



THE HONG KONG
POLYTECHNIC UNIVERSITY

香港理工大學

Pao Yue-kong Library

包玉剛圖書館

Copyright Undertaking

This thesis is protected by copyright, with all rights reserved.

By reading and using the thesis, the reader understands and agrees to the following terms:

1. The reader will abide by the rules and legal ordinances governing copyright regarding the use of the thesis.
2. The reader will use the thesis for the purpose of research or private study only and not for distribution or further reproduction or any other purpose.
3. The reader agrees to indemnify and hold the University harmless from and against any loss, damage, cost, liability or expenses arising from copyright infringement or unauthorized usage.

IMPORTANT

If you have reasons to believe that any materials in this thesis are deemed not suitable to be distributed in this form, or a copyright owner having difficulty with the material being included in our database, please contact lbsys@polyu.edu.hk providing details. The Library will look into your claim and consider taking remedial action upon receipt of the written requests.

Pao Yue-kong Library, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

<http://www.lib.polyu.edu.hk>

ADVANCING SAFE, MULTIMODAL
CONVERSATIONAL AI:
OUT-OF-DISTRIBUTION DETECTION
AND MEDICAL VISUAL QUESTION
ANSWERING APPROACHES

BO LIU

PhD

The Hong Kong Polytechnic University

2025

The Hong Kong Polytechnic University
Department of Data Science and Artificial Intelligence

Advancing Safe, Multimodal
Conversational AI:
Out-of-Distribution Detection and
Medical Visual Question Answering
Approaches

Bo Liu

A thesis submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy

April 2025

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____ (Signed)

Liu Bo

_____ (Name of student)

Abstract

Conversational AI has seen remarkable progress in recent years, driven by the integration of large language models (LLMs) and multimodal learning. However, ensuring the robustness and usability of conversational systems remains a major challenge, particularly in high-stakes fields like medical applications, where incorrect or misleading responses can have serious consequences. Two crucial factors for improving these systems are out-of-distribution (OOD) detection that recognizes unfamiliar inputs and domain-specific multimodal understanding that integrates diverse data types. This thesis aims to advance both areas by (1) enhancing textual OOD detection with LLMs to better secure dialogue systems and (2) developing approaches to improve understanding and reasoning for medical visual question answering (Med-VQA), a key task in medical dialogues.

For textual OOD detection, we first conduct a pioneering empirical study on OOD detection in LLMs, addressing the gap in existing methods designed for smaller models like BERT, which may not generalize well to LLMs. We evaluate OOD detectors in both zero-shot and fine-tuning settings and propose a generative fine-tuning approach aligned with LLM pre-training objectives. Our results show that the cosine distance-based detector outperforms other ones, leveraging LLMs' isotropic embedding space. Next, we introduce a novel framework to tackle near-OOD detection, where in-distribution (ID)

and OOD inputs share semantic similarities, by leveraging the isotropic embedding space of LLMs. Our framework derives semantic prototypes for each ID class and performs semantic matching for both OOD detection and ID classification. With high-quality textual representations from LLMs, our method demonstrates superior performance, especially in few-shot scenarios with limited data.

For Med-VQA, we first introduce SLAKE, a semantically-labeled knowledge-enhanced dataset with accurate visual and textual annotations and an extendable knowledge base, to overcome the limitations of dataset scarcity. To further mitigate overfitting to small-scale training data, we propose a CPRD framework that distills a lightweight visual feature extractor with various radiological knowledge for Med-VQA. Secondly, when faced with more complex medical questions and images compared to general ones, we propose a conditional reasoning framework that consists of a question-conditioned reasoning component and a type-conditioned reasoning strategy to adaptively learn reasoning skills for different Med-VQA tasks. Finally, we present GEMeX, a large-scale, groundable, and explainable benchmark for chest X-ray diagnosis. This new benchmark addresses key limitations of existing datasets by introducing a multi-modal explainability that enhances answer comprehensibility and four distinct question types that better reflect clinical needs. Our evaluations of 12 representative large vision language models and a fine-tuned baseline model demonstrate the dataset’s challenges and effectiveness.

The results have been published in top AI conferences and journals, including ACMMM-2020 [1], ISBI-21 [2], MICCAI-2021 [3], IEEE TMI-2022 [4], COLING-2024 [5], and ICCV 2025 [6].

Publications

Publications Arising from the Thesis

- Li-Ming Zhan*, **Bo Liu***, Lu Fan, Jiabin Chen, and Xiao-Ming Wu. Medical Visual Question Answering via Conditional Reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2345-2354, 2020. **(Cited 150+ times)** [1]
- **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 *IEEE 18th International Symposium on Biomedical Imaging*, pages 1650-1654, 2021. **(Cited 400+ times)** [2]
- **Bo Liu**, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive Pre-training and Representation Distillation for Medical Visual Question Answering Based on Radiology Images. In *24th International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI*, pages 210-220, 2021. **(Cited 100+ times)** [3]
- **Bo Liu**, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical Visual Question Answering via Conditional Reasoning and Contrastive Learning. In *IEEE Transactions on Medical Imaging, IEEE-TMI*, pages 1532-1545, 2022. (IF: 10.6) [4]
- **Bo Liu**, Li-Ming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. How Good Are LLMs at Out-of-Distribution Detection? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 8211-8222, 2024. [7]

*Equal contribution.

- **Bo Liu**, Liming Zhan, Yujie Feng, Zexin Lu, Chengqiang Xie, Lei Xue, Albert Lam, Xiao-Ming Wu. Diversity-grounded Channel Prototypical Learning for Out-of-Distribution Intent Detection. *arXiv 2024*. [8]
- **Bo Liu**, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. GEMeX: A Large-Scale, Groundable, and Explainable Medical VQA Benchmark for Chest X-ray Diagnosis. *Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025*. [6]

Other Publications during My PhD Study

- Lu Fan, Qimai Li, **Bo Liu**, Xiao-Ming Wu, Xiaotong Zhang, Fuyu Lv, Guli Lin, Sen Li, Taiwei Jin, and Keping Yang. Modeling User Behavior with Graph Convolution for Personalized Product Search. In *Proceedings of the ACM Web Conference 2022*, pages 203-212, 2022.
- Li Xu, **Bo Liu**, Ameer Hamza Khan, Lu Fan, and Xiao-Ming Wu. Multi-modal Pre-training for Medical Vision-Language Understanding and Generation: An Empirical Study with A New Benchmark. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 117-132, 2023.
- Yujie Feng, Zexin Lu, **Bo Liu**, Liming Zhan, Xiao-Ming Wu. Towards LLM-driven Dialogue State Tracking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing EMNLP*, pages, 2023.
- Li-Ming Zhan, **Bo Liu**, and Xiao-Ming Wu. VI-OOD: A Unified Framework of Representation Learning for Textual Out-of-distribution Detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, pages 8211-8222, 2024.
- Xiangyu Zhao, **Bo Liu**, Qijiong Liu, Guangyuan Shi, and Xiao-Ming Wu. Easy-Gen: Easing Multimodal Generation with BiDiffuser and LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages, 1351-1370, 2024.

*equal contribution.

Acknowledgement

First, I would like to express my deepest gratitude to my Ph.D. advisor, Prof. Xiao-Ming Wu, for her invaluable guidance, continuous support, and encouragement throughout my doctoral journey. Her insightful advice, patience, and dedication have been instrumental in shaping my research and academic development. Her unwavering belief in my abilities has been a source of motivation, and I am truly grateful for the opportunity to work under her supervision.

Second, I would also like to extend my sincere appreciation to the physicians who have generously shared their expertise and provided essential support for my studies. In particular, I would like to thank Prof. Lin Ma, Prof. Yan Yang, and Prof. YiDi Chen for their invaluable guidance in medical knowledge and their assistance in providing crucial data, which have been fundamental to my research. Additionally, I am deeply grateful to my external advisor, Dr. Huazhu Fu, for his mentorship and constructive feedback, especially during the final year of my Ph.D journey. His support and insightful advice have significantly contributed to the completion of my research.

I am also grateful to my fellow lab members for their companionship, collaboration, and thought-provoking discussions. In particular, I would like to thank Li-Ming Zhan, Lu Fan, Li Xu, Zexin Lu, Yujie Feng, Xiaoyu Dong, and Xiangyu Zhao for their support,

insightful discussions, and shared experiences throughout my research journey. Their encouragement has made my research journey more enjoyable and intellectually stimulating. The friendships and collaborations in the lab have been a cherished part of my Ph.D. life.

Finally, I would like to express my heartfelt gratitude to my family for their unconditional love, understanding, and unwavering support. Their encouragement has been my greatest source of strength, and I could not have accomplished this milestone without them.

Contents

Abstract	i
Publications	iii
Acknowledgement	v
List of Figures	xv
List of Tables	xx
1 Introduction	1
1.1 Challenges	3
1.1.1 Out-of-distribution Detection	3
1.1.2 Medical Visual Question Answering	4
1.2 Contributions	6
1.2.1 Out-of-distribution Detection	7
1.2.2 Medical Visual Question Answering	7
2 Background and Related Work	11
2.1 Textual Out-of-distribution Detection	11
2.2 Medical Visual Question Answering	13
2.3 Visual Question Answering	17
2.4 Medical Vision-Language Tasks	19
3 OOD Detection with LLMs	21
3.1 Introduction	21
3.2 Method	24

3.2.1	ID Generative Fine-tuning with LLMs	25
3.2.2	OOD Detection with LLMs	26
3.3	Experimental Setup	28
3.3.1	Datasets	28
3.3.2	Evaluation Metrics	29
3.3.3	Implementation Details	29
3.4	Findings	30
3.4.1	Zero-grad OOD Detection with LLMs	30
3.4.2	OOD Detection with Generatively Fine-tuned LLMs	33
4	Few-shot Near-OOD Detection	41
4.1	Introduction	41
4.2	Method	43
4.2.1	Semantic Matching as ID Classification	43
4.2.2	Post-hoc OOD Scoring Functions	45
4.3	Experiment	46
4.3.1	Experimental Setup	46
4.3.2	Main Results	48
4.3.3	Analysis and Ablation Study	48
4.4	Limitation	50
5	Overcoming Small-Scale Data	53
5.1	Introduction	53
5.2	SLAKE Construction	57
5.2.1	Image Acquisition and Annotation	58
5.2.2	Knowledge Graph Construction	58

<i>CONTENTS</i>	ix
5.2.3 Question Generation	60
5.2.4 Dataset Splitting	61
5.3 Contrastive Pre-training and Representation Distillation (CPRD)	62
5.3.1 Teachers: Intra-region Contrastive Pre-training	62
5.3.2 Student: Inter-region Representation Distillation	64
5.3.3 Applying CPRD for Med-VQA	65
5.4 Experiments	67
5.4.1 Evaluation of SLAKE	67
5.4.2 Evaluation of CPRD	69
6 Improving Reasoning Abilities	77
6.1 Introduction	77
6.2 Methodology	82
6.2.1 The paradigm of Med-VQA	83
6.2.2 Contrastive Pre-training (CP)	84
6.2.3 Conditional Reasoning (CR)	87
6.2.4 Proposed Med-VQA Model	92
6.3 Experiments	93
6.3.1 Datasets	94
6.3.2 Implementation Details	94
6.3.3 Comparison with the State-of-the-arts	96
6.3.4 Ablation Study and Analysis	100
6.3.5 Qualitative Evaluation	107
6.3.6 Efficiency Evaluation	109
7 Introducing Multimodal Explainability	113

7.1	Introduction	113
7.2	Construction of GEMeX	116
7.2.1	Re-grounding Reports	117
7.2.2	Groundable and Explainable VQA Generation	121
7.3	Statistics of GEMeX	123
7.4	Evaluation of GEMeX	125
7.4.1	Experimental Details	125
7.4.2	Evaluation Metrics	127
7.4.3	Results and Analysis	129
7.4.4	Qualitative Evaluation	133
8	Conclusion and Future Works	137

List of Figures

1.1	Research accomplishments	6
3.1	Illustration of two types of OOD instances compared to ID samples: far-OOD where ID and OOD data come from different domains and near-OOD where ID and OOD data come from the same domain but with different classes.	22
3.2	Our proposed evaluation framework for LLMs at OOD detection, taking three aspects into consideration: ① distribution of OOD samples (near or far), ② impact of model tuning on OOD detection, and ③ diverse OOD detectors (score functions).	26
3.3	UMAP [9] visualization of representations generated from the penulti- mate layer of the zero-grad (snowflake) and fine-tuned (flame) LLaMA- 7B models. (a) Far-OOD: 20NG is treated as ID while SST-2, RTE, and TREC are treated as OOD. (b) Near-OOD: the banking domain of CLINC150 is selected, of which 50% of the classes are treated as ID, and the rest are treated as OOD.	31
3.4	Impact of fine-tuning on logits-based OOD scores (MSP at the top row and Energy at the bottom row). We plot SST-2 (ID) vs. TREC-10 (OOD) for visualization.	32

3.5	Performance changes for ID classification and OOD detection as training progresses with the different number of training shots. Top row: 20NG is ID training task; Bottom row: banking domain of CLINC150 is selected where 50% classes are used as ID training task and the rest are OOD samples. The \star indicates the selected results whose epoch corresponds to the best ID performance on the validation set.	34
3.6	Impact of different ID training objectives, discriminative vs. generative. SST-2 dataset with full data is used as the ID training task.	35
4.1	Proposed semantic matching framework. We prompt class name into a sequence of learnable tokens and forward them into LLMs to generate class prototypes. Further training between prototypes and input representations with matching loss and diversity loss, better ID classification and OOD detection can be performed.	42
4.2	Performance of various class prototypes in the 5-shot scenario using CLINC-Banking dataset.	49
4.3	UMAP [9] visualization of representations of test set and OOD data from 5-shot CLINC-Bank. The purple means the OOD data. The star indicates the learned prototype of each class.	50
5.1	Exemplar image and questions of our SLAKE dataset.	55
5.2	Left: proportions of images of five body parts. Right: distribution of the content types of questions.	57

5.3	Our proposed CPRD framework for Med-VQA. (a) Train a teacher model T_θ by self-supervised contrastive learning on the chest region. (b) Distill three teacher models into one student model S_ϕ . (c) Apply the student model S_ϕ for Med-VQA.	63
5.4	The Med-VQA framework on our SLAKE dataset.	66
5.5	(Left) t-SNE visualization of the representations learned by the student model; (Right) Grad-CAM maps from the visual modules of Med-VQA methods. MEVF+BAN is the baseline. ✓ and ✗ indicate the correctness of the answer given by each method.	74
6.1	Examples of Med-VQA tasks. For closed-ended questions, the answers are limited, e.g., “yes” or “no”. For open-ended questions, the answers can be free-form text.	78
6.2	Our proposed method for training a Med-VQA model. In stage I, we pre-train a visual feature extractor for Med-VQA by contrastive self-supervised learning. In stage II, we solve Med-VQA tasks by introducing a conditional reasoning mechanism.	82
6.3	Our proposed Med-VQA model with conditional reasoning. To prevent overfitting, we freeze the visual model F_θ (pre-trained in stage I) and append a non-linear layer P_ξ for fine-tuning on medical images. First, the TCR module classifies the question as open-ended or closed-ended and chooses the corresponding branch for reasoning. Then, the question features extracted by the textual model Q_ψ will be fused with the image features by the multimodal feature fusion module (e.g., BAN) and our QCR module. Finally, the answer is obtained by an MLP classifier.	86

- 6.4 t-SNE [10] visualization of the multimodal features (input to the classifier layer) of Med-VQA tasks in the test set of VQA-RAD [11] learned by CP+BAN(2) and CP+BAN(2-1) respectively. The TCR module in CP+BAN(2-1) disentangles the representations of open-ended and closed-ended tasks. 106
- 6.5 Visual comparison of the prediction results for an open-ended task in VQA-RAD dataset by variants of CP+BAN. The Grad-CAM maps [12] of the visual model are plotted based on the predicted answers. ✓ and ✗ indicate the correctness of the answer given by each model. 106
- 6.6 Impact of the prediction accuracy of the type classifier on our model CP+BAN+CR. Note that the prediction accuracy of our model refers to the overall metric. 107
- 6.7 The Grad-CAM maps of the visual modules of our methods and baseline MEVF+BAN. The attention map of our QCR module is displayed in the right column, and darker color indicates higher weight. ✓ and ✗ indicate the correctness of the answer given by each model. † indicates the test image comes from SLAKE [2], and ‡ indicates it comes from VQA-RAD [11]. The last row shows a failure case of our method with the conditional reasoning module, which is caused by the misclassification of question type. 108

6.8 Time efficiency of the proposed conditional reasoning mechanism (i.e., TCR and QCR modules). BASE represents the base model MEVF+BAN [13]. BASE+QCR does not differentiate the question type, and hence there is only one reasoning module. Our QCR module is imposed on the basic reasoning module BAN to enhance reasoning ability. BASE+TCR differentiates the question type and chooses different reasoning module correspondingly. However, only the basic reasoning module BAN is used for reasoning, without our QCR enhancement. Deep Blue denotes training time (seconds) per epoch. Cyan denotes test time (seconds) per epoch. 110

7.1 Our GEMeX stands out from existing medical VQA datasets by providing diverse question types and comprehensive multimodal explanations, including textual reasoning and visual grounding, to improve real-world applicability. 114

7.2 Illustration of the pipeline for constructing our GEMeX, with two main stages. In Stage I (left), medical LLM performs re-grounding on the original reports based on the pathological regions and clinical guidance specified by the radiologists, generating more precise sentence-region correspondence. In Stage II (right), the well-crafted prompt enables GPT-4o to generate a high-quality, large-scale Med-VQA dataset with both textual and visual explanations, leveraging the re-grounded reports. 117

7.3 The distribution of normality and abnormality contained in images from the test set of our GEMeX. 124

7.4 The distribution of question content in our GEMeX. 125

List of Tables

2.1	Comparison of medical VQA Datasets. † indicates a composition of multiple body parts (e.g., head, chest, abdomen) and various imaging modalities (e.g., CT, MRI, X-ray, pathology). In the # Question Types‡ column, O., C., S., and M. represent “Open-ended”, “Closed-ended”, “Single-choice”, and “Multi-choice”, respectively.	16
3.1	Batch size configuration for each dataset.	29
3.2	OOD detection performance of zero-grad LLaMA models. We use the full validation set to calculate each OOD score. The results are averaged over five seeds.	30
3.3	The performance of the fine-tuned LLaMA-7B model for OOD detection and ID classification. “Shot” denotes the number of examples in the ID training or validation set. We report the average results of five seeds.	33
3.4	Comparison of large and small PTMs under zero-grad and fine-tuned settings for OOD detection. † denotes the results we reproduce due to different calculating methods, while ‡ indicates results cited from the original paper.	36
3.5	Average sentence anisotropy of model’s last layer.	38
4.1	Data composition of each dataset.	46

4.2	The performance of LLaMA-7B fine-tuned with different methods for OOD detection and ID classification. “Shot” denotes the number of examples in the ID training or validation set. We report the average results of five seeds. † is cited from the original paper. AUROC, FAR@95, and AUPR are metrics for OOD detection.	47
4.3	Effectiveness of Diversity Learning.	50
5.1	Comparison of SLAKE with VQA-RAD.	55
5.2	Examples of our medical knowledge graph.	59
5.3	Statistics of questions in our SLAKE dataset.	60
5.4	Accuracy for vision-only questions (%).	68
5.5	Accuracy for knowledge-based questions (%).	68
5.6	Test accuracy of our method and baselines.	71
5.7	Comparison of different visual modules in test accuracy and model size on VQA-RAD [11]. The number of parameters is calculated on the visual module only.	73
6.1	Med-VQA Dataset statistics	93
6.2	Test accuracy of methods on VQA-RAD [11] and SLAKE [2]. “Fw.” is the abbreviation of “framework”. * means results cited from the original papers.	98
6.3	Comparison of different visual modules on VQA-RAD [11]. † indicates pre-training on ImageNet with standard supervised classification. ‡ indicates pre-training on ImageNet with contrastive self-supervised learning (MoCo-v2 [14]).	100

6.4	Comparison of performance on VQA-RAD [11] by using different datasets for contrastive pre-training (CP in Section 6.2.2).	102
6.5	Ablation study of our proposed conditional reasoning (CR) mechanism on VQA-RAD [11].	103
6.6	Effect of the number of reasoning steps for open-ended and closed-ended questions in VQA-RAD [11].	105
7.1	Anatomical regions transformation from the Chest ImaGenome to our refined version with detailed operations.	118
7.2	Our proposed prompt for refining sentence-region pairs.	120
7.3	Our designed prompt for generating groundable and explainable medical VQA, using a grounded report as input.	122
7.4	Distribution statistics of question types (T) and number of bounding boxes (B) across data splits.	123
7.5	Input format for fine-tuning LLaVA-Med.	126
7.6	Performance of representative LVLMs on our GEMeX across four question types. The AR-score combines answer and reason to evaluate textual outputs. † denotes the GPTScore value (%). The A-score indicates answer or choice accuracy (%), and the V-score represents mIoU (%). The best results are bolded, and the second-best are underlined in each column.	129
7.7	Performance of representative LVLMs evaluated using various natural language generation metrics for AR-score, including BERTScore, ROUGE-L, and BLEU-1. The best results are bolded, and the second-best are underlined in each column.	131
7.8	Performance of LVLMs on SLAKE-CXR.	133

7.9 Challenging questions from GEMeX answered by GPT-4o-mini and our LLaVA-Med-GEMeX, **highlighting the limitations of LVLMs like GPT-4o-mini in visual reasoning, which can be enhanced by integrating knowledge from GEMeX**. Colors in outputs highlight pathological descriptions that are similar to the ground truth, though they may be either correct (✓) or incorrect (✗). 135

Chapter 1

Introduction

Over the past decade, conversational AI has undergone significant breakthroughs, transforming the landscape from basic dialogue systems to sophisticated multi-modal agents. At the early stage, approaches relied on LSTM-based architectures [15], which offered limited contextual understanding and generation capabilities. The introduction of BERT [16] in 2019 marked a significant milestone, demonstrating that pre-training with a large-scale corpus could dramatically improve language understanding. This was quickly followed by a broader exploration of scaling law, which guided the design of models as parameters, data, and computing resources increased. Recently, the emergence of large language models (LLMs) such as GPT-4o [17] and Deepseek [18] have fundamentally redefined human-machine interaction. These advanced systems seamlessly integrate multiple modalities—text, images, audio, and video—within a unified conversational framework, exhibiting exceptional reasoning abilities, contextual awareness, and human-like natural language generation. This evolution represents not merely incremental improvements but a paradigm shift in conversational AI.

However, ensuring the robustness and usability of conversational systems remains a major challenge, particularly in high-stakes fields like medical applications, where incorrect or misleading responses can have serious consequences. Two critical aspects require urgent attention. First, system safety is paramount, especially regarding the detection of out-of-distribution (OOD) textual inputs. Current models often struggle to recognize when queries fall outside their training distribution or task definitions, potentially leading to hallucinated or unsafe responses instead of appropriately refusing to answer or presenting uncertainty. Second, the system’s usability depends on its ability to generate accurate answers when handling domain-specific interactions, like medical visual question answering (Med-VQA). While recent models show promising capabilities, their performance becomes less reliable when faced with complex medical scenarios that demand both visual understanding and specialized knowledge. Overall, in high-stakes applications such as medical dialogue, OOD detection enhances safety by identifying unfamiliar inputs, while Med-VQA strengthens domain-specific utility. Together, they form a complementary foundation for building safe and reliable conversational AI.

To enhance both robustness and usability, textual OOD detection and Med-VQA have received increasing research attention in recent years. OOD detection is typically formulated as a binary classification task: identifying whether an input belongs to the in-distribution (ID) set and executing the corresponding task, or recognizing it as OOD and rejecting the input [19]. Meanwhile, Med-VQA aims to answer medical vision-and-language inquiries by analyzing relevant medical images alongside textual questions and even utilizing an external knowledge graph (base) [3].

Despite ongoing progress, existing textual OOD detection methods have predominantly focused on small-scale models (e.g., BERT), and their effectiveness in the era of LLMs

remains underexplored. Furthermore, OOD detection algorithms specifically designed for LLMs are still lacking. To fill this gap, this thesis conducts a comprehensive evaluation of traditional OOD detection methods and proposes novel, LLM-based OOD detection algorithms. In the Med-VQA domain, development has been relatively constrained. In its early stages, the field had only a single dataset, and most algorithms were directly adapted from the general domain without considering the unique challenges posed by medical data. To advance the field, this thesis introduces two widely used Med-VQA datasets and several algorithms, addressing data scarcity and improving both the reasoning ability and explainability of Med-VQA models. By enhancing robustness and usability, this thesis aims to bridge critical gaps in both OOD detection and Med-VQA research, pushing the boundaries of conversational AI.

1.1 Challenges

1.1.1 Out-of-distribution Detection

Challenge I: OOD detection in the era of LLMs. LLMs have achieved remarkable progress across various cognitive tasks, yet their capabilities in OOD detection remain largely underexplored. Most earlier approaches mainly used small-scale pre-trained language models (PLMs) for OOD detection, such as BERT and RoBERTa, which were primarily encoder-based with sentence classification objectives. In contrast, modern LLMs use auto-regressive training for next-token prediction, creating fundamentally different hidden representations that likely impact OOD detection performance. Additionally, prior works have focused on improving OOD detection by fine-tuning PLMs

with discriminative objectives. However, a more natural approach would leverage LLMs’ pre-trained knowledge through generative fine-tuning, maintaining alignment with their auto-regressive training objectives. Given these distinct characteristics, it is essential to conduct an in-depth investigation into the OOD detection capabilities of LLMs to fully understand their strengths and limitations.

Challenge II: Near-OOD detection in a low-resource setting. Obtaining a large number of data labels is significantly time-consuming and resource-demanding, and data acquisition in certain domains (such as medicine, finance, etc.) is particularly challenging. Therefore, text OOD detection in real-world scenarios frequently encounters low-resource situations, where the labeled ID dataset contains only a few examples per class. Meanwhile, in some scenarios (like the banking system), some inputs that are treated as “near-OOD” share significant semantic similarities with in-distribution (ID) data but contain critical differences that should trigger caution. Distinguishing between these fine-grained cases requires a sophisticated understanding of semantic relationships and context, especially in situations where limited training data is available. However, former PLMs-based methods achieved suboptimal performance. Therefore, it is necessary to explore and design methods powered by LLMs to better solve near-OOD detection in a low-resource setting.

1.1.2 Medical Visual Question Answering

Challenge I: Data scarcity. Med-VQA represents a particularly demanding task that requires systems to interpret medical images, understand clinical queries, and generate

accurate responses that integrate visual and textual information with domain knowledge. However, at the initial stage, there was only one dataset named VQA-RAD [11] that only contained hundreds of images along with questions. This small-scale dataset severely restricted the development of the field. At the same time, due to the lack of proper datasets for designing task-specific methods, most approaches directly transferred general models for solving Med-VQA tasks. However, since visual models in the general domain are typically large, such as ResNet-101 [20], this leads to significant overfitting on the small-scale VQA-RAD dataset, resulting in poor generalization performance. Thus, it is worth designing methods to overcome the scarcity of data and the issues associated with it.

Challenge II: Insufficient reasoning capabilities of the models. Compared to general VQA, Med-VQA is particularly challenging due to the diverse clinical questions requiring different reasoning skills and the complexity of medical images, which differ significantly from natural images. Nevertheless, previous methods mostly addressed the Med-VQA problem by directly transferring models from the general domain, such as relying on reasoning modules like BAN [21] and SAN [22] and general visual extractors like ResNet [20], to solve medical VQA. Because these general modules fail to effectively model the relationship between medical images and textual inquiries, the overall performance is not satisfactory. This makes it crucial to enhance the reasoning capabilities of models in order to address the Med-VQA problem effectively.

Challenge III: Lack of explainability. As the attention on Med-VQA increases, more and more models and datasets have been proposed to advance the field. However, current Med-VQA systems do not provide explanations, especially visual ones, for the

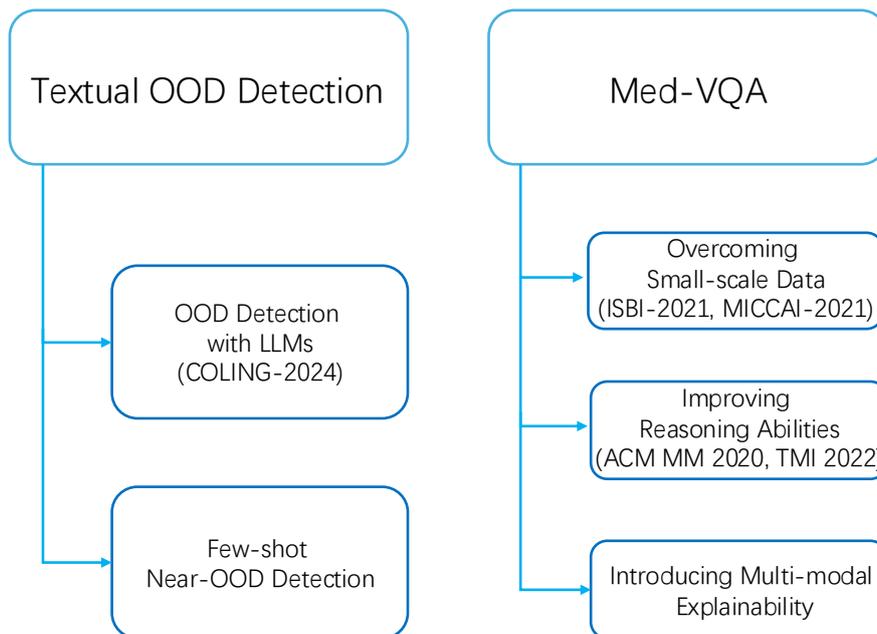


Figure 1.1: Research accomplishments

answers they offer, which greatly reduces the understanding of the question-answer pairs by patients. Explanations are as essential as the answers themselves in general VQA systems. This holds even stronger in medical VQA, where the domain-specific nature of the task amplifies the need for clarity [23]. At the same time, existing systems also struggle to handle choice-based questions effectively. Therefore, the Med-VQA field should address these two issues to improve the practicality of conversational systems.

1.2 Contributions

The contributions of this thesis are organized into two main areas: improving OOD detection and Med-VQA by solving the aforementioned challenges. An overview is shown in Figure 1.1. The following sections will provide a detailed introduction to these contributions.

1.2.1 Out-of-distribution Detection

Contribution I: Comprehensive evaluation of OOD detection with LLMs. To tackle the challenge I in OOD detection, we conduct a systematic empirical evaluation of traditional OOD detection methods for modern LLMs, addressing a significant gap in prior research. Our evaluation spans both zero-shot and fine-tuning settings, focusing on the LLaMA series ranging from 7B to 65B in size, and introduces a generative fine-tuning paradigm that better aligns with LLMs’ pre-training objectives. Our findings reveal that cosine distance-based detection methods outperform other OOD detectors by effectively leveraging the isotropic embedding space of LLMs. The results were published in COLING-24 [5].

Contribution II: A data-efficient near-OOD detection framework. To overcome the extremely challenging scenario of low-resource near-OOD detection, we propose a framework leveraging the isotropic characteristics of LLMs, which first develops semantic prototypes for each ID class from its name using a method based on diversity-grounded prompt tuning. Then, this framework performs semantic matching for both OOD detection and ID classification. For a thorough assessment, we benchmark our method against the prevalent fine-tuning approaches. The results reveal that this framework demonstrates superior performance in few-shot near-OOD detection tasks.

1.2.2 Medical Visual Question Answering

Contribution I: Overcoming small-sale data. First of all, we introduce SLAKE, a semantically-labeled knowledge-enhanced dataset with accurate visual and textual an-

notations and an extendable knowledge base, to overcome data scarcity. This dataset enables contextual understanding of medical images and supports complex questions requiring both visual comprehension and medical reasoning. Compared to the prior dataset VQA-RAD, our SLAKE contains more vision-question pairs involving more imaging modalities and body parts. Next, to overcome the overfitting caused by small-scale datasets, we first gather a large collection of unlabeled images and propose a pre-training framework that learns transferable feature representations from these radiology images. Then, this framework distills a lightweight visual feature extractor for Med-VQA, effectively mitigating overfitting to small-scale training data. These two works are accepted in ISBI-2021 [2] and MICCAI-2021 [3], respectively.

Contribution II: Improving reasoning abilities of Med-VQA models. To enhance the reasoning abilities of Med-VQA models, we propose a novel conditional reasoning mechanism that comprises two key components: the question-conditioned reasoning (QCR) module and the type-conditioned reasoning (TCR) strategy. The QCR module enables the model to develop question-specific reasoning capabilities by utilizing question information to guide multimodal fusion features. Additionally, Med-VQA tasks can be broadly categorized into two types: closed-ended and open-ended questions. Open-ended tasks, in general, are more challenging than closed-ended ones, and existing Med-VQA models often underperform on open-ended tasks. Recognizing that these task types require distinct reasoning capabilities, TCR handles open-ended and closed-ended tasks separately by developing specialized reasoning skills. This research has been accepted in ACMM MM-2020 [1] and TMI-2022 [4].

Contribution III: Introducing multi-modal explainability. To overcome the challenge of current models lacking explainability for their answers, we present GEMeX, a large-scale, groundable, and explainable Med-VQA benchmark for chest X-ray diagnosis, featuring several innovative components: (1) a multi-modal explainability mechanism that offers detailed visual and textual explanations for each question-answer pair, thereby enhancing answer comprehensibility; (2) four question types—open-ended, closed-ended, single-choice, and multiple-choice—to better reflect practical needs. With 151,025 images and 1,605,575 questions, GEMeX is the currently largest chest X-ray VQA dataset. We conduct an evaluation of 12 representative large vision language models on our GEMeX and propose a strong model by fine-tuning an existing LVLm on the training set. The results underscore the dataset’s complexity and effectiveness.

Thesis organization. The remainder of this thesis is organized as follows: Chapter 2 provides a comprehensive review of the literature related to textual OOD detection and Med-VQA. Chapter 3 presents our empirical investigation of OOD detection with LLMs. Chapter 4 introduces a novel framework for few-shot near-OOD detection in LLMs. Chapter 5 constructs the SLAKE dataset and proposes a representation distillation framework to address the challenge of data scarcity in Med-VQA. Chapter 6 designs a conditional reasoning framework to improve the reasoning abilities of Med-VQA models. Chapter 7 describes the GEMeX benchmark for empowering multimodal explainability and evaluates representative models on this new dataset. Finally, Chapter 8 concludes the thesis with a summary of contributions, limitations, and directions for future research. Through these contributions, this thesis aims to enhance the safety and reliability of conversational AI systems.

Chapter 2

Background and Related Work

This chapter begins by introducing related work on textual out-of-distribution (OOD) detection. It then provides an in-depth discussion of medical visual question answering (Med-VQA) methods, followed by a review of various datasets currently available. Additionally, it explores related research areas, including general visual question answering and medical vision-language tasks.

2.1 Textual Out-of-distribution Detection

Out-of-Distribution (OOD) detection has a long history in machine learning and is highly related to research topics like outlier detection, anomaly detection and novelty detection [24]. In the task setting of OOD detection, the *in-distribution* is characterized by the labeled training dataset and the *out-of-distribution* refers to anything else that possesses distributional shifts. Note that OOD detection differs from outlier detection in that it requires accurate classification of both ID and OOD data [25].

The significance of textual OOD detection in ensuring the robustness of NLP applications, such as dialogue systems, has led to a surge in research interest. Pre-trained Transformers have shown intrinsic superiority in handling OOD detection [19, 26, 27].

Several works have further evaluated the OOD performance of PLMs with respect to commonly used OOD detectors, including MSP [28], Mahalanobis distance (MD) [29], and Energy score [30]. For example, Podolskiy et al. [31] show that the Gaussian distribution assumption of MD better matches the representation space of BERT and can yield the best OOD performance in intent OOD detection benchmarks. Zhou et al. [32] show that a contrastive regularizer can further improve the sentence representation of Transformers for OOD detection.

More recently, Uppaal et al. [33] present a thorough analysis on the fine-tuning strategies for OOD detection with RoBERTa and show that RoBERTa [34] without fine-tuning can achieve near-perfect far-OOD detection performance. Similarly, we find that LLMs can also achieve perfect far-OOD detection performance without fine-tuning. Cho et al. [35] explores the OOD detection capability of medium-sized PLMs (such as GPT-2 [36]), as well as the impact of various ID fine-tuning techniques. While they also assess decoder-based models, the models they evaluated are not as extensive as this work, and they neglect to undertake generative ID tuning, a crucial step to fine-tune decoder-based models for downstream ID tasks. Furthermore, the models they examine remain at relatively moderate scales, and an exploration of the possible data-efficient characteristics of the model is lacking.

Recently, large language models (LLMs) have been leading a paradigm shift in the field of natural language processing (NLP) [37–45]. The use of LLMs to solve NLP tasks in

a generative way has become widespread. These LLMs commonly adopt the decoder-based architecture and are trained with the autoregressive objective. In this thesis, we focus on the OOD performance of open-source LLMs and anticipate that our work can provide useful insights for OOD detection under this paradigm.

2.2 Medical Visual Question Answering

Applying general VQA for Med-VQA. Most studies mainly applied popular VQA models on Med-VQA tasks [46–52]. Specifically, the visual features of medical images are extracted by deep pre-trained networks (e.g., ResNet [20] or VGGNet [53]), and the textual features of clinical questions are obtained through stacked RNN-based layers. For multimodal feature fusion, [46] adopted a simple concatenation operation, [47] exploited stacked attention networks (SAN) [22] and compact bilinear pooling (MCB) [54], [49] and [48] employed multimodal factorized bilinear pooling (MFB) [55], and [52] proposed to query an image by means of a written question based on the multimodal low-rank bilinear (MLB) module.

However, because of the large difference between radiology images and general images and the small scale of training datasets for Med-VQA, such straightforward adaptation suffers from severe overfitting. Moreover, VQA methods such as UpDn [56], which leverage algorithms like Fast R-CNN [57] for object detection, cannot be applied to Med-VQA due to the lack of annotation in existing datasets. To conquer the difficulty of data scarcity, [13] proposed mixture of enhanced visual features (MEVF) that pre-trained a small visual feature extractor (several convolutional layers) on VQA-RAD dataset [11] and an undisclosed external medical dataset, by using an auto-encoder and meta-learning

method MAML [58]. While this work achieves good performance on VQA-RAD dataset, the pre-training approach is dataset-dependent and requires extra annotation effort. Besides, it simply employs a bilinear attention mechanism for multimodal feature fusion, which lacks multi-level reasoning ability. Some recent works, including MMBERT [59], MedViLL [60], and RGC [61], try to pre-train a multimodal Transformer on medical vision-language datasets and then fine-tune it on Med-VQA tasks. However, due to the small scale of existing Med-VQA training datasets, large models could easily overfit. In this thesis, we explore lightweight models like MEVF [13] and aim to improve both the reasoning module and the visual feature extraction module.

Contrastive Pre-training. Lately, there is a growing interest in learning data representations with deep neural networks in an unsupervised or self-supervised manner, to reduce the need for laborious annotation work. Several recent studies have shown promise in learning image representations by designing proper pretext tasks and loss functions. [62] proposed to randomly rotate an image by 0, 90, 180, or 270 degrees and train a neural network to predict the rotation angle. CPC [63] pioneered in using a contrastive objective (InfoNCE loss) to learn data representations with a context auto-encoding task and achieved promising results in various domains including speech, image, and text. Recent development of contrastive self-supervised learning includes MoCo [64] that utilizes a queue to efficiently store a large number of negative samples to remove the restriction of mini-batch size, SimCLR [65] that uses stronger data augmentation and a very large batch to accommodate more negative samples, and MoCo-v2 [14] that combines the design improvements of SimCLR with MoCo. In the medical domain, contrastive self-supervised pre-training methods have also gained much attention recently [66, 67]. In this thesis, we utilize MoCo-v2 to pre-train useful representations of radiology images for

Med-VQA to overcome data scarcity challenges. Besides image-image contrastive self-supervised learning, with the rise of contrastive language-image pre-training (CLIP) [68], methods [69–75] start to focus on applying CLIP to Med-VQA. A promising way is to fine-tune CLIP’s joint embeddings to better handle specific medical domains, enhancing the model’s understanding of clinical questions and visual features [76].

Large Vision Language Models. Recently, the explosion of large vision language models (LVLMs) has further pushed the boundaries of medical domain [77–83]. LLaVA-Med-v1 [80] is designed specifically for medical applications based on LLaVA-v1 and LLaVA-Med-v1.5 [80] is an advanced version; MiniGPT-Med [79] is a medical version of Mini-GPT4 [84]; XrayGPT [77] is a specialized GPT model for interpreting chest X-rays; RadFM [78] is a radiology foundation model. Generally, they first pre-train models on a large-scale image-text dataset (like PMC-OA [76], PMC-15M [70]) to map visual features into language model’s embedding space and then further tune with instruction data for medical consultation [79, 85] or disease diagnosis [77, 86–88]. These models are now leveraged for Med-VQA tasks to provide richer, more context-aware answers, extending beyond simple text-image alignment to incorporate broader knowledge-based reasoning. Despite these advances, current methods are limited by the size and diversity of available datasets. These LVLMs fail to provide detailed answer explanations, especially on the visual side, which reduces patients’ and junior doctors’ understanding of the questions.

Datasets. In the initial stage, there is only one Med-VQA dataset, VQA-RAD [11], that offers over 3,000 question-answer pairs focused on radiology images. To advance the development of this field, we introduce SLAKE [2] which is the first manually created dataset with over 14,000 QA pairs across CT, MRI, and X-ray images and a medical knowledge graph, enabling models to handle complex scenarios by combining

Dataset	# Images	# QA Pairs	# Modalities	# Question Types‡	# Groundable	# Explainable
VQA-RAD [11]	0.315K	3.5K	Diverse†	O. & C.	✗	✗
SLAKE (Ours) [2]	0.642K	14K	Diverse†	O. & C.	✗	✗
OmniMedVQA [89]	118.010K	128K	Diverse†	O. & C. & S.	✗	✗
PMC-VQA [90]	149.075K	227K	Diverse†	O. & C. & S.	✗	✗
VQA-Med [91]	4.5K	4.5K	Diverse†	O. & C.	✗	✗
PathVQA [92]	149K	33K	Pathology	O. & C.	✗	✗
RadGenome-Chest CT [93]	50.188K	1.3M	Chest CT	O. & C.	✓	✗
MIMIC-Diff-VQA [94]	164.324K	700K	Chest X-ray	O. & C.	✗	✗
MIMIC-CXR-VQA [95]	142.797K	377K	Chest X-ray	O. & C.	✗	✗
GEMeX (Ours)	151.025K	1.6M	Chest X-ray	O. & C. & S. & M.	✓	✓ (Vision & Language)

Table 2.1: Comparison of medical VQA Datasets. † indicates a composition of multiple body parts (e.g., head, chest, abdomen) and various imaging modalities (e.g., CT, MRI, X-ray, pathology). In the # Question Types‡ column, O., C., S., and M. represent “Open-ended”, “Closed-ended”, “Single-choice”, and “Multi-choice”, respectively.

visual and textual information. As the field continues to develop, various datasets have been created to advance medical VQA research, each tackling specific challenges across clinical domains. A detailed comparison can be seen in Table 2.1. VQA-Med [91] is a key dataset for Med-VQA competitions, with 4,500 radiology images and paired QA sets for training, validation, and testing. OmniMedVQA [89] provides more data and more imaging modalities, which cover the entire body, to encourage model generalization. PMC-VQA [90] generates VQA data by prompting a large language model to decompose captions of biomedical figures, enabling academic knowledge extraction. PathVQA [92] supplies over 32,000 QA pairs on histopathological images for fine-grained pathology analysis.

For specialized tasks, RadGenome-Chest CT [93] supports chest CT diagnostics, while MIMIC-Diff-VQA [94] emphasizes differential diagnosis reasoning between two X-rays.

MIMIC-CXR-VQA [95] expands MIMIC-CXR [96] with diverse question templates to generate thoracic radiology QA pairs, aiding in chest abnormality detection. However, all current datasets lack explainability and diverse question formats. They do not provide detailed visual and textual explanations for answers, which limits the usability for patient and junior doctor in understanding the answers. Moreover, the restricted range of question types reduces their ability to simulate the diverse inquiries encountered in real-world medical practice. Therefore, we introduce a new dataset named GEMeX [6] that covers four categories of varying difficulty levels: open-ended, closed-ended, single-choice, and multiple-choice questions. Each question-answer pair is enriched with explicit reasoning and corresponding visual region annotations. Currently, this is the largest VQA dataset for chest X-rays that comprises 151,025 images and 1,605,575 questions, and the first medical VQA dataset simultaneously includes both textual and visual explanations.

2.3 Visual Question Answering

A typical VQA system consists of 4 basic components: (1) a visual feature extractor to obtain the visual image features; (2) a textual feature extractor to obtain the textual question features; (3) a multimodal feature fusion module to aggregate both the visual and textual features to produce a joint representation; (4) a classifier to predict the final answers based on the joint representation. Various VQA systems differ in how they extract and combine multimodal features.

Early studies mainly employed VGGNet [53] and LSTM [15] to extract visual and textual features respectively and combined them by simple mechanisms such as concatenation and pooling [97]. In recent years, a lot of effort has been devoted to studying inter-

modality relation by exploring the connection between visual and textual semantics [21, 22, 56, 98, 99]. Stacked attention network (SAN) was proposed in [22] to progressively search for related image regions using question semantic representations. Based on low-rank bilinear pooling, bilinear attention network (BAN) [21] was proposed to generate bilinear attention maps to fuse multimodal features, which is also employed in this work. UpDn [56] utilized Faster R-CNN [57] to extract regions of interest (ROI) at object level and aggregated region features with weights generated under the guidance of the question. Based on ROI features, some methods such as LXMERT [100] and Pythia [101] achieved promising results. Very recently, Transformer [102] based vision-and-language pre-training (VLP) becomes a popular paradigm. A typical process is to extract question features with BERT [103] and fuse them with visual features via a self-attention mechanism. According to the different ways of visual feature extraction, VLP can be divided into object-based methods such as ViLBERT [104], VL-BERT [99], and VisualBERT [105], convolution-based methods such as Pixel-BERT [106], and image-patch-based methods such as ViLT [107]. Since object-based methods rely on visual object labels, which are not available in existing Med-VQA datasets, in this thesis, we only explore convolution-based and image-patch-based methods.

Besides, a recent line of research [108–111] focused on developing VQA systems with higher-level reasoning skills. [108] proposed to split questions into a series of semantic segments, which would accordingly activate pre-specified neural network modules. However, it is difficult to train the network due to complex pre-defined structures and annotations. [111] focused on solving VQA tasks in the few-shot setting by generating additional normalization parameters from questions to control the visual feature extractor’s inner layers. To solve the highly challenging compositional questions [112–114],

MAC [109] proposed to use various recurrent cells such as memory, attention, and composition for reasoning, while neural-symbolic (NS) approaches [115, 116] exploited executable symbolic programs to mimic human reasoning process.

2.4 Medical Vision-Language Tasks

Medical vision-language tasks have gained increasing attention in recent years, driven by the need for automated and efficient interpretation of medical images. Besides Med-VQA, there is also report generation and medical image-report retrieval, each addressing different aspects of integrating visual and textual information for clinical decision-making.

Medical Report Generation. Medical report generation aims to automatically produce structured and coherent radiology reports from medical images, such as chest X-rays [96, 117] and CT scans [93]. Early approaches [118] relied on template-based methods or rule-based systems, which often lacked flexibility and generalization capabilities. With advancements in deep learning, encoder-decoder architectures, particularly those based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have become popular for this task [119, 120]. More recently, transformer-based models (e.g., R2Gen [121], Maira-1 [122]) have demonstrated superior performance by capturing long-range dependencies between visual and textual features. However, a key challenge in medical report generation remains the hallucination problem, where generated reports may include incorrect or clinically implausible content. To address this, contrastive decoding [123] and knowledge-enhanced method [124] have been introduced to improve factual correctness and clinical consistency.

Medical Image-Report Retrieval. Medical image-report retrieval involves retrieving relevant radiology reports given a medical image or vice versa. This task is critical for clinical decision support, allowing radiologists to reference similar cases. Traditional methods relied on hand-crafted features and similarity metrics, but deep learning-based approaches now dominate, leveraging joint vision-language embeddings. Cross-modal retrieval methods often employ dual-encoder frameworks based on CNNs and transformers, aligning image and text representations in a shared latent space [125]. Recently, contrastive pre-training (e.g., CLIP-based models [126, 127]) has further improved retrieval performance by learning robust vision-language associations from large-scale medical datasets. However, challenges such as domain shift across imaging modalities and incompleteness in textual annotations still pose limitations in real-world applications.

Chapter 3

OOD Detection with LLMs

3.1 Introduction

Out-of-distribution (OOD) detection has attracted significant attention due to its crucial role in ensuring AI safety [128]. The objective is to identify and raise an alarm for inputs that exhibit distributional shifts compared to the in-distribution (ID) training data. Given that the test distribution can dynamically change over time, OOD detection has become indispensable in high-stakes applications, such as healthcare and self-driving cars. Its ability to detect anomalous inputs and adapt to evolving scenarios makes it a vital component in ensuring the reliability and robustness of AI systems in real-world, dynamic environments.

Utilizing sentence representations yielded by pre-trained language models (PLMs) to derive OOD confidence scores has been the *de facto* method for textual OOD detection. Specifically, PLMs are first fine-tuned on the ID data and then OOD detectors are applied

In-Distribution (ID)	Out-of-Distribution (OOD)
(I) ID: Sentiment Analysis <i>Nelson's intentions are good.</i> SST-2	(I) Far-OOD: Question Classification <i>Who was the first American in space ?</i> TREC
(II) ID: Transfer-related Intent <i>I would like to make a transfer.</i> CLINC150-BANK (Transfer)	(II) Near-OOD: Balance-related Intent <i>What is in my bank accounts?</i> CLINC150-BANK (Balance)

Figure 3.1: Illustration of two types of OOD instances compared to ID samples: far-OOD where ID and OOD data come from different domains and near-OOD where ID and OOD data come from the same domain but with different classes.

on the sentence representations generated by PLMs. Compared to ID data, there are two types of OOD instances: far-OOD where ID and OOD data come from different domains and near-OOD where ID and OOD data come from the same domain but with different classes, as shown in Figure 3.1. Typically, near-OOD samples are harder to recognize. A body of works [26, 27, 31, 33] have shown that Transformer-based models can produce better sentence representations for OOD detection. However, these studies have mainly focused on evaluating the OOD detection performance of small-scale encoder-based Transformers, such as RoBERTa and BERT.

Recently, large language models (LLMs) have made significant strides in various cognitive tasks, yet their capabilities on OOD detection remain largely unexplored. Unlike relatively small-scale PLMs used by prior studies, LLMs often display notable differences. In particular, the majority of previously prominent PLMs utilized for OOD detection adopt the encoder-based architecture, such as BERT and RoBERTa. These models are

predominantly designed with a pre-training objective that focuses on sentence classification. However, recent LLMs [37, 39–42] exclusively adopt an autoregressive training objective during pre-training. Consequently, the hidden states of LLMs are specialized for next token prediction, which could influence their performance in OOD detection. Moreover, previous works test changes in OOD detection when adapting PLMs to downstream tasks through discriminative fine-tuning, even for decoder-based models [35]. However, a more intuitive approach is to probe the pre-training knowledge of LLMs through generative fine-tuning, which better aligns LLMs’ pre-training objective with downstream tasks. Thus, it is imperative to extensively investigate the OOD detection capabilities of LLMs to gain deeper insights into their potential and limitations.

This work aims to fill this gap by offering a comprehensive and structured assessment of OOD detection with LLMs across varying scales (ranging from 7B to 65B). Notably, our evaluation process is specifically designed to consider the scaling laws of LLMs with commonly utilized OOD detection detectors, ensuring broader and more generalized findings. In summary, our analysis has revealed the following new insights:

1. Discriminative vs. generative fine-tuning. We have observed that generative fine-tuning demonstrates greater resilience to the issue of ID overfitting when compared to discriminative fine-tuning. As highlighted by Uppaal et al. [33], there exists a trade-off between achieving higher accuracy on ID tasks and ensuring effective OOD detection. It has been shown that OOD detectors progressively lose efficacy as the training of ID tasks continues. However, our findings indicate that adopting a generative approach to fine-tuning LLMs can effectively mitigate this issue, potentially resulting in stable OOD performance even as training progresses and ID accuracy improves.

2. LLM-based far- vs. near-OOD detection. Our results consistently demonstrate that LLMs are natural far-OOD detectors. Remarkably, LLMs of all scales achieve near-perfect OOD performance in far-OOD scenarios without requiring any fine-tuning. However, when it comes to near-OOD detection, only the 65B model is able to achieve satisfactory performance without any fine-tuning. Despite that, we discover that fine-tuning significantly improves the near-OOD detection capability of LLMs.

3. Anisotropy vs. isotropy. Our experimental results suggest that the cosine distance function, when used as a straightforward OOD detector, performs exceptionally well. This observation leads to an intriguing discovery: the embedding spaces of LLMs exhibit a desirable isotropic property, which is not possessed by the BERT family models. The sentence embeddings produced by the BERT family models have been noted to possess an undesirable characteristic of being concentrated within a narrow cone, a phenomenon referred to as anisotropic representations [129], which negatively affects tasks involving semantic relationships and is commonly known as representation degeneration [130]. The issue is resolved through the isotropic representations generated by LLMs, which allow the cosine distance to excel in OOD detection and may potentially benefit a broad spectrum of tasks.

3.2 Method

Problem statement. The objective of OOD detection is to effectively differentiate between instances that belong to a specific distribution (in-distribution \mathcal{D}_{ID}) and those falling outside of that distribution (out-of-distribution \mathcal{D}_{OOD}). To better and fairly evaluate the capabilities of LLMs for OOD detection compared to prior smaller models (e.g.,

RoBERTa [34]) [33], we utilize the same sentence classification task as the ID training task. In practical application scenarios, undesired inputs (e.g., a severe distribution shift towards ID data) may occur, and an OOD confidence scoring function f_{OOD} can be used to reject whether outputting results for such inputs or not.

3.2.1 ID Generative Fine-tuning with LLMs

For the ID sentence classification task, we align with the nature of LLMs and adopt a generative approach (referred to as open-ended classification) [131]. Concretely, given an input sentence \mathbf{X}_s , we first expand it with a simple template: “### Input:\n \mathbf{X}_s ### Output:\n”, to facilitate the extraction of outputs by identifying the section following the “Output” symbol. Subsequently, we maximize the probability of generating the target label \mathbf{X}_a with L tokens by:

$$\max p(\mathbf{X}_a|\mathbf{X}_s) = \prod_{i=1}^L p_{\theta}(x_i|\mathbf{X}_s, \mathbf{X}_{a,<i}), \quad (3.1)$$

where θ represents the model parameters and $\mathbf{X}_{a,<i}$ are partial label tokens that come before the current prediction token x_i .

Parameter-efficient fine-tuning. To improve the performance of LLMs in the in-distribution sentence classification task, we employ a parameter-efficient fine-tuning (PEFT) approach, to minimize the usage of additional parameters. Specifically, we utilize the low-rank adaptation (LoRA) [132] technique which freezes the pre-trained LLMs’ weights and inserts trainable rank decomposition matrices into each Transformer layer. We perform PEFT with answer predictions, i.e., only class label tokens are utilized to compute the auto-regressive loss. During the test stage, we use *strict matching* to determine whether the generated labels are identical to the ground truth.

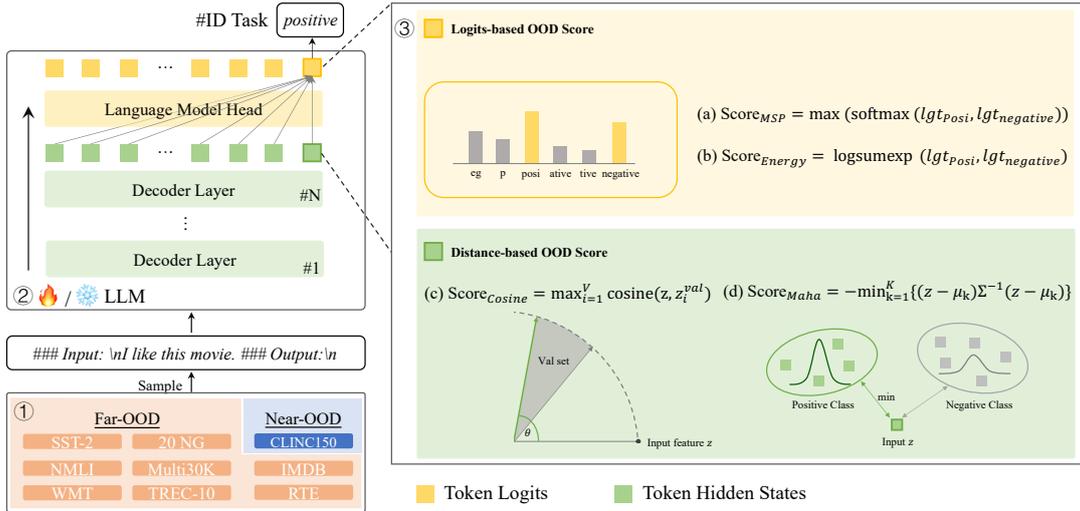


Figure 3.2: Our proposed evaluation framework for LLMs at OOD detection, taking three aspects into consideration: ① distribution of OOD samples (near or far), ② impact of model tuning on OOD detection, and ③ diverse OOD detectors (score functions).

3.2.2 OOD Detection with LLMs

The overview of our OOD detection framework is illustrated in Figure 3.2. Our primary focus is on decoder-like LLMs, such as LLaMA [37], as they have demonstrated excellent performance when their model size scales up [133, 134]. To obtain a comprehensive observation, we conduct OOD detection experiments on two different semantic distribution settings [135, 136]: far-OOD and near-OOD (*cf.* Figure 3.1). Regarding OOD detection methods, we focus on the prevailing post-hoc paradigm [25]. In the following, we elaborate on how to integrate post-hoc OOD detectors into decoder-style LLMs, which has not been addressed in existing literature.

Customized post-hoc methods. According to prior studies [26, 32], there mainly exist two categories of post-hoc methods: logits-based OOD score functions and distance-based ones. Since previous works used these methods for language models accompanied

by a classifier, we here customize them for decoder-type LLMs with only a language model head (as shown in Figure 3.2) in the following:

Logits-based OOD score functions operate on the final class-related logits. In generative classification, the generated class name is usually composed of several tokens, e.g., “positive” consists of “posi” and “tive”. Instead of calculating the probability (logits) for the entire ID class name, we simplify the process by considering the probability assigned to the first token of its class name. For instance, in a sentiment analysis task with classes like “positive” and “negative” as depicted in Figure 3.2, we only need to identify the probability corresponding to the tokens “posi” and “negative” respectively. Considering that different class names may have common prefixes, such as “positive” and “position”, we will rephrase the conflicting class names at the beginning of ID training, such as replacing “position” with “location”. In practice, we observe this re-translation has no impact on the ID task. Overall, there are mainly two logits-based functions:

- Maximum softmax probability (MSP) [29] utilizes the maximum softmax probability corresponding to each class, i.e, score $\mathcal{S}(x) = \max\{p(y_i|x)\}_i^K$, where K is the number of classes, and ID samples always exhibit higher probability scores while OOD ones correspond to lower scores.
- Energy score (Energy) [30, 137] computes confidence score $\mathcal{S}(x) = \log \sum_i^K e^{(w^T \cdot z)_i}$ where w^T is the weight of the language model head and z is all word embeddings. Note that for both MSP and Energy, we only select the probability and logits corresponding to the first token of each class name, as mentioned above.

Distance-based OOD score functions apply to sentence representations. Prior studies using encoder-based PLMs treated the embeddings of special token “cls” as sentence

representations. For LLMs, we employ the embeddings of the last token of the input sentence as the representation. There are mainly two functions considered for evaluation: Mahalanobis distance (Maha) [29] and Cosine distance (Cosine) [32]*.

3.3 Experimental Setup

3.3.1 Datasets

To draw universal conclusions, we conduct a comprehensive evaluation of two kinds of dataset distribution settings [138] as illustrated in Figure 3.1 and Figure 3.2. **Far-OOD.** In this paradigm, ID and OOD samples come from different distributions (datasets), exhibiting significant semantic differences. Following Hendrycks et al. [26] and Zhou et al. [32], we evaluate 8 datasets, including 20 Newsgroups (20NG) [139] for topic classification, RTE [140] and MNLI [141] for nature language inference, TREC-10 [142] for question classification, SST-2 [143] and IMDB [144] for sentiment analysis, and the English side of Multi30K [145] and WMT16 [146] for machine translation. Among them, we choose 20NG and SST-2 as two separate *in-distribution* tasks and the remaining ones are recognized as *out-distribution*. Note that when SST-2 is used as the ID, we do not consider IMDB as an OOD dataset since both of them are sentiment analysis tasks.

Near-OOD. We also test on a more challenging scenario, where ID and OOD samples come from the same domain but with disjoint label sets. A well-researched domain is OOD intent detection [147]. Specifically, we use CLINC150 dataset and choose *Banking* and *Travel* domains. Within each domain, 50% of the classes are chosen as ID, and the

*We refer authors to original papers for more details.

remaining classes as OOD.

3.3.2 Evaluation Metrics

We employ three commonly used metrics for OOD detection: (1) AUROC (area under the receiver operating characteristic curve). (2) FAR@95 (false alarm rate at 95% recall). It represents the probability of incorrectly classifying a negative sample as positive when the Recall or True Positive Rate (TPR) is 95%. We treat the OOD class as negative. (3) AUPR (area under the precision-recall curve). Additionally, we use accuracy as a metric for ID classification task.

Dataset	Full-shot	10-shot	5-shot	1-shot
SST-2	16	8	4	2
20NG	8	8	8	8
CLINC150 (Banking or Travel)	16	16	16	8

Table 3.1: Batch size configuration for each dataset.

3.3.3 Implementation Details

All experiments are conducted on a workstation with 4 NVIDIA A100 80G GPUs. For zero-grad OOD detection, LLaMA-7B, -13B, -30B, and -65B are deployed on 1, 1, 2, and 4 A100 GPUs, respectively. When further fine-tuning LLMs on ID tasks, the LoRA configurations (Section 3.2.1) are that rank r is 16, scaling α is 16, and query/key/-value/output projection matrices $\{W_q, W_k, W_v, W_o\}$ in each self-attention module need to be updated. We train the network for 50 epochs with early stop criteria that if the

ID Dataset	LLM	Maha			Cosine			MSP			Energy			
		AUROC \uparrow	FAR@95 \downarrow	AUPR \uparrow	AUROC \uparrow	FAR@95 \downarrow	AUPR \uparrow	AUROC \uparrow	FAR@95 \downarrow	AUPR \uparrow	AUROC \uparrow	FAR@95 \downarrow	AUPR \uparrow	
Far-OOD	SST-2	LLaMA-7B	0.991	0	0.993	0.990	0.006	0.990	0.905	0.318	0.811	0.368	0.930	0.380
		LLaMA-13B	0.992	0	0.993	0.990	0.005	0.989	0.939	0.213	0.818	0.571	0.778	0.478
		LLaMA-30B	0.994	0.003	0.993	0.991	0.009	0.990	0.881	0.361	0.742	0.651	0.738	0.540
		LLaMA-65B	0.991	0.007	0.991	0.990	0.007	0.992	0.776	0.621	0.646	0.544	0.821	0.485
	20NG	LLaMA-7B	0.997	0	0.995	0.998	0	0.996	0.441	0.929	0.391	0.571	0.784	0.417
		LLaMA-13B	0.996	0.006	0.989	0.993	0.004	0.990	0.622	0.754	0.482	0.491	0.932	0.362
		LLaMA-30B	0.995	0.005	0.987	0.995	0.002	0.993	0.533	0.847	0.424	0.491	0.906	0.362
		LLaMA-65B	1	0	0.998	0.999	0	0.997	0.616	0.764	0.421	0.508	0.925	0.369
Near-OOD	CLINC-Banking	LLaMA-7B	0.896	0.568	0.921	0.891	0.587	0.916	0.720	0.814	0.763	0.722	0.818	0.758
		LLaMA-13B	0.905	0.408	0.922	0.903	0.514	0.922	0.739	0.769	0.760	0.713	0.831	0.743
		LLaMA-30B	0.895	0.472	0.913	0.910	0.424	0.923	0.733	0.813	0.746	0.724	0.795	0.735
		LLaMA-65B	0.951	0.255	0.964	0.956	0.200	0.964	0.823	0.604	0.834	0.826	0.614	0.834
	CLINC-Travel	LLaMA-7B	0.895	0.680	0.932	0.887	0.738	0.927	0.584	0.921	0.640	0.637	0.912	0.674
		LLaMA-13B	0.942	0.485	0.964	0.922	0.730	0.955	0.639	0.834	0.696	0.633	0.909	0.695
		LLaMA-30B	0.926	0.458	0.950	0.928	0.523	0.950	0.650	0.911	0.697	0.653	0.888	0.698
		LLaMA-65B	0.959	0.182	0.971	0.976	0.076	0.986	0.739	0.745	0.753	0.755	0.681	0.768

Table 3.2: OOD detection performance of zero-grad LLaMA models. We use the full validation set to calculate each OOD score. The results are averaged over five seeds.

model’s performance on the validation set continuously drops for 6 consecutive epochs and the current epoch number exceeds 15, training will be terminated. We use AdamW optimizer with learning rate 1×10^{-4} , further decayed by linear schedule. Due to the varying lengths of sentences in different ID datasets, we configure different batch sizes shown in Table 3.1. All experiments are conducted over five seeds (1, 2, 3, 4, 5).

3.4 Findings

3.4.1 Zero-grad OOD Detection with LLMs

In this section, we evaluate the *zero-grad* OOD performance of LLMs. The objective is to examine how well OOD detection performs when utilizing the knowledge acquired by

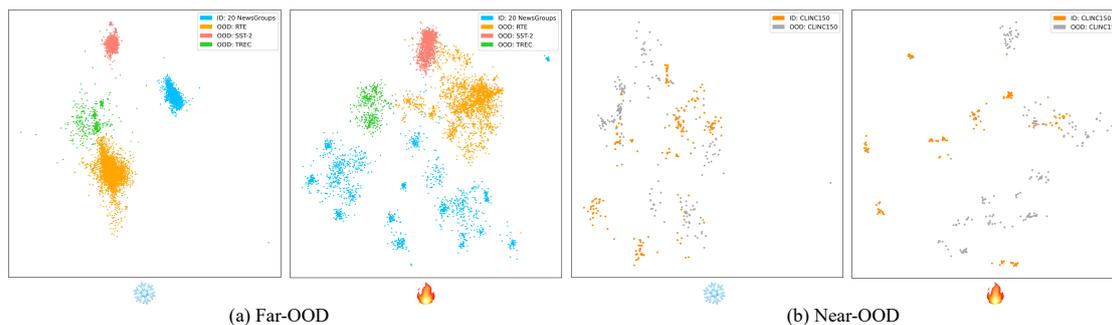


Figure 3.3: UMAP [9] visualization of representations generated from the penultimate layer of the zero-grad (snowflake) and fine-tuned (flame) LLaMA-7B models. (a) Far-OOD: 20NG is treated as ID while SST-2, RTE, and TREC are treated as OOD. (b) Near-OOD: the banking domain of CLINC150 is selected, of which 50% of the classes are treated as ID, and the rest are treated as OOD.

LLMs during pre-training. The results are summarized in Table 3.2 and all LLMs are frozen in this setting. Note that we use full-shot validation set to calculate each OOD score.

LLMs are natural far-OOD detectors. As shown in Table 3.2, when applying distance-based OOD detection methods, such as Maha and Cosine, all LLMs can achieve near-perfect results (e.g., AUROC and AUPR approach 1 while FAR@95 approaches 0). To better understand why distance-based OOD detectors are so effective, we visualize the corresponding sentence representations yielded by the penultimate layer (before the top head layer), as shown in Figure 3.3 (a ). It can be found that representations from the same dataset are tighter, while ID and OOD sentences have clear boundaries, indicating the profound semantic discrimination prowess exhibited by LLMs. However, both MSP and Energy generate poor results. This is foreseeable, as both of them condition on the first token generated from the input sentence. When the model has not been fine-tuned, it often struggles to accurately output class names, leading to inferior OOD performance.

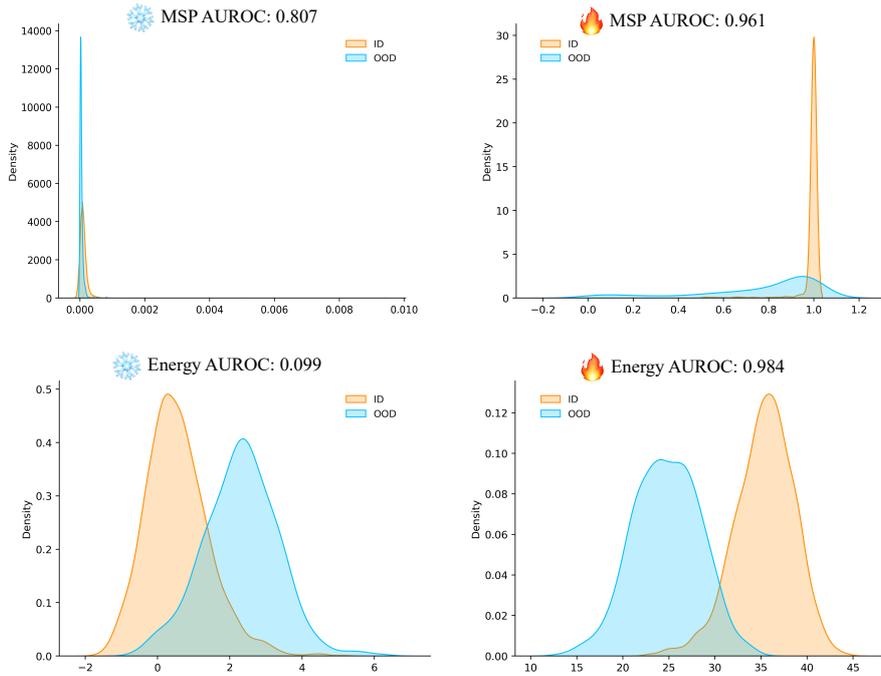


Figure 3.4: Impact of fine-tuning on logits-based OOD scores (MSP at the top row and Energy at the bottom row). We plot SST-2 (ID) vs. TREC-10 (OOD) for visualization.

Moreover, from the probability density of Figure 3.4 (❄️), it can be found that there is a significant overlap between ID and OOD, leading to a decrease in OOD detection performance.

The capability of LLMs for near-OOD detection improves with their scale. We present the zero-grad near-OOD results in Table 3.2 (CLINC-Banking and CLINC-Travel). For the near-OOD setting, as the number of model parameters increases, the OOD detection performance will also be improved. Remarkably, when the model has an exceedingly large number of parameters (i.e., LLaMA-65B), we can observe a dramatic performance surge [38] to detect OOD inputs, especially with distance-based OOD methods. In particular, the AUROC values for Maha and Cosine both surpass 95%, and FAR95 is enhanced by **at least 30%** in comparison to the 7B model.

Furthermore, it is evident that the near-OOD performance of LLMs is notably inferior

ID Dataset	Shot	Maha			Cosine			MSP			Energy				
		ID ACC \uparrow	AUROC \uparrow	FAR@95 \downarrow	AUPR \uparrow	AUROC \uparrow	FAR@95 \downarrow	AUPR \uparrow	AUROC \uparrow	FAR@95 \downarrow	AUPR \uparrow	AUROC \uparrow	FAR@95 \downarrow	AUPR \uparrow	
Far-OOD	SST-2	1	0.535	0.5	1.0	0.422	0.954	0.250	0.934	0.664	0.581	0.587	0.716	0.589	0.637
		5	0.664	0.878	0.625	0.843	0.973	0.045	0.971	0.768	0.493	0.674	0.885	0.408	0.794
		10	0.857	0.967	0.204	0.962	0.991	0.009	0.987	0.771	0.514	0.693	0.896	0.379	0.803
		Full	0.976	0.993	0.004	0.992	0.993	0.005	0.991	0.947	0.298	0.888	0.961	0.189	0.907
	20NG	1	0.463	0.5	1	0.380	0.991	0.047	0.985	0.756	0.779	0.670	0.850	0.681	0.824
		5	0.713	0.983	0.074	0.975	0.991	0.023	0.989	0.868	0.503	0.799	0.947	0.283	0.918
		10	0.796	0.992	0.042	0.987	0.996	0.013	0.991	0.893	0.438	0.840	0.951	0.215	0.924
		Full	0.944	0.995	0.003	0.991	0.993	0.007	0.991	0.959	0.207	0.939	0.968	0.114	0.945
Near-OOD	CLINC-Banking	1	0.589	0.5	1	0.533	0.905	0.510	0.926	0.846	0.696	0.860	0.870	0.658	0.897
		5	0.882	0.863	0.614	0.879	0.962	0.255	0.968	0.873	0.556	0.878	0.903	0.463	0.916
		10	0.949	0.937	0.424	0.956	0.968	0.157	0.974	0.902	0.422	0.902	0.919	0.346	0.929
		Full	0.973	0.958	0.231	0.969	0.964	0.147	0.970	0.936	0.269	0.945	0.930	0.225	0.931
	CLINC-Travel	1	0.526	0.5	1	0.533	0.910	0.481	0.925	0.767	0.756	0.771	0.780	0.733	0.793
		5	0.964	0.897	0.644	0.925	0.974	0.148	0.983	0.886	0.415	0.886	0.875	0.420	0.872
		10	0.984	0.975	0.137	0.983	0.982	0.078	0.988	0.930	0.3	0.931	0.933	0.231	0.933
		Full	0.991	0.980	0.045	0.988	0.978	0.049	0.987	0.942	0.121	0.933	0.948	0.112	0.953

Table 3.3: The performance of the fine-tuned LLaMA-7B model for OOD detection and ID classification. “Shot” denotes the number of examples in the ID training or validation set. We report the average results of five seeds.

compared to their performance on far-OOD instances. To understand this, we provide a visualization for this setting as illustrated in Figure 3.3 (b ). The embeddings of ID and OOD samples are mixed up since their labels come from the same domain (i.e., travel or banking). Consequently, detecting near-OOD instances becomes notably more challenging than far-OOD instances.

3.4.2 OOD Detection with Generatively Fine-tuned LLMs

In this subsection, we study the influence of fine-tuning LLMs on OOD detection. Specifically, we conduct an in-depth examination of how the OOD detection performance evolves with the progression of ID task training.

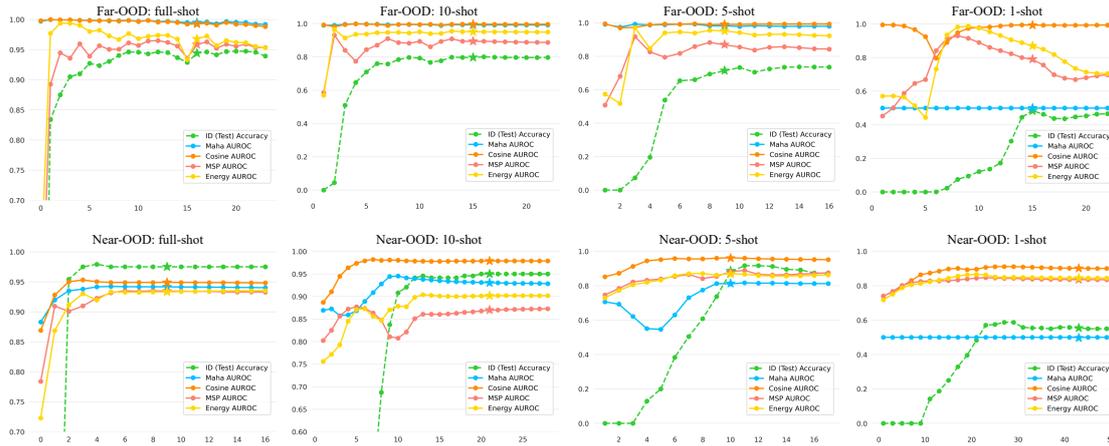


Figure 3.5: Performance changes for ID classification and OOD detection as training progresses with the different number of training shots. Top row: 20NG is ID training task; Bottom row: banking domain of CLINC150 is selected where 50% classes are used as ID training task and the rest are OOD samples. The \star indicates the selected results whose epoch corresponds to the best ID performance on the validation set.

ID fine-tuning can boost OOD detection. We fine-tune LLMs in a generative manner in both few-shot and full-shot scenarios. The results are summarized in Table 3.3. Likewise, we present both far- and near-OOD results comparable to the zero-grad configuration. Clearly, fine-tuning LLMs on in-distribution tasks can notably augment the models’ capacity to detect OOD instances, surpassing the performance of the zero-grad setting by a significant margin in most cases like in near-OOD setting and with logits-based functions (in both full-shot scenarios with LLaMA-7B model).

In Figure 3.5, we present the fine-tuning curves. It can be observed that as the ID accuracy increases, almost all OOD detectors are improved accordingly. To study how fine-tuning impacts the ID vs. OOD separability, we plot their density distributions in Figure 3.4. Clearly, fine-tuning can improve the separability between ID and OOD instances. A similar effect can be cross-validated in Figure 3.3 (b ) in which the embedding of

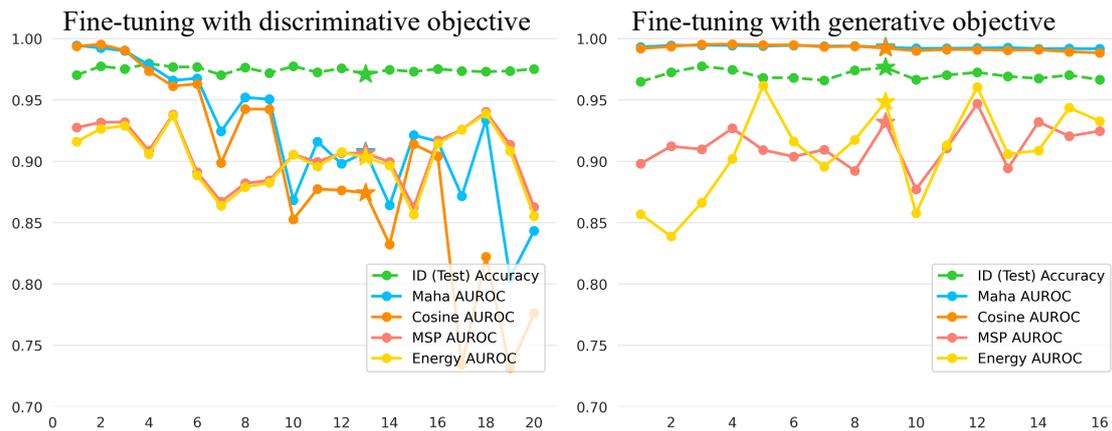


Figure 3.6: Impact of different ID training objectives, discriminative vs. generative. SST-2 dataset with full data is used as the ID training task.

different classes within the ID becomes more compact, while the separation between ID and OOD becomes clearer. However, it is important to highlight that as the training continues, there is a possibility of encountering overfitting, which could result in inferior OOD performance, especially for logits-based methods as illustrated in Figure 3.5 for both full-shot and 1-shot far-OOD scenarios. This observation is similar to the findings in [33].

Generative fine-tuning generalizes better. In addition to generative fine-tuning, we also explore discriminative fine-tuning by appending a classifier after LLMs (replacing the language model head[†]) to conduct ID task. The comparison of the trend charts presented in Figure 3.6 reveals that generative fine-tuning tends to be less overfit on the ID task and all OOD detectors consistently perform better than discriminative fine-tuning, especially for distance-based OOD detectors. To better understand this effect, based on the transformations of embeddings illustrated in Figure 3.3 (a), it becomes evident that throughout the generative training process, while the ID’s distribution shifts

[†]We use `LlamaForSequenceClassification` provided by Huggingface [148]

ID Dataset	PTM	Maha		Cosine		MSP		Energy		
		AUROC \uparrow	FAR@95 \downarrow							
<i>Zero-grad</i>										
SST-2	RoBERTa-L \dagger [33]	0.971	0.152	0.919	0.414	-	-	-	-	
	LLaMA-7B	0.991	0	0.990	0.006	0.905	0.318	0.368	0.930	
	<i>Fine-tuned</i>									
	RoBERTa-L \ddagger [32]	0.969	0.183	0.962	0.236	0.889	0.613	0.877	0.632	
	LLaMA-7B	0.993	0.004	0.993	0.005	0.947	0.298	0.961	0.189	
<i>Zero-grad</i>										
20NG	RoBERTa-L \dagger [33]	0.998	0.002	0.998	0.002	-	-	-	-	
	LLaMA-7B	0.997	0	0.998	0	0.441	0.929	0.571	0.784	
	<i>Fine-tuned</i>									
	RoBERTa-L \ddagger [32]	0.983	0.073	0.978	0.107	0.946	0.305	0.965	0.158	
	LLaMA-7B	0.995	0.003	0.993	0.007	0.959	0.207	0.968	0.114	

Table 3.4: Comparison of large and small PTMs under zero-grad and fine-tuned settings for OOD detection. \dagger denotes the results we reproduce due to different calculating methods, while \ddagger indicates results cited from the original paper.

into class-specific clusters, a distinct separation continues to exist between these clusters and the OOD samples. This preserves the effectiveness of distance-based OOD detection methods. Prior study [33] pointed out that discriminative tuning the small models (e.g., RoBERTa [34]) negatively impacts the performance of distance-based OOD detection methods. This issue also exists in discriminative tuning LLMs but has been solved in the generative tuning.

Besides, in Table 3.4, we compare encoder-based and decoder-based Transformers and observe impressive improvement on decoder-based LLMs.

Cosine distance is a data-efficient OOD detector. To further investigate whether LLMs possess data-efficient OOD detection capabilities, we configure the training samples of the ID as few-shot instances (e.g., 1, 5, and 10). Please note that we also set the

number of validation sets to be the same shot, since all OOD detection methods rely on the validation set. Results presented in Table 3.3 convey that as the number of shots increases, the OOD detection capability of the LLMs also improves. Moreover, distance-based OOD detection methods are superior to logits-based ones, and they can achieve good performance even with only 10-shot samples. Particularly, cosine distance is a data-efficient OOD detector that can provide effective detection by requiring only **1-shot** instance. For example, it achieves AUROC of **99.1%** (near-perfect) on 20NG (ID) and over 90% on others. Besides, in the 1-shot setting, the Mahalanobis distance loses its efficacy since it’s unfeasible to model the necessary Gaussian distribution when there’s only a single sample for each class.

Isotropy vs. anisotropy. By examining Table 3.2 and Table 3.3, it becomes evident that Cosine distance, as a simple OOD detector, consistently delivers superior performance and ranks among the top performers in both the zero-grad and generative fine-tuning settings. We provide an explanation of this phenomenon from the perspective of representation learning. In the past few years, the anisotropic issue, also known as the representation degeneration problem, of BERT family models has garnered considerable attention [129, 130]. Researchers have highlighted that BERT’s sentence representations are concentrated within a narrow cone, resulting in substantial challenges for tasks involving semantic matching. Nevertheless, we discover that this concern does not apply to LLMs. The representations generated by off-the-shelf LLMs inherently exhibit isotropy, enabling Cosine distance to excel in OOD detection.

To quantify anisotropy, we adopt the methodology introduced by Ethayarajh [129] to measure sentence-level anisotropy. Let \mathbf{X}_i be a sentence that appears in the corpus. The

PTMs	Data Corpus		
	CLINC150	CLINC150-Banking	CLINC150-Travel
LLaMA-7B	0.4731	0.5529	0.5312
RoBERTa-L	0.9991	0.9992	0.9989

Table 3.5: Average sentence anisotropy of model’s last layer.

anisotropy value can be calculated by:

$$\text{anisotropy} = \frac{1}{n^2 - n} \left| \sum_i \sum_{j \neq i} \cos(z(\mathbf{X}_i), z(\mathbf{X}_j)) \right|, \quad (3.2)$$

where \cos is the cosine similarity and $z(\cdot)$ is the sentence embedding from the last layer. A higher anisotropy value suggests that the sentence embeddings are less distinguishable by Cosine distance. The quantitative results presented in Table 3.5 show that the anisotropy values of LLMs are considerably lower in comparison to those of RoBERTa.

Chapter Review

In this chapter, we have delved into the critical realm of OOD detection within the context of LLMs. The growing utilization of LLMs across various natural language processing tasks has underscored the need to understand their capabilities and limitations, especially in scenarios involving distribution shifts. Our work deepens the comprehension of OOD detection capabilities of LLMs. Through meticulous analysis, we have showcased the effectiveness of LLMs for OOD detection under various settings, including zero-grad and generative fine-tuning scenarios. Our findings reveal that a simple OOD detector utilizing the cosine similarity function outperforms other sophisticated OOD detectors, especially in the few-shot setting. Our work may serve as a foundational stepping stone for future advancements in effectively and responsibly harnessing the potential of LLMs in diverse environments.

Chapter 4

Few-shot Near-OOD Detection

4.1 Introduction

Task-oriented conversational systems have become widespread across numerous sectors, including banking and travel [149, 150], where they provide a variety of services. Intent recognition [151], a key element of these systems, is crucial for enabling automated assistance to address customer needs. Given the wide range of user inquiries, out-of-distribution (OOD) intent detection [19, 26] seeks to protect the intent recognition system by identifying and alerting to malformed inputs.

Recent advancements in large language models (LLMs) [37, 133, 152] have significantly improved the detection of semantically distinct out-of-distribution (far-OOD) intents [7]. However, identifying semantically similar (near-OOD) intents remains challenging [7, 153]. We focus on few-shot near-OOD detection, due to its importance in practical applications [154, 155]. In a few-shot learning context, the training model receives

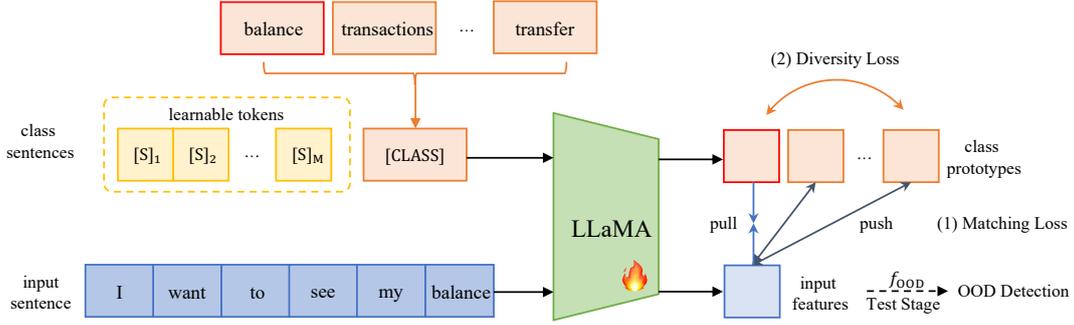


Figure 4.1: Proposed semantic matching framework. We prompt class name into a sequence of learnable tokens and forward them into LLMs to generate class prototypes. Further training between prototypes and input representations with matching loss and diversity loss, better ID classification and OOD detection can be performed.

only a few in-distribution (ID) examples per class. Obtaining discriminative information from limited ID examples for OOD detection is inherently challenging [156]. Previous strategies have involved using extra unlabeled data [157] and generating pseudo-ID examples [158]. We, however, focus on leveraging the inner knowledge of LLMs.

We propose a novel channel prototypical learning framework for few-shot near-OOD detection, shown in Figure 4.1. Our method employs a set of learned class prototypes to conduct semantic matching for ID and OOD classification. To effectively leverage the pre-existing knowledge in LLMs and the limited ID examples, our strategy involves using both the ID class names and utterances to develop the class prototypes. Drawing inspiration from channel models [159, 160] and prompt tuning [161], We feed into the LLMs the ID class names, preceded by a series of learnable continuous prompt embeddings, to generate a semantic prototype for each ID class. Furthermore, to ensure a wider variety of class representations—a factor known to improve OOD detection [162]—we adopt a diversified learning strategy aimed at reducing the mutual information between classes.

For evaluation, we construct few-shot near-OOD intent detection tasks by sample classes from the same domains in CLINC for ID and OOD classes. Since our method is based on LLMs, we compare it with widely used fine-tuning methods: generative Tuning [7] and discriminative tuning*. Our diversity-grounded channel prototypical learning approach generally achieves the best performance compared with these two ways.

4.2 Method

Problem Statement A complete intent recognition system (e.g., banking system) typically encompasses two tasks: ID intent classification and OOD intent detection. Formally, given an ID training set $\{x_i\}_{i=1}^N \subset \mathcal{X}_{\text{ID}}$ with N samples, an ID classifier f_{ID} is trained, mapping each utterance into ID intent label set $\{y_j\}_{j=1}^K \subset \mathcal{Y}_{\text{ID}}$ with K categories. In practical application, due to a possible distribution between the practical and training data, the ID classifier f_{ID} may meet OOD samples ($x_j \notin \mathcal{X}_{\text{ID}}$). Therefore, an OOD confidence scoring function f_{OOD} is applied to accept or reject such inputs.

4.2.1 Semantic Matching as ID Classification

Inspired by Liu et al. [7], LLMs (like LLaMA [37]) have shown impressive isotropy [129, 130], whereby the sentence embeddings produced by LLMs are distinguishable by Cosine distance. As such, we transform the ID classification task into a semantic matching task, shown in Figure 4.1.

*We use LlamaForSequenceClassification provided by Huggingface [148]

4.2.1.1 Class Diversity-grounded Channel Prototypical Learning

Class Prototypes. Our key idea is to push the input sentence closer to its corresponding class prototypes (treated as class-center representations [163]). Thus, we first prompt each class into a learnable sequence $\{\mathbf{c}_i\}_{i=1}^K$ based on its category name [159] and then generate prototypes $\{\mathbf{p}_i\}_{i=1}^K$ with LLM model $g(\cdot)$ through:

$$\mathbf{c}_i = [\mathbf{S}]_1[\mathbf{S}]_2\dots[\mathbf{S}]_M[\text{NAME}], \quad (4.1a)$$

$$\mathbf{p}_i = g(\mathbf{c}_i), \quad (4.1b)$$

where $[\mathbf{S}]_j$ ($j \in \{1, \dots, M\}$) is a learnable token with the same dimension as word embeddings (i.e., 4096 for LLaMA), M is the number of learnable tokens, and “NAME” corresponds to token embeddings of category names. In practice, the initialization of learnable tokens $[\mathbf{S}]_j$ is derived from token embeddings of prompts for specific scenarios, e.g., “[SCENARIO] intent of” where SCENARIO is “banking” for a banking system and is “travel” for a travel system.

Diversity Learning. To employ semantic matching as classification, the semantics between class prototypes \mathbf{p}_i must be distinguishable with satisfied mutual exclusion. Therefore, a diversity loss is proposed to help the class prototypes focus on diversified independent semantics:

$$\mathcal{L}_{\text{Diversity}} = \frac{1}{K^2} \sum_{i=1}^K \sum_{j=1, j \neq i}^K (\cos(\mathbf{p}_i, \mathbf{p}_j))^2, \quad (4.2)$$

where \cos denotes the Cosine similarity. Obviously, the semantic similarities between different class prototypes are expected to be 0.

4.2.1.2 Semantic Matching

In representation space, for input sentence \mathbf{x} , we push its representation closer to the corresponding class prototype \mathbf{p}_j and farther away from other ones via a contrastive manner:

$$\mathcal{L}_{\text{Match}} = -\log \frac{\exp(\cos(g(\mathbf{x}), \mathbf{p}_j)/\tau)}{\sum_{i=1}^K \exp(\cos(g(\mathbf{x}), \mathbf{p}_i)/\tau)}, \quad (4.3)$$

where τ is the temperature [164].

4.2.1.3 Optimization

We optimize the LLM model $g(\cdot)$ as well as learnable tokens $\{[S]_j\}_{j=1}^M$ with the joint training objectives:

$$\mathcal{L}_{\text{Joint}} = \lambda \mathcal{L}_{\text{Diversity}} + \mathcal{L}_{\text{Match}}, \quad (4.4)$$

where λ is a balancing hyper-parameter. Note that different from Lester et al. [161] that froze the backbone LLM, we utilize the low-rank adaptation (LoRA) [132] technique to parameter-efficiently tune it (detailed in Section 4.3.1).

4.2.2 Post-hoc OOD Scoring Functions

Following OOD detection with LLMs [7], we mainly focus on distance-base methods in post-hoc paradigm, i.e., Cosine distance (Cosine) [32], due to its exceptional performance. In particular, the Cosine confidence score is defined as the highest cosine similarity between the test input representation and the representations in the validation set:

$$\mathcal{S}(x) = \max\{\cos(z, z_i^{\text{val}})\}_{i=1}^V, \quad (4.5)$$

Datasets	Classes	Training	Validation	Test
<i>CLINC150-Banking</i>	15	1500	300	450
<i>CLINC150-Travel</i>	15	1500	300	450

Table 4.1: Data composition of each dataset.

where representation $z = g(x)$ is the output obtained from the penultimate layer of LLMs.

4.3 Experiment

4.3.1 Experimental Setup

Datasets. We explore a more challenging scenario, termed the *near-OOD* setting, where ID and OOD samples come from the same domain but with disjoint label sets. Following Liu et al. [7], we use CLINC150 [147] intent dataset and choose *Banking* and *Travel* domains. Within each domain, 50% of the classes are designated as ID, and the remaining classes as OOD. Detailed dataset statistics are provided in Table 4.1.

Evaluation Metrics. In accordance with Liu et al. [7], we adopt three widely accepted measures for OOD detection: (1) AUROC (area under the receiver operating characteristic curve), (2) FAR@95 (false alarm rate at 95% recall), which indicates the likelihood of erroneously labeling a negative sample as positive when recall or true positive rate (TPR) is 95%; Here, we consider the the OOD class as negative, and (3) AUPR (area under the precision-recall curve). Furthermore, accuracy serves as a metric for the ID classification task.

Shot	Tuning Method	CLINC-Banking				CLINC-Travel			
		ID ACC \uparrow	AUROC \uparrow	FAR@95 \downarrow	AUPR \uparrow	ID ACC \uparrow	AUROC \uparrow	FAR@95 \downarrow	AUPR \uparrow
5	Generative \dagger	0.882	0.962	0.255	0.968	0.964	0.974	0.148	0.983
	Discriminative	0.919	0.924	0.390	0.939	0.981	0.967	0.176	0.976
	Semantic Matching	0.964	0.970	0.215	0.970	0.990	0.981	0.103	0.989
10	Generative \dagger	0.949	0.968	0.157	0.974	0.984	0.982	0.078	0.988
	Discriminative	0.959	0.955	0.266	0.966	0.991	0.984	0.075	0.989
	Semantic Matching	0.967	0.970	0.149	0.974	0.993	0.984	0.063	0.993
Full	Generative \dagger	0.973	0.964	0.147	0.970	0.991	0.978	0.049	0.987
	Discriminative	0.988	0.964	0.197	0.969	0.996	0.996	0.013	0.997
	Semantic Matching	0.983	0.971	0.141	0.980	0.997	0.994	0.006	0.996

Table 4.2: The performance of LLaMA-7B fine-tuned with different methods for OOD detection and ID classification. ‘‘Shot’’ denotes the number of examples in the ID training or validation set. We report the average results of five seeds. \dagger is cited from the original paper. AUROC, FAR@95, and AUPR are metrics for OOD detection.

Implementation Details. Following prior work for OOD detection with LLMs [7], we employ the LLaMA-7B. All experiments are conducted over five seeds (1, 2, 3, 4, 5) on an NVIDIA A100 80G GPU card. For both class prototypes and input representations, we use the representation of the last token. τ in Eq. 4.3 is set to 0.01 and λ in Eq. 4.4 is set to 0.2 through cross-validation. We train the whole network for 25 epochs, and the model that achieved the lowest loss on the validation set will be used for testing. We use AdamW optimizer with a learning rate of 1×10^{-4} , further decayed by linear schedule. The LoRA configurations are that rank r is 16, scaling α is 16, and query/key/value/output projection matrices $\{W_q, W_k, W_v, W_o\}$ in each self-attention module need to be updated. For discriminative tuning, we use the same configurations.

4.3.2 Main Results

The main results are shown in Table 4.2. On one hand, the proposed semantic matching method consistently outperforms other tuning methods in most cases, especially for the intent classification task with low-resource training data. Take the results of 5-shot classification on CLINC-Banking as an example, semantic matching outperforms discriminative and generative tuning by 4.5% and 8.2% respectively, demonstrating the high data-efficient ability of our method. Meanwhile, the OOD detection performance of Cosine distance improves accordingly due to more accurate representations of classes. On the other hand, in the full-shot setting, while semantic matching yields slightly lower ID classification results on CLINC-Banking compared to the discriminative approach, its OOD detection performance is superior, and it also remains competitive on the CLINC-Travel dataset.

4.3.3 Analysis and Ablation Study

Variants of Class Prototypes. As elaborated in Sec 4.2.1, we use learnable vectors initialized with scenario-specific prompts plus a class name embedding to generate prototypes. A question worth exploring is what are the effects of other choices? Here, we additionally try 3 types: (1) randomly initialized learnable vectors plus class name embeddings; (2) only randomly initialized learnable vectors; (3) only class name embeddings, e.g., “transactions” in the CLINC-Banking dataset. Performance changes for ID intent classification as training progresses are shown in Figure 4.2. It reveals that (1) Our design (green curve) not only obtains the best performance but only requires a few iterations to achieve optimally. (2) Overall, learnable tokens are very important and can

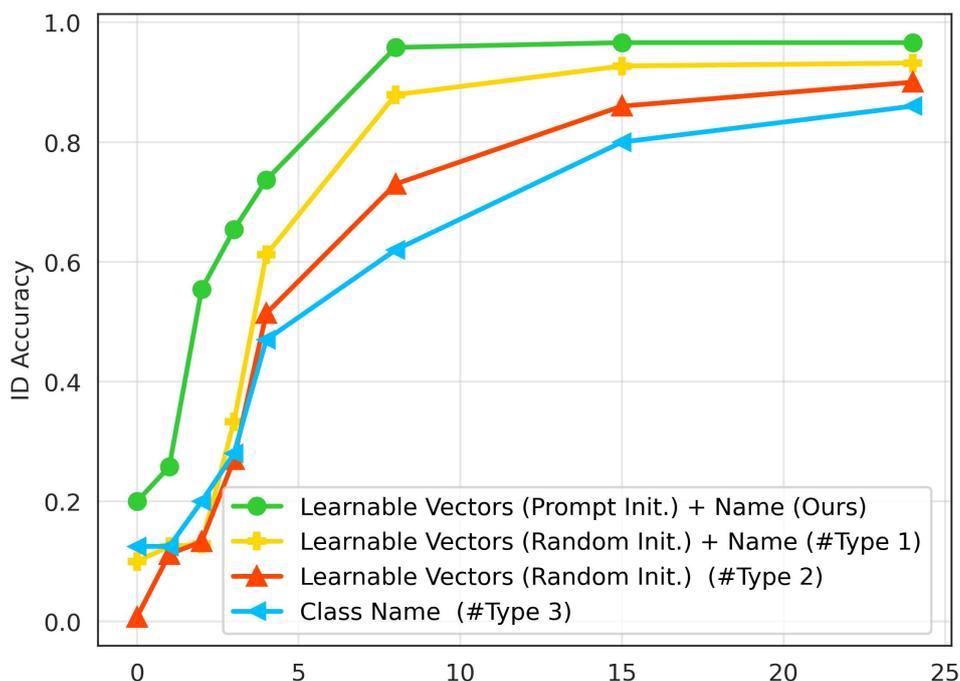


Figure 4.2: Performance of various class prototypes in the 5-shot scenario using CLINC-Banking dataset.

bring better results than fixed embeddings of class names (Type 3).

Effectiveness of Diversity Learning. We investigate the effectiveness of each training objective in the semantic matching framework with results shown in Table 4.3. We can find that (1) Incorporating diversity learning can boost the overall performance on both tasks. Notably, it enhances about 11% on the FAR@95 metric, showing strong effectiveness. (2) To better understand why diversity learning provides an improvement, we visualize the corresponding sentence representations in Figure 4.3. Intuitively, diversity learning makes the separation between classes clearer, thereby enhancing the classification performance of ID data. Simultaneously, it could implicitly distinguish potential OOD data, clarifying the boundaries between ID and OOD data and thus improving OOD detection accuracy.



Figure 4.3: UMAP [9] visualization of representations of test set and OOD data from 5-shot CLINC-Bank. The purple means the OOD data. The star indicates the learned prototype of each class.

CLINC-Banking (5-shot)				
Tuning Method	ID ACC \uparrow	AUROC \uparrow	FAR@95 \downarrow	AUPR \uparrow
Generative	0.882	0.962	0.255	0.968
Discriminative	0.919	0.924	0.390	0.939
only $\mathcal{L}_{\text{Match}}$	0.943	0.950	0.324	0.961
$\mathcal{L}_{\text{Match}} + \mathcal{L}_{\text{Diversity}}$	0.964	0.970	0.215	0.970

Table 4.3: Effectiveness of Diversity Learning.

4.4 Limitation

This work mainly has two limitations: (1) we only utilized the LLaMA-7B model, without investigating other open-source large models such as OPT [165], LLaMA-2 [37], Mistral [166], etc. Furthermore, the impact of larger models like LLaMA-Chat after instruct-tuning on this task was also left unexplored. In the proposed semantic matching framework, strong backbone models could potentially yield superior overall results.

(2) Presently, this study predominantly builds upon prior research by Liu et al. [7], focusing on the CLINC dataset. Therefore, intent datasets such as BANKING77 [167] and StackOverflow [168] remain unexplored in this context. Further exploration of these datasets would be beneficial.

Chapter Review

In this chapter, we have proposed a semantic matching paradigm for intent classification and OOD intent detection. By prompting class names into learnable vectors and pushing their representations yielded by LLMs closer to belonged sentence representations, semantics-awareness classification has been conducted. Extensive experiments compared to generative and discriminative tuning, as well as detailed analyses, were presented, demonstrating the effectiveness and data efficiency of our proposal.

Chapter 5

Overcoming Small-Scale Data

5.1 Introduction

Medical visual question answering (Med-VQA) has gained increasing attention over the past few years. Given a medical image and a clinical question about the image, it aims to find the correct answer by analyzing the visual information of the image. Med-VQA technology has great potential in medical and healthcare services. It can be used for computer-assisted diagnosis, intelligent medical guidance, clinical education and training, etc., which can help to significantly improve the quality of medical services and meet the increasing demand of the general public for medical resources. Specifically, with a reliable Med-VQA system, patients can easily acquire information about their health and be more engaged in the process of decision-making. For doctors, Med-VQA systems can be used to assist diagnosis by providing them a second medical opinion. The systems can also be used in clinical education to train medical professionals. Besides, Med-VQA technology can be potentially integrated into many conversational AI platforms to bring

enormous benefits to healthcare industry.

However, the research of Med-VQA is at an early stage. Unlike VQA in the general domain, where large-scale high-quality datasets [97, 113] are available, there is a lack of publicly-available and well-annotated datasets for training and evaluating Med-VQA systems. To correctly answer a clinical question about a radiology image, it requires clinical expertise and domain-specific medical knowledge, which makes it difficult to construct a realistic and accurate dataset for Med-VQA. VQA-RAD [11] is a first step in this direction. To our knowledge, it is the only available dataset with manual annotation, based on which several Med-VQA models have been proposed [169, 170]. VQA-RAD is a diverse dataset containing a variety of different types of clinical questions, with each question type sufficiently represented. But it does not provide semantic labels, e.g., labeled segmentations of organs and tumors or bounding boxes on objects, which are essential for training a Med-VQA model to find the region of interest in an image to answer complex clinical questions. Moreover, a practical Med-VQA system needs to exploit external knowledge apart from visual content to answer complex compositional questions involving inquiries such as “the functionality of an organ”, “the cause of a disease”, or “the treatment of a disease”, which is also not supported in VQA-RAD.

Meanwhile, the visual feature extraction module of existing Med-VQA models usually employs deep architectures and needs to be trained on a large collection of annotated radiology images, which however are often unavailable and costly to collect. To address this issue, a pioneering work [13] proposes mixture of enhanced visual features (MEVF) to pre-train the visual feature extraction module by constructing an auxiliary organ disease classification task on the radiology images of VQA-RAD[11] and observes positive effect. However, this approach cannot be transferred to other datasets, since the auxiliary pre-

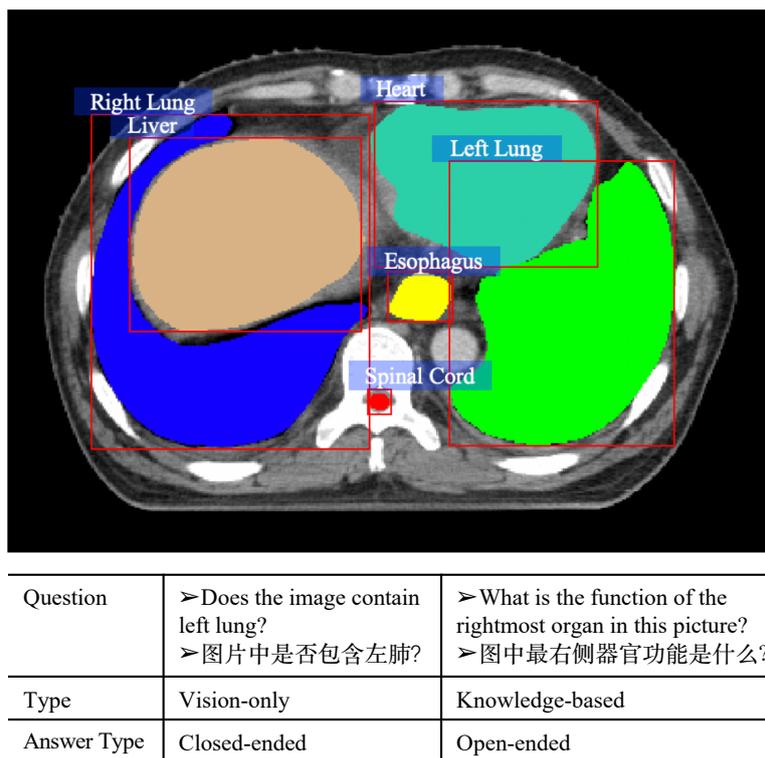


Figure 5.1: Exempler image and questions of our SLAKE dataset.

training task is designed based on the VQA-RAD dataset and requires extra effort for annotation.

Dataset	# Images	# QA Pairs	Question Type	Language	Knowledge Graph
VQA-RAD [11]	315	3.5K	Vision-only	EN	✗
SLAKE (Ours)	642	14K	Knowledge-based & Vision-only	Bilingual (EN & ZH)	✓

Table 5.1: Comparison of SLAKE with VQA-RAD.

To overcome the abovementioned challenges caused by small-scale data, we first construct a **semantically-labeled knowledge-enhanced (SLAKE)** dataset with accurate visual and textual annotations and an extendable knowledge base for Med-VQA. As shown in Figure 5.1, for each radiology image, we provide two kinds of visual annotations: *masks* for semantic segmentation and *bounding boxes* for object detection. Besides basic clinical questions, we also design compositional questions that require multiple reasoning steps,

and knowledge-based questions like [171] that involve external medical knowledge. In general, questions in SLAKE can be categorized as vision-only questions and knowledge-based questions. We provide detailed annotations to distinguish the two types of questions and guide the Med-VQA model to search for answers on the knowledge graph. Besides these new features, SLAKE is designed to be an English-Chinese bilingual dataset to broaden its application range. Further, SLAKE covers more body parts (e.g., neck and pelvic cavity) and more types of questions (e.g., shape and KG-related) than VQA-RAD. A comparison between our SLAKE and VQA-RAD is provided in Table 5.1.

Next, we tackle the data scarcity challenge by utilizing easily-available unannotated radiology image datasets for pre-training and representation distillation. First, we observe that the radiology images in current Med-VQA benchmarks mainly involve three human body regions – brain, chest, and abdomen, and there are large amounts of open-source unlabelled radiology images available for each region. Therefore, we propose to pre-train a visual feature extraction model (*teacher*) for each region respectively via contrastive learning. Second, to obtain a general and lightweight feature extractor, we distill the three teacher models into a small *student* model by contrastive representation distillation. The distilled model can be readily fine-tuned on any training dataset to facilitate the training of a Med-VQA system, without requiring further annotating process. Moreover, the small size of the distilled model can prevent overfitting to the training data, which typically only contains hundreds of radiology images.

To summarize, our contributions are as follows:

- We create SLAKE, a large-scale, semantically annotated, and knowledge-enhanced bilingual dataset for training and testing Med-VQA systems.

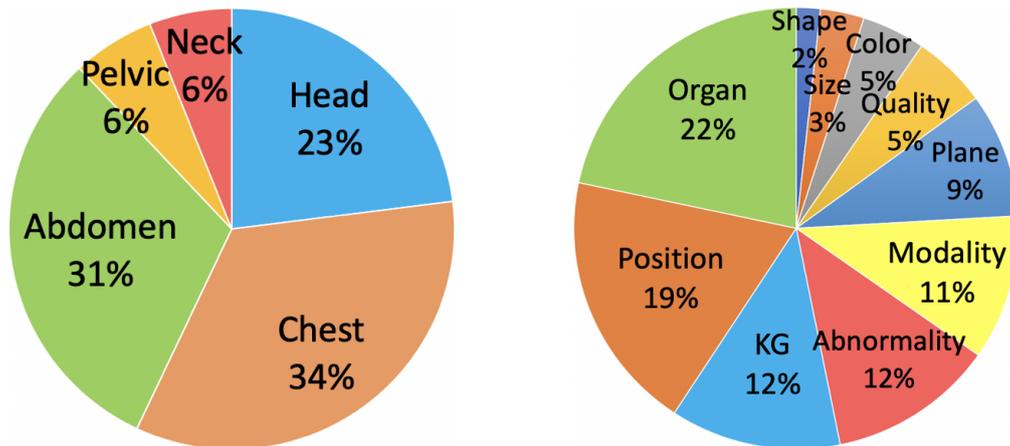


Figure 5.2: Left: proportions of images of five body parts. Right: distribution of the content types of questions.

- We propose a new pre-training framework that leverages easily-acquired unannotated radiology images to pre-train and distill a general and lightweight visual feature extractor for Med-VQA, which can be easily adapted to small-scale training datasets.
- We conduct extensive experiments with strong Med-VQA methods on two benchmarks VQA-RAD [11] and our SLAKE [2] to demonstrate the usefulness and benefits of the pre-trained model.

5.2 SLAKE Construction

In this section, we elaborate on the construction of our SLAKE dataset. In general, we ensure the diversity of the dataset in terms of modalities (e.g., CT, MRI, and X-Ray), covered body parts (e.g., head, neck, and chest), and question types (e.g., vision-only, knowledge-based, and bilingual).

5.2.1 Image Acquisition and Annotation

We select radiology images, covering healthy and unhealthy cases, from three open source datasets [172]* [173]† [174]‡. From [173], we randomly select 179 chest X-Ray images and keep the original disease labels. From [172] and [174], we randomly choose 463 single-slice images from 3D volume cases. Then, experienced physicians label organs and diseases as detailed as possible with ITK-SNAP [175]§ as shown in Figure 5.1.

In total, we annotate 642 images, including 12 diseases and 39 organs of the whole body. The diseases mainly include cancer (e.g., brain, liver, kidney, lung, etc.), and thoracic diseases (e.g., atelectasis, effusion, mass, pneumothorax, etc.). The images include 140 head CTs or MRIs, 41 neck CTs, 219 chest X-Rays or CTs, 201 abdomen CTs or MRIs, and 41 pelvic cavity CTs. The distribution is shown in Figure 5.2 (Left). Among these images, there are 282 CTs, 181 MRIs, and 179 X-Rays. All CTs and MRIs are axial single-slice. The number of images for each body part is set based on the complexity of the body part. For example, the number of diseases and organs in abdomen is much more than that in neck, so there are more images of abdomen than neck in the dataset.

5.2.2 Knowledge Graph Construction

To answer questions that require external medical knowledge, we construct a medical knowledge graph centered on organs and related diseases, which are the main objects of radiology images. We extract a set of 52.6K triplets $\langle head, relation, tail \rangle$ with

*<http://medicaldecathlon.com>

†<https://nihcc.app.box.com/v/ChestXray-NIHCC>

‡<https://doi.org/10.5281/zenodo.3431873>

§<http://www.itksnap.org>

	Examples
Organ	<Heart, Function, Promote blood flow>
	<Kidney, Belong to, Urinary System>
	<Duodenum, Length, 20-25cm>
Disease	<Pneumonia, Location, Lung>
	<Lung Cancer, Cause, Smoke>
	<Brain Tumor, Symptom, Visual impairment>
	<Cardiomegaly, Treatment, Medication>
	<Atelectasis, Prevention, Exercise>

Table 5.2: Examples of our medical knowledge graph.

medical knowledge from OwnThink[¶], a large-scale knowledge base built on Wikipedia. Here, *head* and *tail* are entities such as organ, disease, etc., and *relation* represents the relationship between entities, such as function or treatment. Then, we traverse the set to retrieve triplets related to organs and the corresponding diseases. We further clean the data by manually filtering out some entities that are not presented in medical images such as gastritis and nephritis.

Next, in order to extensively cover frequently referenced knowledge, we refine the filtered triplets with the following rules: (1) The triplets about an organ must describe its function or body system; (2) The triplets about a disease must describe the symptoms, locations, causes, treatment or prevention methodologies. Some examples are shown in Table 5.2.

Finally, we make the triplets bilingual and obtain 2603 triplets in English and 2629 triplets in Chinese.

[¶]<https://www.ownthink.com>

	Training set	Validation set	Test set
Plane	931	173	176
Quality	535	109	118
Modality	1072	203	217
Position	1876	412	390
Organ	2125	462	454
KG	1202	278	260
Abnormal	1230	245	221
Color	424	108	115
Shape	157	42	46
Size	297	77	73
Total	9849	2109	2070

Table 5.3: Statistics of questions in our SLAKE dataset.

5.2.3 Question Generation

Questions are proposed by experienced doctors. To accelerate this process, we develop an annotation system. In this system, we first pre-define a question template for each body part (i.e., head, neck, chest, abdomen, and pelvic cavity). Then, we define ten different content types (e.g., *modality*, *position*, *color*) for the questions, as shown in Table 5.3 and Figure 5.2 (Right). In each template, we provide many candidate questions for each content type. For example, the candidate question for a head image with the content type *organ* may be “Is this a study of the head?” or “What organ system is imaged?”. Physicians could choose those candidate questions or amend or even rewrite them entirely based on their personal clinical experience. The flexibility of our annotation system ensures the question diversity of SLAKE. Note that because we provide different candidates for bilingual questions, the number and content of them in our dataset are not

the same.

Moreover, we provide semantic label for each question. Specifically, we use $\langle vhead, _ , _ \rangle$ ($vhead$ is a placeholder) to denote vision-only questions. For a knowledge-based question like “Which organs in this image belong to the digestive system?”, we denote it as $\langle vhead, belong\ to, digestive\ system \rangle$. Such labeling helps to distinguish question type and identify the part of the question involving external knowledge.

Besides, recent studies [176, 177] have shown that VQA models may be susceptible to the statistical bias of answer distribution of the datasets. To mitigate the inherent bias of SLAKE, we make the answers balanced in general such that the VQA model will not be biased to the most popular answer in the dataset. For example, for the question “Is this a study of the abdomen?”, we make sure this question is asked with abdomen images and non-abdomen images with 50 – 50 chance, thereby keeping the numbers of “Yes” and “No” balanced.

5.2.4 Dataset Splitting

Here, we describe how to divide the obtained 642 images with 14,028 question-answer pairs and 5232 medical knowledge triplets for the training and evaluation of Med-VQA models.

In general, the splitting aims to provide a reliable measure of the generalization ability of the model trained on our dataset. Specifically, we split the dataset into training (70%), validation (15%), and test (15%) sets at the image level. The images in our dataset are split with the 70:15:15 ratio in each of the 8 categories: “head CT”, “head MRI”, “neck

CT”, and “chest X-Ray”, “chest CT”, “abdomen CT”, “abdomen MRI”, and “pelvic cavity CT”. Note that we only divide the images but the questions associated with each image are not split.

Besides, since VQA is usually formulated as a classification task [22, 169, 170], we follow the convention and make sure answers in the test set must appear in the training set. Finally, the images are split into 450 for training, 96 for validating, and 96 for testing. The number of questions of different type in each set is shown in Table 5.3.

5.3 Contrastive Pre-training and Representation Distillation (CPRD)

In current Med-VQA benchmarks, the radiology images mainly involve three human body regions: brain, chest, and abdomen. For each region, unlabeled images can be easily obtained from many large-scale open-source datasets. Motivated by this observation, we propose to train three specialized teacher models to focus on different body region respectively and then teach a student model to learn both intra- and inter-region features for Med-VQA, as illustrated in Figure 5.3.

5.3.1 Teachers: Intra-region Contrastive Pre-training

Let \mathcal{D}_{brain} , \mathcal{D}_{chest} , $\mathcal{D}_{abdomen}$ denote the set of radiology images for the three body regions respectively. Radiology images in each region have large diversity in terms of different organs and versatile imaging modalities, e.g., liver MRI, liver CT, and intestine CT in the

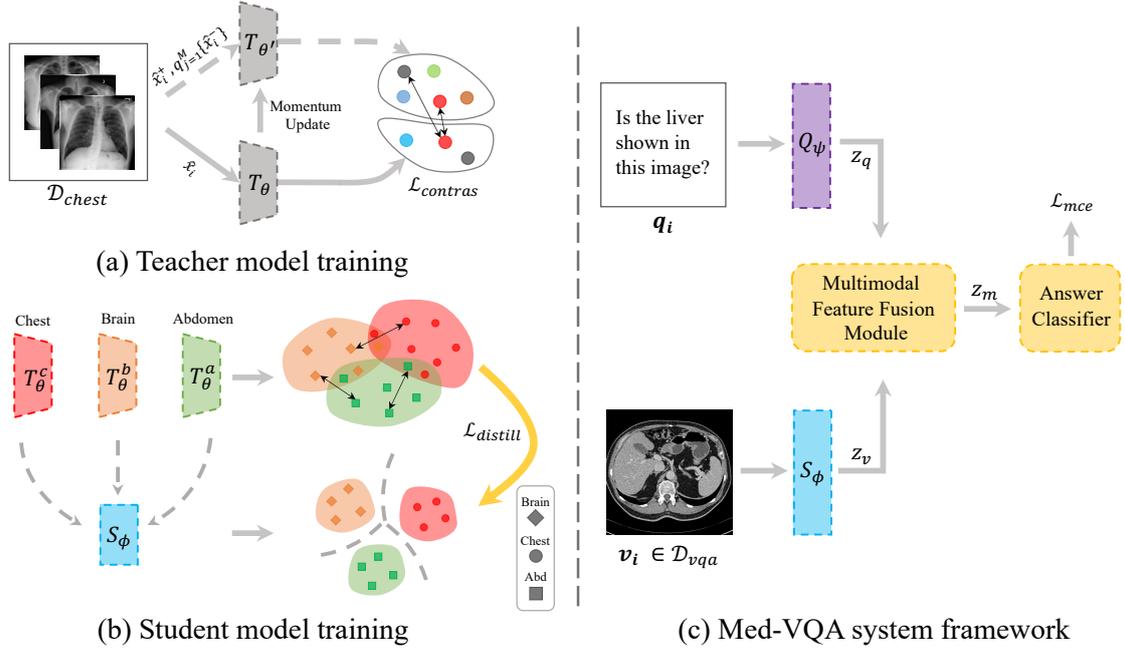


Figure 5.3: Our proposed CPRD framework for Med-VQA. (a) Train a teacher model T_θ by self-supervised contrastive learning on the chest region. (b) Distill three teacher models into one student model S_ϕ . (c) Apply the student model S_ϕ for Med-VQA.

abdomen region. Therefore, we employ Momentum Contrast [14], a self-supervised contrastive learning method, to train a *Teacher* model for each region with the corresponding dataset \mathcal{D}_r ($r \in \{brain, chest, abdomen\}$) to implicitly model these differences. As shown in Figure 5.3 (a), we sample an image x_i and a queue $q = \{x_j^-\}_{j=1}^M$ of M images different from x_i from \mathcal{D}_r . Then, data augmentation (such as resize, crop, color distort, and Gaussian blur), denoted as Aug , is applied on all the sampled images and produce:

$$\hat{x}_i = Aug(x_i), \hat{x}_i^+ = Aug(x_i), \hat{q} = \{\hat{x}_j^- = Aug(x_j^-)\}_{j=1}^M, \quad (5.1)$$

where \hat{x}_i and \hat{x}_i^+ are two different views of x_i , generated by applying random augmentation on x_i twice. An encoder T_θ is used to learn the feature representation of \hat{x}_i , i.e., $z_i = T_\theta(\hat{x}_i)$. Another momentum encoder $T_{\theta'}$ is used to produce the representations of \hat{x}_i^+ and \hat{q} , i.e., $\{z_i^+, z_1^-, z_2^-, \dots, z_M^-\}$. Since z_i and z_i^+ are the representations of differ-

ent views of x_i , z_i should be similar to z_i^+ but dissimilar to the other M representations in \hat{q} . The learning process can be guided by the InfoNCE contrastive loss [63]:

$$\mathcal{L}_{z_i, z_i^+, \{z_j^-\}} = -\log \frac{\exp(z_i \cdot z_i^+ / \tau)}{\exp(z_i \cdot z_i^+ / \tau) + \sum_{j=1}^M \exp(z_i \cdot z_j^- / \tau)}, \quad (5.2)$$

where τ is a temperature parameter [178] and \cdot stands for dot product. In practice, the length of the queue q is usually much larger than the mini-batch size, making it costly to update $T_{\theta'}$ by gradient back-propagation. Following [14], we update it in an efficient way: $\theta' \leftarrow \beta\theta' + (1 - \beta)\theta$, where β is the momentum coefficient. By optimizing the loss in Eq. (5.2), we obtain the teacher model T_{θ} for the region.

5.3.2 Student: Inter-region Representation Distillation

After obtaining the three teacher models: T_{θ}^a for $\mathcal{D}_{abdomen}$, T_{θ}^b for \mathcal{D}_{brain} , and T_{θ}^c for \mathcal{D}_{chest} , we design a lightweight *Student* model S_{ϕ} to distill representations of the teacher models, as shown in Figure 5.3 (b). Let $\mathcal{D}_{all} = \{\mathcal{D}_{brain}, \mathcal{D}_{chest}, \mathcal{D}_{abdomen}\}$. Inspired by the idea of contrastive representation distillation [179], for each region $\mathcal{D}_r \in \mathcal{D}_{all}$, for any image $x_i^r \in \mathcal{D}_r$, we randomly sample K images x_j^o ($j = \{1, \dots, K\}$) from the other two datasets $\mathcal{D}_o = \mathcal{D}_{all} \setminus \mathcal{D}_r$. First, we make the student model inherit knowledge of each teacher by enforcing its representation of x_i^r , $S_{\phi}(x_i^r)$, to be similar to that of the corresponding teacher model, $T_{\theta}^r(x_i^r)$, by minimizing the loss function

$$\mathcal{L}_{sim} = -\frac{1}{N} \sum_{r=1}^3 \sum_{i=1}^{L_r} \log \left(\frac{e^{T_{\theta}^r(x_i^r) \cdot S_{\phi}(x_i^r) / \tau}}{e^{T_{\theta}^r(x_i^r) \cdot S_{\phi}(x_i^r) / \tau} + \frac{K}{N}} \right), \quad (5.3)$$

where τ is the temperature parameter, L_r is the size of \mathcal{D}_r , and N is the size of \mathcal{D}_{all} ($1 < K < N$). Meanwhile, we enable the student model to acquire the ability to distinguish the three regions by enforcing $S_{\phi}(x_i^r)$ to be dissimilar to $T_{\theta}^o(x_j^o)$, the representation of x_j^o

(image of other regions) produced by the corresponding teacher model, by minimizing the loss function

$$\mathcal{L}_{dissim} = -\frac{1}{N \times K} \sum_{r=1}^3 \sum_{i=1}^{L_r} \sum_{j=1}^K \log\left(1 - \left(\frac{e^{T_{\theta}^o(x_j^o) \cdot S_{\phi}(x_i^r)/\tau}}{e^{T_{\theta}^o(x_j^o) \cdot S_{\phi}(x_i^r)/\tau} + \frac{K}{N}}\right)\right). \quad (5.4)$$

Further, we train the student model to produce more discriminative representations by learning to identify the body region R of x_i^r . Note that the images are already grouped by regions in open-source databases so the region labels can be automatically generated. This is achieved by minimizing the classification loss

$$\mathcal{L}_{class} = -\frac{1}{N} \sum_{i=1}^N \log P(R = r | W S_{\phi}(x_i^r)), \quad (5.5)$$

where W is a linear classification layer, and P is the prediction probability of the target region. Finally, by combining Eqs. (5.3), (5.4) and (5.5), the student model is trained with the loss function

$$\mathcal{L}_{distill} = \alpha(\mathcal{L}_{dissim} + \mathcal{L}_{sim}) + (1 - \alpha)\mathcal{L}_{class}, \quad (5.6)$$

where α is a balancing parameter.

5.3.3 Applying CPRD for Med-VQA

The distilled student model can be used as a universal visual feature extractor for any Med-VQA system based on radiology images. Figure 5.3 (c) shows a typical Med-VQA pipeline. Given a radiology image v_i and a question q_i as inputs, the student model S_{ϕ} is applied on v_i to extract the visual features $z_v = S_{\phi}(v_i)$, and a text encoder (e.g., LSTM [15] network) is used to extract the textual features q_i , i.e., $z_q = Q_{\psi}(q_i)$. Then, z_v and z_q will be fused by some attention-based module (e.g., BAN [21]) to produce multimodal features z_m .

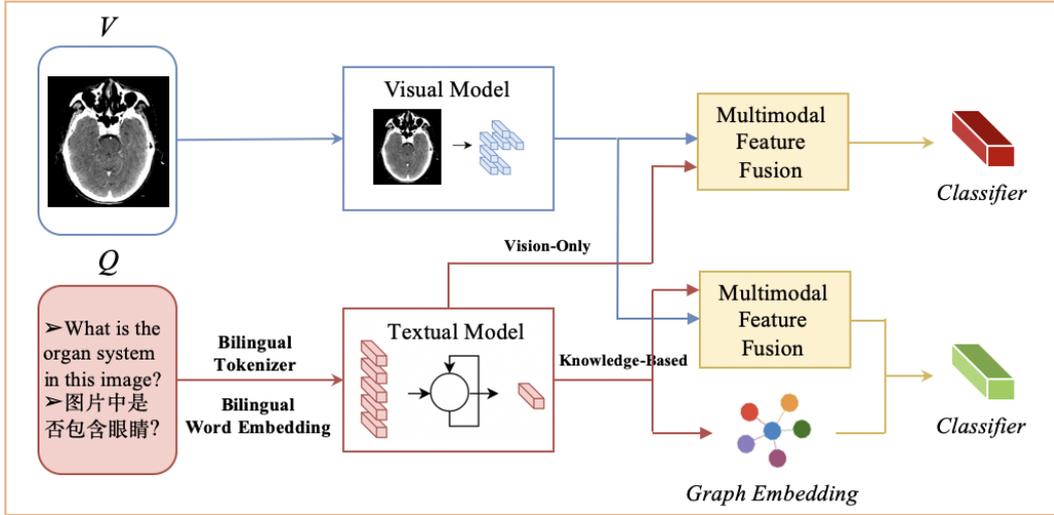


Figure 5.4: The Med-VQA framework on our SLAKE dataset.

Similar to general VQA, Med-VQA is also formulated as a classification problem [97]: predicting an answer from C fixed candidate answers in the training dataset. Note that there might be multiple correct answers for one question. As such, the multimodal features z_m will be fed to a classifier $\Phi(\cdot)$ (e.g., multilayer perceptron), to predict the probability of each candidate answer. All the model parameters, including those of the visual extractor S_ϕ , the text encoder Q_ψ , the feature fusion module and the classifier, are optimized in an end-to-end manner by minimizing the multi-label cross-entropy loss:

$$\mathcal{L}_{mce} = -\frac{1}{I} \sum_{i=1}^I \sum_{c=1}^C [l_i^c \log(\sigma^c(\Phi(z_m))) + (1 - l_i^c) \log(1 - \sigma^c(\Phi(z_m)))], \quad (5.7)$$

where l_i is the multi-hot encoding of the answers for the current (v_i, q_i) pair, σ is the sigmoid function, and I is the size of the training dataset.

5.4 Experiments

5.4.1 Evaluation of SLAKE

In this section, we conduct extensive experiments to comprehensively evaluate our SLAKE dataset. The pipeline of our experiments is illustrated in Figure 5.4. We experiment with a commonly used Med-VQA framework, stacked attention network (SAN) [22], on SLAKE. We use VGG16 [53] to extract visual features from radiology images. For bilingual questions, we first design a bilingual tokenizer to create bilingual word embeddings for the English questions and Chinese questions respectively. Then, a 1024D - LSTM is applied to extract textual semantics from these embeddings and classify types of questions. There are two sub pipelines in Figure 5.4. Given the extracted visual and textual features, vision-only tasks will be directed to the multimodal fusion module of SAN to create fused features for classification. For knowledge-based tasks, question-related embeddings extracted from the knowledge graph will be combined with the multimodal fused features for classification.

5.4.1.1 Dataset Analysis

We report the results for vision-only and knowledge-based questions in Table 5.4 and Table 5.5 respectively. Answers of “closed-ended” questions are limited multiple-choice options, while answers of “open-ended” questions are free-form texts. Open-ended questions are generally harder to answer than closed-ended ones.

Vision-only questions. In Table 5.4, we report the results in accuracy for vision-only

Language	Models	Overall	Open-ended	Closed-ended
English	VGG+SAN	72.73	70.34	76.13
	VGG _{seg} +SAN	75.36	72.20	79.84
Chinese	VGG+SAN	74.27	73.64	75.20

Table 5.4: Accuracy for vision-only questions (%).

Language	Models	Overall
English	VGG+SAN	70.27
	VGG+SAN+KG	72.30
Chinese	VGG+SAN+KG	75.01

Table 5.5: Accuracy for knowledge-based questions (%).

questions in both English and Chinese. Compared with VQA in the general domain, clinical questions in Med-VQA need to be answered as accurate as possible because they relate to health and safety. It can be seen that the baseline models achieve accuracy of around 73% which is still far away from practical use in the medical domain. There is a wide gap between this and clinical standard, which shows that SLAKE is challenging. Moreover, it can be seen that the overall accuracy is roughly the average of those of open-ended and closed-ended questions, proving that the question distribution of SLAKE is balanced.

Besides, to demonstrate the usefulness of the semantic visual annotations elaborated in Section 5.2.1, we design another model, VGG_{seg}+SAN. First, we pretrain a fully convolutional network (FCN) with VGG backbone by the segmentation task of radiology

images with respect to the *mask* labels in the training set. Then, we initialize the VGG backbone in the Med-VQA model with the pretrained parameters. The overall accuracy increases from 72.73% to 75.36% with a 2.6% improvement, which shows that our semantic visual annotations could improve the reasoning abilities of the model.

Knowledge-based questions. We leverage the self-built medical knowledge graph to answer knowledge-based questions. First, we randomly initialize an embedding for each entity in the knowledge graph and use the TransE [180] method to enforce the embeddings of the entities in each triplet, $\langle head, relation, tail \rangle$, to satisfy: $head + relation \approx tail$. Then, based on the semantic textual annotations (Section 5.2.3), we train two LSTMs to predict the words for the “relation” and “tail” of a question separately. Next, we find the corresponding entity embeddings of the relation and tail from the graph and use them to obtain the head entity embedding based on the above approximate equation, which is then combined with the fused multimodal features for final prediction. The result is reported in Table 5.5. For comparison, we also try to predict answers without using the knowledge graph. The result is 2.0% lower, indicating that the constructed knowledge graph is informative and it is helpful to leverage external structural knowledge to tackle knowledge-based questions.

5.4.2 Evaluation of CPRD

In this section, we extensively evaluate the effectiveness of the visual feature extractor pre-trained by our proposed CPRD framework on the only two available manually-annotated Med-VQA datasets. We experiment with common Med-VQA methods and show that the pre-trained feature extractor can be used to significantly improve their performance.

5.4.2.1 Datasets

VQA-RAD [11] consists of 315 radiology images and 3,515 question-answer pairs. We follow the data splitting in [13]. For our **SLAKE** [2], we use the English version, referred to as SLAKE-EN, which contains 642 radiology images and 7,033 question-answer pairs. We use the original data splitting. Besides, questions in VQA-RAD and SLAKE are both categorized into “closed-ended” questions whose answers are in limited choices, and “open-ended” questions whose answers are free-form text.

5.4.2.2 Experimental Setup

To train the teacher and student models, we randomly sample 22,995 unlabelled radiology images from open-resource databases^{||}, including 7,811 chest X-Rays, 7,592 abdomen CTs, and 7,592 brain CTs and MRIs. Our experiments are conducted on a Ubuntu server with 8 NVIDIA TITAN 12GB Xp GPUs. All the hyper-parameters of the teacher and student models are chosen by cross validation via observing the loss in Eq. (5.2) and Eq. (5.6).

Teachers. For each region-focused teacher model, we use ResNet-50 to instantiate T_θ and $T_{\theta'}$ (Sec. 5.3.1) and train for 800 epochs with 4 GPUs for about 7 hours. In each epoch, the mini-batch size is 128, and the queue length M is 1,024. The temperature parameter τ is set to be 0.2, 0.1, and 0.1 for brain, chest and abdomen respectively. For model optimization, we use SGD optimizer with $1.5e^{-2}$ initial learning rate decayed by cosine schedule.

^{||}<http://medicaldecathlon.com/>

Models	VQA-RAD [11]			SLAKE-EN [2]		
	Overall	Open	Closed	Overall	Open	Closed
MFB fw. [181]	50.6	14.5	74.3	73.3	72.2	75.0
SAN fw. [22]	54.3	31.3	69.5	76.0	74.0	79.1
BAN fw. [21]	58.3	37.4	72.1	76.3	74.6	79.1
MEVF+SAN [13]	64.1	49.2	73.9	76.5	75.3	78.4
MEVF+BAN [13]	66.1	49.2	77.2	78.6	77.8	79.8
CPRD+BAN (ours)	67.8	52.5	77.9	81.1	79.5	83.4
MEVF+BAN+CR [170]	71.6	60.0	79.3	80.0	78.8	82.0
CPRD+BAN+CR (ours)	72.7	61.1	80.4	82.1	81.2	83.4

Table 5.6: Test accuracy of our method and baselines.

Student. We use ResNet-8 as the student model (Sec. 5.3.2) and train for 240 epochs with 1 GPU. We use SGD optimizer to minimize the loss $\mathcal{L}_{distill}$ with 0.05 initial learning rate decayed by cosine schedule. Besides, the queue length K is 8192, the temperature parameter τ is 0.07, and α in Eq. (5.6) is 0.9.

Med-VQA. After training the student model, we use the weights in the last epoch as initialization and fine-tune the model on a Med-VQA dataset for 100 epochs. We use Adamax optimizer with initial learning rate $2e^{-3}$ for model optimization. We use accuracy as the evaluation metric.

5.4.2.3 Comparison with the State-of-the-arts

We use our pre-trained model CPRD as the visual feature extractor, combined with the BAN attention mechanism [21] with or without the CR reasoning module [170] for Med-

VQA. To demonstrate the necessity of domain-specific pre-training, we compare with general VQA frameworks including MFB [181], SAN [22], and BAN [21].** Further, we compare with MEVF [13], which is the only baseline that uses a small model and pre-trains with medical images.

The results on VQA-RAD [11] and SLAKE [2] are reported in Table 5.6. For a fair comparison, all methods use a 1024-D LSTM network to extract textual features with word embeddings pre-trained by GloVe [182]. For MFB, SAN and BAN, we use ResNet-50 pre-trained on ImageNet as the visual feature extractor. The following observations can be made. (1) Our method CPRD+BAN not only improves upon the performance of the strong baseline MEVF+BAN [13], but also achieves new results on the two benchmarks when further incorporating the CR [170] module. (2) Although MEVF+BAN [13] can significantly outperform the base framework BAN [21] on VQA-RAD, its performance gain on SLAKE is less significant ($\sim 2\%$), far lower than the gain brought by our CPRD+BAN ($\sim 5\%$). This demonstrates the generalization ability of our CPRD model on different datasets.

5.4.2.4 Ablation Analysis

We conduct an ablation study to analyze the impact of different pre-training strategies for the visual feature extraction module of Med-VQA. The results are summarized in Table 5.7. Specifically, we use BAN [21] as the multimodal feature fusion module and LSTM as the textual encoder for all methods in this subsection. Compared with the large models (i.e., VGG-16 and ResNet-50) pre-trained on ImageNet, it can be seen

**MFB, SAN, and BAN stand for the key reasoning module of the respective framework, where the visual and textual modules can be any applicable models.

Visual Modules	Overall(%)	Open(%)	Closed(%)	#Parameters (M)
VGG-16 [53] (ImageNet)	56.8	35.2	71.0	134.8
ResNet-50 [20] (ImageNet)	58.3	37.4	72.1	23.8
MEVF [13]	66.1	49.2	77.2	1.2
ResNet-8 (random init)	63.2	47.2	73.8	0.1
ResNet-8 (our CPRD)	67.8	52.5	77.9	0.1

Table 5.7: Comparison of different visual modules in test accuracy and model size on VQA-RAD [11]. The number of parameters is calculated on the visual module only.

that lightweight models (i.e., MEVF and ResNet-8) perform better. Further, ResNet-8 pre-trained by our CPRD achieves better results than with random initialization, and outperforms the strongest baseline MEVF with much fewer parameters. This again demonstrates the effectiveness and advantages of our CPRD model.

5.4.2.5 Visualization

The t-SNE [10] visualization of the representations learned by the ResNet-8 student model on the images of \mathcal{D}_{all} (Sec. 5.3.2) is shown in Fig. 5.5 (left). It can be clearly seen that the student model learns discriminative representations for different regions. Further, the representations of brain CT and brain MRI are well separated, indicating that the student model also captures the differences among versatile imaging modalities for the same region. To demonstrate the visual evidence used in Med-VQA models for prediction, in Fig. 5.5 (right), we show the Grad-CAM [12] maps for visual modules based on the final predicted answers of our CPRD+BAN and a strong baseline MEVF+BAN. The first row is about a brain MRI image, and the second is about a chest X-Ray image,

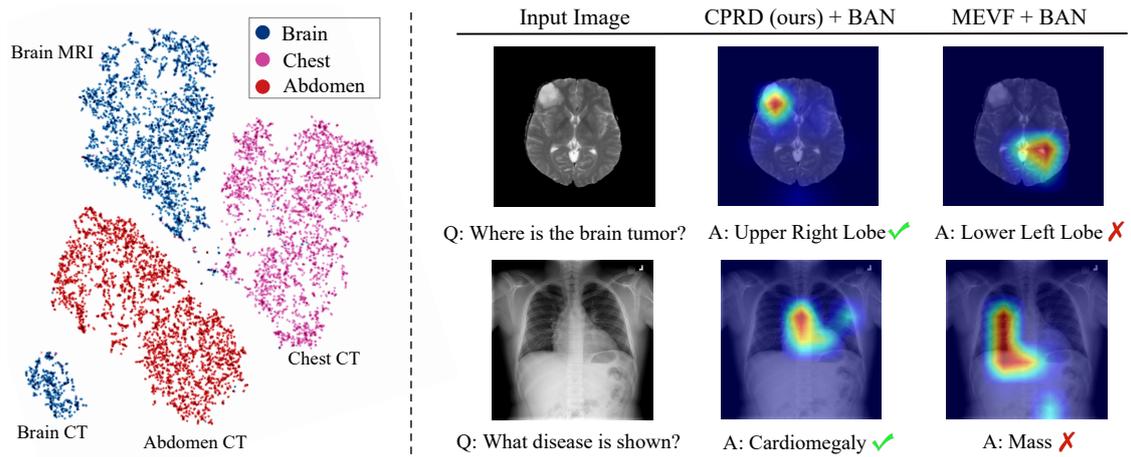


Figure 5.5: (Left) t-SNE visualization of the representations learned by the student model; (Right) Grad-CAM maps from the visual modules of Med-VQA methods. MEVF+BAN is the baseline. ✓ and ✗ indicate the correctness of the answer given by each method.

both from the test set of the SLAKE [2] dataset. It can be seen that our model can correctly answer the questions by locating the right visual evidence about the questions, which demonstrates the effectiveness of our visual module.

Chapter Review

This chapter first introduces SLAKE, a large bilingual dataset designed to advance Med-VQA research by providing diverse and balanced visual and textual annotations, alongside a unique medical knowledge graph. SLAKE facilitates the development of more powerful Med-VQA systems, addressing the challenge of data scarcity in this domain. To further enhance Med-VQA performance, this chapter proposes a two-stage pre-training framework that leverages unannotated radiology images for contrastive learning and representation distillation, resulting in a lightweight visual feature extractor. When integrated into existing Med-VQA methods, this pre-trained model effectively overcomes the small-scale challenges.

Chapter 6

Improving Reasoning Abilities

6.1 Introduction

Medical visual question answering (Med-VQA) considers the problem of taking a medical image and a clinical question in natural language related to the image as input and inferring the correct answer also in natural language. Med-VQA has enormous potential in assisted diagnosis and patient education. As a domain-specific branch of visual question answering (VQA), the research of Med-VQA is still in an early stage, where the literature is rather limited. Hence, we start with introducing VQA, which has recently attracted a great deal of attention from both the computer vision and natural language processing research communities.

VQA focuses on visual perceptual tasks that require common perceptual abilities shared by humans. For example, given a scenery image with a beautiful sunset, either a child or an adult can easily answer the question “what color is the sunset?”. Generally,



Chest X-Ray

	Closed-ended	Open-ended
Question	Are there any abnormalities?	Where is the lesion in this image?
Answer	Yes	Left Lower Lung



Abdomen CT

	Closed-ended	Open-ended
Question	Is this an MRI image?	What is the organ on the left in the picture?
Answer	No	Liver

Figure 6.1: Examples of Med-VQA tasks. For closed-ended questions, the answers are limited, e.g., “yes” or “no”. For open-ended questions, the answers can be free-form text.

visual perceptual tasks consist of easy tasks such as “does the man wear glasses?” and difficult tasks such as “which object in the picture has the same color as the pet dog in front of the man?”. It requires multi-level reasoning skills to solve both kinds of tasks. Easy perceptual tasks require basic skills, e.g., basic-level object recognition and scene understanding, while difficult tasks require higher-level reasoning skills such as counting, comparing, or logical inferring. Nevertheless, most of the existing VQA models are designed for coping with either easy tasks or difficult tasks. Simultaneously solving the two kinds of tasks in a single model is challenging and only considered in the high-data regime [183, 184].

Med-VQA tasks are, however, much more challenging than general VQA tasks. On the one hand, accurate answers are imperative for clinical questions, as they are related to health services and education. To this end, a Med-VQA system should be capable of handling multi-level tasks, including basic perceptual tasks such as identifying the body regions in an image, and difficult tasks such as counting the number of nodes,

locating lesions, or evaluating the health of an organ by its size. Therefore, to infer correct answers, it is essential for the system to acquire domain-specific knowledge and multi-level reasoning skills. On the other hand, well-annotated Med-VQA datasets are extremely lacking, since it requires medical expertise to construct high-quality datasets, which is both costly and time-consuming. To our best knowledge, there are only two manually annotated datasets available - VQA-RAD [11] and our SLAKE [2]. Both of them only contain hundreds of radiology images but include various types of clinical questions. Therefore, it is not effective to train a typical large VQA model from scratch for Med-VQA with the small-scale training datasets. Moreover, it is impossible to apply popular object-detection-based VQA models such as UpDn [56], Pythia [101], and VL-BERT [99] for Med-VQA, due to the lack of visual object labels and the small size of training data.

Previous research tried to apply existing VQA models for Med-VQA. More specifically, they employed deep architectures pre-trained on general datasets such as ImageNet [185] and then fine-tuned the models on small-scale Med-VQA training data [46, 47, 51]. However, due to the large differences in image patterns and language styles of medical data and non-medical data [186], transfer learning provides little benefit [187]. To overcome this problem, [13] proposed mixture of enhanced visual features (MEVF) to learn an initialization for the visual extractor of a Med-VQA model. In particular, they combined an auto-encoder pre-trained with an image reconstruction task on undisclosed external medical datasets and a 4-layer convolutional neural network pre-trained with an auxiliary classification task on the VQA-RAD dataset [11] using the meta-learning algorithm MAML [58]. While this work alleviates the problem to some extent, it cannot be easily applied on other Med-VQA datasets since the auxiliary classification task is

dataset-dependent and requires extra laborious annotations. Besides, it does not explore improving the reasoning module which is of critical importance in solving high-level reasoning tasks. There are some recent attempts [59, 60] to pre-train a multimodal Transformer on large medical vision-language datasets and fine-tune it on Med-VQA tasks. However, large Transformer models tend to overfit on existing small-scale Med-VQA training datasets.

In this work, we explore lightweight models like MEVF [13] and focus on improving both the reasoning module and the visual feature extractor of a Med-VQA system. First, to make the system possess task-adaptive reasoning ability, we design a novel conditional reasoning mechanism, which includes a question-conditioned reasoning (QCR) module and a type-conditioned reasoning (TCR) strategy. QCR enables the model to gain question-specific reasoning skills by leveraging question attention information to modulate multimodal fusion features. Further, it can be seen that Med-VQA tasks mainly consist of two types, closed-ended questions and open-ended questions, as shown in Figure 6.1. For closed-ended questions, the answers are limited choices according to the prompt words, e.g., the answer to the question starting with “Does” can only be “Yes” or “No”. For open-ended questions, the answers are free-form, e.g., no fixed choices for questions starting with “What”. Generally, open-ended tasks are harder to solve than closed-ended ones, and current Med-VQA models usually perform poorly on open-ended tasks. Therefore, motivated by the disparity of the needed reasoning skills for open-ended and closed-ended tasks, we design a TCR strategy to handle the two different types of tasks separately, by learning different sets of reasoning skills.

Second, to address the data scarcity problem, we propose to pre-train a visual feature extractor for Med-VQA in an unsupervised manner without requiring any human anno-

tations. We observe that 1) there involve various types of organs and imaging modalities (e.g., brain MRI, brain CT, chest X-Ray, and abdomen CT) in Med-VQA tasks; and 2) there are many such types of unlabeled radiology images available in open-access sources. Therefore, we propose to leverage these publicly available datasets to pre-train a visual extractor for learning high-level patterns and characteristics of different organs and imaging modalities through contrastive self-supervised learning. Having learned prior knowledge of radiology images, the pre-trained feature extractor can be readily adapted to train Med-VQA systems, even with small-scale training datasets.

The main contributions are summarized as follows:

- We design a novel conditional reasoning mechanism to empower the reasoning ability of Med-VQA models, which contains a question-conditioned reasoning function and a type-conditioned reasoning strategy, by leveraging both question content and task type.
- We propose to leverage publicly available resources to pre-train a generic visual feature extractor for Med-VQA via contrastive self-supervised learning, which can be easily adapted to existing small-scale training datasets.
- We conduct an extensive evaluation on existing Med-VQA benchmarks to validate the effectiveness of the proposed conditional reasoning mechanism and the pre-trained visual feature extractor and observe significant improvements over state-of-the-art methods.

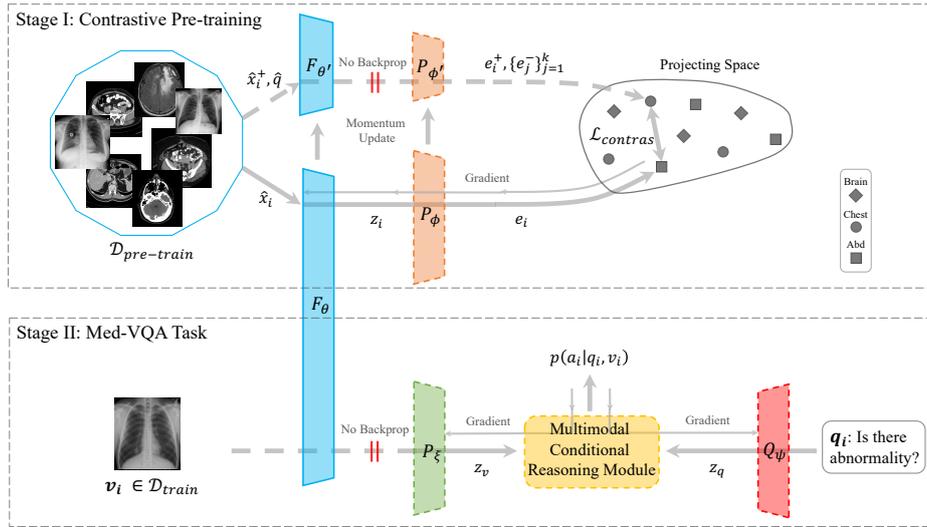


Figure 6.2: Our proposed method for training a Med-VQA model. In stage I, we pre-train a visual feature extractor for Med-VQA by contrastive self-supervised learning. In stage II, we solve Med-VQA tasks by introducing a conditional reasoning mechanism.

6.2 Methodology

In this section, we present our method for training a Med-VQA model, which consists of two stages as shown in Figure 6.2. In stage I, we propose to learn prior knowledge of radiology images for the visual module tailored for Med-VQA from a collected dataset of publicly available unlabeled radiology images by contrastive self-supervised learning [14]. In stage II, we propose a conditional reasoning mechanism with a question-conditioned reasoning component and a type-conditioned reasoning strategy to adaptively learn effective reasoning skills for different Med-VQA tasks.

6.2.1 The paradigm of Med-VQA

The goal of Med-VQA is to automatically answer a clinical question about a radiology image. By convention, it is formulated as a single-label classification task where there are C candidate answers and each answer is a label. Unlike general VQA [56] where a question may have multiple answers, there is only one correct answer for each question in Med-VQA. Denote by $\mathcal{D}_{med-vqa} = \{(v_i, q_i, a_i)\}_{i=1}^N$ the training dataset for a Med-VQA model, where N is the number of training examples, and v , q , and a denote the image, question, and answer of a task respectively. A typical Med-VQA model aims to learn a function f that maps each (v_i, q_i) pair to a score vector $\mathbf{s} \in \mathbb{R}^C$ where the j -th element s_i^j is the score for the j -th answer. The probability for the j -th answer is obtained by the softmax function, i.e., $p(s_i^j) = \frac{e^{s_i^j}}{\sum_{j=1}^C e^{s_i^j}}, 1 \leq j \leq C$. The function f is usually instantiated as a neural network with parameters δ , and optimized by minimizing the cross-entropy loss:

$$\mathcal{L}_{vqa} = -\frac{1}{N} \sum_{i=1}^N a_i \log p(f_{\delta}(v_i, q_i)). \quad (6.1)$$

The function f usually consists of an image feature extractor, a question feature extractor, an attention-based feature fusion module, and an answer classifier, which are trained together in an end-to-end manner. In this work, we focus on designing the visual feature extractor and the feature fusion module. We use long short-term memory network (LSTM) and multi-layer perceptron (MLP) as default choices for the question feature extractor and the answer classifier respectively.

6.2.2 Contrastive Pre-training (CP)

Due to the characteristic of radiology images, it is not effective to directly apply deep models (e.g., ResNet [20]) pre-trained on general datasets such as ImageNet to extract visual features of radiology images. Moreover, due to the small scale of existing Med-VQA datasets (only a few hundred of radiology images available) [2, 11], fine-tuning pre-trained large models on them may lead to severe overfitting [13]. As such, to address the vast diversity of radiology images in terms of different organs and imaging modalities, we propose to pre-train a visual feature extractor by contrastive self-supervised learning on unannotated radiology images. Specifically, we collect a large set of radiology images of different organs and in different modalities, e.g., brain CT, brain MRI, chest X-Ray, and abdomen CT, and train a deep neural network that can pull together similar images and push away the dissimilar ones. Further, to avoid overfitting, after pre-training, we freeze the parameters of the large model and train an additional small network on Med-VQA tasks. We discuss the impact of pre-training datasets and strategies to avoid overfitting in Section 6.3.4.1.

Particularly, denote by $\mathcal{D}_{\text{pre-train}}$ the set of unlabelled radiology images collected for pre-training and $\mathcal{D}_{\text{train}}$ the training set of the Med-VQA dataset respectively. As shown in Figure 5.3 stage I, we randomly sample a radiology image x_i and a queue $q = \{x_j^-\}_{j=1}^K$ of K images disjoint with x_i from $\mathcal{D}_{\text{pre-train}}$. Then, a set of data augmentation operations, denoted as Aug , which includes random crop, color distortion, resize with random flip, and Gaussian blur, is applied to all images:

$$\hat{x}_i = Aug(x_i), \hat{x}_i^+ = Aug(x_i), \hat{q} = \{\hat{x}_j^- = Aug(x_j^-)\}_{j=1}^K, \quad (6.2)$$

where \hat{x}_i and \hat{x}_i^+ are generated by applying Aug on x_i twice and considered as two

different views of x_i . A feature extractor (usually a convolutional neural network such as ResNet) is used to obtain the feature representation of the anchor point \hat{x}_i , i.e., $z_i = F_\theta(\hat{x}_i) : \mathcal{X} \rightarrow \mathcal{F}$, where \mathcal{X} and \mathcal{F} are the input image space and the feature space respectively. Further, a non-linear layer projects the feature representation into the projection space \mathcal{P} , i.e., $e_i = P_\phi(z_i) : \mathcal{F} \rightarrow \mathcal{P}$. Similarly, another two networks $F_{\theta'}$ and $P_{\phi'}$ that share the same structure as F_θ and P_ϕ respectively are used to map \hat{x}_i^+ and \hat{q} to obtain the feature representations $\{z_i^+, z_1^-, z_2^-, \dots, z_K^-\}$ and the projections $\{e_i^+, e_1^-, e_2^-, \dots, e_K^-\}$ respectively. Since e_i and e_i^+ are the projections of different views of x_i , e_i should be similar to e_i^+ (*positive pair*), and dissimilar to $\{e_1^-, e_2^-, \dots, e_K^-\}$ (*negative pairs*).

Following SimCLR [65], we conduct contrastive learning in the projection space using the InfoNCE contrastive loss [63] with dot product similarity:

$$\mathcal{L}_{e_i, e_i^+, \{e_j^-\}} = -\log \frac{\exp(e_i \cdot e_i^+ / \tau)}{\exp(e_i \cdot e_i^+ / \tau) + \sum_{j=1}^K \exp(e_i \cdot e_j^- / \tau)}, \quad (6.3)$$

where τ is a temperature hyper-parameter [178].

Since the length K of the queue q is much larger than the training mini-batch size, it is costly to update $F_{\theta'}$ and $P_{\phi'}$ by gradient back-propagation. Following MoCo-v2 [14], we update them in a momentum-based way:

$$\theta' \leftarrow m\theta' + (1 - m)\theta, \quad (6.4)$$

$$\phi' \leftarrow m\phi' + (1 - m)\phi, \quad (6.5)$$

where m is a momentum coefficient close to 1.

After pre-training, in stage II, we can apply F_θ to extract the visual features of radiology images from \mathcal{D}_{train} . However, we observe in experiments that directly fine-tuning the

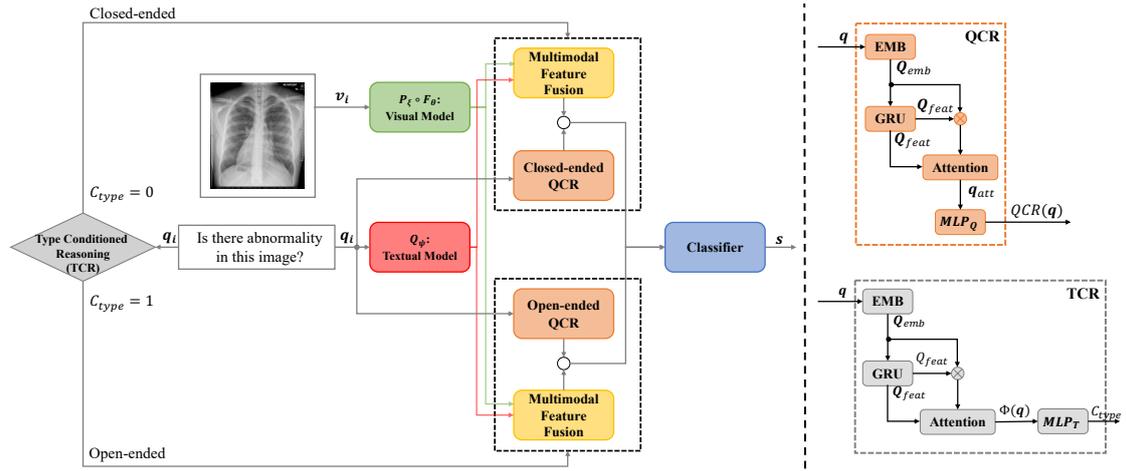


Figure 6.3: Our proposed Med-VQA model with conditional reasoning. To prevent overfitting, we freeze the visual model F_θ (pre-trained in stage I) and append a non-linear layer P_ξ for fine-tuning on medical images. First, the TCR module classifies the question as open-ended or closed-ended and chooses the corresponding branch for reasoning. Then, the question features extracted by the textual model Q_ψ will be fused with the image features by the multimodal feature fusion module (e.g., BAN) and our QCR module. Finally, the answer is obtained by an MLP classifier.

pre-trained large model easily leads to overfitting on the small-scale training set of Med-VQA datasets. Hence, we propose to keep θ fixed and append a non-linear layer P_ξ after F_θ for adaptation on \mathcal{D}_{train} . We set the input and output dimensions of P_ξ to be the same. As such, the visual features are obtained by: $z_v = P_\xi(F_\theta(v_i))$. Notice that unlike MEVF [13] that designs an auxiliary classification task and requires additional annotation effort for pre-training, our method leverages large amounts of unlabeled images to achieve better generalization.

6.2.3 Conditional Reasoning (CR)

Besides improving the feature extraction ability of the Med-VQA model, another key issue is to improve its reasoning ability. Here, we propose a conditional reasoning mechanism, aiming to solve different Med-VQA tasks with task-adaptive reasoning skills, as illustrated in Figure 6.3. It includes a question-conditioned reasoning module and a type-conditioned reasoning module, building on top of a basic multimodal reasoning module (the multimodal feature fusion module indicated by the yellow block in Figure 6.3). We first review the multimodal reasoning module and then elaborate on our proposed reasoning modules.

6.2.3.1 Multimodal Reasoning

In this work, we utilize bilinear attention networks (BAN) [21], a popular model used in general VQA, for multimodal feature fusion and reasoning. Given the extracted visual features $\mathbf{Z}_v \in \mathbb{R}^{d_v \times N}$ (d_v is the dimension of image features and N is the number of channels), textual features $\mathbf{Z}_q \in \mathbb{R}^{d_q \times L}$ (d_q is the dimension of word features and L is the number of words in the question), and the number of reasoning steps – glimpse G , BAN models multimodal feature interaction in the i -th reasoning step via:

$$\mathbf{f}_i = (\mathbf{Z}_v^T \mathbf{W}_v)_j^T \mathbf{M}_i (\mathbf{Z}_q^T \mathbf{W}_q)_j, \quad (6.6)$$

$$\mathbf{M}_i = \text{softmax}(((\mathbf{1} \cdot \mathbf{p}_i^T) \circ \mathbf{Z}_v^T \mathbf{W}_v) \mathbf{W}_q^T \mathbf{Z}_q), \quad (6.7)$$

where $\mathbf{f}_i \in \mathbb{R}^J$ with $J \leq \min(d_v, d_q)$, $i \in \{1, \dots, G\}$ is the index of reasoning step, $j \in \{1, \dots, J\}$ is the index of matrix column, $\mathbf{W}_v \in \mathbb{R}^{d_v \times J}$ and $\mathbf{W}_q \in \mathbb{R}^{d_q \times J}$ are trainable weights, $\mathbf{1} \in \mathbb{R}^N$ is an all-one vector, $\mathbf{p}_i \in \mathbb{R}^J$ is a learnable vector, and \circ is the

element-wise product. We follow MEVF to set the number of channels of visual features to $N = 1$.

After G reasoning steps, the final fused features \mathbf{f} are obtained by:

$$\mathbf{f} = \text{SumPool}\left(\sum_{i=1}^G ((\mathbf{W}_i \mathbf{f}_i) \cdot \mathbf{1}^T + \mathbf{f}_{i-1})\right), \quad (6.8)$$

where $\mathbf{f} \in \mathbb{R}^{d_q}$, $\mathbf{W}_i \in \mathbb{R}^{d_q \times J}$, $\mathbf{1} \in \mathbb{R}^L$, $\mathbf{f}_0 = \mathbf{Z}_q$, and SumPool is the sum operation over the length dimension L . We discuss the impact of the hyper-parameter G in Section 6.3.4.2.

6.2.3.2 Question-Conditioned Reasoning (QCR)

Recent studies [11, 13] have shown that BAN has limited reasoning ability for Med-VQA, especially for open-ended questions. This is because it can not fully capture the interaction between visual and textual features. For example, BAN merely utilizes bilinear matrix multiplication to fuse multimodal features. To equip the Med-VQA model with more powerful reasoning ability, we improve the standard reasoning module by incorporating an additional question-conditioned modulation function. Our motivations are two-fold. First, similar to human reasoning processes, solving different tasks requires corresponding task-specific reasoning skills. Second, the question itself contains rich task information which could be helpful [188].

Hence, the QCR function is designed to extract task information from the question and use it to guide the modulation over multimodal features. In this process, high-level reasoning skills are learned by imposing importance selection over the fusion features.

The details of QCR are illustrated on the right side of Figure 6.3 within the orange dashed

rectangle. First, a question string q , with L words in it, is converted into a sequence of word embeddings pre-trained by Glove [182]. Let $\mathbf{w}_i \in \mathbb{R}^{d_w}$ denote the corresponding word vector for the i -th word:

$$\mathbf{Q}_{emb} = \text{WordEmbedding}(q) = [\mathbf{w}_1, \dots, \mathbf{w}_L]. \quad (6.9)$$

The word embedding sequence $\mathbf{Q}_{emb} \in \mathbb{R}^{d_w \times L}$ is further processed by a d_g -dimensional Gated Recurrent Unit (GRU) to obtain the question embedding:

$$\mathbf{Q}_{feat} = \text{GRU}([\mathbf{w}_1, \dots, \mathbf{w}_L]) = [\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_L], \quad (6.10)$$

where $\mathbf{Q}_{feat} \in \mathbb{R}^{d_g \times L}$, and $\boldsymbol{\eta}_i$ denotes the embedding at the i -th position.

Since the question embedding \mathbf{Q}_{feat} is generated by the GRU network word-by-word sequentially, it may put more emphasis on later words. To further highlight the important words, e.g., “where do nodes locate in the lung?”, we design an attention mechanism to re-calculate attention weights on different words:

$$\tilde{\mathbf{Q}} = \mathbf{Q}_{emb} \otimes \mathbf{Q}_{feat}, \quad (6.11a)$$

$$\mathbf{Y} = \tanh(\mathbf{W}_1 \tilde{\mathbf{Q}}), \quad (6.11b)$$

$$\tilde{\mathbf{Y}} = \sigma(\mathbf{W}_2 \tilde{\mathbf{Q}}), \quad (6.11c)$$

$$\mathbf{g} = \mathbf{Y} \circ \tilde{\mathbf{Y}}. \quad (6.11d)$$

Here, \otimes denotes feature concatenation in the feature dimension, $\tilde{\mathbf{Q}} \in \mathbb{R}^{(d_w+d_g) \times L}$, $\mathbf{W}_1, \mathbf{W}_2 \in \mathbb{R}^{d_g \times (d_w+d_g)}$ are trainable weights, σ and \tanh are the sigmoid activation function and tanh activation function respectively, and \circ is the Hadamard product. $\tilde{\mathbf{Q}}$ can be formed by both context-free embeddings (e.g., Glove) and contextual embeddings (e.g., GRU), which has been demonstrated effective in many NLP tasks [189, 190]. σ and

\tanh (Equations 6.11b–6.11c) make up the gated hyperbolic tangent activation [56, 191], which is a special case of highway networks [192] that outperforms traditional ReLU or tanh layers in many scenarios. \tilde{Y} acts as a gate on the intermediate activation Y to control the output $\mathcal{G} \in \mathbb{R}^{d_g \times L}$ [192].

Then, the attention vector $\alpha \in \mathbb{R}^L$ for the question embedding \mathbf{Q}_{feat} can be obtained by

$$\alpha = \text{softmax}((\mathbf{W}_a \mathcal{G})^T), \quad (6.12)$$

where $\mathbf{W}_a \in \mathbb{R}^{1 \times d_g}$ are trainable weights.

Finally, with the attention vector α , we obtain the final output of QCR as:

$$\mathbf{q}_{att} = \mathbf{Q}_{feat} \alpha, \quad (6.13)$$

$$QCR(q) = MLP_Q(\mathbf{q}_{att}), \quad (6.14)$$

where \mathbf{q}_{att} is the aggregated question representation, and MLP_Q is a multilayer perceptron network that provides additional non-linear transformation for importance selection.

In this work, we propose to impose the proposed QCR module on the multimodal feature fusion module A_{δ_m} by an element-wise multiplication between their outputs: $QCR(q)$ and $A_{\delta_m}(\mathbf{Z}_v, \mathbf{Z}_q)$. The final representations are then fed to the classifier D_{δ_c} , and the prediction scores are given by

$$\mathbf{s} = D_{\delta_c}(A_{\delta_m}(\mathbf{Z}_v, \mathbf{Z}_q) \circ QCR(q)), \quad (6.15)$$

where \circ denotes element-wise product.

6.2.3.3 Type-Conditioned Reasoning (TCR)

It has been observed that closed-ended questions are generally easier than open-ended questions. For example, the closed-ended question “Is this an MRI image?” can be correctly answered with a simple image understanding process, but the open-ended question “What is the abnormality of the patient’s right brain in this radiology image?” needs multi-step reasoning, since the model must locate the abnormality in the right brain first and then diagnose the type of abnormality, e.g., brain tumor. Thus, Med-VQA systems need to be empowered with multi-level reasoning abilities, which are lacking in present VQA models [183].

To this end, we propose to use a separate reasoning component for closed-ended questions and open-ended questions respectively, in which the proposed QCR module is applied on top of the multimodal feature fusion module, as shown in the two black dash rectangles on the left of Figure 6.3. Particularly, we want to train a task type classifier C_{type} that takes a question as input and outputs the question type, i.e., closed-ended or open-ended. We observe that different types of questions put emphasis on different words. For example, closed-ended questions usually start with “Do\Are\Is\etc.”, and open-ended questions often start with “What\How many\Where\etc”. The differences between the two types of questions can be captured by question embeddings, which makes it possible to train a reliable and light-weight classifier that divides Med-VQA tasks into two subbranches, as shown by the rhombus module in Figure 6.3.

Similar to Section 6.2.3.2, we use Equations (6.9) - (6.13) to compute the question embedding and denote the mapping as Φ . We then employ a multilayer perceptron MLP_T to map question embedding into classification scores. The binary classification

probabilities are computed by $\mathbf{p}^t = \text{softmax}(MLP_T(\Phi(q)))$, and p_0^t and p_1^t are the probabilities for closed-ended and open-ended respectively. The binary question type classifier C_{type} is then formulated as:

$$C_{type}(q) = \begin{cases} 0, & \text{if } p_0^t > p_1^t, \\ 1, & \text{else.} \end{cases} \quad (6.16)$$

Hence, the predicted scores \mathbf{s} of candidate answers for a task (v, q) can be obtained by

$$\mathbf{s} = \begin{cases} D_{\delta_c}(A_{\delta_m}^{cl}(\mathbf{Z}_v, \mathbf{Z}_q) \circ QCR^{cl}(q)), & \text{if } C_{type}(q) = 0, \\ D_{\delta_c}(A_{\delta_m}^{op}(\mathbf{Z}_v, \mathbf{Z}_q) \circ QCR^{op}(q)), & \text{if } C_{type}(q) = 1, \end{cases} \quad (6.17)$$

where *cl* and *op* stand for closed-ended and open-ended respectively. For the basic multimodal reasoning module, we use a different number of reasoning steps for the open-ended branch $A_{\delta_m}^{op}$ and the closed-ended branch $A_{\delta_m}^{cl}$ and conduct an ablation study in Section 6.3.4.2.

6.2.4 Proposed Med-VQA Model

The pre-trained visual feature extractor and the conditional reasoning mechanism can be naturally combined to train an end-to-end Med-VQA model. As depicted in Figure 6.3, our proposed Med-VQA model works as follows. First, the image features \mathbf{Z}_v are obtained by the visual feature extractor $P_\xi \circ F_\theta$ (with F_θ pre-trained and fixed), as indicated by the green rectangle. The question features \mathbf{Z}_q are obtained by the textual feature extractor Q_ψ , as indicated by the red rectangle. Second, the TCR module classifies the question as open-ended or closed-ended and chooses the corresponding reasoning module (black

dash rectangle). Note that the visual and textual feature extractors are shared for both branches. Third, the chosen reasoning module will generate fused features, which is the element-wise multiplication between the output of the basic reasoning module (e.g., BAN or SAN) (yellow block) and the modulation vector produced by the QCR module (orange block). Finally, an MLP classifier gives the prediction score s for candidate answers.

6.3 Experiments

In this section, we conduct comprehensive experiments to evaluate the performance of our proposed framework on the only two available manually-annotated Med-VQA datasets, VQA-RAD [11] and SLAKE [2]. We compare our approach with current state-of-the-art baselines, evaluate the effectiveness of each component of our framework by ablation studies, and present qualitative results by visualizing the attention maps of both the images and questions of some Med-VQA tasks.

Dataset	Images	Answers	Questions		
			Overall	Open	Closed
VQA-RAD [11]	315	458	3,515	1,420	2,095
SLAKE-EN [2]	642	219	7,033	4,252	2,781

Table 6.1: Med-VQA Dataset statistics

6.3.1 Datasets

VQA-RAD [11] and SLAKE [2] are the only two available manually-annotated radiology-based datasets for Med-VQA. The statistics of the two datasets are summarized in Table 6.1.

VQA-RAD [11] contains 315 radiology images (e.g., CT, MRI, and X-Ray) and 3,515 clinical question-answer pairs (tasks), with 3,064 tasks for training and 451 tasks for testing. The number of candidate answers is 458. There may be multiple questions associated with one image. For example, the clinicians may ask different types of questions regarding a radiology image such as “imaging modality”, “abnormality”, or “organ system”.

SLAKE [2] is our bi-lingual Med-VQA dataset. It includes more question types such as “organ shape” and “common fact”. In our experiments, we use the English version of SLAKE, referred to as SLAKE-EN, which contains 642 radiology images, 7,033 question-answer pairs, and 219 candidate answers. We follow the original dataset splitting, where 4,919 tasks about 450 images are used for training, 1,053 tasks about 96 images for validation, and 1,061 tasks about 96 images for testing.

6.3.2 Implementation Details

We conduct all experiments on a Ubuntu 16.04 server with 8 Titan XP GPUs using PyTorch. The implementation details for stages I & II of our method (Figure 6.2) are provided below.

Stage I. We collect 22,995 unlabeled radiology images from an online open-access

resource* to form $\mathcal{D}_{pre-train}$, which contains 7,592 brain MRI and CT images, 7,592 abdomen CT images, and 7,811 chest X-Ray images. We use ResNet-50 as the backbone for F_θ and $F_{\theta'}$, and use a non-linear layer with ReLU activation for P_ϕ and $P_{\phi'}$ to project the representations into a 128-dimensional space. Then we train them with the loss $\mathcal{L}_{contras}$ (Equation (6.3)) for 800 epochs, which takes approximately 23 hours. In each epoch, the mini-batch size is 128, and the model is trained in parallel over 4 GPUs. The length K of the queue q is 4,096, the temperature parameter τ is 0.2, and the momentum coefficient m is 0.999. We utilize SGD optimizer with an initial learning rate of $1.5e^{-2}$ decayed by cosine schedule. After training, we save the weights of ResNet-50 in the last epoch for training in stage II.

Stage II. We use F_θ (ResNet-50) pre-trained in stage I combined with P_ξ (a non-linear layer with ReLU activation) to extract visual features. For a fair comparison, we follow MEVF [13] to set the dimension of visual features to 128, use Glove [182] to initialize word embeddings, and employ a 1024-dimensional LSTM to extract textual features. Moreover, the hidden size of all GRUs in the QCR and TCR modules (Figure 6.3) is 1024. The MLP_Q in Equation (6.14) and MLP_T in Equation (6.16) are instantiated with hidden units 1024 and 64 respectively. For each dataset, we pre-train a task type classifier C_{type} (with about 2.4M parameters) for 150 epochs by using the “answer_type” label in the training set and Adam optimizer [193] with learning rate $1e^{-4}$, and freeze the pre-trained weights during both the training and inference stages. The trained classifiers reach 99.33% and 99.81% classification accuracy on the test set of VQA-RAD and SLAKE respectively. For the training of the Med-VQA model, we use Adamax optimizer with initial learning rate $2e^{-3}$ for 100 epochs. Notice that different from general VQA

*<http://medicaldecathlon.com/>

that formulates open-ended question answering as a multi-label classification task (e.g., open-ended questions in the VQA v2.0 dataset normally have more than one correct answer [194]) or a text generative task [195], in Med-VQA each question has only one correct answer regardless of question type. Hence, we follow previous works [11, 13] to formulate Med-VQA as a single-label classification task and use accuracy as evaluation metric for both open-ended and closed-ended questions.

6.3.3 Comparison with the State-of-the-arts

We compare our method with existing Med-VQA models including general VQA frameworks, vision-and-language Transformers, vision-language contrastive pre-training, and the recently proposed MEVF method [13].

- **General VQA frameworks.** We follow [11] to compare with stacked attention network (SAN) [22] and multimodal compact bilinear pooling (MCB) [54]. In addition, we also compare with other frameworks including bilinear attention network (BAN) [21], multi-modal factorized bilinear pooling with co-attention (MFBCoAtt) [181], multimodal factorized high-order pooling (MFH) [196], and multi-modal tucker fusion (MUTAN) [197]. These VQA frameworks are usually named after their respective reasoning modules.
- **Vision-and-language Transformers.** Transformer-based vision-and-language pre-training (VLP) has achieved impressive results in vision-language tasks. In this section, we compare our method with both general and medical VLP models. Due to the lack of regional object labels in existing Med-VQA datasets, object-based VLP methods cannot be applied. Hence, we compare with a convolution-based

method – Pixel-BERT [106] and a patch-based method – general vision-and-language Transformer (ViLT) [107]. For medical VLP models, we compare with multimodal medical BERT (MMBERT) [59] and medical vision language learner (MedViLL) [60].

- **Vision-Language Contrastive Pre-training** is a self-supervised approach that pre-trains a model by pulling the paired image-text instances closer in the embedding space. The representative method in the medical domain is GLoRIA [126], which uses ResNet-50 and BioClinicalBERT [198] as the visual encoder and text encoder respectively. Since there is no feature fusion module in this paradigm, we use BAN [21] and SAN [22] instead.
- **MEVF** [13] is a recently proposed lightweight model for Med-VQA, which pre-trains a visual module on medical datasets and combines it with different attention reasoning modules such as BAN [21] and SAN [22].

Table 6.2 shows the results of our methods and the baselines. For all general VQA frameworks, ResNet-50 pre-trained on ImageNet and 1024- D LSTM network are used as visual extractor and textual extractor, respectively. Note that these VQA frameworks are usually named after their respective reasoning modules. We re-implement Pixel-BERT [106] (the ResNet-50 version) since both the source code and pre-trained weights are not provided. We use the original implementation of ViLT [106] and fine-tune the model on Med-VQA datasets with or without the pre-trained weights. We re-implement MMBERT, MedViLL, and MEVF+BAN/SAN using the code released by the authors. We cannot reproduce the results of MMBERT as reported in the original paper using the

Models	#Parameters (M)	Accuracy (%) on VQA-RAD [11]			Accuracy (%) on SLAKE-EN [2]		
		Overall	Open-ended	Closed-ended	Overall	Open-ended	Closed-ended
► <i>General VQA Frameworks:</i>							
MFBCoAtt Fw. [181]	58.20	50.6	14.5	74.3	73.3	72.2	75.0
SAN Fw. [11, 22]	36.54	54.3	31.3	69.5	76.0	74.0	79.1
MFH Fw. [196]	72.11	57.9	35.2	72.8	75.9	73.6	79.3
MCB Fw. [11, 54]	36.29	58.1	38.0	71.3	76.1	73.2	80.5
MUTAN Fw. [197]	58.46	58.1	34.1	73.9	76.8	73.6	81.7
BAN Fw. [13, 21]	42.19	58.3	37.4	72.1	76.3	74.6	79.1
► <i>Vision-and-Language Transformers:</i>							
Pixel-BERT-R50 [106]	137.37	61.7	48.2	70.5	77.4	77.1	77.9
ViLT-B/32 (w/o pre-training) [107]	113.12	59.6	38.5	73.5	76.0	75.8	76.2
ViLT-B/32 (w/ pre-training) [107]	113.12	66.5	52.0	76.1	78.1	76.9	80.0
MMBERT* [59]	111.53	72.0	63.1	77.9	-	-	-
MMBERT [59]	111.53	68.5	57.5	75.7	79.0	76.1	83.4
MedViLL* [60]	129.78	70.3	59.5	77.7	-	-	-
MedViLL [60]	129.78	69.6	58.7	76.8	78.4	76.3	81.7
► <i>Vision-Language Contrastive Pre-training:</i>							
GLoRIA+SAN [126]	135.01	67.4	56.4	74.6	76.8	75.2	79.3
GLoRIA+BAN [126]	139.15	69.2	57.5	76.8	79.4	78.1	81.3
► <i>Med-VQA Models:</i>							
MEVF+SAN* [13]	13.99	60.8	40.7	74.1	-	-	-
MEVF+SAN [13]	13.99	64.1	49.2	73.9	76.5	75.3	78.4
MEVF+BAN* [13]	19.64	62.6	43.9	75.1	-	-	-
MEVF+BAN [13]	19.64	66.1	49.2	77.2	78.6	77.8	79.8
CP+BAN (ours)	18.44	68.1	53.1	77.9	80.9	79.1	83.7
MEVF+BAN+CR (ours)	27.21	71.6	60.0	79.3	80.0	78.8	82.0
CP+BAN+CR (ours)	26.01	72.5	60.5	80.4	81.9	80.5	84.1

Table 6.2: **Test accuracy** of methods on VQA-RAD [11] and SLAKE [2]. “Fw.” is the abbreviation of “framework”. * means results cited from the original papers.

configurations provided by the authors[†] (probably some key information is missing). Our re-implementation of MEVF+BAN/SAN achieves much better results than the original paper due to longer training epochs (100 epochs for all methods). For MEVF+BAN/SAN, the visual extractor is MEVF, and the reasoning module is BAN/SAN. To compare our proposed visual extractor CP (Section 6.2.2) with MEVF, we combine it with BAN (denoted as CP+BAN). For a fair comparison, we strictly follow MEVF+BAN to use a 1024-*D* LSTM network to extract textual features with word embeddings pre-trained by GloVe [182], set the dimension of visual features to 128, and set the number of reasoning

[†]<https://github.com/VirajBagal/MMBERT/issues/4>

steps of BAN to 2. When applying the proposed CR mechanism to boost MEVF+BAN and CP+BAN, we set the number of reasoning steps of BAN in open-ended and closed-ended branches to 2 and 1, respectively.

According to the performance of each model, we can make the following observations:

- Our method (CP+BAN+CR) performs much better than general VQA frameworks, vision-and-language Transformers, and vision-language contrastive pre-training, with much fewer parameters, showing the benefit of utilizing small models for solving Med-VQA tasks. Also, it can be seen that large models tend to overfit the small-scale Med-VQA training data.
- Our pre-trained feature extractor (indicated by CP) is more effective than MEVF. CP+BAN achieves 2% and 2.3% absolute improvement over MEVF+BAN [13] in overall accuracy on VQA-RAD and SLAKE-EN respectively. Also, it can be noted that since MEVF is pre-trained on VQA-RAD, combining it with BAN (i.e., MEVF+BAN) only brings $\sim 2.3\%$ improvement over BAN on SLAKE-EN, much lower than the $\sim 4.6\%$ improvement brought by our CP model. Besides, CP+BAN+CR improves over MEVF+BAN+CR by $\sim 2\%$ in absolute overall accuracy on SLAKE-EN dataset, which is comparable to the improvement of CP+BAN over MEVF+BAN. While the improvement on VQA-RAD dataset is smaller, CP+BAN+CR still consistently achieves better performance than MEVF+BAN+CR for each question type. These results demonstrate the effectiveness and generality of our pre-trained model CP.
- Our conditional reasoning (CR) mechanism can further bring consistent and significant performance gains on top of different visual feature extractors, including

Visual Modules	#Parameters (M)	Accuracy (%)		
		Overall	Open	Closed
VGG-16 [†] [53]	134.83	56.8	35.2	71.0
ResNet-50 [†] [20]	23.77	58.3	37.4	72.1
ResNet-50 [‡] [14]	23.77	59.9	34.6	76.5
MEVF [13]	1.22	66.1	49.2	77.2
CP (F_θ)	23.77	62.3	38.5	77.9
CP ($P_\xi \circ F_\theta$)	23.79	61.2	38.5	76.1
CP ($P_\xi \circ F_\theta$, w/ F_θ frozen)	0.02	68.1	53.1	77.9

Table 6.3: **Comparison of different visual modules** on VQA-RAD [11]. [†] indicates pre-training on ImageNet with standard supervised classification. [‡] indicates pre-training on ImageNet with contrastive self-supervised learning (MoCo-v2 [14]).

MEVF and our CP model. It can be seen that CP+BAN+CR significantly outperforms CP+BAN while MEVF+BAN+CR significantly outperforms MEVF+BAN, on both datasets. The best performance is achieved by our CP+BAN+CR method. Remarkably, for open-ended Med-VQA tasks, incorporating CR leads to very large performance gains on VQA-RAD.

6.3.4 Ablation Study and Analysis

In this subsection, we conduct experiments to analyze the effectiveness of our proposed contrastive pre-training (CP) and conditional reasoning (CR) modules. We report results in test accuracy on VQA-RAD [11].

6.3.4.1 Ablation study on contrastive pre-training

I. Comparison of different visual feature extractors. Table 6.3 shows the comparison

of different visual feature extractors for Med-VQA. For all the methods, we use BAN with 2 reasoning steps for multimodal feature fusion and a 1024- D LSTM network as the textual extractor, and set the dimension of extracted visual features to 128. For P_ξ , we keep the input dimension same as the output dimension. It can be seen that our CP module (last row) outperforms MEVF and surpasses general visual backbones VGG-16 and ResNet-50 by large margins (8 ~ 11%), demonstrating its effectiveness.

In addition, the following observations can be made. First, compared with ResNet-50[‡] pre-trained on ImageNet also using MoCo-v2[‡] (row 3), it can be seen that ResNet-50 pre-trained on medical images, i.e., our pre-trained F_θ (row 5) performs better, showing the benefit of domain-specific pre-training. Second, appending a non-linear layer P_ξ to F_θ (row 6) leads to worse performance than F_θ , indicating more overfitting. Finally, freezing the parameters of F_θ and only fine-tuning P_ξ (row 7) leads to significantly better performance. This effective and efficient design helps to avoid overfitting caused by fine-tuning large models on small-scale Med-VQA datasets and greatly reduces the training parameters.

II. Comparison of different pre-training datasets. We observe that existing Med-VQA datasets contain radiology images of different body regions and imaging modalities, e.g., brain MRI, chest X-Ray, and abdomen CT, as shown in Figure 6.7. Therefore, we collect a dataset $\mathcal{D}_{\text{pre-train}}$ (Section 6.2.2) with a similar composition for contrastive pre-training. Specifically, it contains brain CT and MRI images, chest X-ray images, and abdomen CT images. Since the brain images do not have modality labels, we conduct an ablation study of $\mathcal{D}_{\text{pre-train}}$ w.r.t. different body parts and compare it with other pre-training datasets.

[‡]The pre-trained weights can be downloaded from <https://github.com/facebookresearch/moco>. We choose the 800-epoch version, which uses the same number of epochs as in our pre-training.

Dataset	Accuracy (%)
	Overall
Random initialization	60.3
ImageNet (22,995)	65.2
Brain (22,995)	61.4
Chest (22,995)	64.3
Abdomen (22,995)	63.0
Brain (11,500), Chest (11,495)	66.3
Chest (11,500), Abdomen (11,495)	67.2
Abdomen (11,500), Brain (11,495)	64.5
Brain (7,592), Chest (7,811), Abdomen (7,592) (ours)	68.1

Table 6.4: Comparison of performance on VQA-RAD [11] by **using different datasets for contrastive pre-training** (CP in Section 6.2.2).

For a fair comparison, we fix the size of all datasets as 22,995, and use the same visual module $P_\xi \circ F_\theta$ (with F_θ frozen), textual model 1024- D LSTM, and reasoning module BAN (with 2 reasoning steps).

The results in Table 6.4 show the importance of pre-training with images of different body parts. Specifically, for the datasets containing images of only a single body part (rows 3 ~ 5), the pre-trained models perform worse than the model pre-trained on ImageNet (row 2), though they are better than random initialization (row 1). By increasing the diversity of the pre-training datasets (rows 6 ~ 8), the performance of the pre-trained models is significantly improved. The best performance is achieved with $\mathcal{D}_{\text{pre-train}}$ which contains images of three different body parts (row 9).

Base Model	QCR	TCR	#Parameters (M)	Accuracy (%)		
				Overall	Open-ended	Closed-ended
			19.64	66.1	49.2	77.2
MEVF+BAN	✓		23.98	67.8	51.4	78.7
		✓	22.87	70.1	56.7	79.0
	✓	✓	27.21	71.6	60.0	79.3
			18.44	68.1	53.1	77.9
CP+BAN	✓		22.78	69.6	56.4	78.5
		✓	21.67	71.4	58.9	79.7
	✓	✓	26.01	72.5	60.5	80.4

Table 6.5: **Ablation study of our proposed conditional reasoning (CR) mechanism** on VQA-RAD [11].

6.3.4.2 Ablation study on conditional reasoning

I. Effect of QCR and TCR. We first evaluate the effectiveness of our proposed conditional reasoning modules QCR and TCR with two base models, MEVF+BAN and our CP+BAN. For both of them, the number of reasoning steps of BAN is set to 2. When incorporating TCR into the base models, we set the number of reasoning steps of BAN as 2 for open-ended questions and 1 for closed-ended questions. When only QCR is included (e.g., MEVF+BAN+QCR), the model does not differentiate the question type, and hence there is only one reasoning module. Our QCR module is then imposed on the basic reasoning module BAN to enhance reasoning ability. When only TCR is included (e.g., MEVF+BAN+TCR), the model differentiates the question type and chooses the corresponding reasoning module. However, only the basic reasoning module BAN is utilized for reasoning, without the QCR enhancement.

As shown in Table 6.5, QCR improves the overall accuracy of MEVF+BAN and CP+BAN by 1.7% and 1.5% respectively, while TCR improves them by 4% and 3.3% respectively. The results show that both QCR and TCR are useful, indicating the importance of utilizing question information to learn task-adaptive reasoning skills for different Med-VQA tasks. Further, the large performance improvement brought by TCR demonstrates the necessity of learning multi-level reasoning skills for different types of Med-VQA tasks.

When both QCR and TCR are incorporated, we observe further improvements. The overall accuracy of MEVF+BAN and CP+BAN is increased by 5.5% and 4.4% respectively, where the accuracy for open-ended questions is increased by 10.8% and 7.4% respectively and the accuracy for closed-ended questions is increased by 2.1% and 2.5% respectively. The large improvements on open-ended tasks show the effectiveness of our conditional reasoning mechanism in learning higher-level reasoning skills to solve difficult tasks.

II. The reasoning ability required for solving open-ended and closed-ended questions.

We further study the reasoning ability required for solving open-ended and closed-ended questions, including the number of reasoning step G (Equation 6.8) and the proposed QCR module. We use $BAN(G)$ to denote BAN with G reasoning steps and $BAN(G_{open}-G_{closed})$ to denote that in our TCR we use BAN with G_{open} reasoning steps for open-ended questions and BAN with G_{closed} reasoning steps for closed-ended questions.

From the results in Table 6.6, we can make the following observations. (1) As the number of reasoning steps (G , G_{open} , or G_{close}) increases, the corresponding model performance first increases and then decreases, indicating a gradual transition from

Base Model	G	#Parameters (M)	Accuracy (%)			
			Overall	Open	Closed	
MEVF+BAN	1	17.40	63.6	44.7	76.1	
	2	19.64	66.1	49.2	77.2	
	3	21.87	65.9	52.5	74.6	
	4	24.11	66.1	53.1	74.6	
	5	26.34	65.6	52.5	74.2	
	6	28.57	64.5	52.0	72.8	
CP+BAN	1	16.20	64.5	48.6	75.4	
	2	18.44	68.1	53.1	77.9	
	3	20.67	67.4	53.6	76.5	
	4	22.90	66.3	52.5	75.4	
	5	25.14	66.5	53.1	75.4	
	6	27.37	65.2	52.0	73.9	
► Incorporating TCR:						
	G_{open}	G_{closed}				
CP+BAN	1	1	19.43	68.7	53.6	78.7
	2	1	21.67	71.4	58.9	79.7
	1	2	21.67	70.3	55.3	80.1
	2	2	23.90	71.2	58.7	79.4
	3	2	26.14	69.8	57.5	77.9
	2	3	26.14	70.5	56.4	79.8
	3	3	28.38	69.2	55.9	77.9
► Further Incorporating QCR:						
	G_{open}	G_{closed}				
CP+BAN	1	1	23.77	70.3	58.1	78.3
	2	1	26.01	72.5	60.5	80.4
	1	2	26.01	71.4	59.2	79.4
	2	2	28.25	71.6	61.4	78.3
	3	2	30.49	72.1	60.3	79.8
	2	3	30.49	71.1	59.2	79.0
	3	3	32.73	69.6	58.1	77.2

Table 6.6: **Effect of the number of reasoning steps for open-ended and closed-ended questions** in VQA-RAD [11].

underfitting to overfitting. (2) By incorporating TCR, even BAN(1-1) outperforms BAN with any G reasoning steps, showing the benefit of using a separate reasoning module for different types of questions. As shown in Figure 6.4, the TCR module disentangles the representations of open-ended and close-ended tasks. (3) The best performance is achieved by CP+BAN(2-1), indicating that solving open-ended questions requires stronger reasoning ability than closed-ended questions. (4) Incorporating QCR can further improve the reasoning ability on both open-ended and closed-ended questions.

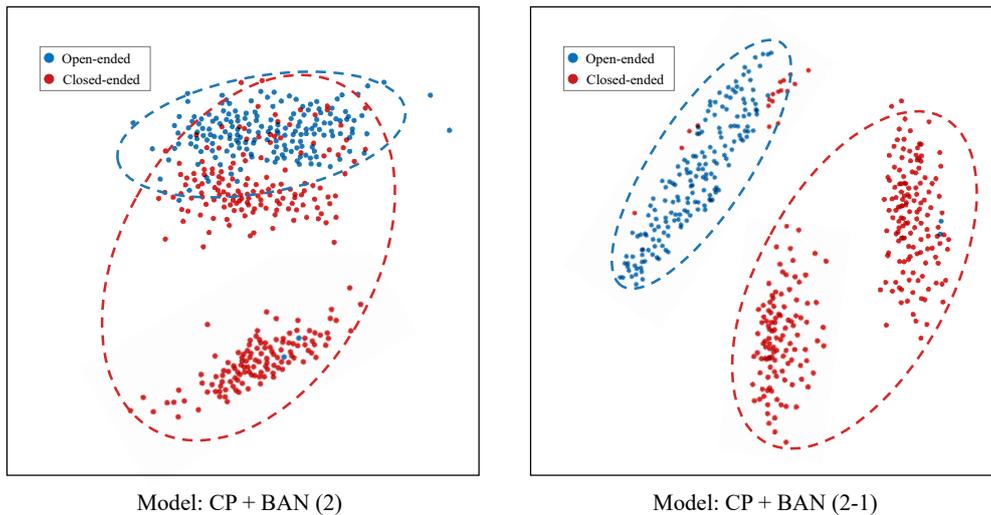


Figure 6.4: t-SNE [10] visualization of the multimodal features (input to the classifier layer) of Med-VQA tasks in the test set of VQA-RAD [11] learned by CP+BAN(2) and CP+BAN(2-1) respectively. The TCR module in CP+BAN(2-1) disentangles the representations of open-ended and closed-ended tasks.

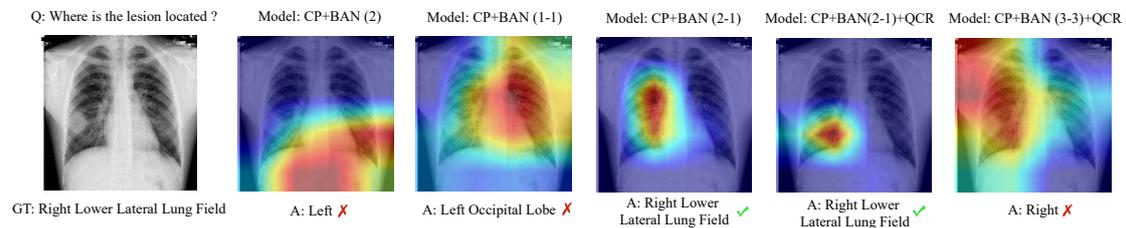


Figure 6.5: Visual comparison of the prediction results for an open-ended task in VQA-RAD dataset by variants of CP+BAN. The Grad-CAM maps [12] of the visual model are plotted based on the predicted answers. ✓ and ✗ indicate the correctness of the answer given by each model.

These observations can also be reflected in the visual comparison in Figure 6.5, where CP+BAN(2-1)+QCR can more accurately find the relevant regions and give the correct answer.

III. Impact of the prediction accuracy of the type classifier. We analyze how the

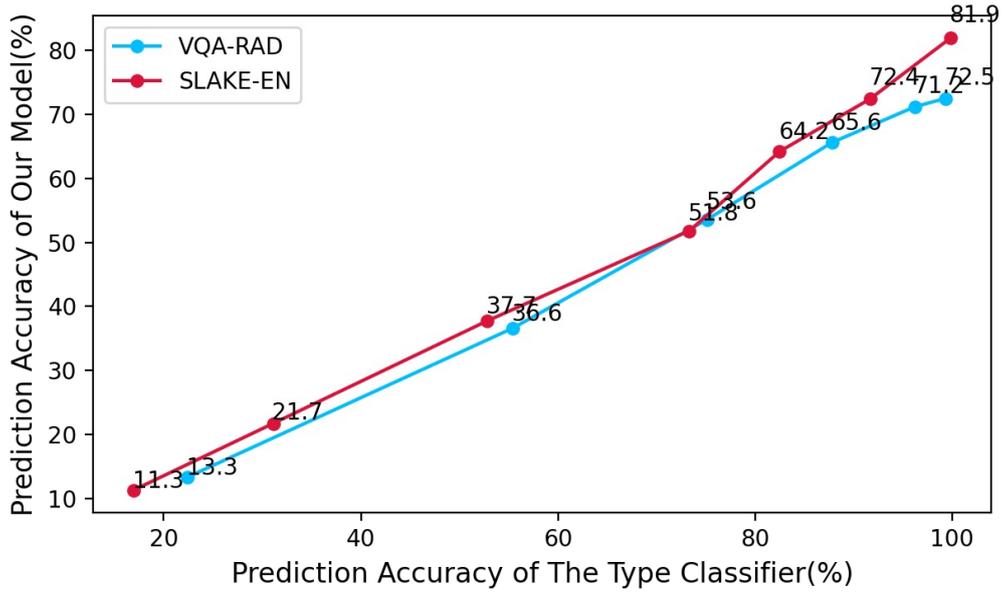


Figure 6.6: Impact of the prediction accuracy of the type classifier on our model CP+BAN+CR. Note that the prediction accuracy of our model refers to the overall metric.

prediction accuracy of the type classifier C_{type} in the TCR module affects the prediction accuracy of our model CP+BAN+CR on both VQA-RAD and SLAKE-EN datasets. As shown in Figure 6.6, the more accurate C_{type} is, the better performance our model can achieve. An inaccurate C_{type} will result in extremely poor performance. Fortunately, with our proposed algorithm (Equations (6.9) - (6.13) & Equation (6.16)), we can easily train a highly accurate C_{type} , achieving 99.33% and 99.81% classification accuracy on the test set of VQA-RAD and SLAKE respectively.

6.3.5 Qualitative Evaluation

We provide a qualitative comparison between our proposed methods and baseline MEVF+BAN.

Figure 6.7 shows the Grad-CAM [12] maps of the visual models based on the predicted

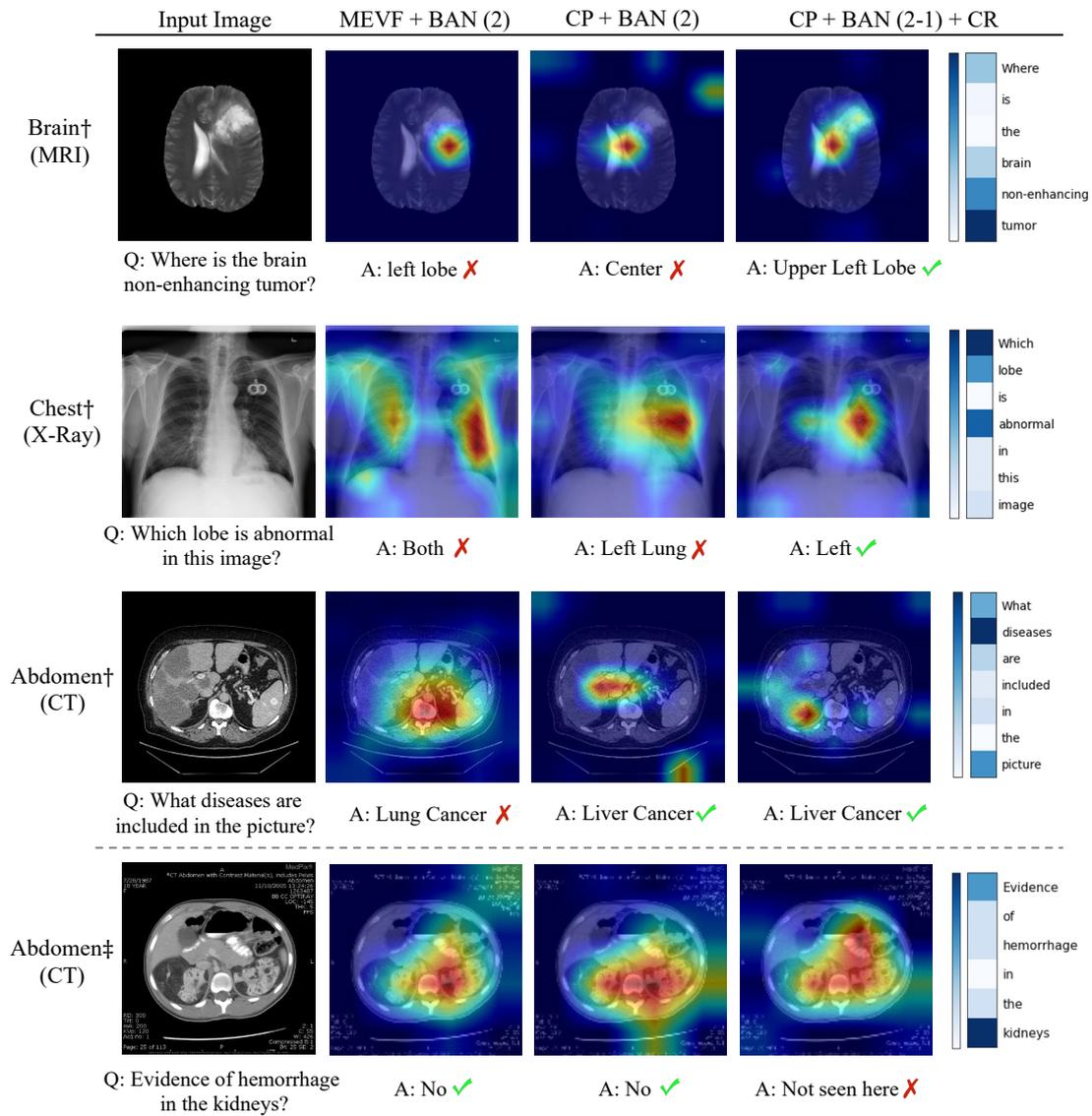


Figure 6.7: The Grad-CAM maps of the visual modules of our methods and baseline MEVF+BAN. The attention map of our QCR module is displayed in the right column, and darker color indicates higher weight. ✓ and ✗ indicate the correctness of the answer given by each model. † indicates the test image comes from SLAKE [2], and ‡ indicates it comes from VQA-RAD [11]. The last row shows a failure case of our method with the conditional reasoning module, which is caused by the misclassification of question type.

answers of four tasks in SLAKE-EN and VQA-RAD, which cover different human body parts and imaging modalities. We also provide a visualization of the attention weights (Equation 6.12) in QCR.

The first task is about a brain MRI image. While MEVF+BAN and our CP+BAN both give wrong answers, our CP+BAN+CR can find the relevant regions and predict the right answer, demonstrating the effectiveness of conditional reasoning. The second task is about a chest X-Ray image. Compared with MEVF+BAN, our CP+BAN can better find more relevant regions, though it also gives a wrong answer. With conditional reasoning, our CP+BAN+CR can find the right answer. The third task is about an abdomen CT image. Both our CP+BAN and CP+BAN+CR give the right answer, but MEVF+BAN still fails. The last task is also about an abdomen CT image. In this case, both MEVF+BAN and our CP+BAN give the right answer, but our CP+BAN+CR fails. This is because the question type classifier of the TCR module gives a wrong prediction, mistaking a close-ended question as an open-ended one. Notice that even though the prediction accuracy of the question type classifier is as high as 99.33%, it may still fail in some rare cases.

6.3.6 Efficiency Evaluation

We progressively evaluate the time efficiency of the proposed QCR and TCR modules on top of the base model MEVF+BAN [13]. The results are provided in Figure 6.8, where we report the average training time and test time of 10 epochs. Compared with the base model, our reasoning modules only increase the training time by a small factor, and the overhead in test time is negligible. It shows that the proposed conditional reasoning mechanism can be efficiently applied to existing Med-VQA systems.

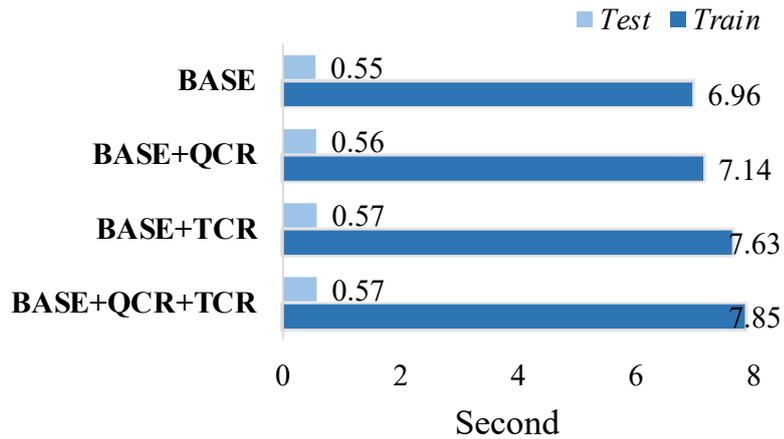


Figure 6.8: Time efficiency of the proposed conditional reasoning mechanism (i.e., TCR and QCR modules). BASE represents the base model MEVF+BAN [13]. BASE+QCR does not differentiate the question type, and hence there is only one reasoning module. Our QCR module is imposed on the basic reasoning module BAN to enhance reasoning ability. BASE+TCR differentiates the question type and chooses different reasoning module correspondingly. However, only the basic reasoning module BAN is used for reasoning, without our QCR enhancement. Deep Blue denotes training time (seconds) per epoch. Cyan denotes test time (seconds) per epoch.

Chapter Review

In this chapter, we have concerned with the design of two key modules of a Med-VQA system – the reasoning module and the visual feature extraction module. For the former, we have proposed an effective conditional reasoning mechanism that endows the system with task-specific reasoning ability, which is lightweight and can be applied to existing Med-VQA models in a plug-and-play manner. For the latter, we have proposed to pre-train a visual feature extractor via contrastive learning to tackle the data scarcity problem, which can be readily used by any Med-VQA model on a small-scale dataset. Empirical evaluation on existing benchmarks demonstrates the high effectiveness of our proposals compared with the state-of-the-arts. We hope this work will serve as a solid step to advance the research of Med-VQA.

Chapter 7

Introducing Multimodal Explainability

7.1 Introduction

Large vision language models (LVLMs) have recently made huge advancements in artificial intelligence [17, 84, 85, 199–203], demonstrating remarkable capabilities in understanding visual content while generating coherent natural language responses. These advancements have driven innovations across various domains [204, 205], with healthcare emerging as an important application. Within this domain, medical visual question answering (Med-VQA) stands out as a crucial task that automatically provides reliable and user-friendly answers [206] to questions about medical images [11], facilitating healthcare professionals in diagnosis, medical education, and clinical decision-making.

To ensure the reliability and user-friendliness of Med-VQA systems, it is crucial to incorporate answer explanations along with a diverse set of question formats. Although significant progress has been made by existing Med-VQA systems [2, 11, 89, 90, 92],

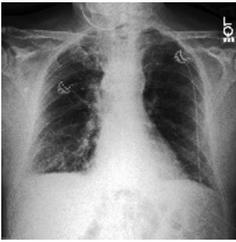
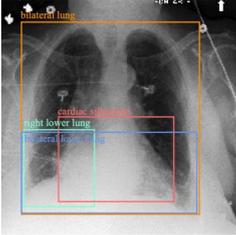
Dataset	VQA Type	Example
 MIMIC-CXR-VQA	Closed-ended	Closed-ended VQA: Q: Is any devices present within the right atrium? A: No.
	Open-ended	Open-ended VQA: Q: What are all the diseases identifiable within the right hilar structures? A: A small right pleural effusion.
 Ours	Closed-ended	Multi-choice VQA: Q: Which regions on the X-ray show signs of abnormalities? C: ["A: Bilateral lungs", "B: Right costophrenic angle", "C: Bilateral lower lung", "D: Cardiac region"] A: ["A", "B", "C", "D"] Reason: Abnormalities are seen in the bilateral lungs (hyperinflation), right costophrenic angle (blunting), bilateral lower lung (atelectasis), and cardiac silhouette (enlargement). Bounding Box: [[28, 40, 167, 190], [28, 124, 85, 184], [28, 126, 165, 189], [57, 114, 147, 180]]
	Open-ended	
	Single-choice	
	Multi-choice	

Figure 7.1: Our GEMeX stands out from existing medical VQA datasets by providing diverse question types and comprehensive multimodal explanations, including textual reasoning and visual grounding, to improve real-world applicability.

none have yet integrated answer explanations (*cf.* Table 2.1). As emphasized by [207], explanations are as essential as the answers themselves in general VQA systems. This holds even stronger in medical VQA, where the domain-specific nature of the task amplifies the need for clarity [11, 23]. If a Med-VQA system is to assist in clinical diagnosis, the datasets must be designed to include explanations so as to enhance patient comprehension and support the learning of junior medical practitioners. Additionally, the limited range of question formats, such as the absence of multiple-choice questions, restricts the real-world applicability and impairs the overall user-friendliness of medical AI systems.

To tackle the aforementioned limitations, we develop a large-scale, **Groundable**, and

Explainable Medical VQA benchmark for chest X-ray diagnosis (GEMeX). We first undertake a comprehensive data refinement process upon the Chest Imagenome [208]. By collaborating with radiologists, we systematically redefine anatomical regions and establish more precise vision-text correspondence mappings, resulting in accurate region-grounded reports for each X-ray image. Subsequently, we leverage GPT-4o [17] to generate a diverse set of questions based on these grounded reports, covering four categories of varying difficulty levels: open-ended, closed-ended, single-choice, and multiple-choice questions. Each question-answer pair is enriched with explicit reasoning and corresponding visual region annotations, as shown in Figure 7.1. The resulting dataset comprises 151,025 images and 1,605,575 questions. Currently, this is the largest VQA dataset for chest X-rays and the first medical VQA dataset that simultaneously includes both textual and visual explanations.

We evaluate 12 representative LVLMs, including 7 from the general domain (*e.g.*, LLaVA [85], DeepSeek-VL [209], GPT-4o-mini [17]), and 5 from the medical domain (*e.g.*, LLaVA-Med [80], XrayGPT [77], RadFM [78]). The experimental findings underscore the challenging characteristics of our dataset. Additionally, we propose a simple instruction-tuning strategy that derives a task-specific LVLm. The impressive performance improvement highlights the effectiveness of our dataset. Overall, we develop three metrics for measuring the accuracy of model outputs in terms of answers, reasoning, and visual grounding (localization generation). Notably, we apply both semantics-level score and gram-based metrics of natural language generation (*e.g.*, BLEU and ROUGE) for textual parts. Results indicate that for models without GEMeX fine-tuning, semantics-level scoring is more reliable. After fine-tuning, however, the natural language generation metrics can better reflect the model’s understanding of the dataset.

Overall, we make the following key contributions:

- We present GEMeX, a large-scale medical VQA dataset for chest X-rays, designed to support diverse question types and provide enhanced explainability for medical VQA systems. To our knowledge, it is the largest chest X-ray VQA dataset and the first Med-VQA dataset to embody the concept of multimodal explainability.
- We systematically benchmark 12 representative LVLMS using GEMeX, introducing multiple evaluation metrics to comprehensively demonstrate the performance of current popular LVLMS on the Med-VQA task.
- Our method shows that integrating precise vision-text explainability enhances the visual reasoning of LVLMS, addressing a key limitation in many models. We emphasize the need for a large-scale, groundable, and explainable VQA benchmark to advance LVLMS development and deployment in healthcare.

7.2 Construction of GEMeX

We will elaborate on the construction of the proposed *GEMeX* dataset, accompanied by a schematic overview in Figure 7.2. The structure of this section is organized as follows: In Section 7.2.1, we introduce the initial step of dataset construction, focusing on refining anatomical regions and grounding reports; Section 7.2.2 covers the generation process for four distinct types of Med-VQA, incorporating multimodal explainability across both visual and textual dimensions.

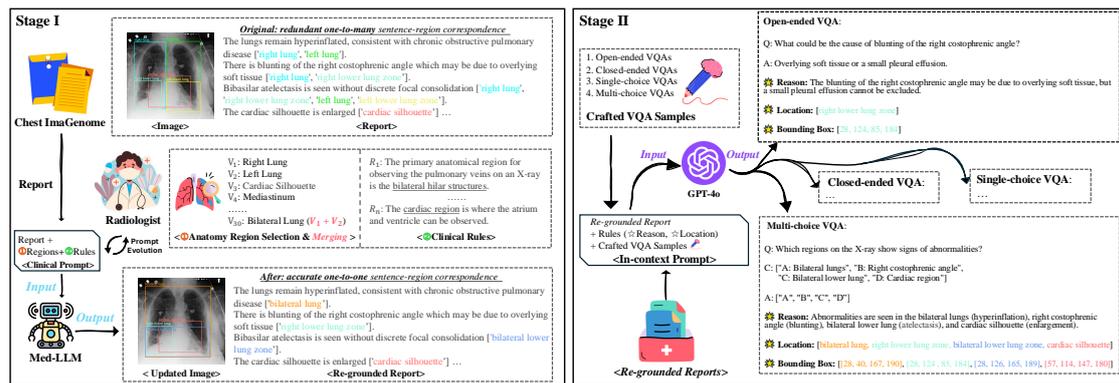


Figure 7.2: Illustration of the pipeline for constructing our GEMeX, with two main stages. In Stage I (left), medical LLM performs re-grounding on the original reports based on the pathological regions and clinical guidance specified by the radiologists, generating more precise sentence-region correspondence. In Stage II (right), the well-crafted prompt enables GPT-4o to generate a high-quality, large-scale Med-VQA dataset with both textual and visual explanations, leveraging the re-grounded reports.

7.2.1 Re-grounding Reports

As shown in stage I of Figure 7.2, we build upon Chest ImaGenome [208] to construct our dataset, but with an emphasis on the mapping precision between visual regions and textually described entities. After consulting radiologists, we find that the anatomical descriptions of regions of interest in ImaGenome are imprecise and redundant, introducing ambiguity into clinical diagnoses. Specifically, a single sentence can be associated with multiple anatomical regions, *e.g.*, “There is blunting of the right costophrenic angle which may be due to overlying soft tissue” corresponding to “right lung” and “right lower lung zone”. This redundancy poses challenges in training models to precisely visual grounding. Hence, we perform re-grounding to ensure each sentence is linked to a single, precise region.

	Chest ImaGenome	Ours
Reserve	right lung, right mid lung zone, right hilar structures, right hemidiaphragm, left lung, left mid lung zone, left hilar structures, left hemidiaphragm, trachea, spine, abdomen, svc	right lung, right mid lung zone, right hilar structures, right hemidiaphragm, left lung, left mid lung zone, left hilar structures, left hemidiaphragm, trachea, spine, abdomen, svc
	left upper lung zone left apical zone	left upper lung zone
Incorporate	right upper lung zone right apical zone	right upper lung zone
	mediastinum upper mediastinum	mediastinum
	right lower lung zone right costophrenic angle	right lower lung zone
	left lower lung zone left costophrenic angle	left lower lung zone
	cardiac silhouette cavoatrial junction right atrium	cardiac silhouette
	carina, right clavicle left clavicle, aortic arch	-
Merge	left lung + right lung	bilateral lung
	left upper + right upper	bilateral upper lung zone
	left mid + right mid	bilateral mid lung zone
	left lower + right lower	bilateral lower lung zone
	left hilar + right hilar	bilateral hilar structures
	left hemidiaphragm + right hemidiaphragm	bilateral hemidiaphragm
	left mid + left lower	left mid-to-lower lung zone
	right mid + right lower	right mid-to-lower lung zone
	left mid + left upper	left mid-to-upper lung zone
	right mid + right upper	right mid-to-upper lung zone
	left ... + right mid-to-lower	bilateral mid-to-lower lung
	left ... + right mid-to-upper	bilateral mid-to-upper lung

Table 7.1: Anatomical regions transformation from the Chest ImaGenome to our refined version with detailed operations.

7.2.1.1 Anatomical Region Selection and Merging

In the original Chest ImaGenome, there are 29 significant pathological regions (with bounding boxes). However, in alignment with radiologists’ practices, our dataset focuses on retaining core regions that are crucial for diagnosing diseases through X-rays, such as the “left lower lung” and “mediastinum”. Less significant or marginal areas are excluded to streamline the diagnostic training process and enhance clinical relevance, like “carina”

which is not considered a core region, and “clavicle” which accounts for only about 2% of the total regional frequency. Furthermore, to enhance clarity and ensure that each sentence corresponds to a single pathological region with finer granularity, semantically similar regions are merged. For instance, the “left lower lung zone” and “right lower lung zone” are combined into a “bilateral lower lung zone”. This aligns with conditions like “bibasilar atelectasis”, as illustrated in Stage I of Figure 7.2, where the condition is described as “Bibasilar atelectasis is seen without discrete focal consolidation”. In total, we define 30 anatomical regions eventually, as shown in Table 7.1.

7.2.1.2 Report Re-grounding Using Medical LLM.

Then, we utilize OpenBioLLM-70B*, known for its outstanding performance across various medical NLP tasks, to *re-ground reports by re-selecting a region for each sentence*. To test the effectiveness of the prompt, we begin by randomly selecting 100 pairs from the Chest ImaGenome test set, which includes approximately 367 sentences. Initially, the performance of the LLM is suboptimal due to: (1) inner knowledge about X-ray disease observation areas is not sufficiently precise, as OpenBioLLM is an NLP model that lacks clinical expertise, and (2) when a sentence indeed involves multiple regions that cannot be merged, the model may either output multiple regions or arbitrarily select one.

Iterative Prompt Refinement via Radiologist Feedback: To address these limitations, we employ an iterative approach, gradually incorporating clinical guidance from radiologists and manually-labeled such cases (in-context learning) to refine the prompt, facilitating (1) more accurate region selection; and (2) more effective sentence splitting and rewrit-

*<https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B>

```

messages = [{"role": "system", "content":
f"\"You are a helpful chest X-ray radiologist. Given an input sentence, your task is to map it to
an anatomical region on X-ray for better observation from a predefined list [right lung, cardiac
silhouette,..., bilateral lower lung zone].
Here are some rules:
(1) If there is no corresponding region for this sentence, leave it out.
(2) If the sentence describes the overall anatomical characteristics without specifying a particular
region, you can choose \"bilateral lung\" as its region. For example, \"No focal consolidation, pleural
effusion or pneumothorax is present\": \"bilateral lung\".
(3) One sentence can only correspond to one region. If a sentence's main symptom involves
several anatomical regions, rephrase it into multiple sentences with corresponding regions. Note
that all derived sentences must be syntactically complete, not phrases (i.e., containing a subject
and a predicate at least). For example: \"The cardiomediastinal silhouette is normal.\" can be
segmented into \"The cardiac silhouette is normal.\": \"cardiac region\", \"The mediastinal silhouette
is normal.\": \"mediastinum\", where \"cardiomediastinal\" corresponds to the \"cardiac region\" and
\"mediastinum\".
(4) Small (tiny) pleural effusion (fluid) usually appears in the lower lung zone, a moderate pleural
effusion appears in the mid-to-lower lung zone, and a large (substantial) pleural effusion can even
occupy the entire lung. If the severity (like small, moderate and large) is not indicated, output the
left lung or right lung directly.
(5) The main anatomical region for observing pulmonary venous is the bilateral hilar structures on
the X-ray.
(6) The region where the atrium and ventricle can be observed is the cardiac region.
Here are some cases: (1)... (2)... (3)... (4)...
Organize your output in a json formatted as Dict{Str(sentence):Str(region)}, without other
words.\""}]

messages += [{"role": "user", "content": "Input: \"Bibasilar atelectasis is seen without discrete focal
consolidation.\""}]

```

Table 7.2: Our proposed prompt for refining sentence-region pairs.

ing. For example, “The cardiomediastinal silhouette is normal.” is converted into {“The cardiac silhouette is normal.”:“cardiac silhouette”, “The mediastinal silhouette is normal.”:“mediastinum”}, where “cardiomediastinal” corresponds to the “cardiac silhouette” and “mediastinum”. This approach ensures the output clauses align one-to-one

with the respective regions. Ultimately, the final prompt is determined with an accuracy of approximately 98.4% on the aforementioned test set, provided in Table 7.2. Figure 7.2 presents an example of a re-grounded report and its corresponding updated image.

7.2.2 Groundable and Explainable VQA Generation

Although there are many Med-VQA datasets [94, 95] available, some even generated using MIMIC-CXR or Chest ImaGenome, they all have two weaknesses that diminish their practicality: (1) lacking strong explainability, especially the visual guidance, that hinders the user’s understanding; (2) a restricted range of question types, typically confined to open-ended or closed-ended formats with no inclusion of choice-based questions, reducing the flexibility and comprehensiveness. In general, these issues highlight the necessity for more versatile and explainable Med-VQA datasets to enhance their utility in clinical settings.

Data Generation with Quality Control. As shown in Stage II of Figure 7.2, we generate our VQA dataset based on re-grounded reports. Here, we employ GPT-4o (2024-08-06) [17] as a generator due to its remarkable capabilities in understanding and generating long texts. *We ensure the quality of the generated dataset by:* (1) to ensure a diverse range of question content [2], like “abnormality” and “location”, we identified 7 distinct categories through discussions with radiologists, as illustrated in Figure 7.4. Then, we manually craft questions covering these 7 types for 30 images, which serve as good demonstrations in the prompt to enhance the generation accuracy and better align with our objectives; (2) we also design specific rules (like not generating questions that need to be answered by comparing two images) to ensure the generated VQAs are

```

messages = [{"role": "system", "content":
f"“You are a chest X-ray AI assistant, and you are seeing a frontal view chest X-ray image, described
by several phrases with visual regions. Generate 3 open-ended questions, 2 closed-ended questions,
3 single-choice questions, and 3 multi-choice questions about this chest X-ray. Format your output
in JSON format.
Here are some rules:
(1) Include questions asking about the visual content of the image, containing abnormality, disease,
location, severity, cause of disease, etc. For a CXR, the types of questions generated need to be
diverse. Do not ask any questions that cannot be answered confidently.
(2) For each question, provide the answer, explain the reason for obtaining such answer, and output
the corresponding visual regions as a visual clue.
(3) For open-ended questions, the answers must be concise. You should generate detailed reasons
based on the provided CXR phrases and your medical knowledge. Do not refer to the text description
in your questions or answers.
(4) Avoid questions that cannot be answered by looking at the given CXR image itself, such as
asking about changes/comparisons from previous scans, asking about staff notifications, or asking
about view types or other scans.
Here is one example:
Chest X-ray: {...}, One open-ended question can be: {...}, One closed-ended question can be:
{...}, One single-choice question can be: {...}, One multi-choice question can be: {...}
”"}]
messages += [{"role": "user", "content": “Chest X-ray: There is also fullness of the right hilum
which is new. [visual location: right hilar structures] ...”}

```

Table 7.3: Our designed prompt for generating groundable and explainable medical VQA, using a grounded report as input.

answerable; (3) similar to the re-grounding process, we extract 50 samples to observe the effect of the prompt. Only when the overall performance meets our expectations do we proceed with the final dataset generation. The detailed prompt is elaborated in Table 7.3.

For each image-report sample, we instruct the GPT-4o to generate a total of **11** questions: 3 open-ended VQAs, 2 closed-ended VQAs, 3 single-choice VQAs, and 3 multi-choice

	Open. (T/B)	Closed. (T/B)	Single. (T/B)	Multi. (T/B)
Train	441,471/466,725	272,323/277,249	441,114/448,810	434,067/861,635
Valid	3,524/3,704	2,145/2,184	3,520/3,599	3,451/6,955
Test	1,144/1,189	543/552	1,300/1,310	973/1,870
Total	446,139/471,618	275,011/279,985	445,934/453,719	438,491/870,460

Table 7.4: Distribution statistics of question types (T) and number of bounding boxes (B) across data splits.

VQAs, culminating in approximately 1.6 million VQA pairs. The reports containing less than three sentences will not be used to generate QAs. The generated location is provided as an anatomical region in text format (*e.g.*, “left lower lung zone”). To enable the VQA model to identify the specific location on the image, a post-processing step is required to map the region to bounding box coordinates. An example demonstrating multi-modal explainability is presented in Figure 7.2 (right).

7.3 Statistics of GEMeX

Dataset Split. GEMeX is partitioned in accordance with the distribution of MIMIC-CXR. Specifically, we have 149,535 images with 1,588,975 QA pairs for training, 1,190 images with 12,640 QA pairs for validation, and 300 images with 3,960 QA pairs for testing. Detailed statistics, including question type distribution and the number of bounding boxes, are shown in Table 7.4.

Data Quality. To establish a golden test set, we selected 300 images from the MIMIC-CXR’s test set, initially accompanied by 3,291 questions auto-generated by GPT-4o. *Radiologists meticulously reviewed these questions, correcting around 10 incorrect answers and adjusting 3 inaccurate location annotations. This minimal revision rate*

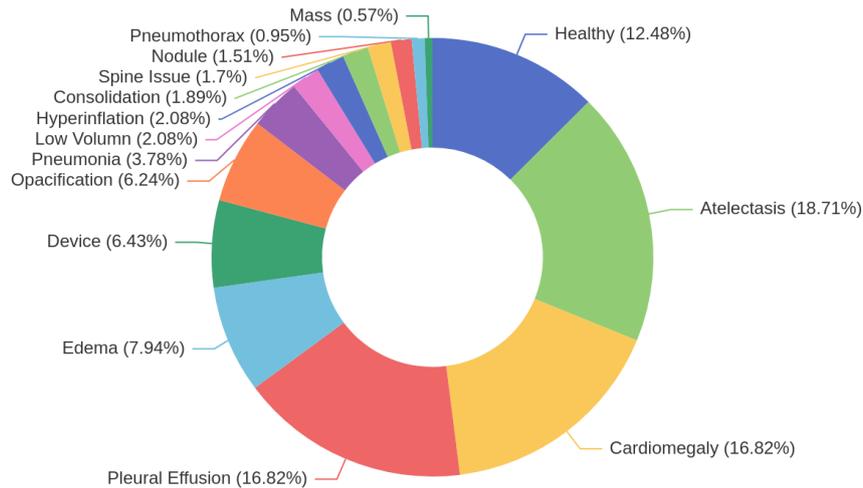


Figure 7.3: The distribution of normality and abnormality contained in images from the test set of our GEMeX.

demonstrates the high quality of the generated dataset. Additionally, the radiologists contributed approximately 600 new questions, thereby creating a comprehensive golden test set for benchmarking large vision language models.

Distribution of Clinical Observations. The X-ray image selection for the test set is according to the clinical observations. In Figure 7.3, we plot the distribution of normality and abnormality contained in images. The original distribution of MIMIC-CXR, characterized by a high proportion of healthy samples, introduces a significant bias affecting model performance [210]. To mitigate this issue, we preserve only a small proportion of healthy samples (around **12.48%**) during manual cleaning. Meanwhile, we ensure that clinically important observations occupy a large proportion, such as “atelectasis”, “cardiomegaly”, “edema”, and “pleural effusion” [211]. Additionally, common diseases or observations like “pneumonia”, “opacification”, and “pneumothorax” are also included.

Distribution of Question Content. We show the distribution of the question content categories that GPT-4o itself summarizes according to demonstrations during VQA gen-

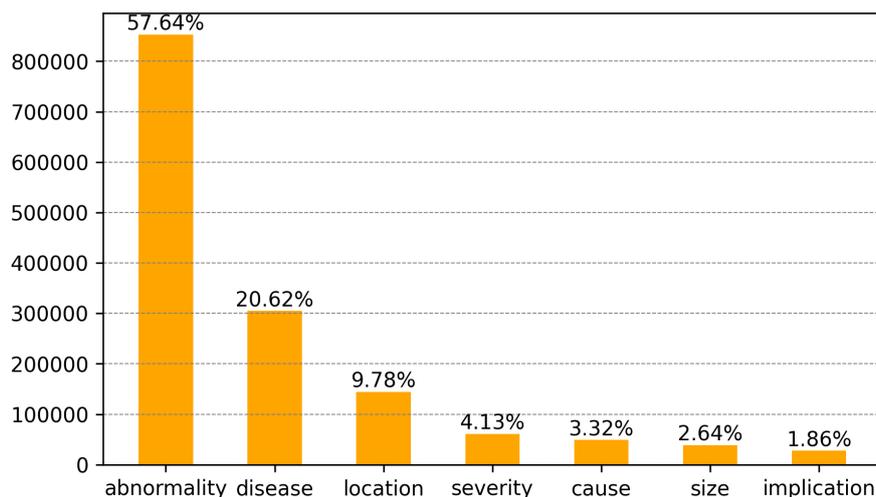


Figure 7.4: The distribution of question content in our GEMeX.

eration. Figure 7.4 shows the corresponding results where we can find that “abnormality”, “disease”, and “location” account for over 88%, while the remaining categories mainly include “cause”, “size”, “severity”, and “implication”, which highlights the diversity of questions.

7.4 Evaluation of GEMeX

7.4.1 Experimental Details

7.4.1.1 A Strong Baseline Fine-Tuned on GEMeX

To validate the effectiveness of the dataset, especially the auto-generated training set, we propose a question-type-aware instruction tuning to fine-tune LLaVA-Med-v1-7B [80] on the training set of GEMeX, termed as LLaVA-Med-GEMeX, serving as a strong baseline. Specifically, for each VQA sample from our GEMeX, we add a type prompt X_{Type} after the original system prompt and a question X_{Question} with its answer X_{Answer} , textual

reason $\mathbf{X}_{\text{Reason}}$, and corresponding visual location $\mathbf{X}_{\text{Location}}$, constructing a single-turn dialogue shown in Table 7.5. In particular, \mathbf{X}_{Type} is “Input an {Type} question, and the assistant will output its answer {Supplement} with a detailed reason and corresponding visual location.” where {Type} refers to “open-ended/closed-ended/single-choice/multi-choice” and {Supplement} is replaced by “none/(yes or no)/(an option)/(some options)”, respectively.

<pre> <Ori.System.Prompt> \mathbf{X}_{Type} <STOP> Human: <image>\n$\mathbf{X}_{\text{Question}}$ $\mathbf{X}_{\text{Choices}}$ (if any) <STOP> Assistant: <answer>$\mathbf{X}_{\text{Answer}}$ <reason>$\mathbf{X}_{\text{Reason}}$ <location>$\mathbf{X}_{\text{Location}}$ <STOP> </pre>
--

Table 7.5: Input format for fine-tuning LLaVA-Med.

7.4.1.2 Training Details

We fine-tune both the visual projection layers and the LLM components of LLaVA-Med-v1 (after stage II) by calculating the auto-regressive loss to predict the assistant’s responses and the dialogue termination token <STOP>. Particularly, the model is fine-tuned for 3 epochs on four NVIDIA H100 GPUs with a batch size of 64, taking around 54 hours. The network is warmed up in the first 0.03 epochs with a linear learning rate from $3e-7$ to $2e-5$, which further decays by cosine schedule. The optimizer is AdamW. To enhance training efficiency, we utilize the Fully Sharded Data Parallel (FSDP) mechanism, the bf16 (Brain Floating Point) precision format, and gradient checkpointing.

7.4.1.3 LVLMS Benchmarks

Besides fine-tuning a task-oriented model, we perform a zero-shot evaluation on our GEMeX dataset across the other 12 LVLMS, with 7 in the general domain and the other 5 in the medical domain:

In the General Domain: we test LLaVA-v1 [85], Mini-GPT4-v1 [84], mPLUG-Owl [212], LLaVA-v1.5 [213], DeepSeek-VL [209], Qwen-VL-Chat [214], and GPT-4o-mini (2024-07-18). Note that we did not test GPT-4o because its safety protection policy prohibits it from analyzing medical images, especially when asking about the condition of patients.

In the Medical Domain: we evaluate LLaVA-Med-v1 [80], LLaVA-Med-v1.5 [80], MiniGPT-Med [79], XrayGPT [77], and RadFM [78].

To make a fair comparison, the evaluated models (except GPT-4o-mini and RadFM (with MedLLaMA-13B [215])) are based on 7B-LLMs in this section. Specifically, LLaVA-v1, LLaVA-Med-v1, and Mini-GPT4-v1 are based on Vicuna-v0-7B [152] while LLaVA-v1.5 and XrayGPT are based on Vicuna-v1-7B; LLaVA-Med-v1.5 is built upon Mistral-7B-Instruct-v0.2 [166]; mPLUG-Owl is using LLaMA-7B [37]; Deepseek-VL [209] is based on DeepSeek-LLM-7B; and Qwen-VL-Chat [216] is based on Qwen-7B. All models' configurations are set according to their open-source codes.

7.4.2 Evaluation Metrics

In GEMeX, each question has a corresponding answer, textual reason, and visual location. Ideally, we aim to evaluate all these three aspects with designed metrics as follows:

Answer-Reason Score (AR-score): In reality, most LVLMs struggle to output correctly in terms of format. This doesn't mean these models lack the knowledge to answer the questions but rather simply lack the ability to follow instructions properly. To ensure a fair comparison, we introduce the Answer-Reason score (AR-score) as an evaluation metric for the textual output, where the answer and reason parts from each test sample are merged as a reference (ground truth), and the evaluated LVLM's output serve as a candidate. We use GPTScore [80] to calculate the AR-Score from a semantic perspective, where GPT-4o is utilized to quantify the correctness. Specifically, we start by treating the aforementioned reference as a textual response from assistant #1, while the candidate as the response from assistant #2. With both responses, the original question, and the X-ray report, GPT-4o assesses the accuracy, relevance, and helpfulness of each assistant's answer and assigns a comprehensive rating on a 1 to 10 scale, with higher values reflecting better performance. We then calculate the relative score using GPT-4o's reference score for normalization. Besides, we also employ common NLG metrics (*e.g.*, BERTScore [217], BLEU, ROUGE) to evaluate AR-score.

Answer Score (A-score): For responses where the model can output specific answers (such as yes/no for closed-ended questions or options for single-/multiple-choice questions), we calculate the accuracy by comparing with the ground truth. It is worth noting that although some models cannot directly output the answer, we still attempt to match it from their responses.

Visual Score (V-score): For models capable of visual grounding (*i.e.*, outputting visual locations), we calculated mean intersection over union (mIoU) as a metric. For a VQA case, considering there might be multiple corresponding locations (commonly seen in multi-choice questions), we use the Hungarian algorithm [218] to match the predicted

Models	Open-ended		Closed-ended			Single-choice			Multi-choice			Avg. †
	AR-score †	V-score	A-score	AR-score †	V-score	A-score	AR-score †	V-score	A-score	AR-score †	V-score	
Random	-	-	48.80	-	-	25.85	-	-	7.50	-	-	-
GPT-4o-mini [17]	97.68	<u>18.05</u>	59.30	71.14	<u>28.64</u>	<u>59.00</u>	<u>77.47</u>	<u>23.62</u>	<u>49.13</u>	<u>82.91</u>	<u>19.19</u>	<u>82.30</u>
LLaVA-v1 [85]	76.14	-	30.76	38.02	-	-	50.47	-	-	66.52	-	57.79
LLaVA-v1.5 [213]	77.62	-	58.93	57.00	-	47.00	57.05	-	-	65.17	-	64.21
Mini-GPT4-v1 [84]	55.32	-	26.33	31.09	-	-	37.63	-	-	46.65	-	42.67
mPLUG-Owl [212]	76.73	-	27.26	36.70	-	32.00	46.89	-	-	67.92	-	57.06
DeepSeek-VL [209]	79.30	11.00	57.10	59.86	8.28	51.69	62.03	8.57	17.99	70.35	12.98	67.89
Qwen-VL-Chat [214]	78.36	3.17	23.02	45.79	12.25	44.69	59.15	16.69	7.30	67.21	2.26	62.63
LLaVA-Med-v1 [80]	90.34	-	62.62	69.91	-	-	61.74	-	-	68.14	-	72.53
LLaVA-Med-v1.5 [80]	94.43	-	<u>71.82</u>	<u>76.54</u>	-	-	66.04	-	-	67.28	-	76.07
MiniGPT-Med [79]	86.12	-	55.24	65.25	-	-	55.61	-	-	64.33	-	67.83
XrayGPT [77]	81.17	-	-	68.17	-	-	48.33	-	-	55.10	-	63.19
RadFM [78]	88.57	-	58.01	67.91	-	-	57.82	-	-	62.41	-	69.18
LLaVA-Med-GEMeX	<u>97.05</u>	51.47	77.35	80.72	53.20	73.08	81.42	54.57	67.42	84.98	47.99	86.04

Table 7.6: Performance of representative LVLMs on our GEMeX across four question types. The AR-score combines answer and reason to evaluate textual outputs. † denotes the GPTScore value (%). The A-score indicates answer or choice accuracy (%), and the V-score represents mIoU (%). The best results are bolded, and the second-best are underlined in each column.

bounding boxes with the actual ones.

7.4.3 Results and Analysis

7.4.3.1 Overall Performance

The comprehensive results are shown in Table 7.6. The first 7 rows indicate the performance of general LVLMs, while the last 6 rows present the results of medical ones and our fine-tuned version of LLaVA-Med-v1 (termed as LLaVA-Med-GEMeX). It can be found that:

- **Most available LVLMS exhibit suboptimal performance when tested on GEMeX.**

The only exception is GPT-4o-mini, which achieves an average AR-score above 80 across all tasks. When considering specific question types, LLaVA-Med (both versions 1 and 1.5) stands out for its strong performance on open-ended questions, scoring above 90 on the AR-score. However, all models show poor results on the other three categories of tasks.

- **When faced with choice-based questions, most models, particularly those in the medical domain, struggle to provide definitive answers, despite their ability to analyze each option.** This difficulty accounts for why many models have an associated AR-score but lack an A-score, highlighting the importance of introducing these types of questions.
- **Powerful LVLMS, such as GPT-4o-mini, often rely on shortcut reasoning rather than real multimodal reasoning.** Although these models can sometimes answer questions to a certain extent (as indicated by the AR-scores), they often fail to accurately achieve visual grounding. This suggests that these models tend to address Med-VQA tasks using shortcut knowledge, such as retrieving information from their pre-training memory, instead of engaging in real multimodal reasoning [219]. However, real multimodal reasoning is essential for the explainability of Med-VQA systems.
- **Through simple question-type-aware instruction tuning, the proposed baseline model achieves a significant performance improvement,** with an approximate **13.5%** increase in average AR-score compared to LLaVA-Med-v1. Notably, it surpasses GPT-4o-mini on most metrics, demonstrating the reliability of the training

Models	Open-ended			Closed-ended			Single-choice			Multi-choice		
	BERTScore	ROUGE-L	BLEU-1									
GPT-4o-mini [17]	<u>30.43</u>	<u>22.67</u>	<u>18.25</u>	40.02	25.63	19.10	<u>48.34</u>	<u>39.17</u>	<u>30.82</u>	<u>46.58</u>	<u>39.20</u>	<u>28.65</u>
LLaVA-v1 [85]	20.09	15.22	11.57	22.42	13.10	8.01	20.25	14.97	10.61	19.69	17.35	11.15
LLaVA-v1.5 [213]	21.49	16.11	12.20	32.59	15.37	6.69	17.42	17.53	1.49	23.74	21.20	8.95
Mini-GPT4-v1 [84]	15.03	14.66	11.46	13.83	9.65	6.31	6.50	6.79	4.60	5.31	5.79	3.22
mPLUG-Owl [212]	22.52	17.03	13.22	32.23	20.20	13.92	39.64	33.69	30.32	26.09	24.97	16.68
DeepSeek-VL [209]	24.06	18.62	15.94	26.12	23.27	13.83	26.16	30.46	18.79	22.10	27.32	20.74
Qwen-VL-Chat [214]	23.31	18.48	14.63	33.18	22.43	17.19	25.47	22.25	6.95	22.03	22.88	12.26
LLaVA-Med-v1 [80]	25.14	19.63	15.93	38.04	29.08	19.74	34.89	30.10	25.84	28.63	26.51	20.99
LLaVA-Med-v1.5 [80]	26.42	21.38	17.28	<u>44.48</u>	<u>36.73</u>	<u>26.35</u>	36.62	30.32	25.44	28.11	24.49	16.53
MiniGPT-Med [79]	23.47	19.20	16.03	34.31	29.47	19.13	30.11	28.51	22.13	26.51	24.42	15.98
XrayGPT [77]	22.57	18.30	15.73	21.35	14.55	10.17	16.31	12.17	9.23	12.15	10.30	6.22
RadFM [78]	24.96	20.71	17.73	37.43	27.95	20.56	32.30	27.02	24.39	25.81	20.02	13.80
LLaVA-Med-GEMeX	42.69	32.75	25.28	54.44	38.39	33.99	56.35	53.23	47.31	54.95	50.85	43.99

Table 7.7: Performance of representative LVLMS evaluated using various natural language generation metrics for AR-score, including BERTScore, ROUGE-L, and BLEU-1. The best results are bolded, and the second-best are underlined in each column.

set. However, a substantial gap remains for practical application, highlighting the challenges associated with GEMeX.

Limitation. Note that the proposed baseline model is inherently task-specific, which may result in reduced accuracy on other tasks or a diminished capacity for conversational engagement. The true potential of our GEMeX lies in its integration into multi-task training frameworks, such as the second training stage of LLaVA-Med. The baseline model primarily serves to demonstrate the dataset’s effectiveness while also providing a robust benchmark.

7.4.3.2 More Metrics

As we mentioned in Section 7.4.2, we also calculate NLG metrics to measure AR-score. Detailed results are presented in Table 7.7. Overall, the NLG metrics generally share the same trend as GPTScore (AR-score in Table 7.6), but there are some minor differences. (1) Models with high NLG scores do not always correlate with good performance, as seen with mPLUG-Owl compared to LLaVA-v1.5. Essentially, LLaVA-1.5 demonstrates higher performance, such as achieving an answer accuracy rate (A-score) in single-choice tasks that is **15%** higher than that of mPLUG-Owl. However, since LLaVA-v1.5’s output mostly consists of the answer without reason, the shorter output results in a lower NLG score, with its BLEU-1 approximately **28.8%** lower than mPLUG-Owl; (2) NLG metrics can better reflect the performance improvement after fine-tuning. For example, the fine-tuned model shows only about a 3.7% average improvement over GPT-4o-mini on GPTScore but brings around a **12.1%** improvement on average NLG metrics. This more significant improvement better demonstrates the model’s learning effect on the dataset. Overall, we believe that for models that have not participated in GEMeX training, using GPTScore is more reasonable, as semantic understanding can be used to judge the difference between the model’s output and the true label. For fine-tuned models, NLG metrics are preferred because they can better reflect the model’s alignment with the dataset.

7.4.3.3 Transferability

To further validate the effectiveness of GEMeX, we assess it from a transfer learning perspective by treating GEMeX as a pre-training dataset, and then testing our LLaVA-

Table 7.8: Performance of LLMs on SLAKE-CXR.

Models	Open-ended AR-Score	Closed-ended	
		A-Score	AR-Score
LLaVA-Med-v1	73.31	56.17	62.35
LLaVA-Med-GEMeX	82.78	69.79	75.06

Med-GEMeX on other X-ray VQA datasets. Given the potential overlap in data sources, we utilize the CXR part from the test set of SLAKE [2], which includes 505 open-ended questions and 235 closed-ended questions. The zero-shot performance of LLaVA-Med-GEMeX and LLaVA-Med-v1 on this dataset is reported in Table 7.8. The results reveal that our LLaVA-Med-GEMeX still outperforms LLaVA-Med-v1 on both tasks by a large margin, demonstrating the effective knowledge regarding X-ray learned from GEMeX.

7.4.4 Qualitative Evaluation

In Table 7.9, we show some questions with outputs from both GPT-4o-mini and our LLaVA-Med-GEMeX for qualitative comparison. In CASE I, although GPT-4o-mini can generate a very detailed answer, it provides answers without reasoning on the visual content, resulting in a significant difference from the ground truth. In contrast, the LLaVA-Med-GEMeX offers relatively accurate visual clues and is able to provide partially correct answers (“the compensatory mechanism”), while there is a false mention of “possible chronic lung disease” when considering the patient’s condition. In CASE II, although GPT-4o-mini can analyze images, its limited capabilities can select only one option and provide a vague reason. In contrast, the LLaVA-Med-GEMeX outputs the correct options but gives an incorrect reason for one option (*i.e.*, answer “A”). In CASE III and CASE IV, GPT-4o-mini cannot both visually reason and output answers cor-

rectly, while the fine-tuned model can give better outputs from these two aspects. From these examples, we can conclude that some LVLMs still lack sufficient understanding of medical images. Meanwhile, while the proposed simple fine-tuning method improves performance, it is still far from fully accurate, leaving much room for further exploration. More cases can be explored by using the display system we designed, as detailed in the project page.

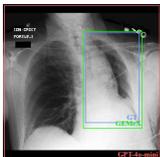
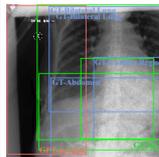
Challenging examples from GEMeX:		
	<p>CASE I</p>  <p>(Bbox predicted by each model)</p>	<p>CASE II</p>  <p>(Bbox predicted by each model)</p>
<i>Question</i>	What might be the cause of the overexpansion of the right lung?	What regions on the X-ray indicate abnormal findings? Choices: [A: Bilateral lung, B: Cardiac region, C: Abdomen, D: Spine]
<i>Ground Truth</i>	<Answer>: Compensatory hyperinflation due to left lung removal. <Reason>: After a pneumectomy, the remaining lung may overexpand to compensate for the loss of the left lung.	<Answer>: [A, B, C]. <Reason>: The bilateral lung shows pleural effusion, the cardiac region shows enlargement, and the abdomen shows the nasogastric tube. The spine is normal.
<i>GPT-4o-mini</i>	<Answer>: The overexpansion of the right lung may be caused by conditions such as a right-sided pneumothorax or chronic obstructive pulmonary disease (COPD). <Reason>: Overexpansion can occur when air enters the pleural space (as in pneumothorax) or when there is an obstruction that prevents air from leaving the lung, causing it to become hyperinflated. In COPD, the lung tissue loses elasticity, leading to overinflation. <GPTScore>: 44.44	<Answer>: [A: Bilateral lung]. <Reason>: The bilateral lung region shows potential abnormalities such as opacities or other irregularities that may indicate pathology (✗). <GPTScore>: 37.50
<i>LLaVA-Med-GEMeX</i>	<Answer>: Possible chronic lung disease or compensatory mechanism (✓). <Reason>: Overexpansion of the lung can be due to chronic lung diseases or compensatory mechanisms, especially after a left lobectomy (✓). <GPTScore>: 66.67	<Answer>: [A, B, C]. <Reason>: The bilateral lung shows interstitial opacities (✗), the cardiac region shows enlargement (✓), and the abdomen shows the feeding tube location (✓). <GPTScore>: 87.50
	<p>CASE III</p>  <p>(Bbox predicted by each model)</p>	<p>CASE IV</p>  <p>(Bbox predicted by each model)</p>
<i>Question</i>	What is the overall condition of the lungs based on the CXR? Choices: [A: Clear, B: Consolidated, C: Effused, D: Thickened]	Are there signs of pneumonia in this CXR?
<i>Ground Truth</i>	<Answer>: A. <Reason>: The CXR indicates that the lungs are clear of consolidation or effusion, although right pleural-based thickening at the base.	<Answer>: No. <Reason>: The CXR explicitly states that there is no evidence of pneumonia, indicating clear lung fields.
<i>GPT-4o-mini</i>	<Answer>: B. <Reason>: The presence of increased opacity in the lung fields suggests consolidation, indicating that the lungs are not clear and may have fluid or solid material in the alveoli (✗). <GPTScore>: 33.33.	<Answer>: Yes. <Reason>: The chest X-ray shows areas of increased opacity, which may indicate the presence of fluid or consolidation (✗) typically associated with pneumonia. <GPTScore>: 37.50
<i>LLaVA-Med-GEMeX</i>	<Answer>: A. <Reason>: The CXR indicates that the lungs are otherwise clear except for the noted pleural thickening (✓). <GPTScore>: 88.89	<Answer>: No. <Reason>: The CXR indicates no focal consolidation, which rules out pneumonia (✓). <GPTScore>: 100

Table 7.9: Challenging questions from GEMeX answered by GPT-4o-mini and our LLaVA-Med-GEMeX, highlighting the limitations of LVLMS like GPT-4o-mini in visual reasoning, which can be enhanced by integrating knowledge from GEMeX. Colors in outputs highlight pathological descriptions that are similar to the ground truth, though they may be either correct (✓) or incorrect (✗).

Chapter Review

In this chapter, we construct a benchmark, GEMeX, aiming to advance the field of medical VQA with two primary advantages: multimodal explainability and diverse question types. GEMeX not only provides more accessible medical explanations for patients and junior doctors but also functions as an important dataset for developing next-generation medical LVLMs with enhanced instruction-following capabilities. We demonstrate the effectiveness and difficulty of the dataset through comprehensive testing of representative LVLMs as well as task-specific fine-tuning, hoping GEMeX can promote medical VQA development and AI-assisted medical care.

Chapter 8

Conclusion and Future Works

This thesis has made significant contributions to addressing critical challenges in conversational AI, focusing on two essential aspects: textual out-of-distribution (OOD) detection and medical visual question answering (Med-VQA). Through comprehensive research and novel methodologies, we have advanced both the safety and usability of conversational systems, particularly for high-stakes applications.

Our research on OOD detection in large language models (LLMs) has made significant advancements, including the first comprehensive evaluation of traditional OOD detection methods across zero-shot and fine-tuning settings with LLMs [7]. The results show that cosine distance-based detector outperforms others by effectively utilizing the isotropic embedding space of LLMs. We also introduce a generative fine-tuning paradigm that aligns with LLMs' pre-training objectives, improving OOD detection capabilities while keeping performance on in-distribution tasks. To tackle near-OOD detection, we develop a novel framework leveraging LLMs' isotropic characteristics, generating semantic prototypes for ID classes and using semantic matching for OOD detection and ID clas-

sification. This framework shows superior performance in few-shot learning, making it particularly valuable for domains like medicine and finance, where data acquisition is limited.

In the field of Med-VQA, we make significant advancements to address data scarcity, enhance reasoning, and improve explainability. We introduced SLAKE [2], a semantically-labeled, knowledge-enhanced dataset with accurate visual and textual annotations, enabling contextual understanding of medical images and supporting complex questions requiring both visual comprehension and medical reasoning. This dataset fills the gap in the medical VQA field where there is a lack of suitable datasets. To mitigate overfitting on small-scale datasets, we develop a pre-training framework CPRD [3] that learns transferable features from unlabeled radiology images and distills a lightweight visual feature extractor. Additionally, we propose a novel conditional reasoning mechanism CR [4] with the QCR and TCR modules to improve reasoning on both closed- and open-ended tasks. Finally, to enhance explainability, we introduced GEMeX [6], the largest chest X-ray VQA benchmark with multi-modal explanations and over 1.6 million questions across 151,025 images, offering a comprehensive foundation for future research in clinical VQA.

The significance of this thesis is two-fold:

- The thesis pioneers in evaluating the capabilities of LLMs for textual OOD detection, pointing out the limitations and effectiveness of existing detection methods. Based on this, a framework is designed for efficient few-shot near-OOD detection. These contributions ensure the safety and robustness of conversational AI.
- This thesis greatly advances the development of Med-VQA, contributing two im-

portant datasets to fill gaps and introduce explainability and a series of methods to enhance visual representation and multimodal reasoning capabilities. These contributions enhance the usability and applicability of conversational AI.

Future work. The following issues are left for future exploration.

1) More LLMs for textual OOD detection. As demonstrated in Chapters 3 and 4, using LLMs for textual OOD detection proves that among existing post-hoc OOD detection methods, cosine distance is the simplest and most effective. Additionally, a near-OOD detection framework based on cosine distance is proposed, achieving excellent results in few-shot settings. Despite these promising results, one worthwhile direction for future research is to validate and generalize these findings across more types of LLMs. The current work primarily focuses on the LLaMA series; however, other recent LLM models, such as Qwen [216] and DeepSeek [209] series, have also shown outstanding performance, making exploration of OOD detection on these models part of future plans. Furthermore, the work in this thesis mainly concentrates on non-chat models. Exploring how well current OOD detection methods perform on chat-based models, how to design efficient prompts, and how to develop methodologies specifically for chat-based models, also represents future work.

2) A large vision-language model for Med-VQA. Recently, the development of LVLMs (Large Vision-Language Models) has received widespread attention. In combination with GEMeX, a large-scale Med-VQA dataset introduced in Chapter 7 of this thesis, one potential future direction is to design a more comprehensive LVLM. Although in Chapter 7 we fine-tune an LVLM, it is task-specific, meaning it loses its ability to engage in general dialogue or solve other downstream tasks. Therefore, future work could involve training

a more general LVLM by incorporating multi-modal medical image-text data with our GEMeX and applying techniques like instruct-tuning for the medical field.

3) Exploration of more medical vision-language tasks. In this thesis, we primarily focus on the Med-VQA task. However, current LVLMs still show insufficient performance for other medical image-text tasks, such as image-text retrieval and report generation. Therefore, a future direction is to explore these tasks and integrate them into a single LVLM to enhance the usability of conversational AI. Additionally, an AI agent system [220] could be created, where each agent is responsible for a specific image-text task, utilizing them as engines for conversational AI to improve overall usability.

Bibliography

- [1] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *ACM MM*, 2020.
- [2] 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, 2021.
- [3] Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24*, pages 210–220. Springer, 2021.
- [4] Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE transactions on medical imaging*, 42(5):1532–1545, 2022.
- [5] Bo Liu, Li-Ming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. How good are llms at out-of-distribution detection? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8211–8222, 2024.

- [6] Bo Liu, Ke Zou, Liming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. *arXiv preprint arXiv:2411.16778*, 2024.
- [7] Bo Liu, Liming Zhan, Zexin Lu, Yujie Feng, Lei Xue, and Xiao-Ming Wu. How good are large language models at out-of-distribution detection? *arXiv preprint arXiv:2308.10261*, 2023.
- [8] Bo Liu, Liming Zhan, Yujie Feng, Zexin Lu, Chengqiang Xie, Lei Xue, Albert Lam, and Xiao-Ming Wu. Diversity-grounded channel prototypical learning for out-of-distribution intent detection. *arXiv preprint arXiv:2409.11114*, 2024.
- [9] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- [10] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 2008.
- [11] 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):1–10, 2018.
- [12] 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 *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

- [13] Binh D. Nguyen, Thanh-Toan Do, Binh X. Nguyen, Tuong Do, Erman Tjiputra, and Quang D. Tran. Overcoming data limitation in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2019 - 22nd International Conference, Part IV*, volume 11767 of *Lecture Notes in Computer Science*, pages 522–530, Shenzhen, China, 2019. Springer.
- [14] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020.
- [15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, pages 1735–1780, 1997.
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.
- [17] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [18] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- [19] Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Lam.

- Out-of-scope intent detection with self-supervision and discriminative training. *arXiv preprint arXiv:2106.08616*, 2021.
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [21] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, NeurIPS*, pages 1571–1581, Montréal, Canada, 2018. NeurIPS.
- [22] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.
- [23] Zihao Zhao, Sheng Wang, Jinchun Gu, Yitao Zhu, Lanzhuju Mei, Zixu Zhuang, Zhiming Cui, Qian Wang, and Dinggang Shen. Chatcad+: Towards a universal and reliable interactive cad using llms. *IEEE Transactions on Medical Imaging*, 2024.
- [24] Dan Hendrycks and Mantas Mazeika. X-risk analysis for ai research. *arXiv preprint arXiv:2206.05862*, 2022.
- [25] Jingkan Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.
- [26] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness.

- In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, 2020.
- [27] Li-Ming Zhan, Bo Liu, and Xiao-Ming Wu. Vi-ood: A unified framework of representation learning for textual out-of-distribution detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17371–17383, 2024.
- [28] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL <https://openreview.net/forum?id=Hkg4TI9xl>.
- [29] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.
- [30] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475, 2020.
- [31] Alexander Podolskiy, Dmitry Lipin, Andrey Bout, Ekaterina Artemova, and Irina Piontkovskaya. Revisiting mahalanobis distance for transformer-based out-of-domain detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13675–13682, 2021.
- [32] Wenxuan Zhou, Fangyu Liu, and Muhao Chen. Contrastive out-of-distribution

- detection for pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1100–1111, 2021.
- [33] Rheeya Uppaal, Junjie Hu, and Yixuan Li. Is fine-tuning needed? pre-trained language models are near perfect for out-of-domain detection. *arXiv preprint arXiv:2305.13282*, 2023.
- [34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- [35] Hyunsoo Cho, Choonghyun Park, Junyeop Kim, Hyuhng Joon Kim, Kang Min Yoo, and Sang-goo Lee. Probing out-of-distribution robustness of language models with parameter-efficient transfer learning methods. *arXiv preprint arXiv:2301.11660*, 2023.
- [36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [37] 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.
- [38] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean,

- and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL <https://openreview.net/forum?id=yzkSU5zdWd>. Survey Certification.
- [39] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.
- [40] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, 2022.
- [41] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.
- [42] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.
- [43] Zexin Lu, Jing Li, Yingyi Zhang, and Haisong Zhang. Getting your conversation on track: Estimation of residual life for conversations. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1036–1043. IEEE, 2021.
- [44] Yujie Feng, Zexin Lu, Bo Liu, Liming Zhan, and Xiao-Ming Wu. Towards llm-driven dialogue state tracking. *arXiv preprint arXiv:2310.14970*, 2023.

- [45] Yujie Feng, Bo Liu, Xiaoyu Dong, Zexin Lu, Li-Ming Zhan, Albert Lam, and Xiao-Ming Wu. Continual dialogue state tracking via reason-of-select distillation. *arXiv preprint arXiv:2408.09846*, 2024.
- [46] Yangyang Zhou, Xin Kang, and Fuji Ren. Employing inception-resnet-v2 and bi-lstm for medical domain visual question answering. In *CLEF Workshop*, 2018.
- [47] Asma Ben Abacha, Soumya Gayen, Jason J. Lau, Sivaramakrishnan Rajaraman, and Dina Demner-Fushman. NLM at imageclef 2018 visual question answering in the medical domain. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, volume 2125 of *CEUR Workshop Proceedings*, Avignon, France, 2018. CEUR-WS.org.
- [48] Lei Shi, Feifan Liu, and Max P. Rosen. Deep multimodal learning for medical visual question answering. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, volume 2380 of *CEUR Workshop Proceedings*, Lugano, Switzerland, 2019. CEUR-WS.org.
- [49] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu. Zhejiang university at imageclef 2019 visual question answering in the medical domain. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, volume 2380 of *CEUR Workshop Proceedings*, Lugano, Switzerland, 2019. CEUR-WS.org.
- [50] Bogdan Ionescu, Henning Müller, Mauricio Villegas, Alba Garcia Seco de Herrera, Carsten Eickhoff, Vincent Andrearczyk, Yashin Dicente Cid, Vitali Liauchuk, Vassili Kovalev, Sadid A. Hasan, Yuan Ling, Oladimeji Farri, Joey Liu, Matthew P. Lungren, Duc-Tien Dang-Nguyen, Luca Piras, Michael Riegler, Liting Zhou, Mathias Lux, and Cathal Gurrin. Overview of imageclef 2018: Challenges,

- datasets and evaluation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF*, volume 11018 of *Lecture Notes in Computer Science*, pages 309–334, Avignon, France, 2018. Springer.
- [51] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, volume 2380 of *CEUR Workshop Proceedings*, Lugano, Switzerland, 2019. CEUR-WS.org.
- [52] Minh H Vu, Tommy Löfstedt, Tufve Nyholm, and Raphael Sznitman. A question-centric model for visual question answering in medical imaging. *IEEE transactions on medical imaging*, 2020.
- [53] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [54] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016.
- [55] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *IEEE International Conference on Computer Vision, ICCV*, pages 1839–1848, Venice, Italy, 2017. IEEE Computer Society.
- [56] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image

- captioning and visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 6077–6086, Salt Lake City, UT, USA, 2018. IEEE Computer Society.
- [57] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.
- [58] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.
- [59] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: multimodal bert pretraining for improved medical vqa. In *ISBI*, pages 1033–1036. IEEE, 2021.
- [60] Jong Hak Moon, Hyungyung Lee, Woncheol Shin, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *arXiv preprint arXiv:2105.11333*, 2021.
- [61] Li Xu, Bo Liu, Ameer Hamza Khan, Lu Fan, and Xiao-Ming Wu. Multi-modal pre-training for medical vision-language understanding and generation: An empirical study with a new benchmark. *arXiv preprint arXiv:2306.06494*, 2023.
- [62] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.
- [63] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

- [64] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020.
- [65] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.
- [66] Yi Zhou, Tianfei Zhou, Tao Zhou, Huazhu Fu, Jiacheng Liu, and Ling Shao. Contrast-attentive thoracic disease recognition with dual-weighting graph reasoning. *IEEE Transactions on Medical Imaging*, 2021.
- [67] Navid Alemi Koohbanani, Balagopal Unnikrishnan, Syed Ali Khurram, Pavitra Krishnaswamy, and Nasir Rajpoot. Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, 2021.
- [68] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [69] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *MICCAI*, pages 679–689. Springer, 2022.
- [70] 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.

- [71] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.
- [72] Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. Towards unifying medical vision-and-language pre-training via soft prompts. In *ICCV*.
- [73] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *MICCAI*, pages 101–111. Springer, 2023.
- [74] Bo Liu, Donghuan Lu, Dong Wei, Xian Wu, Yan Wang, Yu Zhang, and Yefeng Zheng. Improving medical vision-language contrastive pretraining with semantics-aware triage. *IEEE Transactions on Medical Imaging*, 42(12):3579–3589, 2023.
- [75] Bo Liu, Zexin Lu, and Yan Wang. Towards medical vision-language contrastive pre-training via study-oriented semantic exploration. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4861–4870, 2024.
- [76] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *MICCAI*, pages 525–536. Springer, 2023.
- [77] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.

- [78] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*, 2023.
- [79] Asma Alkhalidi, Raneem Alnajim, Layan Alabdullatef, Rawan Alyahya, Jun Chen, Deyao Zhu, Ahmed Alsinan, and Mohamed Elhoseiny. Minigpt-med: Large language model as a general interface for radiology diagnosis. *arXiv preprint arXiv:2407.04106*, 2024.
- [80] 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. *NIPS*, 36, 2024.
- [81] Ke Zou, Yang Bai, Zhihao Chen, Yang Zhou, Yidi Chen, Kai Ren, Meng Wang, Xuedong Yuan, Xiaojing Shen, and Huazhu Fu. Medrg: Medical report grounding with multi-modal large language model. *arXiv preprint arXiv:2404.06798*, 2024.
- [82] Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. *arXiv preprint arXiv:2305.11490*, 2023.
- [83] Bo Liu, Xiangyu Zhao, Along He, Yidi Chen, Huazhu Fu, and Xiao-Ming Wu. Gemex-thinkvg: Towards thinking with visual grounding in medical vqa via reinforcement learning. *arXiv preprint arXiv:2506.17939*, 2025.
- [84] 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.

- [85] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [86] Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, and Xin Gao. Skingpt-4: an interactive dermatology diagnostic system with visual large language model. *arXiv preprint arXiv:2304.10691*, 2023.
- [87] Xiaolan Chen, Ziwei Zhao, Weiyi Zhang, Pusheng Xu, Le Gao, Mingpu Xu, Yue Wu, Yinwen Li, Danli Shi, and Mingguang He. Eyegpt: Ophthalmic assistant with large language models. *arXiv preprint arXiv:2403.00840*, 2024.
- [88] Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, Jietian Jin, Jiayu Zhang, Ruijiang Li, Hongping Tang, Kanran Wang, Yu Li, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, pages 1–9, 2024.
- [89] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *CVPR*, pages 22170–22183, 2024.
- [90] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.
- [91] Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes*. 21-24 September 2021, 2021.

- [92] 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.
- [93] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Jiayu Lei, Ya Zhang, Yanfeng Wang, and Weidi Xie. Radgenome-chest ct: A grounded vision-language dataset for chest ct analysis. *arXiv preprint arXiv:2404.16754*, 2024.
- [94] Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *ACM SIGKDD*, pages 4156–4165, 2023.
- [95] Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Advances in Neural Information Processing Systems*, 36, 2024.
- [96] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- [97] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

- [98] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *NeurIPS*, 2016.
- [99] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VI-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint*, 2019.
- [100] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.
- [101] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0.1: the Winning Entry to the VQA Challenge 2018. *arXiv e-prints*, page arXiv:1807.09956, 2018.
- [102] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.
- [103] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- [104] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. VIlbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NIPS*, 32, 2019.
- [105] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

- [106] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixelbert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.
- [107] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*. PMLR, 2021.
- [108] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, 2016.
- [109] Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. In *ICLR*, 2018.
- [110] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.
- [111] Xuanyi Dong, Linchao Zhu, De Zhang, Yi Yang, and Fei Wu. Fast parameter adaptation for few-shot image captioning and visual question answering. In *ACM MM*, 2018.
- [112] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *ICLR*, 2020.
- [113] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

- [114] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.
- [115] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In *NeurIPS*, 2018.
- [116] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *ICLR*, 2019.
- [117] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.
- [118] Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. Clinically correct report generation from chest x-rays using templates. In *International Workshop on Machine Learning in Medical Imaging*, pages 654–663. Springer, 2021.
- [119] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 721–729. Springer, 2019.
- [120] Pablo Pino, Denis Parra, Pablo Messina, Cecilia Besa, and Sergio Uribe. Inspecting state of the art performance and nlp metrics in image-based medical report generation. *arXiv preprint arXiv:2011.09257*, 2020.

- [121] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.
- [122] Stephanie L Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023.
- [123] Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. Mitigating hallucinations of large language models in medical information extraction via contrastive decoding. *arXiv preprint arXiv:2410.15702*, 2024.
- [124] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19809–19818, 2023.
- [125] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10394–10403, 2019.
- [126] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *ICCV*, pages 3942–3951, 2021.
- [127] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired

- images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR, 2022.
- [128] Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out of-distribution detection: Solutions and future challenges. *Trans. Mach. Learn. Res.*, 2022, 2022. URL <https://openreview.net/forum?id=aRtjVZvbpK>.
- [129] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *arXiv preprint arXiv:1909.00512*, 2019.
- [130] Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. *arXiv preprint arXiv:1907.12009*, 2019.
- [131] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.
- [132] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [133] OpenAI. Gpt-4 technical report, 2023.
- [134] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

- Askeel, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [135] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10051–10059, 2022.
- [136] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 15313–15323, 2021.
- [137] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fugie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.
- [138] Udit Arora, William Huang, and He He. Types of out-of-distribution texts and how to detect them. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10687–10701, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.835. URL <https://aclanthology.org/2021.emnlp-main.835>.
- [139] Ken Lang. Newsweeder: Learning to filter netnews. In *Machine learning proceedings 1995*, pages 331–339. Elsevier, 1995.
- [140] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355,

- Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL <https://aclanthology.org/W18-5446>.
- [141] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*, 2017.
- [142] Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.
- [143] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.
- [144] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.
- [145] Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, 2016.
- [146] Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie Névéal, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, Karin Verspoor, and Marcos Zampieri. Findings of the 2016 conference on machine

- translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2301. URL <https://aclanthology.org/W16-2301>.
- [147] Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. An evaluation dataset for intent classification and out-of-scope prediction. *arXiv preprint arXiv:1909.02027*, 2019.
- [148] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL <https://aclanthology.org/2020.emnlp-demos.6>.
- [149] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. Building task-oriented dialogue systems for online shopping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.
- [150] Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027, 2020.

- [151] Haode Zhang, Yuwei Zhang, Li-Ming Zhan, Jiaxin Chen, Guangyuan Shi, Xiao-Ming Wu, and Albert Lam. Effectiveness of pre-training for few-shot intent classification. *arXiv preprint arXiv:2109.05782*, 2021.
- [152] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023.
- [153] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.
- [154] Yiwei Jiang, Maarten De Raedt, Johannes Deleu, Thomas Demeester, and Chris Develder. Few-shot out-of-scope intent classification: analyzing the robustness of prompt-based learning. *Applied Intelligence*, pages 1–23, 2024.
- [155] Jianguo Zhang, Kazuma Hashimoto, Yao Wan, Zhiwei Liu, Ye Liu, Caiming Xiong, and S Yu Philip. Are pre-trained transformers robust in intent classification? a missing ingredient in evaluation of out-of-scope intent detection. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 12–20, 2022.
- [156] Jianguo Zhang, Kazuma Hashimoto, Wenhao Liu, Chien-Sheng Wu, Yao Wan, S Yu Philip, Richard Socher, and Caiming Xiong. Discriminative nearest neighbor few-shot intent detection by transferring natural language inference. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5064–5082, 2020.

- [157] Pei Wang, Keqing He, Yutao Mou, Xiaoshuai Song, Yanan Wu, Jingang Wang, Yunsen Xian, Xunliang Cai, and Weiran Xu. APP: Adaptive prototypical pseudo-labeling for few-shot OOD detection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL <https://openreview.net/forum?id=mrD5HN7ZNR>.
- [158] Li-Ming Zhan, Haowen Liang, Lu Fan, Xiao-Ming Wu, and Albert YS Lam. A closer look at few-shot out-of-distribution intent detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 451–460, 2022.
- [159] Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Noisy channel language model prompting for few-shot text classification. *arXiv preprint arXiv:2108.04106*, 2021.
- [160] Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, Robert L Mercer, et al. The mathematics of statistical machine translation: Parameter estimation. 1993.
- [161] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.
- [162] Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Nataraajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.
- [163] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

- [164] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.
- [165] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [166] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- [167] Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*, 2020.
- [168] Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. Short text clustering via convolutional neural networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 62–69, 2015.
- [169] Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 522–530. Springer, 2019.
- [170] Li-Ming Zhan, Bo Liu, Lu Fan, Jiabin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 2345–2354, New York,

- NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885.
doi: 10.1145/3394171.3413761. URL <https://doi.org/10.1145/3394171.3413761>.
- [171] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2018.
- [172] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.
- [173] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.
- [174] Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, April 2019. URL <https://doi.org/10.5281/zenodo.3362844>.
- [175] Paul A. Yushkevich, Joseph Piven, Heather Cody Hazlett, Rachel Gimpel Smith, Sean Ho, James C. Gee, and Guido Gerig. User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability. *Neuroimage*, 31(3):1116–1128, 2006.

- [176] Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*, 2016.
- [177] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022, 2016.
- [178] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *arXiv preprint arXiv:1805.01978*, 2018.
- [179] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*, 2019.
- [180] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795, 2013.
- [181] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering, 2017.
- [182] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543, Doha, Qatar, 2014. ACL.
- [183] Robik Shrestha, Kushal Kafle, and Christopher Kanan. Answer them all! toward universal visual question answering models. In *CVPR*, 2019.

- [184] Liang Peng, Yang Yang, Zheng Wang, Xiao Wu, and Zi Huang. Cra-net: Composed relation attention network for visual question answering. In *ACM MM*, 2019.
- [185] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012.
- [186] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 2017.
- [187] Maithra Raghu, Chiyuan Zhang, Jon M. Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *NeurIPS*, 2019.
- [188] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 2017.
- [189] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. Recurrent convolutional neural networks for text classification. In *AAAI*, 2015.
- [190] Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, and Lawrence Carin. Baseline needs more love: On simple word-embedding-based models and associated pooling mechanisms. In *ACL*, 2018.

- [191] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *ICML*, 2017.
- [192] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway Networks. *arXiv preprint arXiv:1505.00387*, 2015.
- [193] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [194] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017.
- [195] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*. PMLR, 2021.
- [196] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE transactions on neural networks and learning systems*, 2018.
- [197] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *ICCV*, 2017.
- [198] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.
- [199] Xiangyu Zhao, Bo Liu, Qijiong Liu, Guangyuan Shi, and Xiao-Ming Wu. Easygen: Easing multimodal generation with bidiffuser and llms. In *Proceedings of the 62nd*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1351–1370, 2024.

- [200] 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. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- [201] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.
- [202] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL <https://arxiv.org/abs/2305.06500>.
- [203] Jiaqi Cui, Xinyi Zeng, Pinxian Zeng, Bo Liu, Xi Wu, Jiliu Zhou, and Yan Wang. Mcad: Multi-modal conditioned adversarial diffusion model for high-quality pet image reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 467–477. Springer, 2024.
- [204] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-eHub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*, 2023.
- [205] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language

- models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1932–1940, 2024.
- [206] Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611, 2023.
- [207] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *ECCV*, pages 552–567, 2018.
- [208] Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021.
- [209] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.
- [210] Fenglin Liu, Shen Ge, Yuexian Zou, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. *arXiv preprint arXiv:2206.14579*, 2022.
- [211] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilicus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

- [212] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [213] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024.
- [214] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL <https://arxiv.org/abs/2308.12966>.
- [215] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, page ocae045, 2024.
- [216] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.
- [217] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [218] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

- [219] Jiawei Chen, Ding kang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*, 2024.
- [220] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, et al. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*, 2024.