 60.21%, outperforming all fully trained ML models. Similarly, on the MIMIC-III dataset, EHR-CoAgent obtains an F1 score of 73.88%, comparable to the fully trained Decision Tree model and superior to Logistic Regression and Random Forest.
- Compared with the few-shot setting with a single LLM predictor, EHR-CoAgent improves significantly on all four metrics. This can be attributed to the feedback instructions provided by the critic agent, which analyzes the outputs and identifies issues and biases in LLM's reasoning process, such as overly relying on conservative thinking or neglecting certain key factors. The feedback instructions generated by the critic agent help to correct these issues, dynamically refining the predictor agent's reasoning process, thus improving the accuracy of the prediction.

3.4.6 Generated Instructions

Figure 3.17 showcases examples of the criteria and instructions generated by the critic agent. These examples demonstrate the critic agent's ability to identify potential issues in the predictor agent's prediction and reasoning process and provide targeted instructions to address them. For instance, the first instruction for the CRADLE dataset, "Avoid bias towards predicting a positive CVD endpoint based

Instructional Feedback Examples from the Critic LLM Agent for the MIMIC dataset (GPT-4)
The diagnosis of conditions directly related to lipid metabolism, such as "Disorders of lipoid metabolism," in the patient's medical history, requires ongoing management and monitoring rather than assumptions of worsening or new diagnoses without recent lipid profile assessments.
Risk factors and underlying conditions: Consider the presence of risk factors and underlying conditions that are commonly associated with disorders of lipid metabolism disease, such as diabetes mellitus, obesity, hypertension, and cardiovascular diseases.
The performance of procedures related to cardiovascular health, such as hemodialysis or cardiac catheterization, without direct evidence of unmanaged lipid metabolism issues, should not be solely used to predict future disorders of lipid metabolism disease.
Pharmacological interventions consideration: Incorporate an evaluation of prescribed drugs, focusing on their relevance to managing the risk factors of the disorders of lipoid metabolism.
Instructional Feedback Examples from the Critic LLM Agent for the CRADLE dataset (GPT-4)
Avoid bias towards predicting a positive CVD endpoint based on conservative thinking when the patient is actively monitored and managed for known risk factors. Evaluate the effectiveness of the interventions in place.
The presence of type 2 diabetes mellitus without complication does not necessarily lead to a cardiovascular disease (CVD) endpoint within a year of the initial diagnosis.
The presence of symptoms such as chest pain, dyspnea, and edema, especially when combined with diagnoses like hypertension and hyperlipidemia, increases the likelihood of developing a cardiovascular disease (CVD) endpoint within a year of the initial diagnosis.
Essential hypertension alone, without evidence of poor management or additional cardiovascular risk factors, does not necessarily lead to a cardiovascular disease (CVD) endpoint within a year of the initial diagnosis.

Figure 3.17: Examples of instructional feedback generated by the GPT-4-based critic agent, which aims to refine the predictor agent's reasoning process and improve the accuracy of its prediction.

on conservative thinking when the patient is actively monitored and managed for known risk factors. Evaluate the effectiveness of the interventions in place" highlights a possible prediction bias of the predictor agent. This instruction encourages the predictor agent to avoid relying on conservative assumptions when making predictions, as such assumptions may be a result of the over-alignment of advanced AI models. By explicitly addressing this issue, the critic agent aims to guide the predictor agent toward more objective and comprehensive reasoning. Another example for the MIMIC dataset, "Pharmacological Interventions Consideration: Incorporate an evaluation of prescribed drugs, focusing on their relevance to managing the risk factors of the disorders of lipoid metabolism" suggests that the predictor agent should take into account the role of prescribed medications in managing the patient's condition. By analyzing the relevance and potential impact of these drugs on the risk factors associated with disorders of lipoid metabolism, the predictor agent can make more informed predictions. These examples illustrate how the critic agent's feedback can guide the predictor agent towards more comprehensive and nuanced reasoning, ultimately leading to improved disease prediction performance.

3.4.7 Conclusions

In this study, we investigated the application of Large Language Models (LLMs) to Electronic Health Record (EHR) based disease prediction tasks. We evaluated the zero-shot and few-shot diagnostic performance of LLMs using various prompting strategies and proposed a novel collaborative approach combining a predictor agent and a critic agent. This approach enables the system to learn from its mistakes and adapt to the challenges of EHR-based disease prediction. Our work highlights the potential of LLMs as a tool for clinical decision support and contributes to the development of efficient disease prediction systems that can operate with minimal training data.

3.4.8 Ethical Considerations

To ensure the ethical use of credential data with GPT-based services, we have signed and strictly adhered to the PhysioNet Credentialed Data Use Agreement[9]. We follow the guidelines[10] for responsible use of MIMIC data in online services, including opting out of human review of the data through the Azure OpenAI Additional Use Case Form[11], to prevent sensitive information from being shared with third parties.

3.5 Multimodal Foundation Models with Augmented Inference for Adapting Generic Models to the Healthcare Domain

Recent advancements in multimodal foundation models have shown remarkable capabilities in understanding and reasoning visual and textual information simultaneously.

[9]https://physionet.org/about/licenses/physionet-credentialed-health-data-license-150
[10]https://physionet.org/news/post/gpt-responsible-use
[11]https://aka.ms/oai/additionalusecase

However, adapting such foundation models to specialized domains like biomedicine remains challenging due to the lack of large-scale, high-quality datasets for model instruction tuning. In this work, we propose a novel framework that combines the scalability of GPT-4V and expert knowledge to create clinician preference-aligned instruction-following datasets for biomedical visual instruction tuning. Specifically, we first curate a set of clinician preference data to guide GPT-4V in generating large-scale, biomedical-specific instruction-following datasets. Then we train a distilled scoring model to further filter out low-quality or irrelevant samples, ensuring the quality and relevance of the data used for the actual instruction-tuning. We perform extensive evaluations with the instruction-tuned model on biomedical tasks, demonstrating our method's potential to advance the state-of-the-art in biomedical multimodal reasoning. Our approach presents a scalable and efficient framework for adapting general-purpose multimodal models to the biomedical domain, aiming to facilitate the development of intelligent systems that can assist healthcare professionals and enhance patient care.

3.5.1 Introduction

Recent advanced large pre-trained multimodal models have achieved impressive performance in applications that require understanding and reasoning about visual and textual information simultaneously. These models, such as CLIP [199], ALIGN [108], LLaVA [152], and MiniGPT4 [326], demonstrate remarkable capabilities in various tasks, including image captioning, visual question answering, image-text retrieval, etc. Also, the recently introduced GPT-4V [1] from OpenAI has shown remarkable zero-shot performance on various visual understanding and reasoning tasks. Although these models have demonstrated great potential in general domains, their successful adaptation to specialized domains, such as biomedicine, is still not well-developed. Adapting these robust foundation models to biomedicine can potentially have signif-

icant benefits for many real-world applications. For instance, it could enable precise medical image analysis, facilitate more effective patient communication, and support clinical decision-making processes.

Instruction tuning has emerged as a promising approach for large pre-trained models to perform specific tasks by providing them with explicit, natural language instructions. It is a powerful technique where only one single model is trained in a multi-task manner with specified instructions, leading to the model being natural and easy to generalize to new tasks in zero-shot settings. This approach has been successfully applied to models like GPT-3 [19] and InstructGPT [189], enabling them to follow instructions and generate human-like responses. However, instruction tuning for the biomedical domain is not a straightforward task. Creating large-scale instructional datasets in the biomedical domain can be expensive and time-consuming, often requiring the expertise of medical professionals.

Recently, self-instruct tuning has been proposed to further enhance the efficiency and scalability of instruction tuning, where large language models are leveraged to generate instruction-following data with their strong generative capabilities. For example, Stanford Alpaca [236] achieves impressive performance with self-instructed data, demonstrating the potential to reduce the reliance on manually annotated datasets for effective model instruction-tuning.

However, although large pre-trained models bring the scalability and efficiency needed to generate large amounts of data for model adaptation, directly applying self-instruct tuning in the biomedical domain has its limitations. The generated instruction-following data may contain errors, inconsistencies, or irrelevant information, which can be particularly problematic in the critical context of healthcare. Furthermore, the rigorous nature of biomedical knowledge and the complexity of medical reasoning require the involvement of expertise to ensure the validity and relevance of the generated instructions for effective instruction tuning and alignment with real

practice.

In this paper, we propose a novel approach that leverages GPT-4V's generative capabilities and clinician expertise to create high-quality, instruction-following datasets aligned with clinician preferences, which leads to effective visual instruction tuning for the biomedical domain. As is shown in Figure 3.18, the proposed framework consists of three main steps: (1) clinician-guided multimodal instruction-following data generation using GPT-4V, (2) data filtering with a distilled scoring model to ensure data quality and relevance, and (3) instruction tuning to adapt a general-domain multimodal model to the medical domain using the filtered dataset. The proposed framework enables the efficient adaptation of general-purpose multimodal models to acquire domain-specific reasoning abilities in the biomedical field through self-instruction tuning. The main contributions of this work are as follows:

- We introduce an integrated framework for biomedical visual instruction tuning with clinician preference alignment. Specifically, the framework comprises clinician preference-guided data generation, a distilled scoring model from clinicians and GPT-4V rating for data filtering, and instruction tuning with the generated data to adapt large multimodal pre-trained models for biomedicine.
- We curate a set of clinician preference data and rating factors, which reflect the nuanced differences in real-world biomedical practice. These serve as few-shot demonstrations or foundations to guide large-scale, clinical-aligned instruction-following dataset generation.
- We evaluate various difficulty levels of biomedical tasks with our instruction-tuned model. The results showcase that by effectively harnessing clinician expertise, we offer a scalable and efficient solution for creating high-quality, domain-specific instructional datasets that can effectively improve the reasoning abilities of multimodal models in the biomedical domain.

3.5.2 Background

Instruction-Tuning and Self-Instruct

Instruction tuning has emerged as a powerful technique for adapting pre-trained language models to perform specific tasks by providing them with task-specific instructions and examples. The concept of instruction tuning was first introduced by OpenAI's InstructGPT [189], which demonstrated the ability to follow instructions and generate human-like responses. This approach has been further explored and refined in subsequent studies, such as FLAN-T5 [37], LLaMA [240], and LLaMA2 [241], indicating the effectiveness of instruction tuning in various domains and tasks. Instruction tuning leverages the knowledge and capabilities of large pre-trained language models by fine-tuning them on instruction-following examples. This enables the model to understand and follow task-specific instructions, allowing for flexibly solving new tasks without extensive task-specific fine-tuning.

Recently, a new approach called *self-instruct* has been proposed to further enhance the efficiency and scalability of instruction tuning. Self-instruct leverages the generative capabilities of large language models to generate their own instruction-following data [262]. By prompting the language model with a set of seed instructions and examples, the model generates additional instruction-following pairs, effectively expanding the training dataset. This self-generated data can then be used to fine-tune the language model, further improving its ability to follow instructions and perform various tasks. Self-instruct has shown promising results in reducing the reliance on manually annotated instructional datasets and enabling more efficient adaptation of language models to new domains and tasks. For example, Stanford Alpaca [236] fine-tuned the large language model LLaMA using self-generated instructions and achieved competitive performance on various natural language processing tasks. In our work, we aim to take advantage of the self-instruct technique to generate scal-

able, clinician-preference-aligned instruction-following datasets for biomedical visual instruction tuning.

Vision-Language Foundation Models in Biomedical Domain

Vision-language foundation models, such as CLIP [199], ALIGN [108], and LLaVA [152], have achieved remarkable success in understanding and generating multimodal content across various domains. These models are trained on large-scale datasets of image-text pairs, learning to align visual and textual representations in a shared embedding space. By capturing the semantic relationships between images and text, these models can perform tasks that require multimodal reasoning, such as image captioning, visual question answering, and image-text retrieval.

In the biomedical domain, researchers have been actively exploring the adaptation of vision-language foundation models to tackle domain-specific tasks. These efforts have yielded promising results for various applications. For instance, Huang et al.[104] developed a pathology visual-language foundation model that achieved state-of-the-art performance in classifying new pathology images and retrieving similar cases using either image or natural language search. Similarly, Elias et al.[17] recently demonstrated that an AI model outperformed radiologists in detecting cardiac pathology, highlighting the potential of vision-language models in medical diagnosis.

However, adapting vision-language foundation models to the biomedical domain presents challenges, particularly in terms of limited training data due to the high cost of annotation. To this end, our work aims to overcome this challenge by efficiently infusing domain expertise into a self-instructing tuning framework. In particular, we build upon the generative capabilities of large pre-trained vision-language foundation models and establish a practical visual instruction-tuning framework that can capture the nuanced real-world clinician preference for biomedical tasks.

3.5.3 Clinician-Aligned Biomedical Multimodality Instruction Tuning Model

The goal of this work is to instruction-tune a biomedical model $\mathcal{M}$ with a set of instruction-following dataset $\mathcal{D} = \{(I_i, C_i, \mathbf{Q}_i, \mathbf{A}_i)\}_{i=1}^{N}$, where I_i represents the i-th biomedical image; C_i represents the caption and inline-mentions associated with the i-th image; $\mathbf{Q}_i = \{\mathcal{Q}_j\}_{j=1}^{n_i}$ represents the set of n_i instructional questions for the i-th image-text sample, where each question asks to analyze a specific aspect of the information presented in the given image and text; $\mathbf{A}_i = \{\mathcal{A}_j\}_{j=1}^{n_i}$ represents the set of n_i answers corresponding to the questions in $\mathbf{Q}_i$; N is the total number of samples in the dataset.

Figure 3.18 presents an overview of the proposed framework for self-instruction tuning biomedical multimodal models with clinician preference alignment. The framework consists of three main steps: (1) data generation with diverse expert-selected demonstration, (2) data filtering with a distilled scoring model, and (3) instruction tuning to adapt a general-domain multimodal model to the biomedical domain. We describe the technique designs of each step in the following sections.

Step 1: Data Generation with Diverse Expert-Selected Demonstration

Large pre-trained models have shown strong in-context learning capabilities by learning from a few presented examples and mimicking the behavior when generating responses. To guide the instruction-following data generation process effectively with clinician expertise, we first select a diverse set of samples for clinician annotation. The annotated samples are then used as few-shot demonstrations for the GPT-4V (vision) model to generate an instruction-following dataset at scale.

◇ *Diverse Few-Shot Demonstration Selection* To ensure the diversity and representatives of the few-shot demonstrations used for model prompting, we employ a strategic sampling approach, which starts by extracting image and text representations

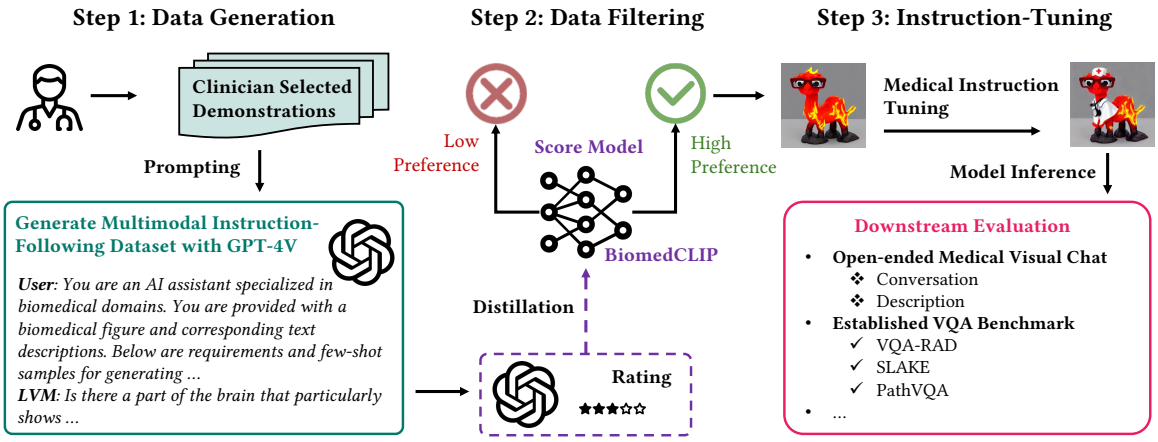


Figure 3.18: Overview of our proposed framework for biomedical visual instruction tuning with clinician preference alignment. The framework consists of three main steps: (1) clinician-guided multimodal instruction-following data generation, (2) data filtering with a distilled scoring model to ensure data quality and relevance, and (3) visual instruction tuning to adapt a general-domain pre-trained model to biomedical with the filtered dataset with high preference.

for each sample (I_i, C_i) in the dataset $\mathcal{D}$ using BiomedCLIP [317]. K-means clustering is then performed on these representations to categorize the samples into K distinct categories, denoted as $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_K$. From these categories, we manually select M samples with complex captions and inline mentions, resulting in a subset $S = (I_m, C_m)_{m=1}^{M}$. For each selected sample $(I_m, C_m) \in S$, we use GPT-4V (`gpt-4-vision-preview`) to generate a set of instructional questions $\mathbf{Q}_m = \{\mathcal{Q}_j\}_{j=1}^{n_m}$, and two candidate answers A_j^1, A_j^2 for each question $\mathcal{Q}_j \in \mathbf{Q}_m$.

Clinicians are then presented with the image I_m, the corresponding caption and inline mentions C_m, and are asked to choose the better answer A_j^{pref} between the two candidate answers A_j^1, A_j^2 for each question $\mathcal{Q}_j \in \mathbf{Q}_m$, select both if the answers are equally good, or deselect both to drop this question. This preference selection process is performed for all questions in $\mathbf{Q}_m$ and for all samples in S. The resulting demonstration set $\mathcal{D}_{pref} = \left\{(I_m, C_m, \mathbf{Q}_m, \mathbf{A}_m^{pref})\right\}_{m=1}^{M}$ contains the image-text pairs and clinician-preferred instructional QA pairs. This strategic sampling and preference annotation approach offers several advantages: (1) it ensures diversity in the few-shot demonstrations by selecting samples from distinct categories, (2) it emphasizes

informative samples within each category, thereby providing a representative set of demonstrations, and (3) it incorporates expert knowledge through clinician preference annotation, which guides the large pre-trained models towards generating clinically relevant instructional dataset.

◇ *Instruction-Following Data Generation with GPT-4V(ision)* Building upon the diverse clinician-selected demonstrations in $\mathcal{D}_{pref}$, we employ the GPT-4V(ision) model to generate a large-scale instruction-following dataset $\mathcal{D}_{gen}$. GPT-4V is a state-of-the-art foundation model that has shown remarkable performance in understanding language and vision contexts and exhibits the emerging capability of in-context learning by observing just a few demonstration examples.

To generate $\mathcal{D}_{gen}$, we make API calls to `gpt-4-vision-preview` by including a set of randomly selected samples from $\mathcal{D}_{pref}$ as demonstrations in the prompts of each call. The GPT-4V model then generates a large number of instruction-following examples based on these prompts. The detailed prompt used for the pairwise win-rate evaluation is shown in Figure D.1 of Appendix D.1. The generated dataset is denoted as $\mathcal{D}_{gen} = \{(I_i, C_i, \mathbf{Q}_i, \mathbf{A}_i)\}_{i=1}^{N}$, where I_i represents the image, C_i is the image caption, $\mathbf{Q}_i$ denotes the question, $\mathbf{A}_i$ is the answer, and N represents the total number of generated samples. In our experiments, we generate datasets with samples of $N = 10,000$ and $N = 60,000$ for model instruction tuning. By providing such clinician-selected few-shot demonstrations, we further align the generation process with expert preference while harnessing the generative capabilities of GPT-4V(vision), leading to a large-scale biomedical instruction-following dataset that enables instruction tuning.

Step 2: Data Filtering with Distilled Scoring Model

While $\mathcal{D}_{gen}$ contains a vast amount of valuable biomedical information, it may also still include low-quality samples that can introduce noise, bias, or inconsistencies in the training data, leading to suboptimal instruction tuning and potentially harm

the model's ability to provide accurate and reliable biomedical insights. This step aims to address this challenge and ensure the quality of the instruction-tuning data by developing a model that can automatically assess the quality and relevance of the generated data. Specifically, we introduce a scoring model to serve as the data filter. This scoring model helps to identify and remove low-quality samples from the generated dataset, preserving only the most relevant, accurate, and clinically meaningful examples for instruction tuning.

◇ *Rating Collection through Clinician-Guided GPT-4V Self-Evaluation* To curate the rating data for the scoring model training, we employ GPT-4V to perform self-evaluation on the generated instructional-following datasets $\mathcal{D}_{gen}$ under the guidance of clinician-curated rating factors. First, we collect a set of clinician-curated factors for data quality rating, such as missing information, recognition errors, lack of medical precision, insufficient depth, valueless questions, etc. With these criteria set, we prompt GPT-4V to assign a score $\mathcal{R}_i$ on a scale of 0 to 10 for each question-answer pair in $\mathcal{D}_i$ to reflect the overall quality of the sample. Such self-evaluation harmonizes the advanced language understanding of large models with the provided clinician guidelines to provide an overall quality assessment of the generated question-answer pairs effectively. The resulting self-evaluated ratings, $\mathcal{R}_i$, serve as valuable references for training the scoring model, enabling it to distinguish the patterns and characteristics of high-quality and low-quality samples effectively.

◇ *Scoring Model Distillation* Directly employing clinician expertise or GPT-4V for data filtering can both be expensive. To create an effective and scalable scoring model for data filtering, we propose distilling the rating ability of GPT-4V and clinician experts into a local scoring model. Specifically, BiomedCLIP is used as the backbone, followed by an MLP head to construct the scoring model. With the curated rating data obtained from GPT-4V's self-evaluation, the local scoring model is trained to mimic the rating ability of GPT-4V and human experts. This distillation

process essentially allows knowledge to be transferred from large pre-trained models and human experts into a local model with limited annotated data, enabling it to assess the quality of biomedical data samples at scale.

A pairwise ranking task is modeled to train the scoring model: given a pair of candidate samples x_i and x_j, along with their corresponding rating scores $\mathcal{R}_i$ and $\mathcal{R}_j$ from GPT-4V annotation under clinician-provided factors, the training objective is formulated as a pairwise classification loss:

$$\mathcal{L}_Q = -z_i \log \sigma\left(f(x_i)\right) - z_j \log \sigma\left(f(x_j)\right),$$

where σ represents the sigmoid function, and $f(\cdot)$ denotes the scoring function learned by the model. The values of z_i and z_j are determined by comparing the rating scores:

$$(z_i, z_j) = \begin{cases} (1,0), & \mathcal{R}_i \geq \mathcal{R}_j \\ (0,1), & \mathcal{R}_i < \mathcal{R}_j \end{cases}.$$

By minimizing the pairwise classification loss, the scoring model learns to assign higher scores to samples with higher GPT-4V ratings and lower scores to samples with lower ratings. This training process enables the scoring model to capture the patterns and characteristics of high-quality and low-quality biomedical data samples as determined by GPT-4V under clinician guidance. This approach not only improves the efficiency of the data filtering process but also allows for flexibility and customization in adapting the scoring model to the specific requirements of biomedical applications.

◇ *Applying the Scoring Model for Data Filtering* Once the scoring model is trained, we apply it to the entire generated dataset $\mathcal{D}_{gen}$, where the scoring model assigns a quality score to each sample in the dataset. We then draw the Precision@K curve to determine the threshold for data filtering, which is selected based on the massive performance drop @K. After the data filtering, we obtain a high-quality preferred dataset with relatively high ratings, denoted as $\mathcal{D}_{filtered}$, which contains the most informative, accurate, and clinically relevant examples. This process enhances the

overall quality of the data used for real instruction tuning, potentially leading to improved performance and reliability of the final instruction-tuned model and providing more accurate and useful insights for real-world applications.

Step 3: Instruction-Tuning

Finally, we perform visual instruction tuning with $\mathcal{D}_{filtered}$ based on a general-domain multimodal conversation model LLaVA [152, 151] as the initial model, which has shown impressive performance in understanding and generating multimodal content. Similar to LLaVA-Med [137], we continue training the LLaVA model on our curated instruction-following dataset $\mathcal{D}_{filtered}$, which contains the high-preference biomedical image-text pairs, instructional questions, and clinician-preference-aligned answers from the proposed pipeline. The instruction tuning objective is to minimize the negative log-likelihood of the target $\mathbf{A}_i$ given the corresponding image I_i, caption C_i, and question $\mathbf{Q}_i$:

$$\mathcal{L}_{IT} = - \sum_{i=1}^{|\mathcal{D}_{filtered}|} \log p(\mathbf{A}_i | I_i, C_i, \mathbf{Q}_i, \theta),$$

where θ represents the parameters of the LLaVA model. By minimizing this objective, the model learns to generate answers that align with the preferences of clinicians, as captured in the filtered dataset $\mathcal{D}_{filtered}$.

3.5.4 Evaluation Plan and Preliminary Results

To evaluate the effectiveness of the proposed method, we perform evaluations in two scenarios: (1) open-ended biomedical visual chat, which indicates the vision-language understanding and generation of the instruction-tuned model, and (2) performance benchmark of the instruct-tuned model on multiple standard visual question-answering datasets spanning various image modalities. These evaluations provide a comprehensive understanding of the proposed framework's potential to advance biomedical

multimodal reasoning aligned with clinician expertise.

Scenario 1: Open-ended Medical Visual Chat

To evaluate the open-ended multimodal understanding and generation ability of different models, we adopt the multi-round visual chat task, where the trained language models (LMMs) are prompted to answer several questions (potentially with progressive relations) given input contexts of images and texts. This evaluation scenario assesses the model's ability to engage in dialogue-like interactions and provide accurate, relevant, and coherent responses based on the given visual and textual information.

◇ *Dataset and Evaluation Paradigm* For the evaluation dataset, we use 50 unseen image and caption pairs with 193 question-answer pairs collected by the LLaVA-Med [137] authors. The dataset is designed to assess the model's ability to engage in multi-round visual chat, where the model is prompted to answer several questions based on the given input context, including images and texts. These data correspond to five domains, including CXR, MRI, Histology, Gross, and CT. The questions are categorized into two types: (1) Conversation questions: These questions require the model to engage in a dialogue-like interaction, where the model needs to understand the context of the conversation and provide relevant and coherent responses. For example, given an image of a chest X-ray, a conversation question might ask, "What abnormalities do you see in this X-ray image?" (2) Description questions: These questions focus on eliciting detailed descriptions or explanations from the model based on the provided visual and textual input. For instance, a description question for a histology image could be, "Describe the morphological features of the cells in this histology slide."

To evaluate the quality of the model's responses, we leverage powerful large language models (LLMs) and large multimodal models (LVMs) as evaluators. These evaluators first generate a reference prediction based on the input context and the

Table 3.18: Preliminary performance comparison of the instruction-tuned models on open-ended biomedical visual chat. We utilize the relative score with two large model evaluators, namely GPT-3.5 and GPT-4V. The number followed by "#: " represents the number of testing samples in this category.

Evaluator	Model	Question Types		Domains					Overall
		Conversation (#:143)	Description (#: 50)	CXR (#: 37)	MRI (#: 38)	Histology (#: 44)	Gross (#: 34)	CT (#: 40)	(#: 193)
GPT-3.5	LLaVA-Med	34.11	24.93	42.67	23.83	32.58	27.93	31.41	31.73
	LLaVA-Med-VGen	**35.44**	**32.50**	**43.82**	**26.87**	**41.21**	**28.39**	**31.79**	**34.68**
GPT-4V	LLaVA-Med	54.69	54.40	47.65	48.35	46.42	82.52	52.35	54.61
	LLaVA-Med-VGen	**57.28**	**66.70**	**49.20**	46.00	**62.14**	**98.30**	47.00	**59.71**

given question. Then, they assess the response provided by the trained model by assigning a relative score on a scale from 1 to 10. A higher score indicates that the model's response is more accurate, relevant, and coherent with respect to the reference prediction from the evaluator. By using different large models as evaluators, we can obtain a more reliable and multi-faceted assessment of the model's performance. This evaluation assesses the framework's effectiveness in enhancing model capabilities to engage in open-ended biomedical visual chat.

◇ *Preliminary Results* We conducted a preliminary evaluation on a baseline model (LLaVA-Med-VGen), which is instruction-tuned on a 10,000-sample GPT-4V generated instruction-following dataset, by comparing it with the LLaVA-Med model trained on 10,000 samples. The preliminary results are shown in Table 3.18. The model trained with the generated instructional-following dataset, which is presented with additional visual input during generation, demonstrates improvements over both types of questions. When using GPT-3.5 as the evaluator, LLaVA-Med-VGen surpasses LLaVA-Med in all domains, indicating the strength of its reasoning ability in capturing textual input. When using GPT-4V as the evaluator, where the reference answer captures visual and textual information, LLaVA-Med-VGen still outperforms the majority of domains with a significant advantage.

Interestingly, we found that for the small performance gap on the MRI and CT domains, both appear only once in the as few0shot demonstrations for instruction-

Table 3.19: The dataset statistics of the three established biomed-multimodal datasets.

Dataset	VQA-RAD		SLAKE			PathVQA		
	Train	Test	Train	Val	Test	Train	Val	Test
# Images	313	203	450	96	96	2,599	858	858
# QA Pairs	1,797	451	4,919	1,053	1,061	19,755	6,279	6,761
# Open	770	179	2,976	631	645	9,949	3,144	3,370
# Closed	1,027	272	1,943	422	416	9,806	3,135	3,391

following dataset generation, compared to two appearances for the other domains. We conjecture that this discrepancy is due to insufficient in-context learning for the two domains with fewer guidance samples.

To further investigate the effectiveness of our proposed components, we plan to conduct additional experiments comparing the performance of models trained on datasets generated with and without clinician-preference-aligned demonstrations, as well as datasets filtered using the distilled-scoring model. These experiments will provide valuable insights into the impact of these components on the model's ability to understand and generate clinically relevant content in an open-ended setting.

Scenario 2: Performance on Established Benchmarks

◇ *Datasets* We evaluate the performance of our instruct-tuned models on three biomedical multimodal benchmark datasets, which have been used by both LLaVA-Med and Med-PaLM M [243], two representative state-of-the-art multimodal foundation models in the biomedical domain. Table 3.19 presents the detailed statistics for each benchmark dataset.

- VQA-RAD [135] is a dataset containing 3,515 question-answer pairs created by medical professionals, along with 315 radiology images. Each image is linked to several questions, which are categorized into 11 types, including abnormality, attribute, modality, organ system, color, counting, object/condition presence, size, plane, positional reasoning, and others. The dataset features a balanced mix of closed-ended (yes/no) and open-ended (one-word or short phrase) answers.

- SLAKE [150] is a comprehensive medical visual question-answering dataset with knowledge-enhancement features. It contains radiology images and diverse question-answer pairs annotated by experienced physicians. The dataset incorporates external medical knowledge through a provided medical knowledge graph, and the images are supplemented with rich visual annotations, including semantic segmentation masks and object detection bounding boxes. SLAKE covers a wide range of modalities and human body parts, such as the brain, neck, chest, abdomen, and pelvic cavity. We adopt only the English subset of SLAKE in our experiments.
- PathVQA [94] focuses on pathology images. Each image is associated with multiple questions that cover various aspects, such as location, shape, color, and appearance. The questions in PathVQA include open-ended questions (e.g., why, what, how, where) and closed-ended questions.

◇ *Evaluation Paradigm* To assess model performance in both closed-ended and open-ended question-answering tasks, we adopted a reference-guided pairwise win-rate evaluation paradigm. Specifically, we employ GPT-4V as an impartial judge to assess the quality of the responses provided by two AI assistants to a given question, considering both the generated responses and the reference answer. GPT-4V compares the responses from the two models and determines which model provides a more accurate, relevant, and coherent answer. The win rate is then calculated based on the number of times each model's response is preferred by the GPT-4V judge. The prompt used for the pairwise win-rate evaluation is shown in Figure D.2 of Appendix D.2. Such evaluation eliminates the limitations of token-level matching metrics, which may not adequately capture the diversity and nuance of language narratives. We would also include the metric performance by reporting accuracy for binary classification questions and recall (defined as the ratio of ground-truth tokens appearing in the generated sequences) for non-binary questions.

◇ *Preliminary Results* We conducted a preliminary evaluation to compare the perfor-

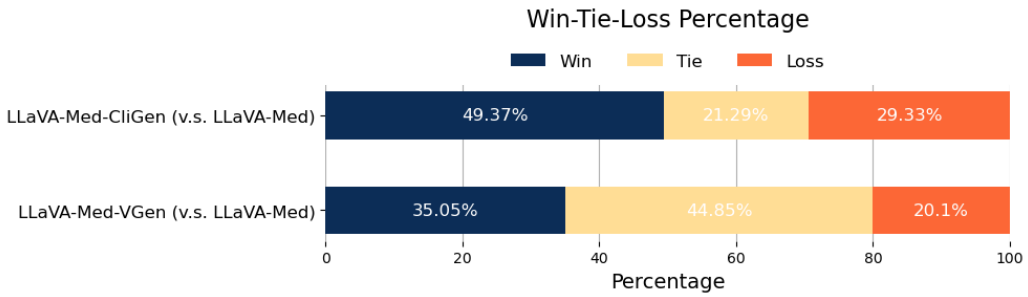


Figure 3.19: Preliminary win-rate evaluation comparing the performance of LLaVA-Med-VGen and LLaVA-Med-CliGen, which are instruction-tuned on generated datasets, against the LLaVA-Med model trained on the original dataset. The results demonstrate the effectiveness of our proposed approach in generating high-quality instruction-following data for biomedical multimodal reasoning.

mance of two of our instructional-tuned models against the baseline LLaVA-Med model. The first model, LLaVA-Med-VGen, was instruction-tuned on a 10,000-sample dataset generated by GPT-4V without any specific guidance. The second model, LLaVA-Med-CliGen, was instruction-tuned on a 10,000-sample dataset generated by GPT-4V under the guidance of clinician preferences. The win-rate comparison results, presented in Figure 3.19, demonstrate that both LLaVA-Med-VGen and LLaVA-Med-CliGen outperform the baseline LLaVA-Med model and achieve higher win rates. Note that LLaVA-Med-VGen, which is instruction-tuned on a dataset generated by GPT-4V without any specific guidance, still demonstrates advantages over LLaVA-Med. This can be attributed to the fact that LLaVA-Med-VGen better incorporates visual information during the data generation process, leading to improved multimodal reasoning capabilities. LLaVA-Med-CliGen, which utilizes the dataset generated with clinician preference guidance, outperforms both LLaVA-Med-VGen and the baseline LLaVA-Med. These results highlight the effectiveness of the proposed method in generating high-quality instruction-following data for biomedical visual instruction tuning. By incorporating clinician preferences during the data generation process, the proposed framework further enhances the biomedical multimodal reasoning abilities of the trained models, ensuring that they learn to align with expert knowledge and clinical preference.

3.5.5 Further Plans

To further validate the effectiveness of the proposed approach and investigate the impact of its key components, we plan to conduct a comprehensive evaluation focusing on the following aspects: (1) Effectiveness of the data filtering with the distilled scoring model: We plan to assess the impact of the distilled scoring model by comparing instruction-tuned models with filtered and unfiltered datasets. This determines the extent to which the scoring model helps remove low-quality or irrelevant examples, leading to improved model reasoning capabilities. (2) Influence of instructional data size: We plan to investigate the performance of models tuned on different sizes of instruction-following datasets (e.g., 10,000 and 60,000 samples) to understand the influence of data size, where the scalability is also a potential key benefit of our proposed framework in advancing the state-of-the-art in biomedical multimodal reasoning. (3) Additional evaluation scenarios: We plan to expand a more complex task, radiology report generation, to reflect the instruction-tuned model's capability for biomedical applications that require long contexts. By further including these evaluation perspectives, we aim to provide insights into the strengths and limitations of the proposed framework and its potential to enhance biomedical multimodal reasoning.

3.5.6 Conclusion

In this work, we propose a novel approach for biomedical visual instruction tuning with clinician preference alignment. We leverage GPT-4V and clinician expertise to create large-scale, high-quality instruction-following datasets tailored to the biomedical domain. The techniques involve curating clinician preference data to guide GPT-4V in generating biomedical-specific datasets and developing a distilled scoring model to filter out low-quality or irrelevant examples. Extensive evaluations on open-ended biomedical visual chat and established benchmarks demonstrate the effectiveness of our approach in advancing biomedical multimodal reasoning. Future work will focus

on further validating the performance of models trained using our generated dataset across a wider range of biomedical tasks and exploring the application of the proposed approach to other domain-specific use cases. We believe our work contributes towards harnessing the power of vision-language foundation models for biomedical applications, with the goal of advancing medical research and clinical practice.

Chapter 4

Conclusion

4.1 Summary of Research Contributions

The rapid advancement of artificial intelligence (AI) has unlocked new opportunities for specialized domains such as healthcare. However, the data heterogeneity and complexity, spanning scientific literature, clinical texts, multi-modality imaging, and electronic health records, pose significant challenges in extracting useful knowledge and leveraging AI models effectively for decision-making in such critical areas.

This thesis has systematically addressed these challenges by focusing on two key themes: multimodal structured knowledge extraction and augmented inference. By developing techniques to integrate knowledge from diverse data sources and pre-trained models, the thesis lays the foundation for comprehensive data understanding in specialized domains. The proposed augmented inference methods, categorized into augmentation through model input, augmentation through the model training process, and augmentation through output feedback, have demonstrated the potential to improve the domain-specific reasoning capabilities and reliability of AI models.

The effectiveness of the core idea "extract-then-augment" has been showcased in various applications, including brain analysis, scientific literature understanding,

visual reasoning, disease prediction, and biomedical reasoning. In summary, the integration of multimodal structured knowledge extraction and augmented inference, as explored in this thesis, opens up exciting opportunities for building AI systems that are accurate, reliable, and grounded in domain expertise. Such systems have the potential to improve decision-making processes in fields like healthcare, leading to improved outcomes and more reliable, data-driven insights.

4.2 Future Work

The ideas on multimodal knowledge extraction and augmented inference presented in this dissertation open up several new opportunities for future research. We highlight three particularly promising directions: (1) *discovering unknown knowledge from known data*, (2) *grounding foundation model evaluation and alignment with domain knowledge*, and (3) *augmenting reasoning through human-AI collaboration.*

4.2.1 Discovering Unknown Knowledge from Known Data

Despite the shortage of labeled data for model training in specialized domains, there is an abundance of available known data left underexplored. The key challenge lies in identifying the relevant data from massive resources and determining how to effectively utilize them for new discoveries. To uncover unknown knowledge from extensive heterogeneous data, it is beneficial to develop new techniques that can identify the most informative data subsets and devise strategies for their optimal use in knowledge discovery. This also emphasizes the significance of studying open knowledge extraction in the scientific domain to improve our understanding and interpretation of diseases, biomedical signals, etc. Such advancements can lead to improved disease treatment, informed policy-making, and accelerated scientific discovery.

4.2.2 Grounding Foundation Model Evaluation and Alignment with Domain Knowledge

Applying foundation models in healthcare requires clinical-grounded evaluation and alignment. Therefore, it is crucial to establish comprehensive evaluation benchmarks for LLMs and LVMs, such as curating new benchmark datasets, standardizing evaluation settings, and automating model evaluation. Domain knowledge can serve as the grounding for designing novel evaluation paradigms and new domain-specific evaluation metrics in such processes. Evaluation frameworks can also derive more meaningful dimensions for evaluation in healthcare settings by incorporating established medical knowledge such as clinical relevance, guidelines, and experiences. Moreover, structured domain knowledge can be leveraged in model training to mitigate factual errors and reduce AI models' hallucinations, thereby improving the factuality of generative AI models.

4.2.3 Augmenting Reasoning by Human-AI Collaboration

While AI models excel in scalability and breadth of knowledge, augmenting complex reasoning through human-AI collaboration is crucial for ensuring their validity and relevance for real-world scenarios. Humans may not necessarily trust all the outputs from AI systems. Therefore, better protocols for human-AI collaboration, especially for complex reasoning scenarios, must be investigated. For example, active learning and human-in-the-loop approaches can be explored to efficiently incorporate domain expertise into the model learning or evaluation processes. With expert feedback and assessment of model outputs, AI models can gather valuable insights for refinement and optimization. This also helps to strike a balance between AI's scalability and the reliable knowledge of domain experts, ultimately leading to improved reliability and trustworthiness of AI models for complex reasoning tasks.

Appendix A

Additional Information for Chapter 3.2

A.1 Details of Data Augmentation with External Knowledge Resources

✔ *Enhance Relation Recognition*: We enriched the relationships between objects parsed from the original knowledge descriptions by leveraging the external resource of ConceptNet. ConceptNet comprises commonly observed entities and their connections, where edge weights signify the reliability and frequency of these relationships. The typical value of edge weights in ConceptNet is 1. To prevent the redundancy of common information and to maintain the validity of the enriched relations, we categorized the relationships based on their weights. Relationships with weights less than 1 were deemed "weak" and those with a weight of 1 were labeled "average". We refrained from using these categories for relation enhancement. Instead, only relationships with weights greater than 1, indicative of high reliability, were employed for augmenting the relations.

✔ *Boost Entity Perception*: On the entity side, we augment complement entities and

descriptive information with two external knowledge resources. On one hand, for descriptions with a high TF-IDF+ score, we enrich related entities of the object from ConceptNet to create additional knowledge descriptions. The relatedness is based on the between-word relatedness score provided by ConceptNet and we take the threshold as 0.85. On the other hand, we employ the Commonsense Transformers (COMET) [18] model to enrich related new objects and descriptive information. The COMET model is a language model designed to generate commonsense knowledge and understand causal relationships between descriptions. It is pretrained using the atomic dataset, which consists of structured, crowd-sourced knowledge about everyday events and their associated causes and effects. The COMET model can provide neighbor descriptions of the given input of nine different categories of relation. We take the `xAttr` and `oEffect` relation categories and augmented the COMET model by formulating the existing knowledge description texts as the input and choose the corresponding category branch during generation for enriching objects and descriptions respectively.

A.2 Dataset Information

Table A.1: Dataset statistics.

split	**#image**	**#descriptor**	**#relation**	**#subject & object**
Train	75,456	832,351	30,241	302,735
Validation	4,871	64,137	5,164	34,177
Test	4,873	62,579	5,031	32,384

The statistic information of our augmented dataset is summarized in Table A.1, where **split** specifies the dataset split, **#image** indicates the number of images in the split, **#descriptor** indicates the total number of relational descriptors of the images, **#relation** is the total number of unique relations in the relational descriptors after deduplication, and **#subject & object** is the total number of subjects and objects contained in the description text.

A.3 Implementation Details

Hyperparameter	Assignment
batch size	4
learning rate optimizer	Adam
Adam epsilon	1e-8
Adam initial learning rate	1e-5
learning rate scheduler	cosine scheduler
Adam decay weight	0.05

Table A.2: Hyperparameters for training open relational region detector.

Hyperparameter	Assignment
batch size	4
learning rate optimizer	Adam
Adam epsilon	1e-8
Adam initial learning rate	1e-5
learning rate scheduler	cosine scheduler
Adam decay weight	0.05
α	0.7
ϕ	0.01

Table A.3: Hyperparameters for training format-free visual knowledge generator.

Open relational region detector. The visual feature extraction backbone is constructed upon a pre-trained ResNet50-FPN. The detector head incorporates a $\text{BLIP}_{\text{base}}$ equipped with the essential ViT-B/16 for text supervision, using multiple fully connected layers to derive region features. For each candidate region, we engage a regressor to conduct boundary regression on these features. The detector undergoes fine-tuning for 20 epochs using the relational region bounding box dataset and an Adam optimizer [161]. The hyperparameters for training are detailed in Table A.2.

Format-free visual knowledge generator. The format-free visual knowledge generator is initialized from $\text{BLIP}_{\text{base}}$, which incorporates the basic ViT-B/16. We fine-tune the generator model for 20 epochs using the same optimizer as the one employed for the region detector. Detailed hyperparameters for the visual knowledge generator can be found in Table A.3.

A.4 Human Evaluation Guidance and Interface

We perform the human evaluation on two of the four in-depth knowledge quality assessment metrics. We build an interface by referring to [266], where raters are

presented with a given image and the corresponding knowledge descriptions and are required to choose one from the multiple choice for two questions on whether the knowledge is valid to humans and whether the knowledge description depicts the image. The detailed scoring criteria for *Validity* and *Conformity* are provided below:

- *Validity* (↑): *whether the generated visual knowledge is valid to humans.*
  - 0 (Invalid): The knowledge description does not conform to human cognition, rendering it unreliable or misleading to humans.
  - 1 (Valid): The knowledge description is valid and accurately conforms to human cognition, providing reliable and meaningful knowledge to humans.
- *Conformity* (↑): *whether the generated knowledge faithfully depicts the scenarios in the images.*
  - 0 (Inconsistent): The knowledge description does not faithfully depict the scenarios in the images, showing significant deviations or discrepancies, making it difficult for users to relate the textual information to the visual context.
  - 1 (Partially Conforming): The knowledge description partially conforms to the scenarios in the images, but there might be minor inconsistencies or missing relevant details.
  - 2 (Moderately Conforming): The knowledge description exhibits a moderate level of conformity with the scenarios in the images, capturing the key aspects and providing coherent descriptions.
  - 3 (Highly Conforming): The knowledge description highly conforms to the scenarios in the images, accurately capturing the details and faithfully representing the visual context.

Agreement/validation We use Cohen's κ as the agreement score to measure potential subjectivity involved in ratings of knowledge quality. Cohen's κ is a statistic that is used to measure inter-rater reliability for qualitative items and is scaled from

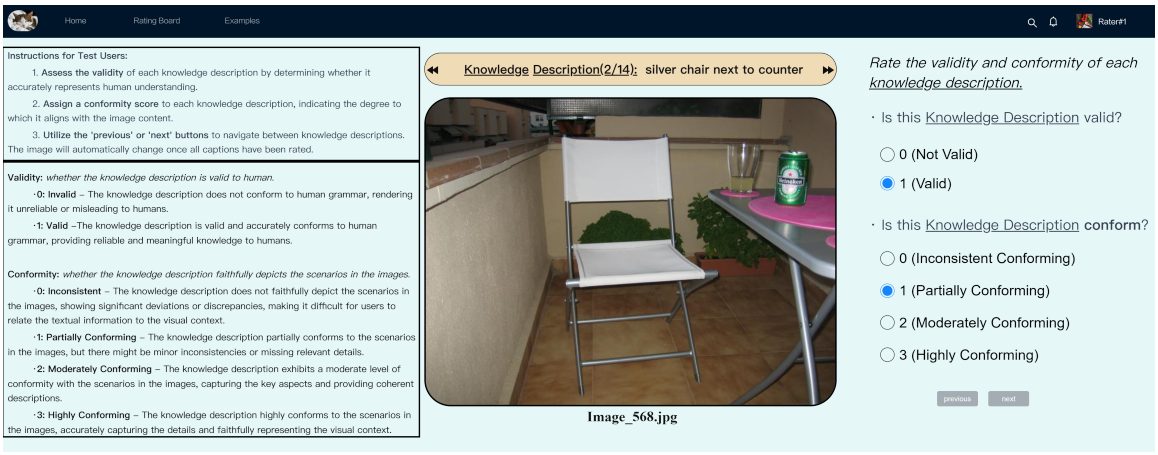


Figure A.1: The human evaluation interface for in-depth knowledge quality evaluation.

-1 (perfect systematic disagreement) to 1 (perfect agreement), where values ≤ 0 as indicating *no agreement* and 0.01-0.20 as *none to slight*, 0.21-0.40 as *fair*, 0.41–0.60 as *moderate*, 0.61-0.80 as *substantial*, and 0.81-1.00 as *almost perfect* agreement. Our calculated average pairwise Cohen's κ on human evaluation results from three different raters is 0.76, which indicates a good agreement.

A.5 Parametric Knowledge Prompting Template

Given an image $\mathcal{I}$ and the corresponding extracted visual knowledge from it based on `OpenVik`, we perform knowledge comparison with parametric knowledge contained in LLM by prompting the gpt-3.5-turbo model with the object information contained in the $\mathcal{I}$. The prompt format is shown in the followings:

```
Suppose you are looking at an image that contains the
following subject and object entities:
Subject list:  [Insert the subject names here]
Object list:  [Insert the object names here]
Please extract 5-10 condensed descriptions that describe
the interactions and/or relations among those entities
in the image.  Try to elucidate the associations and
relationships with diverse language formats instead of
being restricted to sub-verb-obj tuples.
```

A.6 More Case Studies of Open Visual Knowledge from OpenVik

Figure A.2 shows some other cases on the extracted open visual knowledge from OpenVik. In comparison to VG and Relational Caps, OpenVik exhibits superior performance at capturing novel entities, expanding object interactions through diverse relations, and enriching knowledge representation with nuanced descriptive details. For example for the bottom right image, OpenVik can extract novel entities such as "*tracks*", "*shoe*", diverse relations such as "*sticking out of*", and nuanced descriptive details such as "*cold thick*", "*with man feet on it*", "*brave*". The generated knowledge with a more format-free semantic structure is highlighted in red.

A.7 More Qualitative Examples on Applications

A.7.1 Text-to-Image Retrieval

Figure A.3 presents more qualitative examples of OpenVik-based visual knowledge enrichment on captions. The enriched text is based on the objects present in the

Figure A.2: Case studies of open visual knowledge from `OpenVik`.

images themselves, supplemented with additional relationships from our generated visual knowledge in `OpenVik`. It is shown that the introduced relationships often provide new context information that aligns with the visual content of the images. For example, in the image of an old woman sitting on a bench in a park, the enriched context information includes the positional relationship between the "*bench*", "*fence*", and "*park*", which provides a more comprehensive description of the original image.

A.7.2 Grounded Situation Recognition

Figure A.4 presents more qualitative examples of `OpenVik`-based context enrichment in the grounded situation recognition (GSR) task. Our context enrichment setting for the GSR task is to perform enrichment based on verbs like "*shopping*" and "*carrying*". We further restrict the enriched context with the objects contained in the image to avoid noisy enrichment. For example, for the image showing people shopping at a

Original text: Three young men playing Wii on a projection television. Three men laughing at some pictures from a projector. A group of gentleman playing video games in a dimly lit room. Some people chilling on the couch playing with a Nintendo Wii. A group of men playing a game with remote controllers.

Enriched text: men in group. men behind people. men playing. men in room playing video game. group of people. men in group are playing video game. people playing. people watching game. playing game.

Original text: A row of parked motorcycles sitting in front of a tall building. A stone street with bicycles and motor bikes parked on the side and people standing on the sidewalks in front of buildings. Cityscape of pedestrians enjoying an old European city. a row of bikes and mopeds is parked along the street. Motorcycles and mopeds line a side street during the day in a city.

Enriched text: row made of stone leading into city. motor in row. row of people. street made of stone. wall made of stone next to side. stone wall behind people. people in line crossing street. street in city. motor on side. people riding motor in city. motor in line. people in line in city. day at city.

Original text: An elderly woman sitting on the bench resting. An old woman leans on her back while sitting on an ornate bench. A woman is sitting on a bench near a fence. Older woman in dress sitting on a park bench. An old woman sitting on a bench next to a fence.

Enriched text: woman sitting on bench with a ornate. woman behind fence. woman wearing dress. woman in park. bench by fence. bench in park. woman in ornate dress on the bench. fence behind park.

Original text: A herd of cattle is feeding at the river's edge. Many cows next to a body of water in a field. A herd of cows grazes in a field near a river. A herd of cattle standing in grassy area next to water. A herd of cattle is near a flock of birds swimming in the water.

Enriched text: herd of cattle crossing river. herd traveling by water. cattle crossing river. cattle in field. river across field in front of area. water near field. water near area. water next to flock. Birds inside of water. flock in field.

Original text: A man is leaning over a fence offering food to an elephant. A man reaching out to an elephants trunk near a gate. A man is feeding an elephant over a fence. A man handing an elephant a stick in an enclosure at a zoo. A man reaches out to give the elephant something.

Enriched test: man behind fence. man next to trunk preparing food. man holding stick in enclosure. man pointing at something. fence truck behind food. fence wrapped around trunk. fence behind elephant. fence made of stick. fence surrounds enclosure. trunk of elephant. elephant in enclosure.

Original text: A white refrigerator freezer sitting inside of a kitchen. A corner of a kitchen with a big fridge. A kitchen has a plain white fridge in the corner. A refrigerator in the corner of a kitchen just off the dining room a room showing a very big fridge and a dining table.

Enriched text: refrigerator has freezer. refrigerator in corner. refrigerator in bright kitchen. refrigerator in room. refrigerator next to table sitting in kitchen. freezer next to table. corner window in room. corner of table. fridge in kitchen. table in kitchen. fridge table next to table in room.

Figure A.3: Qualitative examples of `OpenVik` context enrichment on text-to-image retrieval.

market, the enriched knowledge contexts could be "*the people shopping at market*", "*standing person shopping for fruit*". The idea is to enrich the original description $\mathcal{T}$: "*An image of ¡verb¿*" with relevant actions and relations with the extracted visual knowledge from `OpenVik`, which can potentially help in drawing-in the matched candidates.

A.7.3 Visual Commonsense Reasoning

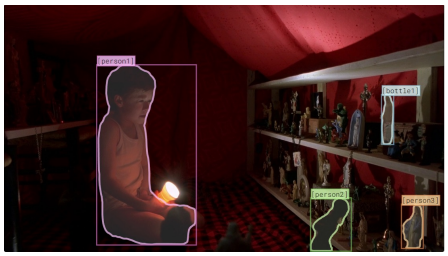
Figure A.5 presents more qualitative examples of `OpenVik`-based context enrichment in the visual commonsense reasoning (VCR) task. The context enrichment on VCR is performed at two-level, incorporating both entities and relations: (1) we parse the question and options to obtain all `(S, O)` pairs and, for each entity pair, apply

Figure A.4: Qualitative examples of `OpenVik` context enrichment on task GSR.

the same relation augmentation as in the image retrieval task; (2) for the `V` in each option, we enrich the visual context using the same method as illustrated in GSR. It is shown that unrelated answers are usually enriched with contexts that are not relevant to the image, thus enlarging the distance between incorrect answers and the question, e.g., the enriched contexts "*squating person fixing handy bathroom*" for example 3 in Figure A.5. At the same time, the knowledge description of the correct answer is enhanced by incorporating information that aligns with the image contents, e.g., the enriched knowledge contexts "*sitting people on red ground*" for example 1 in Figure A.5.

A.8 Full List of Filtered Verbs for GSR

We provide the full list of verbs out of the predefined 504 candidates of GSR [198] that can be accurate-matched or fuzzy-matched to extracted visual knowledge in Table A.4, based on which we compose the testing subset for our evaluation on GSR application in Section 5.2.

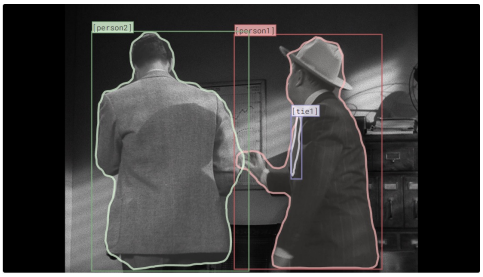


Question: Where is' Person1 sitting?
A He is in a laboratory.
B He is sitting at a bar. the person sitting behind sneaky barrier.
C In a fort in his house. the person walking by light house.
D He is sitting on the ground. sitting person on red ground.
Answer: D He is sitting on the ground.

Question: Where is Person2 going?
A Person2 is going into the store. the person walking into store.
B Person2 is getting into a carriage. sitting person inside carriage.
C Person1 is going to the bathroom. squating person fixing handy bathroom.
D Person1 is going outside to play after the conversation with Person2 is over.
Answer: A Person2 is going into the store.

Question: Why is Person7 in motion?
A Person14 is running desperately.
B Person7 is climbing over the boat. the person standing inside white boat.
C Person7 is walking fast to the bathroom. squating person fixing handy bathroom.
D Person7 is going to try to protect Person10 from a threat. Person7 is moving forward to challenge what ever could be there.
Answer: B Person7 is climbing over the boat.

Question : What will Person2 do next?
A Person2 will speak angrily at diningtable2, then walk off.
B Person2 will sit down on chair1. painting person near giant chair.
C Person2 will feed bowl1. the person skate boarding in a athletic bowl.
D Person2 will open the box. the person holding a box full of oranges.
Answer: B Person2 will sit down on chair1.



Question: Where are Person1 and Person2?
A Person1 and Person2 are sitting outside of a general store. the person walking by store.
B Person1 and Person2 are standing on top of a train car. jumping person on top board. walking person next to white train. the person walking near active car. yellow train sitting atop track. sliced carrot on top counter red car of old train.
C Person1 and Person2 are in an office. walking person outside office.
D Person1 and Person2 are in the kitchen. the person eating in hungry kitchen.
Answer: C Person1 and Person2 are in an office.

Question: What is Person1 doing here?
A He is in prison serving a prison sentence. person writing sentences.
B He is trying to get information. person gaining information.
C Person1 is a waiter. person talking with waiter in restaurant.
D He is existing a building. walking person near large building.
Answer: C Person1 is a waiter.

Figure A.5: Qualitative examples of `OpenVik` context enrichment on task VCR.

Table A.4: The full list of filtered verbs for GSR.

Matching Type	**The Word List of Event Types**
Accurate	putting, butting, bathing, dusting, rearing, turning, skating, placing, carting, staring, biting, mashing, folding, wetting, sprinkling, branching, drying, standing, flaming, taxiing, performing, circling, molding, parachuting, glowing, fishing, drinking, speaking, pawing, blocking, milking, racing, stripping, potting, spinning, eating, making, kicking, catching, lacing, urinating, sleeping, pressing, buttering, shearing, sliding, hiking, glaring, dipping, swimming, shopping, slicing, shelling, wagging, grilling, crafting, raining, clawing, splashing, rubbing, snowing, breaking, guarding, clipping, sewing, braiding, telephoning, buttoning, waiting, serving, picking, camping, leaning, working, kissing, wrapping, trimming, tripping, pasting, soaring, driving, kneeling, pumping, coloring, lighting, training, ducking, bowing, arching, cooking, checking, pushing, flipping, rocking, cresting, cleaning, reading, nailing, stitching, building, climbing, covering, shelving, attaching, calming, selling, gluing, dyeing, lapping, photographing, peeling, sprouting, licking, displaying, combing, stacking, planting, fastening, buying, mopping, burning, erasing, measuring, dining, tattooing, gardening, decorating, clearing, fixing, weeding, pulling, feeding, watering, crowning, shaking, dripping, emptying, typing, chasing, poking, leaping, pouring, hanging, sniffing, piloting, falling, overflowing, resting, crashing, carving, ballooning, wading, loading, shaving, boarding, pinning, rowing, juggling, shoveling, hugging, throwing, calling, singing, carrying, walking, writing, crouching, floating, painting, opening, tying, riding, strapping, dialing, saying, bubbling, signing, camouflaging, operating, leading, laughing, parading, skiing, drawing, gnawing, celebrating, spreading, filling, giving, running, smelling, plowing, helping, brushing, scooping, adjusting, wrinkling, steering, biking, smiling, spraying, boating, paying, chewing, stuffing, clinging, landing, wheeling, talking, scoring, teaching, jogging, pitching, flapping, tipping, scrubbing, sitting, surfing, stirring, competing, drumming, jumping, filming, dancing, waxing, hitting, recording, baking, waving, washing, signaling, chopping, stretching, rafting, microwaving, phoning, lifting, swinging, releasing, ramming, towing, packing, hauling, frying (*244 words*)
Fuzzy	educating, marching, spanking, descending, smearing, heaving, cramming, inflating, stooping, inserting, squeezing, tugging, tilting, moistening, swarming, subduing, waddling, winking, flexing, punching, attacking, nuzzling, sprinting, sucking, puckering, sketching, rotting, videotaping, complaining, tuning, locking, hurling, pricking, arranging, constructing, slapping, sweeping, restraining, dousing, frisking, twisting, wringing, hoisting, immersing, shredding, blossoming, igniting, spying, offering, pouting, confronting, docking, assembling, prying, grinning, sharpening, pruning, disciplining, nipping, coaching, nagging, storming, handcuffing, apprehending, bouncing, clenching, taping, distributing, striking, studying, plunging, curling, aiming, sowing, grinding, rinsing, punting, mowing, hitchhiking, skipping, leaking, providing, hunching, spoiling, kneading, burying, foraging, lathering, vaulting, ejecting, mending, pinching, deflecting, ascending, peeing, bothering, repairing, pedaling, ailing, fueling, skidding, scraping, soaking, grimacing, scolding, spitting, knocking, crushing, bandaging, saluting, fording, stumbling, discussing, raking, launching, whirling, fetching, brawling, retrieving, snuggling, exercising, colliding, stroking, whipping, tilling, betting, farming, browsing, examining, dropping, barbecuing, ignoring, asking, flinging, perspiring, embracing, slipping, flicking, smashing, arresting, lecturing, tearing, gasping, applying, counting, spilling, dragging, recovering, practicing, scratching, shooting, packaging, hunting, stinging (*154 words*)

Appendix B

Additional Information for Chapter 3.3

B.1 Implementation Details

Our models are implemented with PyTorch [191] and Huggingface Transformer library and trained on an 8 Tesla V100 GPU node. The model is trained for 10 epochs, where the Item Form dataset takes around 12 hours, the Color dataset takes about 32 hours, and the Pattern dataset needs around 35 hours to run on a single GPU. The overall architecture of PV2TEA consists of 361M trainable parameters, where a ViT_{base} [61] is used as the image encoder and initialized with the pre-trained model on ImageNet of 85M parameters, and the text encoder is initialized from $\text{BERT}_{\text{base}}$ [55] of 123M parameters. We use AdamW [160] as the optimizer with a weight decay of 0.05. The learning rate of each parameter group is set using a cosine annealing schedule [159] with the initial value of $1e$-5. The model is trained for 10 epochs, with both training and testing batch sizes of 8. The memory queue size M is set as 57600 and the temperature τ of in Equation 3.12 is set as 0.07. We performed a grid search for the softness α from [0, 0.2, 0.4, 0.6, 0.8] and used the best-performed 0.4 for

reporting the final results. The K for two-level neighborhood regularization is set at 10. The input textual description is cropped to a maximum of 100 words. The input image is divided into 30 by 30 patches. The hidden dimension of both the visual and textual encoders is set to 768 to produce the representations of patches, tokens, or the whole image/sequence. The epoch E for adding the second-level prediction neighbor regularization to reliability score $s(\mathcal{X}_n)$ is set as 2.

B.2 More Source-Aware Evaluation

Method	Gold Value Source	$\mathcal{D}$: **Color**			$\mathcal{D}$: **Pattern**		
		P	R	F_1	P	R	F_1
$\text{OpenTag}_{\text{cls}}$	Text ✓	85.06	43.28	57.37	85.00	42.96	57.07
	Text ✗ Image ✓	66.28	10.24	17.74	66.23	12.02	20.35
	GAP ↓	18.78	33.04	39.63	18.77	30.94	36.72
PAM	Text ✓	73.20	71.88	72.53	75.00	57.04	64.80
	Text ✗ Image ✓	50.30	45.45	47.75	51.82	36.23	42.64
	GAP ↓	22.90	26.43	24.78	23.18	20.81	22.16
PV2TEA	Text ✓	81.74	74.25	77.82	71.19	61.25	65.85
	Text ✗ Image ✓	71.89	47.19	56.98	54.48	37.26	44.25
	GAP ↓	9.85	27.06	**20.84**	16.71	23.99	**21.59**

Table B.1: Fine-grained source-aware evaluation for the Color and Pattern datasets.

The source-aware evaluation of the Color and Pattern datasets is shown in Table B.1. We can observe that similarly to the discussions in Section 3.3.5, compared with the baselines, the proposed PV2TEA effectively mitigates the performance gap of F_1 when the gold value is not contained in the text. More specifically, we observed that compared with the unimodal method, PV2TEA mainly reduces the recall performance gap across modalities, while compared with the multimodal method, the reduction happens mainly in precision, which all corresponds to the weaker metrics for each type

of method. This indicates the stronger generalizability and more balanced learning ability of PV2TEA.

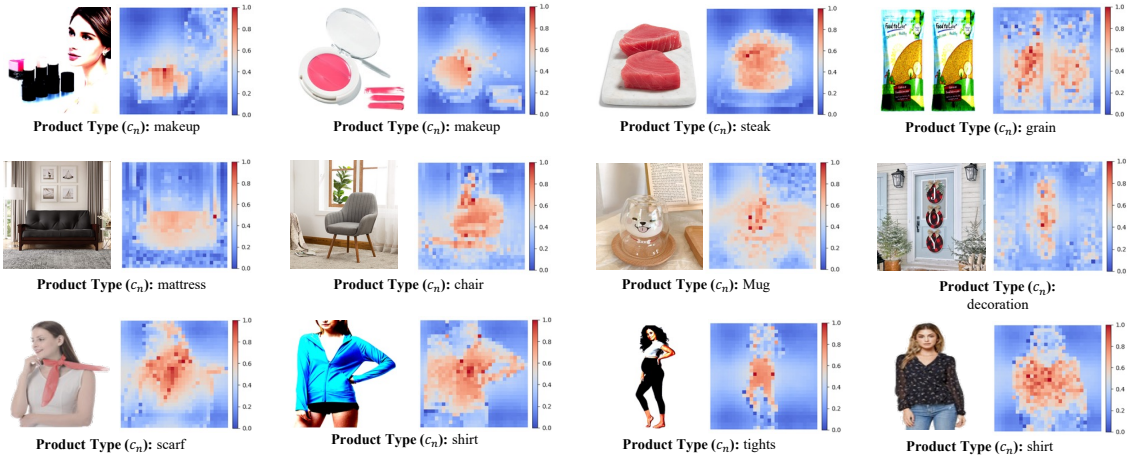


Figure B.1: Visualization examples of the learned category aware attention pruning mask.

B.3 Ablation Studies on Pattern Dataset

We further include the ablation results on the single-value type dataset Pattern for each proposed mechanism in Table B.2, Table B.3, and Table B.4, respectively. The observations are mostly consistent with the discussion in section 3.3.5, where all three proposed mechanisms support improvements in the overall performance of F_1. It is noted that the recall performance with attention-pruning drops a bit compared with that without. This may indicate potential information losses on the challenging dataset such as Pattern with only the selected foreground. We discuss this potential risk in detail in the Limitation section.

Method	Single Value Dataset: Pattern		
	Precision	Recall	F_1
PV2TEA w/o L_{sc}	60.03	45.59	51.82
PV2TEA w/o smooth	61.87	45.72	52.58
PV2TEA	62.10	46.84	53.40

Table B.2: Ablations on the augmented label-smoothed contrast for cross-modality alignment (%).

Method	Single Value Dataset: Pattern		
	Precision	Recall	F_1
PV2TEA w/o L_{ct} & Attn Prun	59.01	46.74	52.16
PV2TEA w/o Attn Prun	60.14	46.98	52.75
PV2TEA	62.10	46.84	53.40

Table B.3: Ablation study on the category supervised visual attention pruning (%).

Method	Single Value Dataset: Pattern		
	Precision	Recall	F_1
PV2TEA w/o NR	59.92	44.92	51.35
PV2TEA w/o Vis-NR	61.59	46.24	52.82
PV2TEA w/o Pred-NR	60.77	45.11	51.78
PV2TEA	62.10	46.84	53.40

Table B.4: Ablations on the two-level neighborhood-regularized sample weight adjustment (%).

B.4 Retrieval Ablation on Pattern Dataset

Similar to Figure 3.14, we also demonstrate the cross-modality retrieval results on the pattern dataset in Figure B.2. The conclusion is consistent with our observations mentioned in Section 3.3.5, where the contrastive objective demonstrates advantages in cross-modal alignment and fusion, and the best smoothness choice peaks at 0.4.

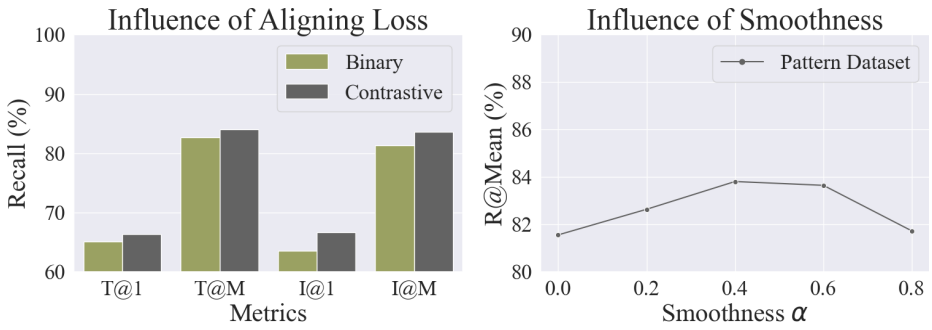


Figure B.2: The influence study of alignment objectives, i.e., binary matching v.s. contrastive, and softness α study via cross-modality retrieval on the Pattern dataset.

B.5 Visualizations of Attention Pruning

Examples of visualization on the learned attention mask are demonstrated in Figure B.1. It is observed that the visual foreground is highlighted under the supervision of category classification, which potentially encourages a higher prediction precision for fine-grained tasks like attribute extraction, as proved by the experimental results.

Category (Product Type)	Candidate Attribute Values Given the Category
cereal	grain, flake, seed, liquid, powder, ground
dishwasher detergent	gel, capsule, pac, liquid, tablet, pod, powder
face shaping makeup	powder, pencil, cream, liquid, stick, oil, spray, gel, cushion, blush, drop, balm, gloss
fish	fillet, chunk, steak, solid, stick, whole, slice, ground
herb	powder, root, leaf, thread, flake, seed, tea bag, stick, oil, slice, pod, ground, bean, paste
honey	jelly, capsule, lozenge, candy, cream, powder, granule, flake, liquid, stick, oil, crystal, butter, drop, syrup, comb
insect repellent	wipe, spray, band, granular, liquid, stick, candle, coil, oil, lotion, gel, capsule, tablet, powder, balm, patch, roll on
jerky	strip, slab, shredded, bite, bar, slice, stick, ground
sauce	puree, jelly, paste, seed, liquid, gravy, ground, oil, powder, cream
skin cleaning agent	powder, capsule, toothpaste, wipe, cream, spray, mousse, bar, flake, liquid, lotion, gel, serum, mask, ground, balm, paste, foam
skin foundation concealer	powder, pencil, cream, mousse, liquid, stick, oil, lotion, spray, cushion, gel, drop, serum, balm, airbrush
sugar	granule, crystal, pearl, liquid, powder, cube, ground
sunscreen	wipe, cream, spray, mousse, liquid, ointment, stick, fluid, oil, lotion, milk, compact, gel, drop, serum, powder, balm, foam, mist
tea	leaf, powder, granule, tea bag, liquid, pod, ground, brick

Table B.5: The annotation candidates provided to annotators given each sample type on the Item Form dataset.

B.6 Human Annotation Instruction

We create source-aware fine-grained datasets with internal human annotators. Below are the instruction texts provided to annotators:

The annotated attribute values are used for research model development of multimodal attribute information extraction and fine-grained error analysis. The datasets are named source-aware multimodal attribute extraction evaluation benchmarks and will be released to facilitate public testing and future studies in bias-reduced multi-

modal attribute value extraction model designs. All the given sample profiles (title, bullets, and descriptions) and images are collected from the public `amazon.com` web pages, so there is no potential legal or ethical risk for annotators. Specifically, the annotation requirements compose two tasks in order: (1) Firstly, for each given sample_id in the given ASINs set, first determine the category of the sample by referring to ID2Category.csv mapping file, then label the gold value for the queried attribute by selecting from the candidates given the category. The annotation answer candidates for the Item Form dataset can be referred to in Table B.5. Note that this gold value annotation step requires reference to both sample textual titles, descriptions, and images; (2) For each annotated ASIN, mark down which modality implies the gold value with an additional source label, with different meanings as below:

- **0**: *the gold attribute value can be found in text.*
- **1**: *the gold attribute value cannot be inferred from the text but can be found in the image.*

The annotated attribute values and source labels are assembled in fine-grained source-aware evaluation.

B.7 Neighborhood Regularization Demos

We provide two more demo examples for illustrating the two-level neighborhood-regularized sample weight adjustment in Figure B.3. The example on the left demonstrates a higher consistency between the green arrows (which point to samples with the same training label as y_n) and red arrows (which point to k-nearest neighbor samples in visual feature and previous prediction space), indicating a higher reliability of y_n. Thus the sample weight of $\mathcal{X}_n$ will be increased in the next training epoch. In contrast, the training label neighbors and visual/prediction neighbors of the right example show a large inconsistency, which implies a relatively lower reliability of y_n.

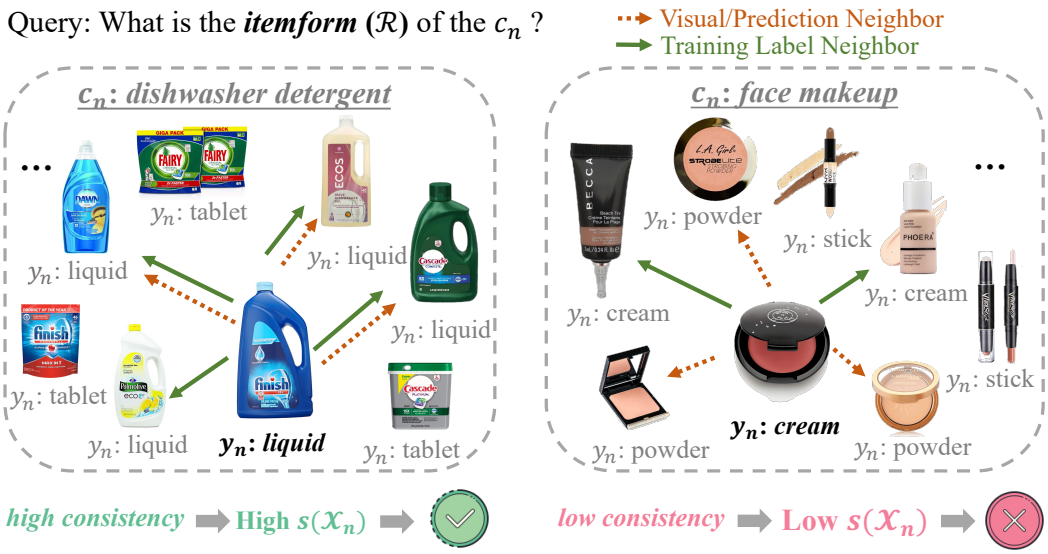


Figure B.3: Demo examples for illustrating S3: two-level neighborhood-regularized sample weight adjustment.

Therefore, the sample weight $s(\mathcal{X}_n)$ of the right $\mathcal{X}_n$ will be degraded in the next epoch.

This regularization process adjusts the sample weights of all the training samples in each epoch.

Appendix C

Additional Information for Chapter 3.4

C.1 Prompt for Predictor Agent

C.2 Prompt for Critic Agent

Prompt for the Predictor LLM Agent	
role: system	**content:** You are a medical expert with a specialization in type 2 diabetes and cardiovascular disease. Your task is to predict whether a patient with type 2 diabetes will develop a cardiovascular disease (CVD) endpoint within a year of their initial diagnosis.
role: user	**content:** Task: Your task is to predict whether a patient with type 2 diabetes will develop a cardiovascular disease (CVD) endpoint within a year of their initial diagnosis based on the provided patient's medical history. You will be presented with a patient's medical history and various resources to aid in your prediction. Please provide your reasoning and make your prediction by learning from the resources. You are presented with the following: 1. [CVD Endpoint Definition] The definition of the prediction target: cardiovascular disease (CVD) endpoint. 2. [Past Medical History] Patient's past medical history, which captures specific diagnoses made, medications prescribed, and procedures performed. 3. [Instructions] Guidelines on how to analyze the patient's medical history, provide reasoning, and make predictions. This includes referring to the demonstration cases and exploring the interaction of various factors and the interplay between diseases, medications, and procedures that the patient has undergone. The reasoning process should support and aid in the final prediction for a CVD endpoint. 4. [Demonstration Cases] Some real and typical cases, including the patient's medical history (diseases, medications, and procedures) and the ground truth result of whether the patients with type 2 diabetes experience cardiovascular disease (CVD) endpoint within a year after the initial diagnosis. 5. [Output Format] The required format for your response. Please ensure that you strictly adhere to the format requirements. You must provide a confirmed prediction by choosing between "Yes" or "No". [CVD Endpoint Definition] A CVD endpoint is identified by the presence of coronary heart disease (CHD), congestive heart failure (CHF), myocardial infarction (MI), or stroke. [Past Medical History] {info_generate(code_1_list)} [Instructions] Based on the patient's past medical history, provide reasoning on whether a patient with type 2 diabetes will develop a cardiovascular disease (CVD) endpoint within a year of their initial diagnosis. You should know that: globally, cardiovascular disease (CVD) affects about 32.2% of people with type 2 diabetes (T2D). CVD is the most common cause of morbidity and mortality in people with T2D. In this specific training set, 20% of the patients develop CVD within a year of their initial diagnosis. Please: (1) use your knowledge; (2) learn from the provided demonstration cases; (3) perform a comprehensive analysis of the patient's medical history; (4) consider the interplay of diseases, medications, and procedures, make sure to: Identify and weigh both risk factors and protective factors evident in the patient's medical history; Consider the presence of any comorbid conditions that may independently increase or decrease the risk of CVD; Examine the patient's medication profile to discern any pharmacological interventions that may alter the course of disease progression; Evaluate any medical or surgical procedures the patient has undergone that could impact their cardiovascular health; (5) Utilize a multihop and step-by-step reasoning approach to systematically analyze the data. [Demonstration Cases] {few_shots_generate(few_shots_label_0, few_shots_label_1)} [Output Format] Your final response should include: 1. Prediction: Conclude your analysis with a clear and concise prediction. This prediction must be a single word, either "Yes" or "No", indicating whether you believe the patient is likely to develop a CVD endpoint within a year of their initial diabetes diagnosis. This prediction should be the first line of your response, to facilitate easy parsing. 2. Reasoning: Provide a detailed reasoning process. Ensure that your analysis is thorough and based on the information provided, leading logically to your final prediction.

Figure C.1: Prompt for Predictor Agent in EHR-CoAgent for the CRADLE dataset.

Prompt for the Critic LLM Agent	
role: system	**content:** You are an assistant who is good at self-reflection, gaining experience, and summarizing criteria. By reflecting on failure predictions that are given below, your task is to reflect on these incorrect predictions, compare them against the ground truth, and formulate criteria and guidelines to enhance the accuracy of future predictions.
role: user	**content:** Task: You will be given a batch of input data samples, where each sample is composed of three parts: the patient's medical history, the ground-truth result for whether the patient will develop a cardiovascular disease (CVD) endpoint within a year of their initial diagnosis of type 2 diabetes, and a wrong prediction. Please always remember that the predictions above are all incorrect. You should always use the ground truth as the final basis to discover many unreasonable aspects in the predictions and then summarize them into instructions and criteria. You are presented with the following: 1. [Input Data] A batch of input data samples. Each data in the batch includes three parts: (a) the patient's medical history, including diseases that the patient has been diagnosed with, medications that the patient has taken, and procedures the patient has undergone; (b) the ground-truth result for each patient's medical history on whether the patient will develop a cardiovascular disease (CVD) endpoint within a year of their initial diagnosis of type 2 diabetes; (c) a wrong prediction. 2. [Instructions] Instructions on suggesting criteria. [Input Data] {batch_generate(input_data_batch)} [Instructions] 1. Please always remember that the predictions above are all incorrect. You should always use the ground truth as the final basis to discover many unreasonable aspects in the predictions and then summarize them into experience and criteria. 2. Identify why the wrong predictions deviated from the ground truth by examining discrepancies in the medical history analysis. 3. Determine key and potential influencing factors, reasoning methods, and relevant feature combinations that could better align predictions with the ground truth. 4. The instructions should be listed in distinct rows, each representing a criteria or guideline. 5. The instructions should be generalizable to multiple samples, rather than specific to individual samples. 6. Conduct detailed analysis and write criteria based on the input samples, rather than writing some criteria without foundation. 7. Please note that the criteria you wrote should not include the word "ground truth".

Figure C.2: Prompt for Critic Agent in EHR-CoAgent for the CRADLE dataset.

Appendix D

Additional Information for Chapter 3.5

D.1 Data Generation Prompt Design

The Prompt for Generating Instruction-Following Data with GPT-4Vision
You are an AI assistant specialized in biomedical topics. You are provided with a figure image from a biomedical research paper. In some cases, you may have additional text (Figure Context) that mentions the image. Please meticulously extract all possible visual details from the image, and when generating questions and answers, ensure to integrate and consider the provided supplementary textual information. It is crucial to highlight the connections and correlations between the textual content and the visual elements within the picture to capture the full context. Your task is to facilitate a dialogue where a person (User) seeks information about the image, and you (Assistant) provide insightful responses. During this interaction, the conversation should evolve as if both the User and Assistant are observing the image together. It is essential to thoroughly consider and reference the accompanying textual information (Figure Caption and Figure Context) and visual information to ensure a rich and informative exchange that highlights the significance of the visual details present. Below are requirements for generating the questions and answers in the conversation: - Focus on visual aspects of the image that can be inferred without the text information, and extract as much key detailed information from the image as possible.. - Ensure that questions are diverse and cover a range of visual aspects of the image. - The conversation should encompass a minimum of 4-5 exchanges of questions and answers. You may adjust the number of rounds based on the provided image and text. For content with substantial information, employing additional questions and answers may be more appropriate to ensure thorough discussion and understanding. - When the provided textual information is relevant to the question, try to answer using the expertise and specialized terminology contained within the text, rather than with vague, non-specialized descriptions.

Figure D.1: The prompt for generating instruction-following data with GPT-4V(ision).

D.2 Benchmark Evaluation Prompt Design

The Prompt for Reference-Guided Pairwise Win-Rate Evaluation on VQA Benchmarks
Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. Your evaluation should consider correctness and helpfulness. You will be given a reference answer, assistant A's answer, and assistant B's answer. Your job is to evaluate which assistant's answer is better. Begin your evaluation by comparing both assistants' answers with the reference answer. Identify and correct any mistakes. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie. You must begin with [[A]] or [[B]] or [[C]]. Assigning "[[C]]" should be a last resort, used only if you absolutely cannot discern any difference in the quality of the two responses.

Figure D.2: The prompt for reference-guided pairwise win-rate evaluation on VQA benchmarks.

Bibliography

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *EMNLP*, 2022.

[3] Harith Alani, Sanghee Kim, David E Millard, Mark J Weal, Wendy Hall, Paul H Lewis, and Nigel R Shadbolt. Automatic ontology-based knowledge extraction from web documents. *IEEE Intell Syst*, 18:14–21, 2003.

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

[5] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and

Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6816–6826, 2021.

[7] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. 1999.

[8] Bing Bai et al. Why attentions may not be interpretable? In *SIGKDD*, 2021.

[9] Peter J Basser, Sinisa Pajevic, Carlo Pierpaoli, Jeffrey Duda, and Akram Aldroubi. In vivo fiber tractography using dt-mri data. *Magn Reson Med*, 44: 625–632, 2000.

[10] Romain Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. `https://github.com/rom1504/clip-retrieval`, 2022.

[11] Timothy EJ Behrens, Mark W Woolrich, Mark Jenkinson, Heidi Johansen-Berg, Rita G Nunes, Stuart Clare, Paul M Matthews, J Michael Brady, and Stephen M Smith. Characterization and propagation of uncertainty in diffusion-weighted mr imaging. *Magn Reson Med*, 50:1077–1088, 2003.

[12] Timothy EJ Behrens, H Johansen Berg, Saad Jbabdi, Matthew FS Rushworth, and Mark W Woolrich. Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? *NeuroImage*, 34:144–155, 2007.

[13] Tiffany K Bell, Kate J Godfrey, Ashley L Ware, Keith Owen Yeates, and Ashley D Harris. Harmonization of multi-site mrs data with combat. *NeuroImage*, page 119330, 2022.

[14] Pierre Bellec, Carlton Chu, Francois Chouinard-Decorte, Yassine Benhajali, Daniel S Margulies, and R Cameron Craddock. The neuro bureau adhd-200 preprocessed repository. *NeuroImage*, 144:275–286, 2017.

[15] Alexander Bernstein, Renat Akzhigitov, Ekaterina Kondrateva, Svetlana Sushchinskaya, Irina Samotaeva, and Vladislav Gaskin. MRI brain imagery processing software in data analysis. *Trans. Mass Data Anal. Images Signals*, 9:3–17, 2018.

[16] Alaa Bessadok, Mohamed Ali Mahjoub, and Islem Rekik. Graph neural networks in network neuroscience. *ArXiv.org*, 2021.

[17] Shreyas Bhave, Victor Rodriguez, Timothy Poterucha, Simukayi Mutasa, Dwight Aberle, Kathleen M Capaccione, Yibo Chen, Belinda Dsouza, Shifali Dumeer, Jonathan Goldstein, Aaron Hodes, Jay Leb, Matthew Lungren, Mitchell Miller, David Monoky, Benjamin Navot, Kapil Wattamwar, Anoop Wattamwar, Kevin Clerkin, David Ouyang, Euan Ashley, Veli K Topkara, Mathew Maurer, Andrew J Einstein, Nir Uriel, Shunichi Homma, Allan Schwartz, Diego Jaramillo, Adler J Perotte, and Pierre Elias. Deep learning to detect left ventricular structural abnormalities in chest X-rays. *European Heart Journal*, 2024.

[18] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL*, 2019.

[19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[20] Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat. Rev. Neurosci.*, 10:186–198, 2009.

[21] Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds,

Nicole Hamilton, and Gregory N. Hullender. Learning to rank using gradient descent. In *ICML*, 2005.

[22] Chen Cai and Yusu Wang. A Simple Yet Effective Baseline for Non-Attributed Graph Classification. *ArXiv.org*, 2018.

[23] Vince D Calhoun and Jing Sui. Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness. *Biol Psychiatry Cogn Neurosci Neuroimaging*, 1:230–244, 2016.

[24] BJ Casey, Tariq Cannonier, May I Conley, Alexandra O Cohen, Deanna M Barch, Mary M Heitzeg, Mary E Soules, Theresa Teslovich, Danielle V Dellarco, Hugh Garavan, et al. The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Dev Cogn Neurosci*, 32:43–54, 2018.

[25] Julian Caspers, Christian Rubbert, et al. Within-and across-network alterations of the sensorimotor network in parkinson's disease. *Neuroradiology*, 63:2073, 2021.

[26] Aaron Chan, Jiashu Xu, Boyuan Long, Soumya Sanyal, Tanishq Gupta, and Xiang Ren. Salkg: Learning from knowledge graph explanations for commonsense reasoning. In *NeurIPS*, 2021.

[27] Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. In *Conference on Learning Theory*, 2012.

[28] Hila Chefer, Idan Schwartz, and Lior Wolf. Optimizing relevance maps of vision transformers improves robustness. In *NeurIPS*, 2022.

[29] Andrew A Chen, Joanne C Beer, Nicholas J Tustison, Philip A Cook, Russell T Shinohara, Haochang Shou, Alzheimer's Disease Neuroimaging Initiative, et al.

Removal of scanner effects in covariance improves multivariate pattern analysis in neuroimaging data. *bioRxiv*, page 858415, 2020.

[30] Jiayu Chen, Jingyu Liu, Vince D Calhoun, Alejandro Arias-Vasquez, Marcel P Zwiers, Cota Navin Gupta, Barbara Franke, and Jessica A Turner. Exploration of scanning effects in multi-site structural mri studies. *J. Neurosci. Methods*, 230:37–50, 2014.

[31] Nian-Shing Chen, Kinshuk, Chun-Wang Wei, and Hong-Jhe Chen. Mining e-learning domain concept map from academic articles. *Comput. Educ.*, pages 1009–1021, 2008.

[32] Qingyu Chen, Yifan Peng, and Zhiyong Lu. Biosentvec: creating sentence embeddings for biomedical texts. In *ICHI*, 2019.

[33] Xuelu Chen, Ziniu Hu, and Yizhou Sun. Fuzzy logic based logical query answering on knowledge graphs. In *AAAI*, 2022.

[34] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[35] Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. Learning the graphical structure of electronic health records with graph convolutional transformer. In *AAAI*, 2020.

[36] Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. Towards coherent multi-document summarization. In *NAACL*, 2013.

[37] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma,

et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[38] Gabriele Corso et al. Principal neighbourhood aggregation for graph nets. *Advances in Neural Information Processing Systems*, 33:13260–13271, 2020.

[39] Hejie Cui, Zijie Lu, Pan Li, and Carl Yang. On positional and structural node features for graph neural networks on non-attributed graphs. *CoRR*, abs/2107.01495, 2021.

[40] Hejie Cui, Wei Dai, Yanqiao Zhu, Xuan Kan, Antonio Aodong Chen Gu, Joshua Lukemire, Liang Zhan, Lifang He, Ying Guo, and Carl Yang. BrainGB: a benchmark for brain network analysis with graph neural networks. *IEEE Transactions on Medical Imaging*, 2022.

[41] Hejie Cui, Wei Dai, Yanqiao Zhu, Xiaoxiao Li, Lifang He, and Carl Yang. Interpretable graph neural networks for connectome-based brain disorder analysis. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2022.

[42] Hejie Cui, Jiaying Lu, Yao Ge, and Carl Yang. How can graph neural networks help document retrieval: A case study on cord19 with concept map generation. In *European Conference on IR Research (ECIR)*, 2022.

[43] Hejie Cui, Zijie Lu, Pan Li, and Carl Yang. On positional and structural node features for graph neural networks on non-attributed graphs. In *ACM International Conference on Information and Knowledge Management (CIKM)*, 2022.

[44] Hejie Cui, Rongmei Lin, Nasser Zalmout, Chenwei Zhang, Jingbo Shang, Carl Yang, and Xian Li. PV2TEA: Patching visual modality to textual-established

information extraction. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[45] Hejie Cui, Jiaying Lu, Shiyu Wang, Ran Xu, Wenjing Ma, Shaojun Yu, et al. A survey on knowledge graphs for healthcare: Resources, application progress, and promise. *ICML Workshop on Interpretable Machine Learning in Healthcare (ICML-IMLH)*, 2023.

[46] Tian Dai, Ying Guo, Alzheimer's Disease Neuroimaging Initiative, et al. Predicting individual brain functional connectivity using a bayesian hierarchical model. *NeuroImage*, 147:772–787, 2017.

[47] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *EACL*, 2023.

[48] Van Dang, Michael Bendersky, and W Bruce Croft. Two-stage learning to rank for information retrieval. In *ECIR*, 2013.

[49] Tushar K Das, Jyothika Kumar, et al. Parietal lobe and disorganisation syndrome in schizophrenia and psychotic bipolar disorder: A bimodal connectivity study. *Psychiatry Research: Neuroimaging*, 303:111139, 2020.

[50] Aloïs De la Comble, Anuvabh Dutt, Pablo Montalvo, and Aghiles Salah. Multimodal attribute extraction for e-commerce. *arXiv preprint arXiv:2203.03441*, 2022.

[51] Gustavo Deco, Viktor K Jirsa, and Anthony R McIntosh. Emerging concepts for the dynamical organization of resting-state activity in the brain. *Nat. Rev. Neurosci.*, 12:43–56, 2011.

[52] Anup Anand Deshmukh and Udhav Sethi. IR-BERT: leveraging BERT for semantic search in background linking for news articles. *CoRR*, abs/2007.12603, 2020.

[53] Gopikrishna Deshpande and Hao Jia. Multi-level clustering of dynamic directional brain network patterns and their behavioral relevance. *Front. Neurosci.*, 13:1448, 2020.

[54] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31:968–980, 2006.

[55] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[56] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2019.

[57] Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry*, 19:659–667, 2014.

[58] Yang Ding, Jing Yu, Bang Liu, Yue Hu, Mingxin Cui, and Qi Wu. Mukea: Multimodal knowledge extraction and accumulation for knowledge-based visual question answering. In *CVPR*, 2022.

[59] Yifan Ding, Yan Liang, Nasser Zalmout, Xian Li, Christan Grant, and Tim Weninger. Ask-and-verify: Span candidate generation and verification for attribute value extraction. In *EMNLP*, 2022.

[60] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

[61] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[62] Chi Thang Duong, Thanh Dat Hoang, Ha The Hien Dang, Quoc Viet Hung Nguyen, and Karl Aberer. On node features for graph neural networks. *ArXiv.org*, 2019.

[63] Martin Dyrba, Michel Grothe, Thomas Kirste, and Stefan J. Teipel. Multimodal analysis of functional and structural disconnection in Alzheimer's disease using multiple kernel SVM. *Hum. Brain Mapp.*, 36:2118, 2015.

[64] Tugba Akinci D'Antonoli, Arnaldo Stanzione, Christian Bluethgen, Federica Vernuccio, Lorenzo Ugga, Michail E Klontzas, Renato Cuocolo, Roberto Cannella, and Burak Koçak. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagnostic and Interventional Radiology*, 30:80, 2024.

[65] Kathryn A Ellis, Ashley I Bush, David Darby, Daniela De Fazio, Jonathan Foster, Peter Hudson, Nicola T Lautenschlager, Nat Lenzo, Ralph N Martins, Paul Maruff, et al. The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited

for a longitudinal study of alzheimer's disease. *Int Psychogeriatr*, 21:672–687, 2009.

[66] Federico Errica, Marco Podda, Davide Bacciu, and Alessio Micheli. A fair comparison of graph neural networks for graph classification. In *ICLR*, 2020.

[67] Flor A Espinoza, Victor M Vergara, Eswar Damaraju, Kyle G Henke, Ashkan Faghiri, Jessica A Turner, Aysenil A Belger, Judith M Ford, Sarah C McEwen, Daniel H Mathalon, et al. Characterizing whole brain temporal variation of functional connectivity via zero and first order derivatives of sliding window correlations. *Front. Neurosci.*, 13:634, 2019.

[68] James Fan, Aditya Kalyanpur, David C Gondek, and David A Ferrucci. Automatic knowledge extraction from documents. *IBM J Res Dev*, 56:5–1, 2012.

[69] Farzad V Farahani, Waldemar Karwowski, and Nichole R Lighthall. Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review. *Front. Neurosci.*, 13:585, 2019.

[70] Sidali Hocine Farhi and Dalila Boughaci. Graph based model for information retrieval using a stochastic local search. *Pattern Recognit. Lett.*, pages 234–239, 2018.

[71] Joshua Faskowitz, Richard F Betzel, and Olaf Sporns. Edges in brain networks: Contributions to models of structure and function. *ArXiv.org*, 2021.

[72] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. In *RLGM@ICLR*, 2019.

[73] Teresa D Figley et al. Probabilistic white matter atlases of human auditory, basal ganglia, language, precuneus, sensorimotor, visual and visuospatial networks. *Frontiers in human neuroscience*, 11:306, 2017.

[74] Jessica S Flannery, Michael C Riedel, et al. Hiv infection is linked with reduced error-related default mode network suppression and poorer medication management abilities. *medRxiv*, 2021.

[75] Jean-Philippe Fortin, Nicholas Cullen, Yvette I Sheline, Warren D Taylor, Irem Aselcioglu, Philip A Cook, Phil Adams, Crystal Cooper, Maurizio Fava, Patrick J McGrath, et al. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120, 2018.

[76] Egill A. Fridgeirsson, David Sontag, and Peter Rijnbeek. Attention-based neural networks for clinical prediction modelling on electronic health records. *BMC Medical Research Methodology*, page 285.

[77] Mikhail Galkin, Zhaocheng Zhu, Hongyu Ren, and Jian Tang. Inductive logical query answering in knowledge graphs. In *NeurIPS*, 2022.

[78] Giorgio Ganis and Stephen M. Kosslyn. Neuroimaging. In *Encyclopedia of the Human Brain*, pages 493–505. 2002.

[79] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *ACL-IJCNLP*, 2021.

[80] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICML*, 2017.

[81] Matthew F. Glasser, Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R. Polimeni, David C. Van Essen, and Mark Jenkinson. The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, 80:105–124, 2013.

[82] Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John PA Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association: JAMIA*, 24:198, 2017.

[83] Karthik Gopinath, Christian Desrosiers, and Herve Lombaert. Learnable pooling in graph convolution networks for brain surface analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[84] Daniele Grattarola, Daniele Zambon, Filippo Maria Bianchi, and Cesare Alippi. Understanding pooling in graph neural networks. *ArXiv.org*, 2021.

[85] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *INTERSPEECH*, pages 5036–5040, 2020.

[86] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven CH Hoi. From images to textual prompts: Zero-shot vqa with frozen large language models. In *CVPR*, 2023.

[87] Meng-Hao Guo, Tian-Xing Xu, Jiang-Jiang Liu, Zheng-Ning Liu, Peng-Tao Jiang, Tai-Jiang Mu, Song-Hai Zhang, Ralph R. Martin, Ming-Ming Cheng, and Shi-Min Hu. Attention mechanisms in computer vision: A survey. *ArXiv.org*, 2021.

[88] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017.

[89] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem.

Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

[90] Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics, 2023.

[91] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.

[92] Lifang He, Kun Chen, Wanwan Xu, Jiayu Zhou, and Fei Wang. Boosted sparse and low-rank tensor regression. In *NeurIPS*, 2018.

[93] Tao He, Lianli Gao, Jingkuan Song, and Yuan-Fang Li. Towards open-vocabulary scene graph generation with prompt-based finetuning. In *ECCV*, 2022.

[94] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[95] Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *AISTATS*, 2023.

[96] Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, Emily Alsentzer, et al. Do we still need clinical language models? In *Conference on Health, Inference, and Learning*, 2023.

[97] Megan M Herting, Kristina A Uban, et al. Default mode connectivity in youth with perinatally acquired hiv. *Medicine*, 94, 2015.

[98] Ixavier A Higgins, Suprateek Kundu, Ki Sueng Choi, Helen S Mayberg, and Ying Guo. A difference degree test for comparing brain networks. *Hum Brain Mapp*, pages 4518–4536, 2019.

[99] Robert V Hogg, Joseph McKean, et al. *Introduction to mathematical statistics*. 2005.

[100] Andreas Holzinger, Peter Kieseberg, Edgar Weippl, and A Min Tjoa. Current advances, trends and challenges of machine learning and knowledge extraction: from machine learning to explainable ai. In *CD-MAKE*, 2018.

[101] Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, Kirk Roberts, and Hua Xu. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259, 2024.

[102] Yingtian Hu, Mahmoud Zeydabadinezhad, Longchuan Li, and Ying Guo. A multimodal multilevel neuroimaging model for investigating brain connectome development. *J Am Stat Assoc*, 117:1–15, 2022.

[103] Qian Huang, Horace He, Abhay Singh, Ser-Nam Lim, and Austin R. Benson. Combining label propagation and simple models out-performs graph neural networks. *CoRR*, abs/2010.13993, 2020.

[104] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 2307–2316, 2023.

[105] Thomas R Insel and Bruce N Cuthbert. Brain disorders? precisely. *Science*, 348:499–500, 2015.

[106] Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *NAACL-HLT*, 2019.

[107] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput Surv*, 55:1–38, 2023.

[108] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[109] Biao Jie, Mingxia Liu, Xi Jiang, and Daoqiang Zhang. Sub-network based kernels for brain network classification. In *ICBC*, 2016.

[110] Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40:btae075, 2024.

[111] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:1–9, 2016.

[112] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016.

[113] Amir Joudaki, Niloufar Salehi, Mahdi Jalili, and Maria G Knyazeva. Eeg-based functional brain networks: does the network size matter? *PLoS One*, 7:e35673, 2012.

[114] Chris Kamphuis. Graph databases for information retrieval. In *ECIR*, 2020.

[115] Xuan Kan, Hejie Cui, and Carl Yang. Zero-shot scene graph relation prediction through commonsense knowledge integration. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD)*, 2021.

[116] Xuan Kan, Hejie Cui, Lukemire Joshua, Ying Guo, and Carl Yang. Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In *MIDL*, 2022.

[117] Xuan Kan, Hejie Cui, Joshua Lukemire, Ying Guo, and Carl Yang. Fbnetgen: Task-aware gnn-based fmri analysis via functional brain network generation. In *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2022.

[118] Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Yang Carl. Brain network transformer. In *NeurIPS*, 2022.

[119] Giannis Karamanolakis, Jun Ma, and Xin Luna Dong. Txtract: Taxonomy-aware knowledge extraction for thousands of product categories. In *ACL*, 2020.

[120] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049, 2017.

[121] Jeremy Kawahara, Colin J. Brown, et al. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038, 2017.

[122] AT Karagulle Kendi, S Lehericy, et al. Altered diffusion in the frontal lobe in parkinson disease. *American Journal of Neuroradiology*, 29:501, 2008.

[123] Nicolas Keriven and Gabriel Peyré. Universal invariant and equivariant graph neural networks. In *NeurIPS*, 2019.

[124] Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *CVPR*, 2019.

[125] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[126] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[127] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In *NeurIPS*, 2022.

[128] Martin Krallinger, Maria Padron, and Alfonso Valencia. A sentence sliding window approach to extract protein annotations from biomedical articles. *BMC bioinformatics*, 6:1–12, 2005.

[129] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017.

[130] Liqun Kuang, Deyu Zhao, Jiacheng Xing, Zhongyu Chen, Fengguang Xiong, and Xie Han. Metabolic brain network analysis of fdg-pet in alzheimer's disease using kernel-based persistent features. *Molecules*, 24:2301, 2019.

[131] Isotta Landi, Benjamin S Glicksberg, Hao-Chih Lee, Sarah Cherng, Giulia Landi, Matteo Danieletto, Joel T Dudley, Cesare Furlanello, and Riccardo

Miotto. Deep representation learning of electronic health records to unlock patient stratification at scale. *NPJ digital medicine*, 3:96, 2020.

[132] Elmar Wolfgang Lang, Ana Maria Tomé, Ingo R. Keck, Juan Manuel Górriz Sáez, and Carlos García Puntonet. Brain connect analysis: A short survey. *Comput. Intell. Neurosci.*, 2012:412512:1–412512:21, 2012.

[133] Hunter Lang, Monica N Agrawal, Yoon Kim, and David Sontag. Co-training improves prompt-based learning for large language models. In *ICML*, 2022.

[134] Hunter Lang, Aravindan Vijayaraghavan, and David Sontag. Training subset selection for weak supervision. In *NeurIPS*, 2022.

[135] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5:1–10, 2018.

[136] Alex D Leow, Siwei Zhu, Liang Zhan, Katie McMahon, Greig I de Zubicaray, Matthew Meredith, MJ Wright, AW Toga, and PM Thompson. The tensor distribution function. *Magn Reson Med*, 61:205–214, 2009.

[137] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[138] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 2021.

[139] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrap-

ping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[140] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[141] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022.

[142] Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, et al. Gaia: A fine-grained multimedia knowledge extraction system. In *ACL*, 2020.

[143] Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. Connecting the dots: Event graph schema induction with path language modeling. In *EMNLP*, 2020.

[144] Wei Li, Miao Wang, Yapeng Li, Yue Huang, and Xi Chen. A novel brain network construction method for exploring age-related functional reorganization. *Comput. Intell. Neurosci.*, 2016, 2016.

[145] Xiaoxiao Li, Nicha C Dvornek, Yuan Zhou, Juntang Zhuang, Pamela Ventola, and James S Duncan. Graph neural network for interpreting task-fmri biomarkers. In *MICCAI*, 2019.

[146] Xiaoxiao Li, Yuan Zhou, Nicha Dvornek, Muhan Zhang, Siyuan Gao, Juntang Zhuang, Dustin Scheinost, Lawrence H Staib, Pamela Ventola, and James S Duncan. Braingnn: Interpretable brain graph neural network for fmri analysis. *Med Image Anal*, 2021.

[147] Yunfang Li et al. Structural gray matter change early in male patients with hiv. *International journal of clinical and experimental medicine*, 7:3362, 2014.

[148] Rongmei Lin, Xiang He, Jie Feng, Nasser Zalmout, Yan Liang, Li Xiong, and Xin Luna Dong. Pam: understanding product images in cross product category attribute extraction. In *SIGKDD*, 2021.

[149] Martin A Lindquist. The statistical analysis of fmri data. *Stat Sci*, 23:439–464, 2008.

[150] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *International Symposium on Biomedical Imaging (ISBI)*, 2021.

[151] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[152] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[153] Tie-Yan Liu. Learning to rank for information retrieval. 2011.

[154] Ye Liu, Lifang He, Bokai Cao, Philip Yu, Ann Ragin, and Alex Leow. Multi-view multi-graph embedding for brain network clustering analysis. In *AAAI*, 2018.

[155] Ye Liu, Lifang He, Bokai Cao, Philip Yu, Ann Ragin, and Alex Leow. Multi-view multi-graph embedding for brain network clustering analysis. In *AAAI*, 2018.

[156] Ye Liu, Hui Li, Alberto Garcia-Duran, Mathias Niepert, Daniel Onoro-Rubio, and David S Rosenblum. Mmkg: multi-modal knowledge graphs. In *ESWC*, 2019.

[157] Ziqi Liu, Chaochao Chen, Longfei Li, Jun Zhou, Xiaolong Li, Le Song, and Yuan Qi. Geniepath: Graph neural networks with adaptive receptive paths. In *AAAI*, 2019.

[158] Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. Can large language models reason about medical questions? *Patterns*, 5:100943, 2024.

[159] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2016.

[160] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[161] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[162] Haiping Lu, Konstantinos N. Plataniotis, et al. MPCA: Multilinear principal component analysis of tensor objects. *TNN*, 19:18, 2008.

[163] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[164] Jiaying Lu and Jinho D Choi. Evaluation of unsupervised entity and event salience estimation. In *FLAIRS*, 2021.

[165] Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. Quark: Controllable text generation with reinforced unlearning. *NeurIPS*, 2022.

[166] Yuxing Lu, Xukai Zhao, and Jinzhuo Wang. Medical knowledge-enhanced

prompt learning for diagnosis classification from clinical text. In *Clinical Natural Language Processing Workshop*, pages 278–288, 2023.

[167] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *CVPR*, 2022.

[168] Dongsheng Luo, Wei Cheng, et al. Parameterized explainer for graph neural network. In *NeurIPS*, 2020.

[169] Qiong Ma, Xiudong Shi, et al. Hiv-associated structural and functional brain alterations in homosexual males. *Frontiers in Neurology*, 12:757374, 2021.

[170] Luigi A Maglanoc, Tobias Kaufmann, Rune Jonassen, Eva Hilland, Dani Beck, Nils Inge Landrø, and Lars T Westlye. Multimodal fusion of structural and functional brain imaging in depression using linked independent component analysis. *Hum Brain Mapp*, 41:241–255, 2020.

[171] R. Manmatha, Chao-Yuan Wu, Alexander J. Smola, and Philipp Krähenbühl. Sampling matters in deep embedding learning. In *ICCV*, 2017.

[172] Niklas Mannhardt, Elizabeth Bondi-Kelly, Barbara Lam, Chloe O'Connell, Mercy Asiedu, Hussein Mozannar, Monica Agrawal, Alejandro Buendia, Tatiana Urman, Irbaz B. Riaz, Catherine E. Ricciardi, Marzyeh Ghassemi, and David Sontag. Impact of large language model assistance on patients reading clinical notes: A mixed-methods study, 2024.

[173] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *ACL*, 2014.

[174] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini

Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Prog. Neurobiol.*, 95:629–635, 2011.

[175] Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *ICLR*, 2019.

[176] Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the universality of invariant networks. In *ICML*, 2019.

[177] Gustav Martensson, Joana B Pereira, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Kłoszewska, Hilkka Soininen, Simon Lovestone, Andrew Simmons, Giovanni Volpe, et al. Stability of graph theoretical measures in structural brain networks in alzheimer's disease. *Sci. Rep.*, 8:1–15, 2018.

[178] David McClosky, Eugene Charniak, and Mark Johnson. Automatic domain adaptation for parsing. In *NAACL Linguistics*, 2010.

[179] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. In *ICLR*, 2023.

[180] Diego Mesquita, Amauri Souza, and Samuel Kaski. Rethinking pooling in graph neural networks. *NeurIPS*, 2020.

[181] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *EMNLP*, 2004.

[182] Gowtham Krishnan Murugesan, Chandan Ganesh, Sahil Nalawade, Elizabeth M Davenport, Ben Wagner, Won Hwa Kim, and Joseph A Maldjian. Brainnet: Inference of brain network topology using machine learning. *Brain Connect*, 10:422–435, 2020.

[183] Feiping Nie, Zinan Zeng, Ivor W Tsang, Dong Xu, and Changshui Zhang. Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. *IEEE Trans. Neural Netw.*, 22:1796–1808, 2011.

[184] Feiping Nie, Jing Li, Xuelong Li, et al. Parameter-free auto-weighted multiple graph learning: a framework for multiview clustering and semi-supervised classification. In *IJCAI*, pages 1881–1887, 2016.

[185] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[186] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. *ArXiv.org*, 2019.

[187] Rebecca A O'Bryan, Colleen A Brenner, et al. Disturbances of visual motion perception in bipolar disorder. *Bipolar disorders*, 16:354, 2014.

[188] Karol Osipowicz, Michael R Sperling, Ashwini D Sharan, and Joseph I Tracy. Functional mri, resting state fmri, and dti for predicting verbal fluency outcome following resective surgery for temporal lobe epilepsy. *J. Neurosurg.*, 124:929–937, 2016.

[189] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 2022.

[190] Jiaul H Paik. A novel tf-idf weighting scheme for effective ranking. In *SIGIR*, 2013.

[191] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga,

et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019.

[192] Adam Paszke, Sam Gross, et al. PyTorch: an imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[193] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Yi Guo, Jiang Bian, and Yonghui Wu. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6:210, 2023.

[194] Ronald Carl Petersen, PS Aisen, Laurel A Beckett, MC Donohue, AC Gamst, Danielle J Harvey, CR Jack, WJ Jagust, LM Shaw, AW Toga, et al. Alzheimer's disease neuroimaging initiative (adni): clinical characterization. *Neurology*, 74: 201–209, 2010.

[195] Maíra Siqueira Pinto, Roberto Paolella, Thibo Billiet, Pieter Van Dyck, Pieter-Jan Guns, Ben Jeurissen, Annemie Ribbens, Arnold J den Dekker, and Jan Sijbers. Harmonization of brain diffusion mri: Concepts and methods. *Front. Neurosci.*, 14:396, 2020.

[196] Raymond Pomponio, Guray Erus, Mohamad Habes, Jimit Doshi, Dhivya Srinivasan, Elizabeth Mamourian, Vishnu Bashyam, Ilya M Nasrallah, Theodore D Satterthwaite, Yong Fan, et al. Harmonization of large mri datasets for the analysis of brain imaging patterns throughout the lifespan. *NeuroImage*, 208: 116450, 2020.

[197] Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann,

Fran M Miezin, Bradley L Schlaggar, et al. Functional Network Organization of the Human Brain. *Neuron*, 72:665–678, 2011.

[198] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *ECCV*, 2020.

[199] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[200] Ann B Ragin, Hongyan Du, et al. Structural brain alterations can be detected early in hiv infection. *Neurology*, 79:2328, 2012.

[201] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1: 1–10, 2018.

[202] Eric A Reavis, Junghee Lee, et al. Structural and functional connectivity of visual cortex in schizophrenia and bipolar disorder: a graph-theoretic analysis. *Schizophrenia bulletin open*, 1:sgaa056, 2020.

[203] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 2015.

[204] Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen M. Voorhees, Lucy Lu Wang, and William R. Hersh. Searching for scientific evidence in a pandemic: An overview of TREC-COVID. *J. Biomed. Informatics*, 121:103865, 2021.

[205] S. Robertson, S. Walker, Susan Jones, M. Hancock-Beaulieu, and Mike Gatford. Okapi at trec-3. In *TREC*, 1994.

[206] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, 1: 1–20, 2010.

[207] Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *CVPR*, 2015.

[208] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*, 2019.

[209] Saman Sarraf and Jian Sun. Functional brain imaging: A comprehensive survey. *ArXiv.org*, 2016.

[210] Theodore D Satterthwaite, Daniel H Wolf, David R Roalf, Kosha Ruparel, Guray Erus, Simon Vandekar, Efstathios D Gennatas, Mark A Elliott, Alex Smith, Hakon Hakonarson, et al. Linked sex differences in cognition and functional connectivity in youth. *Cereb. Cortex*, 25:2383–2394, 2015.

[211] Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020.

[212] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *ESWC*, 2018.

[213] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[214] Weixiang Shao, Lifang He, and Philip S. Yu. Clustering on multi-source incomplete data via tensor modeling and factorization. In *PAKDD*, 2015.

[215] Ran Shi and Ying Guo. Investigating differences in brain functional networks using hierarchical covariate-adjusted independent component analysis. *Ann Appl Stat*, page 1930, 2016.

[216] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22: 1589–1604, 2017.

[217] William R Shirer, Srikanth Ryali, et al. Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cerebral cortex*, 22:158, 2012.

[218] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *CVPR*, 2022.

[219] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620: 172–180, 2023.

[220] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023.

[221] Sonish Sivarajkumar and Yanshan Wang. Healthprompt: A zero-shot learning paradigm for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2022, page 972, 2022.

[222] Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing. *arXiv preprint arXiv:2309.08008*, 2023.

[223] Stephen M Smith. The future of fmri connectivity. *NeuroImage*, 62:1257–1266, 2012.

[224] Stephen M Smith, Peter T Fox, Karla L Miller, David C Glahn, P Mickle Fox, Clare E Mackay, Nicola Filippini, Kate E Watkins, Roberto Toro, Angela R Laird, et al. Correspondence of the brain's functional architecture during activation and rest. *Proc. Natl. Acad. Sci. U.S.A.*, 106:13040–13045, 2009.

[225] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *NeuroImage*, 54:875–891, 2011.

[226] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017.

[227] Olaf Sporns. Graph theory methods: applications in brain networks. *Dialogues Clin. Neurosci.*, 20:111–121, 2022.

[228] Chang Su, Zhenxing Xu, Jyotishman Pathak, and Fei Wang. Deep learning in mental health outcome research: a scoping review. *Transl. Psychiatry*, 10:1–26, 2020.

[229] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020.

[230] Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. A contrastive framework for neural text generation. In *NeurIPS*, 2022.

[231] Yashar Talebirad and Amirhossein Nadiri. Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*, 2023.

[232] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[233] Haoteng Tang, Lei Guo, Xiyao Fu, Benjamin Qu, Olusola Ajilore, Yalin Wang, Paul M Thompson, Heng Huang, Alex D Leow, and Liang Zhan. A hierarchical graph learning model for brain network regression analysis. *Frontiers in Neuroscience*, 16:1–5, 2022.

[234] Haoteng Tang, Lei Guo, Xiyao Fu, Benjamin Qu, Paul M Thompson, Heng Huang, and Liang Zhan. Hierarchical brain embedding using explainable graph learning. In *ISBI*, pages 1–5, 2022.

[235] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020.

[236] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[237] Komal Teru, Etienne Denis, and Will Hamilton. Inductive relation prediction by subgraph reasoning. In *ICML*, 2020.

[238] Alessandro Tessitore et al. Default-mode network connectivity in cognitively unimpaired patients with parkinson disease. *Neurology*, 79:2226, 2012.

[239] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[240] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[241] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[242] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *NeurIPS*, 2021.

[243] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1:AIoa2300138, 2024.

[244] Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 15: 273–289, 2002.

[245] Thilo van Eimeren, Oury Monchi, et al. Dysfunction of the default mode network in parkinson disease: a functional magnetic resonance imaging study. *Archives of neurology*, 66:877, 2009.

[246] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The human connectome project: a data acquisition perspective. *NeuroImage*, 62:2222–2231, 2012.

[247] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, pages 1–9, 2024.

[248] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[249] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

[250] Petar Veličković, William Fedus, et al. Deep graph infomax. In *ICLR*, 2019.

[251] Petar Veličković et al. Graph attention networks. In *ICLR*, 2018.

[252] Minh N. Vu and My T. Thai. Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. In *NeurIPS*, 2020.

[253] Chenguang Wang, Xiao Liu, and Dawn Song. Language models are open knowledge graphs. *arXiv preprint arXiv:2010.11967*, 2020.

[254] Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. Clinicalgpt: large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv preprint arXiv:2306.09968*, 2023.

[255] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, et al. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL*, 2020.

[256] Meng Wang, Sen Wang, Han Yang, Zheng Zhang, Xi Chen, and Guilin Qi. Is visual context really helpful for knowledge graph? a representation learning perspective. In *ACM MM*, 2021.

[257] Qifan Wang, Li Yang, Bhargav Kanagal, Sumit Sanghai, D Sivakumar, Bin Shu, Zac Yu, and Jon Elsas. Learning to extract attribute value from product via question answering: A multi-task approach. In *KDD*, 2020.

[258] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[259] Xu Wang, Chen Yang, and Renchu Guan. A comparative study for biomedical named entity recognition. *International Journal of Machine Learning and Cybernetics*, 9:373–382, 2018.

[260] Yikai Wang and Ying Guo. A hierarchical independent component analysis model for longitudinal neuroimaging studies. *NeuroImage*, 189:380–400, 2019.

[261] Yikai Wang, Jian Kang, Phebe B Kemmer, and Ying Guo. An efficient and reliable statistical method for estimating functional connectivity in large scale brain networks using partial correlation. *Front. Neurosci.*, 10:123, 2016.

[262] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[263] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. 2022.

[264] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[265] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 35, 2022.

[266] Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training, 2019.

[267] Leanne M Williams. Precision psychiatry: a neural circuit taxonomy for depression and anxiety. *Lancet Psychiatry*, 3:472–480, 2016.

[268] Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *NeurIPS*, 2023.

[269] Jionglin Wu, Jason Roy, and Walter F Stewart. Prediction modeling using ehr data: challenges, strategies, and a comparison of machine learning approaches. *Medical care*, 48:S106–S113, 2010.

[270] Qiang Wu, Christopher J. C. Burges, Krysta M. Svore, and Jianfeng Gao.

Adapting boosting for information retrieval measures. *Inf. Retr.*, 13:254–270, 2010.

[271] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155*, 2023.

[272] Shu Wu, Yuyuan Tang, Yanqiao Zhu, Liang Wang, Xing Xie, and Tieniu Tan. Session-based recommendation with graph neural networks. In *AAAI*, volume 33, pages 346–353, 2019.

[273] Mingrui Xia, Jinhui Wang, and Yong He. Brainnet viewer: a network visualization tool for human brain connectomics. *PloS one*, 8:e68910, 2013.

[274] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25: 1419–1428, 2018.

[275] Kai Yuanqing Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. In *ICLR*, 2021.

[276] Danfei Xu, Yuke Zhu, Christopher Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.

[277] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017.

[278] Huimin Xu, Wenting Wang, Xinnian Mao, Xinyu Jiang, and Man Lan. Scaling up open tagging from tens to thousands: Comprehension empowered attribute

value extraction from product title. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5214–5223, 2019.

[279] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.

[280] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *ICLR*, 2019.

[281] Ran Xu, Yue Yu, Chao Zhang, Mohammed K Ali, Joyce C Ho, and Carl Yang. Counterfactual and factual reasoning over hypergraphs for interpretable clinical predictions on ehr. In *Machine Learning for Health*, 2022.

[282] Ran Xu, Yue Yu, Hejie Cui, Xuan Kan, Yanqiao Zhu, Joyce Ho, Chao Zhang, and Carl Yang. Neighborhood-regularized self-training for learning with few labels. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[283] Ran Xu, Yue Yu, Joyce Ho, and Carl Yang. Weakly-supervised scientific document classification via retrieval-augmented multi-stage training. In *SIGIR*, pages 2501–2505, 2023.

[284] Noriaki Yahata, Jun Morimoto, Ryuichiro Hashimoto, Giuseppe Lisi, Kazuhisa Shibata, Yuki Kawakubo, Hitoshi Kuwabara, Miho Kuroda, Takashi Yamada, Fukuda Megumi, et al. A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nat. Commun.*, 7:1–12, 2016.

[285] Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, Naohiro Okada, et al. Harmonization of resting-state functional mri data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias. *PLOS Biol.*, 17:e3000042, 2019.

[286] Jun Yan, Nasser Zalmout, Yan Liang, Christan Grant, Xiang Ren, and Xin Luna Dong. Adatag: Multi-attribute value extraction from product profiles with adaptive decoding. In *ACL*, 2021.

[287] Carl Yang, Peiye Zhuang, Wenhan Shi, Alan Luu, and Pan Li. Conditional structure generation through graph variational generative adversarial nets. In *NeurIPS*, 2019.

[288] Carl Yang, Aditya Pal, Andrew Zhai, Nikil Pancha, Jiawei Han, Charles Rosenberg, and Jure Leskovec. Multisage: Empowering GCN with contextualized multi-embeddings on web-scale multipartite networks. In *KDD*, 2020.

[289] Carl Yang, Jieyu Zhang, Haonan Wang, Bangzheng Li, and Jiawei Han. Neural concept map generation for effective document classification with interpretable structured summarization. In *SIGIR*, 2020.

[290] Carl Yang, Jieyu Zhang, Haonan Wang, Sha Li, Myungwan Kim, Matt Walker, Yiou Xiao, and Jiawei Han. Relation learning on social networks with multi-modal graph edge variational autoencoders. In *WSDM*, 2020.

[291] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. A large language model for electronic health records. *npj Digital Medicine*, page 194.

[292] Yi Yang, Yanqiao Zhu, Hejie Cui, Xuan Kan, Lifang He, Ying Guo, and Carl Yang. Data-efficient brain connectome analysis via multi-task meta-learning. *KDD*, pages 4743–4751, 2022.

[293] Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. Generative data augmentation for commonsense reasoning. In *Findings of EMNLP*, 2020.

[294] Zeynep Akkalyoncu Yilmaz, Shengjin Wang, Wei Yang, Haotian Zhang, and Jimmy Lin. Applying BERT to document retrieval with birch. In *EMNLP*, 2019.

[295] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *KDD*, 2018.

[296] Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling. In *NeurIPS*, 2018.

[297] Zhitao Ying, Dylan Bourgeois, et al. Gnnexplainer: Generating explanations for graph neural networks. In *NeurIPS*, 2019.

[298] Jiaxuan You, Rex Ying, and Jure Leskovec. Position-aware graph neural networks. In *ICML*, 2019.

[299] Juntao Yu, Mohab El-karef, and Bernd Bohnet. Domain adaptation for dependency parsing via self-training. In *Proceedings of the 14th International Conference on Parsing Technologies*, 2015.

[300] Renping Yu, Lishan Qiao, Mingming Chen, Seong-Whan Lee, Xuan Fei, and Dinggang Shen. Weighted graph regularized sparse brain network construction for mci identification. *Pattern Recognit*, 90:220–231, 2019.

[301] Yue Yu, Kexin Huang, Chao Zhang, Lucas M Glass, Jimeng Sun, and Cao Xiao. Sumgnn: multi-typed drug interaction prediction via efficient knowledge graph summarization. *Bioinformatics*, 2021.

[302] Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In *NAACL*, 2021.

[303] Yue Yu, Rongzhi Zhang, Ran Xu, Jieyu Zhang, Jiaming Shen, and Chao Zhang. Cold-start data selection for few-shot language model fine-tuning: A prompt-based uncertainty propagation approach. *arXiv preprint arXiv:2209.06995*, 2022.

[304] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias. *arXiv preprint arXiv:2306.15895*, 2023.

[305] Hao Yuan, Haiyang Yu, et al. Explainability in graph neural networks: A taxonomic survey. *CoRR*, abs/2012.15445, 2020.

[306] Jiayi Yuan, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. Large language models for healthcare data augmentation: An example on patient-trial matching. In *AMIA Annual Symposium Proceedings*, volume 2023, page 1324, 2023.

[307] Seongjun Yun, Minbyul Jeong, et al. Graph transformer networks. *NeurIPS*, 2019.

[308] Nasser Zalmout and Xian Li. Prototype-representations for training data filtering in weakly-supervised information extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[309] Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Bridging knowledge graphs to generate scene graphs. In *ECCV*, 2020.

[310] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *NeurIPS*, 2022.

[311] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *CVPR*, 2018.

[312] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019.

[313] Liang Zhan, Yashu Liu, Yalin Wang, Jiayu Zhou, Neda Jahanshad, Jieping Ye, Paul M Thompson, and Alzheimer's Disease Neuroimaging Initiative (ADNI). Boosting brain connectome classification accuracy in alzheimer's disease using higher-order singular value decomposition. *Frontiers in neuroscience*, 9:257, 2015.

[314] Liang Zhan, Jiayu Zhou, Yalin Wang, Yan Jin, Neda Jahanshad, Gautam Prasad, Talia M Nir, Cassandra D Leonardo, Jieping Ye, Paul M Thompson, et al. Comparison of nine tractography algorithms for detecting abnormal structural brain networks in alzheimer's disease. *Front. Aging Neurosci.*, 7:48, 2015.

[315] Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. Aser: A large-scale eventuality knowledge graph. In *WWW*, 2020.

[316] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *CVPR*, 2020.

[317] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a

multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[318] Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6:1–9, 2019.

[319] Yilin Zhang and Karl Rohe. Understanding regularized spectral clustering via graph conductance. In *NeurIPS*, 2018.

[320] Yufeng Zhang, Jinghao Zhang, Zeyu Cui, Shu Wu, and Liang Wang. A graph-based relevance matching model for ad-hoc retrieval. In *AAAI*, 2021.

[321] Zhiqiang Zhang, Linan Wang, Xiaoqin Xie, and Haiwei Pan. A graph based document retrieval method. In *CSCWD*, pages 426–432, 2018.

[322] Guineng Zheng, Subhabrata Mukherjee, Xin Luna Dong, and Feifei Li. Opentag: Open attribute value extraction from product profiles. In *KDD*, 2018.

[323] Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *ECCV*, 2020.

[324] Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Prompt consistency for zero-shot task generalization. In *Findings of EMNLP*, 2022.

[325] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022.

[326] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[327] Qi Zhu, Huijie Li, Jiashuang Huang, Xijia Xu, Donghai Guan, and Daoqiang Zhang. Hybrid functional brain network with first-order and second-order information for computer-aided diagnosis of schizophrenia. *Front. Neurosci.*, page 603, 2019.

[328] Tiangang Zhu, Yue Wang, Haoran Li, Youzheng Wu, Xiaodong He, and Bowen Zhou. Multimodal joint attribute prediction and value extraction for e-commerce product. In *EMNLP*, 2020.

[329] Yanqiao Zhu, Hejie Cui, Lifang He, Lichao Sun, and Carl Yang. Joint embedding of structural and functional brain networks with graph neural networks for mental illness diagnosis. In *Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022.

[330] Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In *ICML*, 2022.

[331] Joelle Zimmermann, John D Griffiths, and Anthony R McIntosh. Unique mapping of structural and functional connectivity on cognition. *J. Neurosci.*, 38: 9658–9667, 2018.