

Crossing the Chasm:
Bridging Perception and Generative Models for
Enhanced Vision-Language Understanding

Ying Jin

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Jenq-Neng Hwang, Chair

Linda Shapiro

Rania Hussein

Program Authorized to Offer Degree:
Electrical and Computer Engineering

©Copyright 2026

Ying Jin

University of Washington

Abstract

Crossing the Chasm:
Bridging Perception and Generative Models for
Enhanced Vision-Language Understanding

Ying Jin

Chair of the Supervisory Committee:
Professor Jenq-Neng Hwang
Electrical and Computer Engineering

Perception and generative models are usually treated as opposing paradigms in machine learning. Perception models, such as classification, segmentation, and detection models, extract structured labels that integrate easily into downstream software, whereas generative models, such as diffusion models and large language models, synthesize data and offer broader semantic coverage through large-scale, weakly-supervised training. Yet this opposition is mainly one of task formulation. Both learn the mapping between images and labels but differ in its direction. In this dissertation, we show that the two paradigms can be complementary rather than disconnected, and bridging them improves vision-language understanding in both directions: generative models provide richer supervision for perception, while perception models bring verifiability and controllability to generation.

To examine the generality of this insight, we study both general and medical domains. For Human-Object Interaction (HOI) recognition, we propose a Heterogeneous Teacher-Student (HTS) framework in which a generative image captioner serves as the teacher and a contrastive-initialized classifier serves as the student. Unlike conventional knowledge distillation, the teacher and student perform different tasks and have complementary strengths. By distilling HOI knowledge from noisy image captions, HTS removes the need for expensive human annotations. Trained with the proposed LogSumExp-Sign loss,

HTS achieves 49.6 mAP on HICO without using ground-truth labels, outperforming fully supervised baselines despite being only a fraction of the teacher’s size. In addition, it enables a zero-shot HOI detector, derived by pairing with an off-the-shelf object detector without further training, that surpasses prior methods.

In the medical domain, we propose DAug for medical image classification and retrieval. Medical image understanding requires meticulous examination of fine, localized abnormalities, and it is challenging for models to learn where to look from limited training data with only image-level supervision, because such supervision does not indicate where abnormalities appear. In DAug, we train a diffusion-based image-to-image translation model to generate abnormality heatmaps, which are appended as additional image channels to direct the model’s attention to clinically significant regions. Combined with a novel Image-Text-Class Hybrid Contrastive loss that leverages both text and class labels, DAug achieves state-of-the-art performance on both medical image classification and retrieval benchmarks.

To improve radiology report generation, we introduce QRad, which reframes report generation as a self-directed Visual Question Answering process (Auto-VQA). Instead of using an image captioning pipeline, we train the model to first generate a chain of questions and then answer each question, concatenating the answers to form the report. The QA training data is converted from the reports. By isolating sentence-level linguistic variation (such as the omission or ordering of medical topics) from each diagnostic statement, this reformulation lets the model focus on factual accuracy rather than presentation style. As a result, QRad outperforms existing baselines while using only 13% of their parameters. Moreover, by casting classification as a single-token question-answering problem for a target finding, QRad equips a generative model with perception-style capabilities, including class-specific confidence estimation and ROC-based evaluation. Building on this property, we further propose a framework that combines an image classifier with a

report generation model to enable threshold-controllable report generation, supporting regulatory validation and allowing clinicians to tune sensitivity–specificity trade-offs for different clinical settings.

Taken together, these results establish bridging perception and generation as a general design principle that unites the open-vocabulary generalizability of generative models with the verifiability and controllability of perception models, yielding vision-language models that are at once more capable and more reliable.

TABLE OF CONTENTS

	Page
List of Figures	iv
List of Tables	v
Chapter 1: Introduction	1
1.1 Perception and Generative Models	1
1.2 From Generation to Perception	2
1.3 From Perception to Generation	4
Chapter 2: Related Work	6
2.1 Human-Object Interaction (HOI) Recognition	6
2.1.1 HOI Classification and Detection	6
2.1.2 Language Models and Image-Text Pre-training	7
2.2 Medical Image Retrieval and Classification	7
2.2.1 Image Retrieval and Classification	7
2.2.2 Generation for Perception	8
2.3 Radiology Report Generation	9
2.3.1 Image Captioning	9
2.3.2 Free-form Radiology Report Generation	10
Chapter 3: HTS: Captioning Models Supervise HOI Perception	12
3.1 Overview	12
3.2 Heterogeneous Teacher-Student Framework	15
3.2.1 Teacher and Student Models	15
3.2.2 Bridging Modality Gaps via Pseudo-Label Filtering	16
3.2.3 Label Aggregation from Multiple Regional Captions	17
3.3 LogSumExp-Sign (LSE-Sign) Loss	18

3.3.1	Dynamic Gradient Re-weighting	18
3.3.2	Controlling the Re-weighting Strength	19
3.3.3	A Unified View of Discriminative and Contrastive Learning	20
3.4	Experiments	21
3.4.1	Datasets and Training Details	21
3.4.2	HOI Classification Results	23
3.4.3	Transferring HOI Classification to Detection	24
3.5	Ablation Studies	25
3.5.1	Comparison of Teacher Models	25
3.5.2	Comparison of Classifier Initialization Methods	25
3.6	Qualitative Analysis	27
3.6.1	Visualizations of Localized Image Features	27
3.6.2	Regional Image Captions and Object Sparsification	27
3.7	Summary	32
Chapter 4:	DAug: Generative Channel Augmentation for Medical Perception . .	33
4.1	Overview	33
4.2	DAug: Diffusion-Based Channel Augmentation	35
4.2.1	Dataset and Label Generation	36
4.2.2	Abnormality Heatmap Generation through Classifier-Guided Diffusion	36
4.2.3	Channel-wise Feature Augmentation	39
4.3	Image-Text-Class Hybrid Contrastive Loss	40
4.3.1	Model Architecture and Loss Formulation	40
4.3.2	Unified Inference for Retrieval and Classification	42
4.4	Experiments	44
4.4.1	Implementation Details	44
4.4.2	Results	44
4.5	Ablation Studies	45
4.6	Summary	45
Chapter 5:	QRad: Combining Generation and Perception in Radiology Report	
	Generation	46
5.1	Overview	46
5.2	MedImageInsight: Contrastive Pre-trained Medical Foundation Model . . .	50

5.2.1	Medical Image-Text Contrastive Pre-training	50
5.2.2	Radiology Report Generation Baselines	51
5.3	Auto-VQA Training Data Generation	53
5.3.1	Report Generation Question-Answer Pairs (Open Vocabulary)	53
5.3.2	Image Classification Question-Answer Pairs (Closed Vocabulary) . .	53
5.4	Auto-VQA Pipeline and QRad Model Architecture	55
5.4.1	Question Generation Module	57
5.4.2	Answer Generation Module	57
5.4.3	Training Recipe	57
5.5	Experiments and Ablation Studies for Report Generation	58
5.5.1	Implementation Details and Evaluation Metrics	58
5.5.2	Radiology Report Generation	59
5.5.3	Ablation of Each Component	60
5.6	Class Confidence Score Extraction	63
5.6.1	Image Classification with a Text Generation Model	64
5.6.2	Confidence Score Calibration	64
5.7	Threshold-Controllable Radiology Report Generation	65
5.7.1	Motivation	65
5.7.2	Method	66
5.7.3	Evaluation Protocols and Results	67
5.8	Additional Results and Discussion	70
5.8.1	UniRG-CXR: Reinforcement Fine-tuning for Report Generation . . .	70
5.8.2	Results on IU X-ray	71
5.8.3	Comparison of Hyperparameter Configurations	72
5.8.4	Qualitative Results	73
5.8.5	ROC-Based Evaluation for Regulatory Validation	75
5.8.6	LLM Instructions for Threshold-Controllable Report Generation . .	76
5.8.7	Ethical Considerations	78
5.9	Summary	79
Chapter 6:	Conclusion	81

LIST OF FIGURES

Figure Number	Page
1.1 The training of perception models <i>vs.</i> generative models	3
3.1 Heterogeneous Teacher-Student (HTS) framework	13
3.2 Architecture of the student model	16
3.3 t-SNE visualization of classifier weights	28
3.4 Visualization of CLS attention maps	29
3.5 Examples of pseudo-labels with and without object sparsification	30
3.6 Examples of regional captions	31
4.1 Diffusion-based channel augmentation (DAug) pipeline	34
4.2 Examples of DAug-generated abnormality heatmaps	38
4.3 Model architecture and Image-Text-Class Hybrid Contrastive loss	40
5.1 Overview of QRad’s Captioning-to-VQA reframing approach	47
5.2 Overview of the MedImageInsight foundation model architecture	51
5.3 Examples of the report generation QA dataset	54
5.4 Examples of the image classification QA dataset	55
5.5 QRad’s Auto-VQA pipeline	56
5.6 Threshold-controlled report generation on four representative classes	69
5.7 Threshold-controlled report generation evaluation	70
5.8 Qualitative examples of QRad’s Auto-VQA process	74
5.9 Qualitative examples of QRad-generated answers using oracle questions	75
5.11 LLM rewriting instruction for threshold-controlled report generation	78
5.10 ROC curves of QRad across disease classes	80

LIST OF TABLES

Table Number	Page
1.1 Comparison of perception and generative models	1
3.1 HOI classification on HICO	22
3.2 HOI classification on MPII	23
3.3 Few-shot performance on HICO	24
3.4 Comparison of teacher models on HICO	26
3.5 Ablations on classifier initialization	26
4.1 Retrieval performance on MIMIC-CXR	42
4.2 Image classification performance on MIMIC-CXR	43
4.3 Ablation studies on retrieval and classification	43
5.1 MedImageInsight pre-training & evaluation datasets	52
5.2 Report generation performance on MIMIC-CXR	60
5.3 Performance on the ReXrank benchmark	61
5.4 Ablation studies on MIMIC-CXR	62
5.5 Ablation of the question generator	63
5.6 Expected Calibration Error before and after calibration	65
5.7 Report generation performance on IU-XRay	72
5.8 Comparison of hyperparameter configurations on MIMIC-CXR	73

ACKNOWLEDGEMENTS

As I set down the last words of this dissertation, the excitement of receiving my Ph.D. admission letter and the long nights of work that followed feel as vivid as ever. Over these years, my understanding of what it means to pursue a Ph.D. has changed in more ways than I expected. It began with following those I admired and gradually deepened into a demanding journey of personal growth. Courage, patience, and rational open-mindedness are qualities I came to understand only when tested at this level. Looking back, I am deeply thankful for this transformative journey and for everyone who walked alongside me.

First and foremost, I thank my advisor, Professor Jenq-Neng Hwang, for opening the door to this journey. He expected the best from his students, accommodated my special scheduling needs, and was always the first to lift us up when things went wrong. The most valuable lesson I learned from him was the importance of attitude and consistency, which has profoundly shaped how I approach challenging goals in life. Beyond research, the night during COVID when our lab gathered at his home to celebrate the Mid-Autumn Festival under the moon remains one of my fondest memories. I extend my heartfelt thanks to Professor Hwang and his family.

I am grateful to my dissertation committee members, Professors Linda Shapiro, Rania Hussein, and Chengcheng Zhu, for their time and steady presence throughout my doctoral study. I would also like to thank Professors Mari Ostendorf and Ming-Ting Sun for serving on my Qualifying Exam committee, and Jennifer Huberman, our academic counselor, for helping me navigate every administrative step along the way. I feel fortunate to have shared the Information Processing Lab with many talented peers: Jiarui Cai, Cheng-Yen (Chris) Yang, Yizhou Wang, Haotian Zhang, Renshu Gu, Jie Mei, Zhongyu Jiang, Hung-

Min Hsu, Jen-Hao (Andy) Cheng, Aotian Zheng, Hsiang-Wei Huang, and Huayu Wang. I also thank the University of Washington and the Graduate School for providing an environment that is both productive and caring. These years have been as lovely as the cherry blossoms that line the campus each March.

I began my career in research at Microsoft before starting my Ph.D., and in many ways Microsoft has been a second school for me. I thank Dr. Lijuan Wang and Dr. Zicheng Liu, my earliest mentors, who had faith in me and guided me toward doctoral study. They are exceptional scientists and the finest mentors I could have hoped for, in research and in life. I also thank Dr. Noel Codella, whose mentorship helped me produce some of my best work. Our collaboration remains one of the most rewarding chapters in my research life. I also want to acknowledge the many colleagues and coauthors at Microsoft who have inspired me with their work and leadership: Jamie Shotton, Dinei Florencio, Yu-Ting Kuo, Jianfeng Wang, Lin Liang, Ke-Yun (Kevin) Lin, Yu (Aiden) Gu, Yanbo Xu, Mu Wei, Baoyuan Wang, Boxin Li, Sheng Zhang, Qianchu (Flora) Liu, Guanghui Qin, Lei Zhang, Yumao Lu, Yinpeng Chen, and Quanzeng You.

Beyond Microsoft, I had the opportunity to work closely with entrepreneurs, venture investors, and hedge funds at the intersection of science and business. These experiences encouraged me to tackle hard research problems and to create something valuable that people want. Along the way, I learned the importance of being audacious, thoughtful, and unpretentious. I am fortunate to have met people who opened my eyes to a broader world: Yifei Huang, Yi Ping (Y.P.) Chan, Charlie Wen, Alan (Lin) Li, Faqiang (Fargo) Sheng, Tielin Feng, Li Li, Yingyu Chen, Qi Lu, Peter Mao, Changtai Liu, Jielin Mo, Gary Norden, and many others.

Finally, I thank my family, my extended family, and my close friends. Life would be entirely different without them, and they mean more to me than I can easily put into words. I feel fortunate to have grown up in a family that values knowledge, character, and wisdom

above all else, which shaped my aspirations early on. I thank Mei Sun and Yuan Li, who were my college advisors and have treated me like their own ever since. I am thankful for Shunshun Yao, whose support has been a constant for me throughout my years at UW. I also thank everyone who has filled these years with warmth: Wenlai Zhang, Wei Zhou, Ruby Zhu, Yimo Zhang, Meilin Lyu, Rachel Rui, Chunying Zhang, Yuxing Qiu, Lei Zhang, and Bo Ning. I feel fortunate to have you in my life.

Chapter 1

INTRODUCTION

1.1 Perception and Generative Models

“What I cannot create, I do not understand.”

— Richard Feynman

Perception and generative models are usually treated as opposing paradigms in machine learning. In computer vision, perception tasks such as image classification and object detection typically rely on class-based annotations. The cost of human annotation limits dataset scale and contributes to long-tail [177] and few-shot [142] challenges, which in turn constrain performance and restrict perception models to labeled classes. In contrast, generative models such as diffusion models [55] and large language models [128] can be trained with self-supervision or weak supervision, allowing them to learn from large-scale web data with broader semantic coverage. However, because their outputs are not naturally tied to explicit classes or structured decisions, they are harder to integrate into downstream software, such as rule-based systems. These differences in supervision and output format give rise to complementary strengths and weaknesses, summarized in Table 1.1.

Table 1.1: Comparison of perception and generative models

	Perception Models	Generative Models
Examples	Image classifiers; object detectors	Diffusion models; language models
Training data	Curated datasets; label-intensive	Web-scale; weak/self-supervision
Output format	Closed-vocabulary; structured	Open-vocabulary; unstructured
Strengths	Easy downstream integration	Natural language interface
Weaknesses	Limited to predefined classes	No explicit classes for integration

This complementarity suggest that the two paradigms should not be viewed in isolation. As illustrated in Figure 1.1, a generative model learns the distribution of the input data X by modeling or reconstructing its samples. Such knowledge is useful not only for synthesis but also for perception, the inverse problem that extracts labels from X . Conversely, structured outputs from perception models have been used to guide both text generation [25] and image generation [29]. These examples suggest that perception and generation are not merely different objectives, but can serve as complementary components of a unified learning framework.

Motivated by this view, the central claim of this dissertation is that bridging perception and generative models is a general and effective design principle for enhancing vision-language understanding. We examine this claim through three contributions spanning both general and medical domains. Together, they show that generation can improve perception by supplying richer supervision than fixed class labels, while perception can improve generation by restoring explicit structure, confidence estimation, and controllability.

1.2 From Generation to Perception

Perception models such as image classifiers and object detectors learn to discriminate among a predefined set of classes. Such discriminative training allows the model to predict probabilities for explicit classes, making downstream integration straightforward. However, these models are prone to overfitting to predefined classes, especially when the dataset is limited in size.

Existing work on perception tasks has benefited from generative modeling. One stream of prior work explores using generative models to augment or synthesize training data [122, 137, 38]. Another stream uses a generative objective as the pre-training task to initialize model weights. For example, BERT [28] and Masked AutoEncoders (MAE) [54] are widely used to pre-train text and image encoders, respectively. Both BERT and MAE reconstruct text or images from partial input and can be considered generative pre-training methods. A key advantage of such generative learning over discriminative learning is that

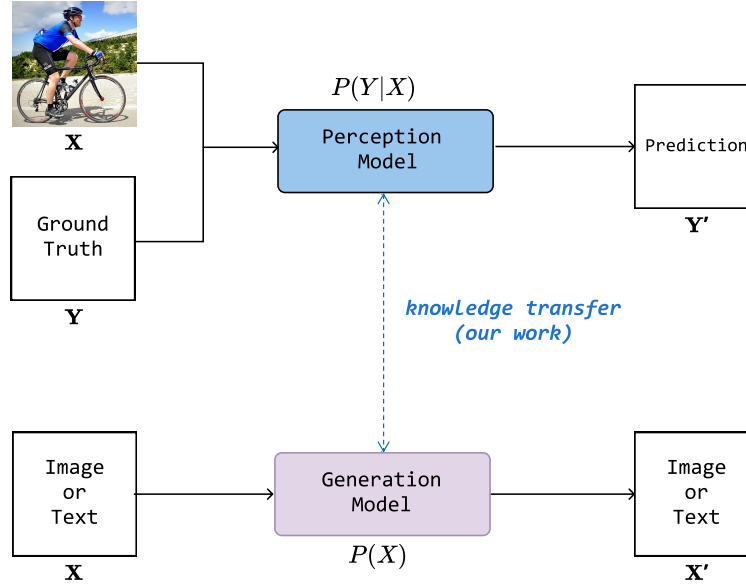


Figure 1.1: The training of perception models *vs.* generative models. A perception model (*e.g.*, an image classifier) requires images X and human-annotated labels Y . Generative models (*e.g.*, large language models) use self-supervised learning, where X alone is largely sufficient, enabling them to leverage large amounts of data from the web. We study how to combine their strengths to improve both perception and generation tasks.

the training signal is not collapsed to the fixed set of classes. As a result, these methods develop a generalizable understanding of text and image semantics via self-supervision and have proven effective in downstream perception tasks. In this dissertation, we propose two new methods for using generative models to enhance perception tasks:

- **Teacher-Student Framework [69, 70].** We propose a Heterogeneous Teacher-Student (HTS) framework in which a generative image captioner and a contrastive-initialized classifier act as a teacher and a student, respectively, with complementary strengths. By distilling knowledge from noisy image captions, HTS enables zero-shot Human-Object Interaction recognition without human annotations and outperforms supervised baselines.
- **Feature Augmentation using Heatmaps [72].** We introduce DAug, a method that uses a classifier-guided diffusion model to generate class-specific abnormality heatmaps and appends them as additional image channels to aid medical image classification and retrieval.

1.3 *From Perception to Generation*

While generative models learn from web-scale data and exhibit strong generalizability, their open-ended outputs lack explicit classes, making it difficult to evaluate their factual correctness in high-stakes scenarios and to integrate them into real-world, rule-based systems. Prior studies have explored using perception models such as image classifiers and object detectors to guide generative models, grounding generated content in structured predictions [25, 159, 29]. These methods shape the behavior of the generative model, but the alignment between the perception model and the generated content is not directly measurable because free-form outputs lack explicit classes. Such limitations hinder the utility of generative models, especially in the medical domain. For example, in existing models for radiology report generation [189, 181, 133], the lack of ROC-based evaluations and threshold control presents obstacles to clinical utility and regulatory approval, such as the FDA review process [37]. In this dissertation, we address these limitations through QRad [71], which enhances radiology report generation by:

- **Framing Classification as Generation.** We unify image classification and report generation into a self-directed Visual Question Answering (Auto-VQA) process, using both report text and class labels to improve factual correctness.
- **Class-Based Confidence Scores.** We enable the extraction of class-based confidence scores from a text generation model, supporting standard sensitivity–specificity evaluations.
- **Threshold Adjustments.** We enable threshold-based adjustments in the text generation process to meet various clinical needs, a capability previously limited to perception models.

The remainder of this dissertation is organized as follows. Chapter 2 surveys related work. Chapter 3 demonstrates how generation improves perception in a general-domain task, where a generative captioner supervises an image classifier without human annotations. Chapter 4 extends this direction to the medical domain, where a diffusion model generates discriminative features to aid image classification and retrieval. Chapter 5

studies the reverse direction, showing how perception-oriented capabilities such as class confidence estimation and threshold control can be incorporated into a report generation model to improve its reliability and clinical utility. Finally, Chapter 6 synthesizes the common principle across all three contributions and discusses future directions.

Chapter 2

RELATED WORK

2.1 Human-Object Interaction (HOI) Recognition

2.1.1 HOI Classification and Detection

HOI recognition includes an image-level classification task and an instance-level detection task. In the classification task, the positions of the HOIs are unknown. Therefore, most existing methods [47, 105, 45, 36] depend on object detectors to extract human and object regions and train with Multiple Instance Learning (MIL) [106]. Representative state-of-the-art methods include PaStaNet [93] and HAKE [92], which infer HOIs by reasoning from body part-level actions, but require additional labels. HOI detection methods fall into three categories: two-stage methods [42, 46, 41, 101, 93, 82, 178], one-stage methods [94, 79, 187, 58, 157], and recent end-to-end methods [80, 190, 145, 176, 31, 126, 188, 63, 81] that follow a DETR-based architecture [11].

A partially zero-shot setting was previously studied under the name “zero-shot HOI recognition”. The classes are split into a *seen* set (with labels) and an *unseen* set (without labels). The unseen classes can be novel combinations of known verbs and object types. The model is trained on the seen set and evaluated on the unseen set. This differs from our zero-shot setting, in which all classes are unseen. For clarity, we call the prior setting “partially zero-shot” and ours “zero-shot”. ReViT [103] studies partially zero-shot HOI classification using auxiliary training objectives and EsViT [89]. Methods for partially zero-shot HOI detection [136, 57, 58] focus on learning features of individual verbs, objects, or object affordances to improve generalization to new HOIs. Later work [156, 95] leverages CLIP [127], which contains HOI knowledge in the image-text pre-training data.

2.1.2 *Language Models and Image-Text Pre-training*

BERT [28] set new benchmarks through unsupervised pre-training with a transformer network. BERT variants [130, 87, 143] optimize the text embedding space to be isotropic for improved performance on text similarity tasks. SimCSE [43] shows that contrastive fine-tuning on BERT, using dropout as augmentation, can achieve state-of-the-art performance.

Recent years have also seen remarkable progress in image-text pre-training. Contrastive and generative models can be pre-trained on millions of image-text pairs. Models such as CLIP [127], ALIGN [67], and Florence [174] apply contrastive learning to align image and text representations, significantly boosting performance on various downstream tasks. Among generative models, GIT [154] achieves state-of-the-art image captioning performance using a contrastively trained image representation. In the HOI scenario, image captioning models have fine-grained knowledge of both verb and object concepts because their pre-training task requires the model to reconstruct each word in the sentence. In contrast, contrastive models are prone to ignore detailed features, *e.g.*, visual relationships and small objects [148].

2.2 *Medical Image Retrieval and Classification*

2.2.1 *Image Retrieval and Classification*

With the rise of transformer models, pre-training on web-scale data has improved performance on downstream tasks such as retrieval and classification. As discussed previously, CLIP [127] aligns image and text representations via contrastive learning. Retrieval methods based on the cosine similarity of CLIP features have become dominant since then. Due to the significant domain gap between medical images and natural images, CLIP shows limited zero-shot performance in the medical domain. A line of work [161, 34] aims to fine-tune CLIP on medical image-text data.

In this dissertation, we address the challenges of medical image retrieval and classification with feature augmentation and an Image-Text-Class Hybrid Contrastive loss.

Related to feature augmentation, prior studies have found that maintaining feature banks of prototypes built from training data is often effective for tasks with limited data, including medical imaging [155, 17]. Recently, X-TRA [150] proposes improving radiology multi-modal retrieval and classification through retrieval-based feature augmentation. It uses a CLIP model to select the top- K most similar samples from the dataset and construct an augmented feature for each input sample. In contrast, DAug leverages generative models for augmentation and outperforms existing methods.

Related to our loss function, UniCL [167] combines image-text and image-class datasets for contrastive pre-training. However, UniCL assumes that each training sample has either a text label or a class label, but not both. This differs from the medical domain, where datasets such as MIMIC-CXR and CheXpert [73, 74, 139] often provide both text and class labels for the same sample. Our Image-Text-Class Hybrid Contrastive loss leverages this structure by combining the image-text contrastive and multi-label classification objectives in a unified cross-entropy formulation. The key idea is to treat image classification as a special case of image-to-class retrieval. This design yields a single model that performs both retrieval and classification and allows the two tasks to reinforce each other during training. As a result, it improves performance and reduces deployment complexity.

2.2.2 *Generation for Perception*

Common image generation methods include Generative Adversarial Networks (GANs) [49], Variational Autoencoders (VAEs) [85], and diffusion models [141]. Diffusion models have become mainstream because of their ease of training and high image quality. During training, a diffusion model learns to remove noise from a noisy input and gradually turn Gaussian noise into images. The denoising steps can be guided by a classifier, whose gradients determine the direction of the denoising process and encourage the output to maximize the probability of a certain class. The result is an image of the chosen class.

Existing work in image understanding has benefited from generative modeling. One

line of work explores using generative models to augment or synthesize training data [122, 137, 38]. Another line uses generative modeling as a pre-training task. For example, Masked AutoEncoders (MAE) [54] reconstruct the original image from a partial input during pre-training and can be fine-tuned on downstream classification tasks. In the medical domain, classifier-guided diffusion models have been used to detect anomalies without bounding box supervision [163]. While this method shows promising performance on simple Chest X-ray findings such as pleural effusion, its reliability is limited when multiple diseases co-exist. Building on [163], we make algorithmic changes that improve its accuracy across a wider range of medical conditions and transform its outputs into discriminative features that guide downstream perception tasks.

2.3 Radiology Report Generation

2.3.1 Image Captioning

Image captioning aims to generate a sentence that describes an image. Recent methods benefit from vision-language pre-training on large-scale web data [16, 32, 153, 83]. These methods commonly use encoder-decoder architectures [90, 154, 110] to support additional vision-language tasks, such as Visual Question Answering.

While radiology report generation commonly follows an image captioning pipeline [24, 152, 165, 171], the task has unique challenges: (1) radiology reports contain multiple sentences covering different medical topics and are much longer than generic image captions, such as those in COCO Captions [96]; (2) factual correctness is critical for radiology reports, requiring close examination of fine visual details; (3) image captioning datasets may provide multiple ground truths per image to capture linguistic variation, but this is usually not available in radiology report datasets.

2.3.2 *Free-form Radiology Report Generation*

Linguistic variations are common in radiology reports. For Chest X-rays, textbook guidelines such as the inside-out order [140] and the ABCDE order (where each letter represents an anatomical region) [102] are available, but they are not strictly followed in practice [9]. For example, medical conditions are sometimes omitted from the report [64]. These valid linguistic variations lead to loss-metric mismatch problems, creating challenges for both training and evaluation [51, 170, 50]. Existing state-of-the-art methods use the original radiology reports as supervision and train models in an image captioning setup, differing primarily in datasets, architectures, and pre-training regimes. Early studies connect a frozen image encoder to a pre-trained language model such as LLaMA [88, 175], while later studies explore mimicking clinical setups [7] and leverage pre-training and fine-tuning techniques [169, 109, 9]. A related line of work uses two-step report generation [114, 98, 166]. Unlike our work, these methods do not supervise the intermediate step with sentence-level topic labels; instead, they rely on unsupervised topics or unordered proxy targets.

Specifically, Liu et al. [98] adopt a hierarchical framework that predicts sentence-level topics as the first step. However, their topic generation module is not supervised with any labels, leaving uncertainty in what the predicted topics represent. Nooralahzadeh et al. [114] first generate high-level context sentences and then refine them into reports. Their first step generates medical keywords per sentence, which are extracted by a text processing model. We differ from them in how this intermediate step is supervised. Yan et al. [166] replace full reports with serialized RadGraph representations (entities and attributes) as supervision, thereby filtering out non-semantic words. However, RadGraph-based supervision still preserves sentence-level style variations, such as omitted findings and different finding orders, and therefore still asks the model to fit these variations. QRad is designed to address this issue by supervising the intermediate generation process with sentence-level topics.

Structured Report Generation [26] takes a different direction by standardizing report formats and organizing content under fixed topics (*e.g.*, lungs, airways, pleura). Following this direction, Rad-ReStruct [124] casts each topic as single- or multi-label classification to enable F1-based evaluation. By contrast, QRad produces free-text reports, where the topic of each sentence is not fixed but is generated dynamically from the input.

Some existing work on Medical Visual Question Answering [116, 184, 60, 134] uses image captioning models to describe the image, allowing a text-only LLM to perform the VQA task. Although QRad also uses a Captioning-to-VQA reframing, its goal is to improve image captioning, with VQA capability as a by-product. Additionally, the Auto-VQA component of our work differs from conventional VQA in that our model learns to predict both the questions and the answers, where the chain of questions specifies the structure of the output text.

Regarding interactive capability, RaDialog [125] adapts a conversational VLM from the general domain for radiology report generation. Compared with QRad, general-domain LLMs and VLMs inherit stronger conversational capabilities, but they show significant gaps in clinical metrics. In the ReXrank leaderboard, QRad outperforms RaDialog by 43.1% in the composite metric.

Chapter 3

HTS: CAPTIONING MODELS SUPERVISE HOI PERCEPTION

3.1 Overview

Human-Object Interaction (HOI) recognition plays a crucial role in scene understanding by identifying all $\langle verb, object \rangle$ pairs that exist in an image. However, annotating HOI is costly due to the large number of possible verb-object combinations (e.g., 600 pairs in HICO [15]) and the need to judge each class individually, since multiple classes can appear simultaneously. Existing methods for HOI recognition rely on human annotations, making it expensive to apply them in new scenarios such as retail and healthcare. To address this issue, we study a zero-shot method for HOI classification and then transfer the classification model to HOI detection.

We introduce a new zero-shot setting in which only the class names and images are provided, with no ground-truth labels available during training. This setting reflects a real-world scenario where users can specify the HOI classes of interest and obtain a customized HOI model without annotation effort. Our zero-shot setting is stricter than existing studies [136, 103] described as zero-shot, since they require labels on a *seen* set of classes (e.g., 480 seen classes and 120 unseen classes). For clarity, we refer to these methods as *partially* zero-shot.

A key challenge in HOI classification is the co-occurrence of multiple humans and objects, which makes it difficult to train a model without knowing the locations of HOIs. Existing methods [47, 105, 45, 36] depend on object and keypoint detections to extract human and object regions and then use Multiple Instance Learning (MIL). For partially zero-shot HOI classification, RelViT [103] proposes auxiliary training objectives based on EsViT [89]. For HOI detection, [136, 57, 58] propose to learn individual verb, object and

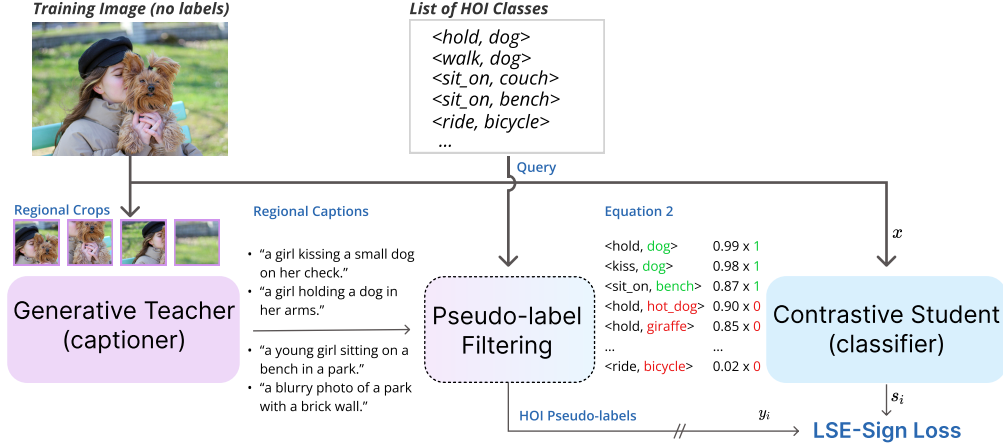


Figure 3.1: **Heterogeneous Teacher-Student framework** for zero-shot HOI classification. Unlike conventional knowledge distillation, our teacher and student perform different tasks (“heterogeneous”) and have complementary strengths. We bridge them by *pseudo-label filtering*, which extracts HOI classes from noisy image captions. Ultimately, the student outperforms the teacher (6 times larger) by 5.5 mAP (+15.4%).

affordance concepts to generalize to new combinations.

Our approach to HOI classification differs from existing methods in that it leverages two types of image-text pre-trained models and does not rely on object and keypoint detectors. Although image-text pre-training has been used in HOI recognition, this direction remains under-explored, with existing work [156, 95] only utilizing CLIP [127], a contrastive pre-trained model, as the distillation source. In contrast, generative pre-trained models such as image captioners have not been used in this field. The proposed method combines the strengths of both types of image-text models: the contrastive model (such as CLIP) has a small backbone, while the generative model (such as an image captioner) has strong discriminability because it is trained to recover each word in the captioning ground truth. We aim to create an HOI classifier that combines these complementary strengths, since CLIP is weak in recognizing visual relationships and small objects [148, 120], while powerful captioning models [154] are much larger to deploy.

To this end, we propose a novel *Heterogeneous Teacher-Student* framework (Figure 3.1)

that incorporates two types of image-text pre-trained models, using GIT [154], a large image captioning model, as the teacher and a CLIP-initialized classifier as the student. Unlike conventional teacher-student methods, our teacher and student perform different tasks and cannot trivially teach each other. Additionally, the teacher is not fine-tuned or supervised on the downstream HOI task, resulting in non-HOI and noisy output. To address these challenges, we introduce *pseudo-label filtering* to extract HOI class probabilities from multiple regional captions per image as the student’s supervision. Since image captions are unconstrained, the process incorporates SimCSE [43] and WordNet [107] to extract HOI knowledge and filter out non-HOI noise. This process is inspired by domain adaptation from web-crawled image-text data to HOI, but it is more challenging because both the image and target domains are different and no ground-truth labels exist.

The HOI classes are highly imbalanced due to the rarity of some combinations of verbs and objects. Therefore, existing methods struggle with rare (few-shot) classes, and while solutions such as balanced data re-sampling and balanced loss functions are available, they are insufficient for multi-label scenarios. For this reason, we propose the LogSumExp-Sign (LSE-Sign) loss, which prioritizes the learning of hard classes through a dynamic gradient re-weighting mechanism optimized for multi-label classification. Experimental results demonstrate significant performance improvements for rare/few-shot classes.

Our zero-shot method achieves strong performance on multiple HOI classification benchmarks, with 49.6 mAP on HICO [15] and 35.2 mAP on MPII [5], outperforming existing supervised methods. Furthermore, the student model outperforms the teacher while being only 14% of the teacher’s size. Our model is trained for HOI classification at the image level and can be extended to HOI detection by pairing with an off-the-shelf object detector without further training. Extensive experiments show the effectiveness of each component as well as the transferability to HOI detection. The main contributions of this work are threefold:

- We introduce a new zero-shot setting for HOI classification and achieve state-of-the-art performance on the public HICO and MPII datasets.

- We demonstrate a novel Heterogeneous Teacher-Student framework that leverages both generative and contrastive image-text pre-trained models and surpasses previous supervised baselines.
- We propose the LSE-Sign loss, which enhances multi-label classification performance on imbalanced HOI datasets.

3.2 *Heterogeneous Teacher-Student Framework*

The motivation for HTS is twofold. First, it leverages the verb and object knowledge in image-text pre-trained models. Second, neither pre-trained model is sufficient for HOI recognition alone, but their strengths are complementary. Our goal is to develop a small student model that combines the strengths of both models.

3.2.1 *Teacher and Student Models*

The teacher is an image captioning model that generates a text sequence τ to describe an image I . We choose GIT [154], a 700-million-parameter model pre-trained on 800 million image-text pairs from the web. The student (see Figure 3.2) is a multi-label image classifier initialized from the image-text contrastive model CLIP [127], which tends to ignore features of visual relationships and small objects [148]. Both GIT and CLIP carry implicit HOI knowledge from image-text pre-training, achieving 35.6 mAP and 25.8 mAP on HICO, respectively. This evaluation is based on the embedding similarity between the model output and HOI prompts (see Section 3.2.2).

Our approach leverages the complementary strengths of both models. The captioning model can recognize fine-grained verb and object concepts since it is trained to reconstruct each word in the caption, while CLIP can initialize a classifier and provide a good starting point to learn from the teacher. HTS bridges the gap between them to combine these strengths.

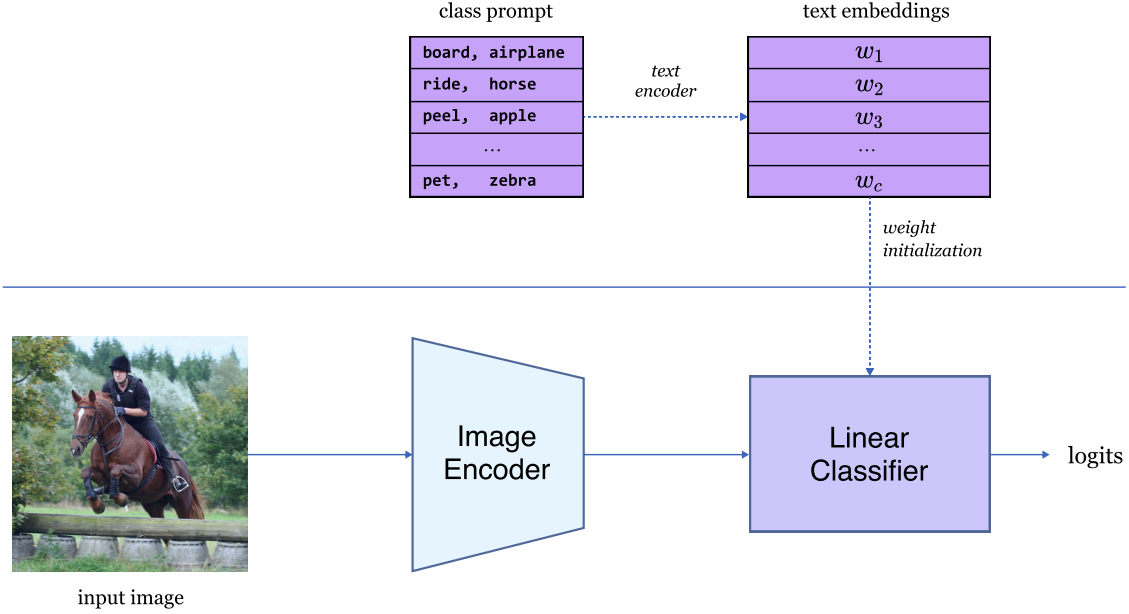


Figure 3.2: **Architecture of the student model.** It has an image encoder and a linear classifier. The weights for the i^{th} class in the linear classifier are initialized with $\mathbf{w}_i$, the text embedding of the corresponding class prompt.

3.2.2 Bridging Modality Gaps via Pseudo-Label Filtering

Our Heterogeneous Teacher-Student framework differs from conventional knowledge distillation in two ways. First, our teacher and student perform different tasks and therefore have different output formats that cannot be directly used for distillation. Second, neither model is supervised on the HOI task, and thus the generated captions are not HOI-centric and can be noisy. To enable learning in this setup, we introduce *pseudo-label filtering* to extract HOI class probabilities from image captions and use them to supervise the student.

In essence, the process queries the probabilities of HOI classes $\mathbf{Y} = \{y_1, y_2, \dots, y_C\}$ from the captions. It starts by using a text-similarity model to evaluate the semantic similarity between the caption and the HOI classes. This approach can handle the vocabulary gap between captions (open vocabulary) and HOI classes. Formally, let τ denote an image caption, and t_i denote the prompt for the i^{th} HOI class (e.g., “a person riding a bicycle” for

<ride, bicycle>). The text is encoded into embeddings using a language model $f(\cdot)$ as:

$$\phi = f(\tau), \quad \mathbf{q}_i = f(t_i). \quad (3.1)$$

Then we estimate HOI pseudo-labels with filtered cosine similarity:

$$P(y_i|\tau) \approx \pi(o_i, o_\tau) \frac{\phi \cdot \mathbf{q}_i}{\|\phi\| \|\mathbf{q}_i\|}, \quad (3.2)$$

where $\pi(o_i, o_\tau)$ represents a filter mask between the object classes in y_i and τ . To extract object classes o_τ from τ , we use a part-of-speech tagging tool¹. π equals 1 if the objects are linked as “synsets” or “hypernyms” in WordNet [107], and 0 otherwise. The filter mask is necessary because text similarity alone reflects semantic similarity rather than the desired visual similarity. For instance, “dog” and “hot dog” are semantically close but visually different, and would become a false positive if not filtered out by π (see Figure 3.1).

The choice of language model $f(\cdot)$ is critical. We use SimCSE [43] for its isotropic embedding space. The resulting pseudo-labels achieve 35.6 mAP on HICO. In comparison, BERT only achieves 1.0 mAP, and the text encoder of CLIP, which shares the same architecture, achieves 17.6 mAP.

3.2.3 Label Aggregation from Multiple Regional Captions

A single caption per image may not cover all HOIs present. This issue can be addressed by generating multiple regional captions per image. For example, the action of “kissing” in Figure 3.1 becomes visible in the second regional crop. We aggregate selected regions using max pooling:

$$P(y_i) = \max(P(y_i|r)), \forall r \in \Phi(R), \quad (3.3)$$

where R denotes all regional crops, and Φ is a sampling function that selects the regions r to include in the max pooling. We find that selecting 10 regions that cover the most detected

¹<https://www.nltk.org/book/ch05.html>

boxes results in the best performance (41.4 mAP with a ViT-B/32 student). Pseudo-labels generated by alternative methods, such as using one caption per image (40.4 mAP) and sampling 10 regions randomly (38.7 mAP), are less complete or less clean. Finally, we normalize P to $[0, 1]$ per image and convert it to binary labels with a threshold of 0.97. The binary labels are used to train the student classifier.

3.3 LogSumExp-Sign (LSE-Sign) Loss

3.3.1 Dynamic Gradient Re-weighting

As a multi-label classification task, HOI classification suffers from severe class imbalance, and existing methods often exhibit a significant performance gap on few-shot subsets (Table 3.3). Although solutions are available for single-label classification, they often assume classes are independent, which does not hold in multi-label scenarios. We propose the LSE-Sign loss for multi-label classification on imbalanced datasets.

The LSE-Sign loss prioritizes the learning of harder classes through a dynamic gradient re-weighting mechanism. These harder classes (indicated by higher loss) usually belong to the few-shot subsets with insufficient training samples. Unlike existing methods such as weighted cross-entropy loss and focal loss, which use fixed hyperparameters to control the re-weighting, our approach is dynamic, as it adjusts on the fly based on the *relative* difficulty of each class over all classes in each sample.

We start from an expected gradient form and then work backward to derive the corresponding loss. Inspired by softmax attention, we adopt Equation (3.5) to redistribute the gradient of each class. Let s_i denote the logit for the i^{th} class, and let the ground-truth label be $y_i \in \{+1, -1\}$ for positive and negative classes, respectively. Let $\mathcal{L}$ denote the loss function. We want:

$$\frac{\partial \mathcal{L}}{\partial s_i} = \frac{-1_{i \in \text{pos}} e^{-s_i} + 1_{i \in \text{neg}} e^{s_i}}{\sum_{j \in \text{pos}} e^{-s_j} + e^0 + \sum_{j \in \text{neg}} e^{s_j}} \quad (3.4)$$

$$= \frac{-y_i e^{-y_i s_i}}{1 + \sum_{j=1}^C e^{-y_j s_j}}, \quad (3.5)$$

where $1_{i \in pos}$ is an indicator function that returns 1 if the i^{th} class is positive and 0 otherwise. The constant 1 or e^0 sets the lower bound of the denominator for numerical stability, and C is the number of classes. Equation (3.5) has a softmax form, which can amplify the gradient for classes whose logits s_i deviate more from their optimal values, effectively increasing their loss. We can then derive the loss function:

$$\mathcal{L} = \sum_i^C \int \frac{\partial \mathcal{L}}{\partial s_i} ds_i \quad (3.6)$$

$$= \log \left(1 + \sum_{i=1}^C e^{-y_i s_i} \right). \quad (3.7)$$

In this formulation, the loss of both positive and negative classes is unified into $e^{-y_i s_i}$, where the label y_i controls the sign. Overall, it is a Log-Sum-Exp (LSE) function, so we name it the LSE-Sign loss. Unlike existing methods that sum the binary cross-entropy loss of each class, the simple modification of placing the sum inside the log enables dynamic gradient re-weighting. This mechanism leads to improved performance on few-shot subsets, as shown in Table 3.3.

3.3.2 Controlling the Re-weighting Strength

In the model, the logit s_i is the dot product of the visual feature $\mathbf{x}$ and the weight $\mathbf{w}$ in the classification head (bias is not used). We can control the strength of the re-weighting by scaling the logits as:

$$s_i = \gamma \frac{\mathbf{x}^\top \mathbf{w}_i}{\|\mathbf{x}\| \|\mathbf{w}_i\|}, \quad (3.8)$$

where the hyperparameter γ acts as the softmax temperature in Equation (3.5). A smaller γ makes the softmax closer to uniform, which minimizes the re-weighting. A larger γ makes the softmax closer to one-hot and maximizes the re-weighting. We can select γ empirically or leave it as a learnable parameter.

3.3.3 *A Unified View of Discriminative and Contrastive Learning*

As shown in Figure 3.2, the student classification model has an image encoder and a linear classifier, a typical architecture for discriminative learning over a fixed set of classes. In such classification models, the logit per class is usually calculated as either $\mathbf{w}_i \mathbf{x} + b$, where b is the optional bias term, or a formulation based on cosine similarity as in Equation (3.8). These two formulations differ in whether the embeddings are normalized, but in both cases the training process encourages the visual feature $\mathbf{x}$ to be close to the weight $\mathbf{w}_i$ of the corresponding class and far from the weights of other classes. Therefore, the weight $\mathbf{w}_i$ can be considered a proxy for the i^{th} class in the feature space, and the training process encourages the model to learn a feature space where samples are clustered around their corresponding class proxies.

In Section 1.2, we discussed that class-based perception models often overfit to pre-defined classes, hindering the model’s generalization. This is because closed-vocabulary classes are overly coarse abstractions of real-world samples and cannot capture the full data spectrum. For example, when flowers of different colors are labeled as the same “flower” class, color captures distinctions among intra-class samples and may be helpful for generalization, but it is not factored into the training loss. As a result, the model is prone to having a collapsed feature space, leading to shortcut learning and poor generalization to unseen samples. A natural solution is to replace the limited number of classes with a comprehensive textual description of the image during training. In this case, a linear classification head is no longer applicable because the classes are now open-vocabulary and cannot be represented by a discrete set of vectors $\mathbf{W}$. One solution is to train a model $\mathbf{w} = f(\text{text})$ that generates the proxy vector $\mathbf{w}$ dynamically given the text description of the image. This model produces the same or similar embeddings for texts with the same or similar meanings, which is the role of a text encoder. This suggests a learning objective that contrasts the visual feature $\mathbf{x}$ with the text embedding $\mathbf{w}$ using cross-entropy loss, which is exactly the contrastive learning objective. From this perspective, class-based

perception models and contrastive models share a unified view, where the former learns a fixed set of class proxies $\mathbf{W}$ and the latter learns an encoder $f(\text{text})$ to generate the proxies. The training objective of both models encourages the visual feature $\mathbf{x}$ to be close to the corresponding proxy $\mathbf{w}$ in the feature space, and far from the other proxies. This unified view provides insight into how to leverage contrastive pre-trained models for discriminative tasks, which is crucial for our student model initialized with CLIP.

In our model, Equation (3.8) is equivalent to using the scaled cosine similarity between $\mathbf{x}$ and $\mathbf{w}_i$ as the logit. Existing methods initialize $\mathbf{W}$ randomly, which ignores the semantic correlations among $\mathbf{w}_i$. We find that for HOI classification, it is crucial to inject class semantics by initializing $\mathbf{w}_i$ with the text embedding of the class it represents. When training the student with ground-truth labels, initializing $\mathbf{W}$ with BERT embeddings leads to a performance boost of 12.8 mAP. The gain increases to 20.1 mAP when the language model is contrastively pre-trained with the image backbone so that their feature spaces are aligned.

Text embeddings are also used in existing work [95, 63, 126] but in different ways, such as using text embeddings as additional features or as a distillation source. In these studies, class proxies $\mathbf{W}$ in the classification head are initialized randomly. In our method, we use text embeddings as weight initialization and update them during training.

3.4 Experiments

3.4.1 Datasets and Training Details

Datasets: We evaluate our approach on two datasets, HICO and MPII. HICO is the largest dataset for HOI classification, consisting of 600 classes that are combinations of 80 object categories (from COCO) and 117 verb categories. Each image has one or multiple labels. HICO contains 38,118 training images and 9,658 testing images. The rare/few-shot subset contains 162 classes that have no more than ten training images. The MPII dataset includes 15,205 training images and 5,708 testing images, with a total of 393 interaction classes. Each

Table 3.1: **HOI classification on HICO**. Classes are divided into 480 non-rare classes and 120 rare classes. Methods use different extra supervision or inputs, including bbox: object detection, pose: human pose, PaSta: part-level action labels, and image-text pairs from the web. RN101 stands for ResNet-101. HTS models are trained without ground-truth labels, and HTS* variants are trained on ground-truth labels for reference. † results are obtained from [103].

Method	Extras	Ground Truth		mAP
		Non-Rare	Rare	
<i>Supervised methods</i>				
R*CNN [47]	Bbox	✓	✓	28.5
Girdhar <i>et al.</i> (RN101) [45]	Pose	✓	✓	34.6
Mallya <i>et al.</i> (RN101) [105]	Bbox	✓	✓	36.1
Pairwise-Part [36]	Bbox, Pose	✓	✓	39.9
PastaNet [93]	Bbox, Pose, PaSta	✓	✓	46.3
HAKE [92]	Bbox, Pose, PaSta	✓	✓	47.1
CLIP-I (ViT-B/32) [127]	Image-text	✓	✓	34.2
HTS* (RN101)	Image-text	✓	✓	<u>53.6</u>
HTS* (ViT-B/32)	Image-text	✓	✓	60.5
<i>Partially zero-shot methods</i>				
VCL [†] [57]		✓		26.7
VCL [†] [57]			✓	21.8
RelViT [103]		✓		37.2
RelViT [103]			✓	23.1
<i>Zero-shot methods</i>				
GIT [154]	Image-text			35.6
CLIP (RN101)	Image-text			25.8
CLIP (ViT-B/32)	Image-text			25.8
CLIP (ViT-L/14)	Image-text			33.8
HTS (RN101)	Image-text			<u>36.5</u>
HTS (ViT-B/32)	Image-text			<u>41.4</u>
HTS (ViT-L/14)	Image-text			49.6

image is labeled with only one interaction. For our zero-shot experiments, we remove the ground-truth labels from the training set and only use the images. Results are evaluated on the test set with ground-truth labels.

Table 3.2: **HOI classification on MPII**. We report performance on the validation set that contains 6,987 images.

Method	Backbone	mAP
<i>Supervised methods</i>		
R*CNN [47]	VGG16	21.7
Girdhar <i>et al.</i> [45]	ResNet101	30.6
Pairwise-Part [36]	ResNet101	32.0
<i>Zero-shot methods</i>		
CLIP [127]	ResNet101	33.7
HTS, ours	ResNet101	35.2

Training: In our main experiment, the student has a ViT-B/32 image backbone connected to a linear classification head. The student is fine-tuned on filtered pseudo-labels with LSE-Sign loss for five epochs. We use the AdamW optimizer and a cosine learning rate scheduler with a base learning rate of $1e^{-5}$. We use an image resolution of 672 and a batch size of 128 on eight V100 GPUs. The teacher model is frozen.

3.4.2 HOI Classification Results

Table 3.1 shows the performance comparison on the HICO dataset. Our best zero-shot model achieves 49.6 mAP, outperforming previous supervised state-of-the-art methods and the image-text pre-trained models. To isolate the gain from CLIP pre-training, we evaluate our model trained with ground-truth labels (HTS*) and compare it with CLIP-I, which has the same pre-trained backbone. HTS* surpasses CLIP-I by a large margin, showing the gain from the LSE-Sign loss and classifier weights initialized with text embeddings. On the MPII dataset, Table 3.2 shows that our method also outperforms supervised baselines.

Table 3.3 presents the few-shot performance on HICO. While the overall performance of HTS (ViT-B/32) is lower than HAKE due to the absence of ground-truth labels, our method outperforms HAKE significantly on the few-shot subsets. We also conduct ablation studies with ground-truth labels (HTS*) to validate the efficacy of the LSE-Sign loss in improving few-shot performance.

Table 3.3: **Few-shot performance** on HICO. Few@i denotes the subset of classes with $\leq i$ training images. Few@1, 5, and 10 contain 49, 125, and 162 classes, respectively. HTS* variants are trained on ground-truth labels to rule out the influence of pseudo-label quality, showing that the LSE-Sign loss outperforms weighted cross-entropy (WBCE) and focal loss, especially on the few-shot subsets.

Method	Loss	mAP	Few@1	Few@5	Few@10
<i>Supervised methods</i>					
Pairwise-Part [36]	WBCE	39.9	13.0	19.8	22.3
PastaNet [93]	WBCE	46.3	24.7	31.8	33.1
HAKE [92]	WBCE	47.1	25.4	32.5	33.7
HTS* (ViT-B/32)	WBCE	54.7	43.2	44.7	44.9
HTS* (ViT-B/32)	Focal	53.2	35.1	38.7	40.0
HTS* (ViT-B/32)	LSE-Sign	60.5	49.6	52.6	52.9
<i>Zero-shot methods</i>					
HTS (ResNet101)	LSE-Sign	36.5	28.9	30.2	31.1
HTS (ViT-B/32)	LSE-Sign	41.4	33.6	36.3	37.2
HTS (ViT-L/14)	LSE-Sign	49.6	41.6	47.1	47.4

3.4.3 Transferring HOI Classification to Detection

We find that a strong HOI classification model trained on image-level labels can effectively aid HOI detection. We connect our frozen classification model with an off-the-shelf object detector to recognize regional HOIs. The two models are trained separately. When combined, they achieve state-of-the-art performance *without* additional fine-tuning on instance-level supervision.

To connect the classifier with the detector, we convert the detected bounding boxes to self-attention masks, Φ . We modify the last transformer layer in the ViT backbone to consume Φ so that the CLS token attends only to the specified region of interest:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\Phi + \frac{QK^T}{\sqrt{d_k}}\right)V \quad (3.9)$$

Equation (3.9) is the attention function [151] of a transformer layer. Φ is a binary mask converted from the bounding boxes. $\Phi_{i,j}$ equals $-\infty$ if i is the CLS token and j is a patch outside the given bounding boxes, and 0 otherwise. d_k is the dimension of Q , K , and V .

Therefore, the CLS token in the last layer attends only to the region of interest specified by the pair of human and object bounding boxes. By doing so, we treat HOI detection as a special case of HOI classification, where the model classifies only the regional HOIs. This pipeline requires no training on instance-level HOI annotations, enabled by the localized, high-quality visual features (see Figure 3.4).

Results: We use an off-the-shelf object detector [41] trained on COCO [97] and multiply the HOI probabilities by the object probability. Our method achieves 11.6 mAP without additional training, outperforming existing state-of-the-art approaches ATL [58] (5.1 mAP) and GEN-VLKT_s [95] (10.5 mAP) in the partially zero-shot setting.

3.5 Ablation Studies

3.5.1 Comparison of Teacher Models

To study the influence of teacher models, we use UFO-B/32 and UFO-L/16 from [153] to replace GIT. UFO-B/32, UFO-L/16, and GIT achieve captioning performance of 122.8, 131.2, and 144.8 CIDEr on COCO [97], respectively. Table 3.4 reveals that the filtering process provides consistent improvements to the pseudo-labels by up to 5.3 mAP. Additionally, the student continues to improve with stronger teachers. It consistently matches or surpasses the performance of its teacher while being much smaller.

3.5.2 Comparison of Classifier Initialization Methods

We use ground-truth labels to ablate the classifier initialization method. In this experiment, the classifier is a single linear layer classification head, $\hat{y} = \mathbf{w}\mathbf{x} + b$, where input $\mathbf{x}$ is the visual feature, output $\hat{y}$ is the logits vector, and $\mathbf{w}$ and b (optional) are the learnable weight and bias parameters, respectively.

To initialize $\mathbf{w}$ and b , we use the standard random initialization implementation [48] as the baseline: $\mathbf{w}, b \sim \mathcal{U}(-\sqrt{k}, \sqrt{k})$, where $k = \text{num_features}^{-1}$. Our proposed method

Table 3.4: **Comparison of teacher models** on HICO. *Cosine similarity*: SimCSE-based cosine similarity used as pseudo-labels without filtering or normalization. *Our approach*: pseudo-labels filtered and normalized as described in Section 3.2. *Student model*: the student after learning from pseudo-labels. Only one caption per image is used for this controlled experiment.

Input	UFO-B/32	UFO-L/16	GIT
Pseudo-label: cosine similarity	25.9	29.6	35.6
Pseudo-label: our approach	28.2	32.7	40.9
Student model: ViT-B/32	35.2	38.4	40.4

Table 3.5: **Ablations on classifier initialization**. Numbers are mAP on the HICO dataset. *CLS1K*, *CLS21K*, *CLIP*: the pre-training of the visual backbone using ImageNet-1K classification, ImageNet-21K classification, and CLIP, respectively. For linear classifier weights, *Random*: the conventional random initialization. *Embedding*: text embedding initialization using BERT or CLIP’s text encoder. The backbone is ViT-B/32 and fine-tuned with LSE-Sign loss using ground-truth class labels.

Pre-training	Classifier Initialization		
	Random Initialization	BERT Embedding	CLIP Embedding
CLS1K	44.1	53.5 _(+9.4)	54.7 _(+10.6)
CLS21K	44.2	53.9 _(+9.7)	55.1 _(+10.9)
CLIP	36.8	51.0 _(+14.2)	60.5 _(+23.7)

uses text embeddings to initialize $\mathbf{w}$. Specifically, for the row vector $\mathbf{w}_i$ and scalar $\mathbf{b}_i$ corresponding to the i^{th} class:

$$\mathbf{e}_i = f(\tau_i) \quad (3.10)$$

$$\mathbf{w}_i = \frac{\mathbf{e}_i}{\|\mathbf{e}_i\|}, \quad \mathbf{b}_i = 0 \quad (3.11)$$

$f(\cdot)$ is a transformer-based text encoder such as BERT [28] or CLIP [127]. We compare both text-only pre-trained BERT and image-text pre-trained CLIP to show the generalizability of our method. τ_i is the input text converted from the i^{th} class label, for example, “a person riding a bicycle” for $\langle \text{ride}, \text{bicycle} \rangle$ and “a person and a car” for $\langle \text{no_interaction}, \text{car} \rangle$. The

output CLS token from $f(\cdot)$ encodes the whole text sequence and is used as the semantic embedding $\mathbf{e}_i$ for an HOI class. $\mathbf{e}_i$ is normalized to the same magnitude. As shown in Table 3.5, both BERT and CLIP embeddings provide significant performance gains over random initialization, with CLIP embeddings achieving the best performance. This is likely because CLIP’s image-text pre-training allows it to learn more visually grounded semantic representations for HOI classes, which are more effective for initializing the classifier weights. Empirically, the t-SNE visualization of the classifier weights in Figure 3.3 shows that CLIP embedding initialization leads to a better clustered structure of class proxies after fine-tuning, which likely contributes to the improved performance.

3.6 Qualitative Analysis

3.6.1 Visualizations of Localized Image Features

We visualize the attention map of the CLS token in the last transformer layer of the student model. As shown in Figure 3.4, the attention map of the CLIP pre-trained backbone without fine-tuning on HICO is more diffused and less focused on the HOI-related objects. After fine-tuning with our method, the attention map becomes more concentrated on the relevant objects and less active elsewhere. This indicates that our method effectively learns to focus on the important regions for HOI classification, which also benefits HOI detection when transferred to that task.

3.6.2 Regional Image Captions and Object Sparsification

Figure 3.5 provides examples of the image captions. It also shows that object sparsification using WordNet effectively reduces the noisy labels coming from semantically related objects, helping ground-truth HOIs surface to the top.

To map the object types from the caption (open-vocabulary) to predefined object categories in the dataset, we use the following information from WordNet. Take “*bicycle*” for example:

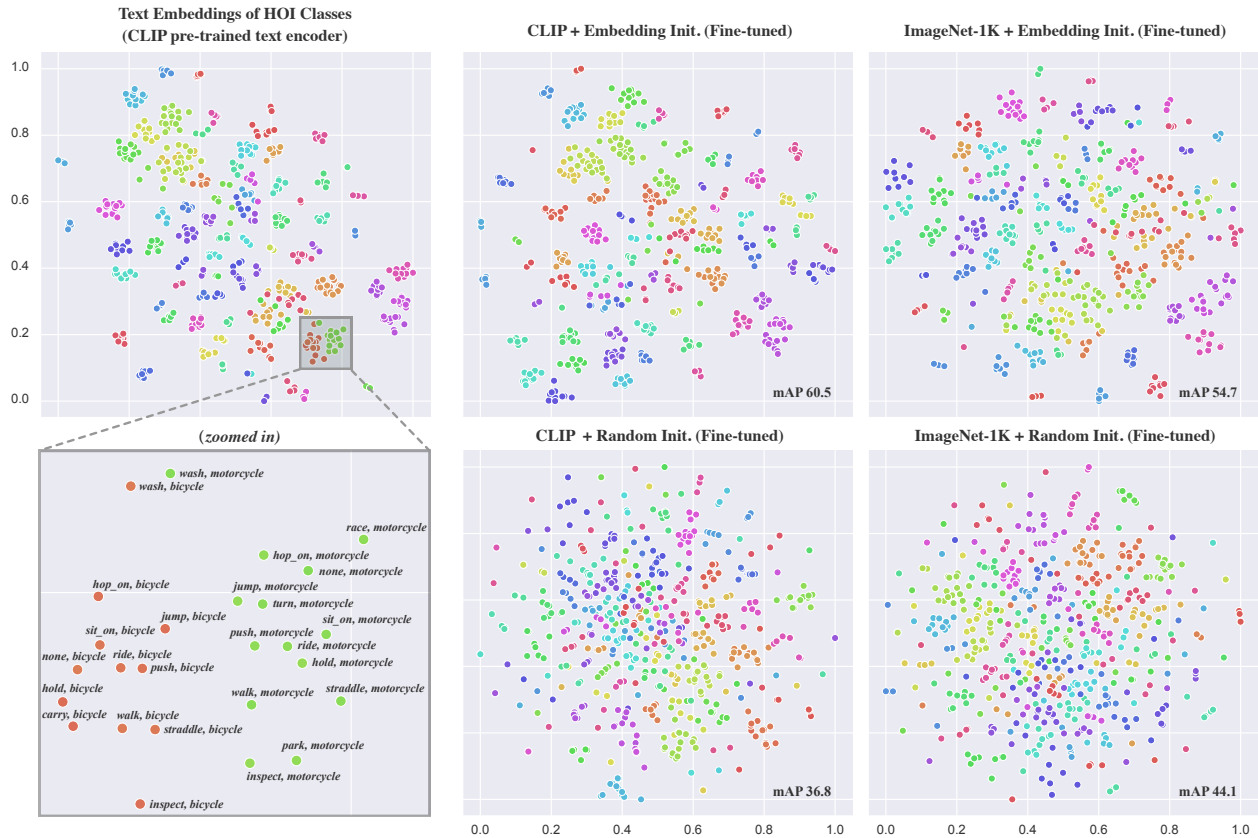


Figure 3.3: **t-SNE visualization of classifier weights.** **Left Column:** CLIP text embeddings of class prompts. **Middle Column:** the weight vectors (proxies) of 600 classes after fine-tuning, using a CLIP pre-trained ViT-B/32 backbone. The top and bottom use text embedding initialization and random initialization, respectively. **Right Column:** the weight vectors after fine-tuning, using an ImageNet-1K pre-trained ViT-B/32 backbone. With text embedding initialization, the clustered structure largely persists after fine-tuning, whereas it is not clearly learned when fine-tuned with random initialization, as reflected in lower model performance (see mAP results in the plots). Fine-tuning is performed with ground-truth labels.

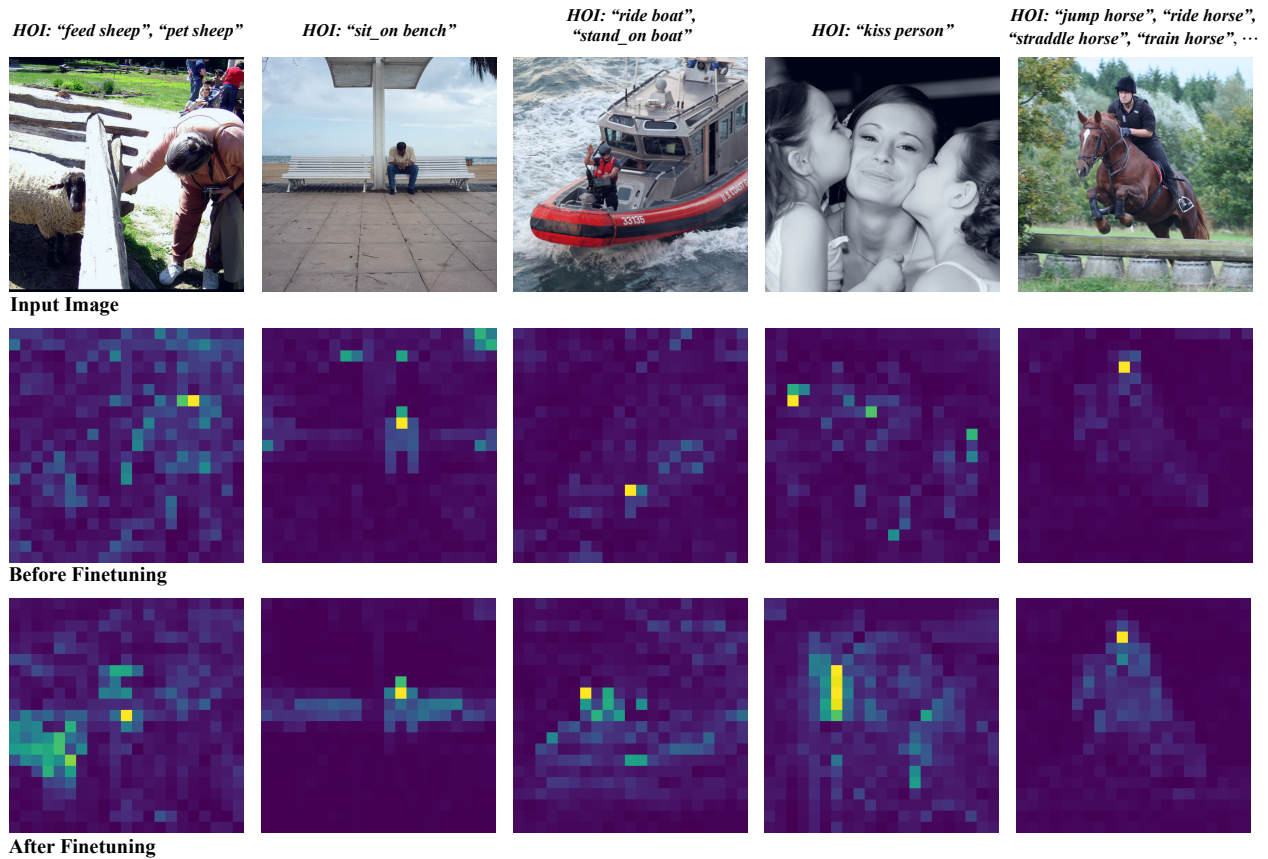
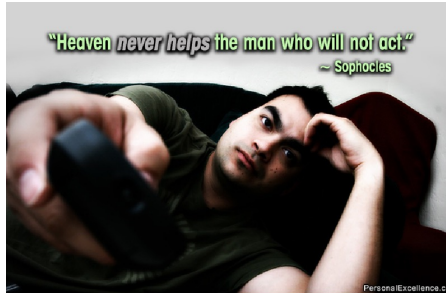


Figure 3.4: **Visualization of CLS attention maps.** First row: the input image. Second row: the attention map of the CLIP pre-trained backbone without fine-tuning on HICO. Third row: the attention map of HTS fine-tuned on HICO. After fine-tuning, attention is more concentrated on HOI-related objects and less active elsewhere. The HOI features learned in the classification task effectively aid HOI detection.

- **Synsets** that provide synonyms, including *bike, cycle*
- **Hyponyms** that include variations of the object, such as *tandem, E-bike*
- **Derived Forms**, for example, *biker, cyclist, bicyclist*

Figure 3.6 compares using a single caption for the whole image *vs.* aggregating HOIs from multiple regional image captions. When the interacting object is small, the captioning model tends to ignore it in the output. However, we can recover the missing object by

**Image Caption:**

a man laying on a couch holding a remote control.

Ground-truth HOI:

<hold, remote>, <point, remote>

**Pseudo-HOI
(w/o sparsification):**

×	1.00	lie_on, couch
×	0.97	none, couch
✓	0.88	hold, remote
×	0.88	sit_on, couch
×	0.87	none, remote
×	0.80	carry, couch
×	0.79	lie_on, chair
×	0.77	control, tv
×	0.76	none, tv
×	0.73	none, bed
×	0.72	swing, remote
×	0.69	none, chair
×	0.68	lie_on, bed
✓	0.67	point, remote

**Pseudo-HOI
(w/ sparsification):**

×	1.00	lie_on, couch
×	0.97	none, couch
✓	0.88	hold, remote
×	0.88	sit_on, couch
×	0.87	none, remote
×	0.80	carry, couch
×	0.72	swing, remote
✓	0.67	point, remote

(a)

**Image Caption:**

a man in a white hat and glasses with a dog in his lap.

Ground-truth HOI:

<hold, dog>, <hug, dog>

**Pseudo-HOI
(w/o sparsification):**

×	1.00	none, dog
✓	0.98	hold, dog
×	0.88	straddle, dog
×	0.84	pet, dog
×	0.83	carry, dog
✓	0.82	hug, dog
×	0.81	inspect, dog
×	0.76	kiss, dog
×	0.75	walk, dog
×	0.72	feed, dog
×	0.72	none, giraffe
×	0.72	hold, hot_dog
×	0.71	groom, dog
×	0.70	none, hot_dog

**Pseudo-HOI
(w/ sparsification):**

×	1.00	none, dog
✓	0.98	hold, dog
×	0.88	straddle, dog
×	0.84	pet, dog
×	0.83	carry, dog
✓	0.82	hug, dog
×	0.81	inspect, dog
×	0.76	kiss, dog
×	0.75	walk, dog
×	0.72	feed, dog
×	0.71	groom, dog

(b)

Figure 3.5: **Examples of pseudo-labels** generated with and without object sparsification. Only top-ranking classes are shown. In both images, the ground-truth HOI classes (highlighted in grey) rank near the top of the 600 classes, showing the effectiveness of the captioning model as a teacher. However, some noisy labels involve objects that are semantically related but visually distant from the ground truth (highlighted in red). We effectively reduce such noise by sparsifying the objects.


Ground-truth HOI:

<hold, book>, <open, book>, <read, book>


Image Caption:

a man and a woman standing in front of a subway.

Regional Image Caption:

a person holding a book and a backpack.

(a)


Ground-truth HOI:

<wear, tie>


Image Caption:

a woman and two children are standing next to each other.

Regional Image Caption:

a boy looking at a woman.

Regional Image Caption:

a boy wearing a tie and tie.

Regional Image Caption:

a person standing outside.

(b)

Figure 3.6: **Examples of regional captions.** (a) shows that the caption for the whole image may not reveal the HOI information. The whole-image caption does not mention the *book*, which is an object category in our HOI list, but a regional crop can recover it. (b) shows that, when using regional captions, the sampling strategy is important for finding the HOI-related region (underlined) while filtering out irrelevant regions. Errors in the captions reflect the original output from the GIT model.

generating multiple regional captions. We then use object-guided sampling to select regions that cover more objects of interest and max-pooling to aggregate regional pseudo-labels.

3.7 *Summary*

This chapter demonstrates a generation-to-perception knowledge distillation approach for bridging open-vocabulary text generation models and closed-vocabulary image classifiers. Specifically, we propose a Heterogeneous Teacher-Student (HTS) framework for learning HOI recognition without ground-truth annotations. HTS achieves clear performance gains by combining the strengths of both generative and contrastive image-text models, along with the LSE-Sign loss. The generative teacher in HTS serves as a knowledge proxy for web-scale image-text data that contains HOI-related knowledge. The pseudo-label filtering process effectively distills HOI-related information into the label space, helping the classification model outperform multiple supervised baselines and transfer to HOI detection without further training.

Chapter 4

DAUG: GENERATIVE CHANNEL AUGMENTATION FOR MEDICAL PERCEPTION

4.1 Overview

Radiologist shortages and burnout are significant problems worldwide and pose risks to patient care [10, 131]. Since training a radiologist takes thirteen to fifteen years, AI models that assist diagnostic workflows offer a scalable way to help address this bottleneck. Among these models, classification and retrieval models are fundamental: classification models can provide a second opinion for clinicians, while retrieval models support comparisons with historical cases for more accurate diagnoses.

Vision models pre-trained on large datasets provide a strong baseline for understanding common natural images. However, when applied to medical images, their capabilities are under-explored for two reasons. First, medical images require meticulous examination of fine details, which differs from natural images where salient objects are often large and have clear boundaries. Second, radiology images such as chest X-rays are monochrome, limiting the use of all three color channels in standard pre-trained vision models. Consequently, vision models trained on limited radiology images struggle to focus on the correct regions for accurate understanding. In this study, we tackle this issue by augmenting image features with abnormality heatmaps as additional channels alongside the original monochrome medical image. The augmented features provide additional information to the model about areas requiring extra attention, mimicking the expertise of experienced radiologists. Our method, dubbed DAUG, is applicable to a wide range of model architectures and leverages their native compatibility with multi-channel (RGB) images.

Specifically, we repurpose a diffusion-based image generation model to generate at-

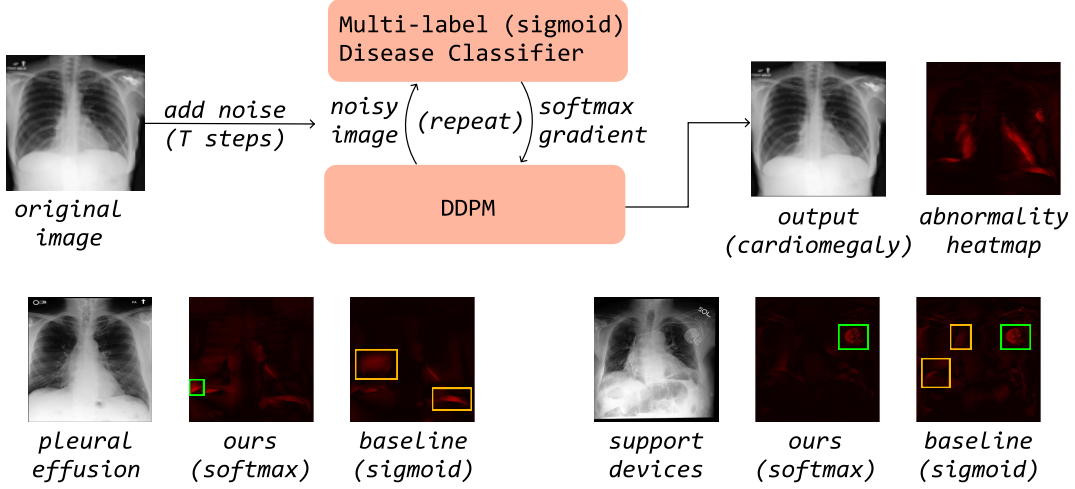


Figure 4.1: Diffusion-based channel augmentation (DAug) pipeline. The original image is translated into diseased versions (or vice versa) with a classifier-guided diffusion model. The differences between the input and output images yield an attention heatmap highlighting the potential region affected by the corresponding disease. The heatmaps are added to the original monochrome radiology image as additional image channels, resulting in augmented features that improve the performance on downstream tasks. The DAug features support multiple disease categories (two examples in the second row), and our softmax-based approach generates more accurate heatmaps than the existing baseline. Green and orange bounding boxes indicate correctly and incorrectly highlighted regions, respectively.

tention heatmaps for predefined disease classes. Each heatmap emphasizes the region potentially affected by the disease. As an additional feature channel, the heatmap explicitly directs the model’s attention to clinically significant regions, a skill difficult to acquire through conventional training. The augmentation module is implemented with a classifier-guided diffusion model [29]. We add Gaussian noise to a chest X-ray image, and the diffusion model, guided by the disease classifier, removes the noise in a direction that either mitigates or exacerbates a disease. The differences between the input and output images yield an attention heatmap that highlights the potential disease region. This generative learning process provides knowledge beyond what is typically learned by classification or retrieval models and improves the performance of these tasks. Because the

heatmaps are produced through generative learning, our method does not require human-annotated heatmaps. Our diffusion-based heatmap generation is inspired by [163], and we improve the method to handle the co-occurrence of multiple diseases and to generate cleaner and more accurate outputs.

In addition, motivated by the synergy of jointly learning the retrieval and classification tasks, we design an Image-Text-Class Hybrid Contrastive learning criterion. During training, the loss function contrasts both image-text and image-class pairs to leverage both text and class labels. This method results in a single model that supports both classification and retrieval tasks, making real-world deployment easier.

To validate the performance gains, we combine both methods and evaluate on the largest Chest X-ray dataset, MIMIC-CXR [73]. Our model outperforms existing state-of-the-art methods on both retrieval and classification tasks.

In summary, the contributions of this work are threefold:

- We propose DAug, a portable feature augmentation method that improves medical image understanding performance by adding abnormality heatmaps as additional image channels.
- We introduce Image-Text-Class Hybrid Contrastive learning, which leverages both image-text and image-class labels to improve performance on both retrieval and classification tasks.
- We deliver a single model that performs both retrieval and classification tasks with state-of-the-art performance. The proposed methods can be applied to standard pre-trained models (such as CLIP ViT [127]), making real-world deployment easier.

4.2 DAug: Diffusion-Based Channel Augmentation

We aim to use generative models to provide helpful visual cues for medical image perception tasks. In this section, we introduce the method for training a diffusion-based generative model to create heatmaps and using these heatmaps as additional input features.

4.2.1 Dataset and Label Generation

For experiments on both retrieval and classification tasks, we use MIMIC-CXR [73], the largest Chest X-ray image-text dataset. As class labels are unavailable, we generate pseudo-labels by converting the text to 14 disease classes with the VisualCheXbert labeler [66], a text classification model.

4.2.2 Abnormality Heatmap Generation through Classifier-Guided Diffusion

Our work is built on [163], which shows promising results on visually salient diseases such as pleural effusion. We extend this method to multiple diseases and improve its reliability for our scenario. First, we train a Denoising Diffusion Probabilistic Model (DDPM) [55] on Chest X-ray images. Given an input image I , we generate a series of noisy images $\{x_0, x_1, \dots, x_T\}$ by gradually adding Gaussian noise for T steps. x_0 is the original image with no noise and $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. A U-Net model is trained to reverse the process by estimating the noise added at any timestep. During evaluation, the model can recover the original image x_0 by removing the estimated noise for T steps from Gaussian noise x_T . We set $T = 1000$.

To translate a radiology image into its healthy or diseased variant, the generated image must retain the same anatomic structure as the input image. For this purpose, following [163], we add noise for only 500 steps to the original input, resulting in x_{500} , where the anatomic structures are still visible. The diffusion model is used to recover the original input from x_{500} . To produce a diseased or healthy version of the input, the denoising process is guided by an image classifier that maps an X-ray to predefined disease classes. The gradients for a selected class are used to condition the denoising process, making the output image maximize or minimize the probability of the class in the classifier.

Several challenges remain unsolved in existing work [163]. For Chest X-ray images, each image may contain multiple diseases, making the binary classifier from prior work unsuitable. In our evaluation, we find that particularly challenging cases involve diseases

that often co-occur (*e.g.*, pleural effusion and lung opacity). In such instances, the generated heatmap highlights areas for both diseases without accurately distinguishing them. This problem arises because the image classifier cannot differentiate diseases that frequently co-occur, where the presence of one disease serves as a strong bias for the presence of the other. To address these challenges, we propose the following improvements:

First, we replace disease classes with disease super-classes. Instead of training the classifier on the common CheXbert classes, we group them into 7 super-classes based on medical knowledge. For example, Consolidation, Edema, and Pneumonia are grouped into one class, as edema and pneumonia are child nodes of consolidation. This approach allows the classifier to focus on distinguishing visual features rather than the underlying causes. The reason is twofold: it eliminates ambiguity in distinguishing between a parent disease and its child diseases, and it acknowledges that some diseases share similar visual features but differ in their causes.

Second, we alter the activation function from training to testing. Specifically, we use sigmoid activation for training and softmax for testing. Sigmoid is required during training to support the multi-label setup. During inference, however, the generated outputs guided by sigmoid gradients highlight false positive regions. This issue can be intuitively explained with an example. Suppose that we are generating a diseased version of the input image where the disease class is cardiomegaly. We want the generated output to contain only cardiomegaly so that the difference heatmap will accurately highlight this particular disease. During the denoising process, we update the noisy image in a direction that increases the sigmoid probability z_i for cardiomegaly. As the sigmoid equation Equation (4.1a) considers only one class, the generated output is not guaranteed to be free of other diseases (*i.e.*, minimized $z_j, j \neq i$). In fact, as some diseases tend to co-occur (*e.g.*, cardiomegaly and pleural effusion), the generated image tends to have both diseases, which introduces false positives. Using softmax as the activation during inference solves this issue because Equation (4.1b) maximizes z_i while minimizing the probability of other classes ($z_j, j \neq i$). Intuitively, by changing to the softmax activation during testing,

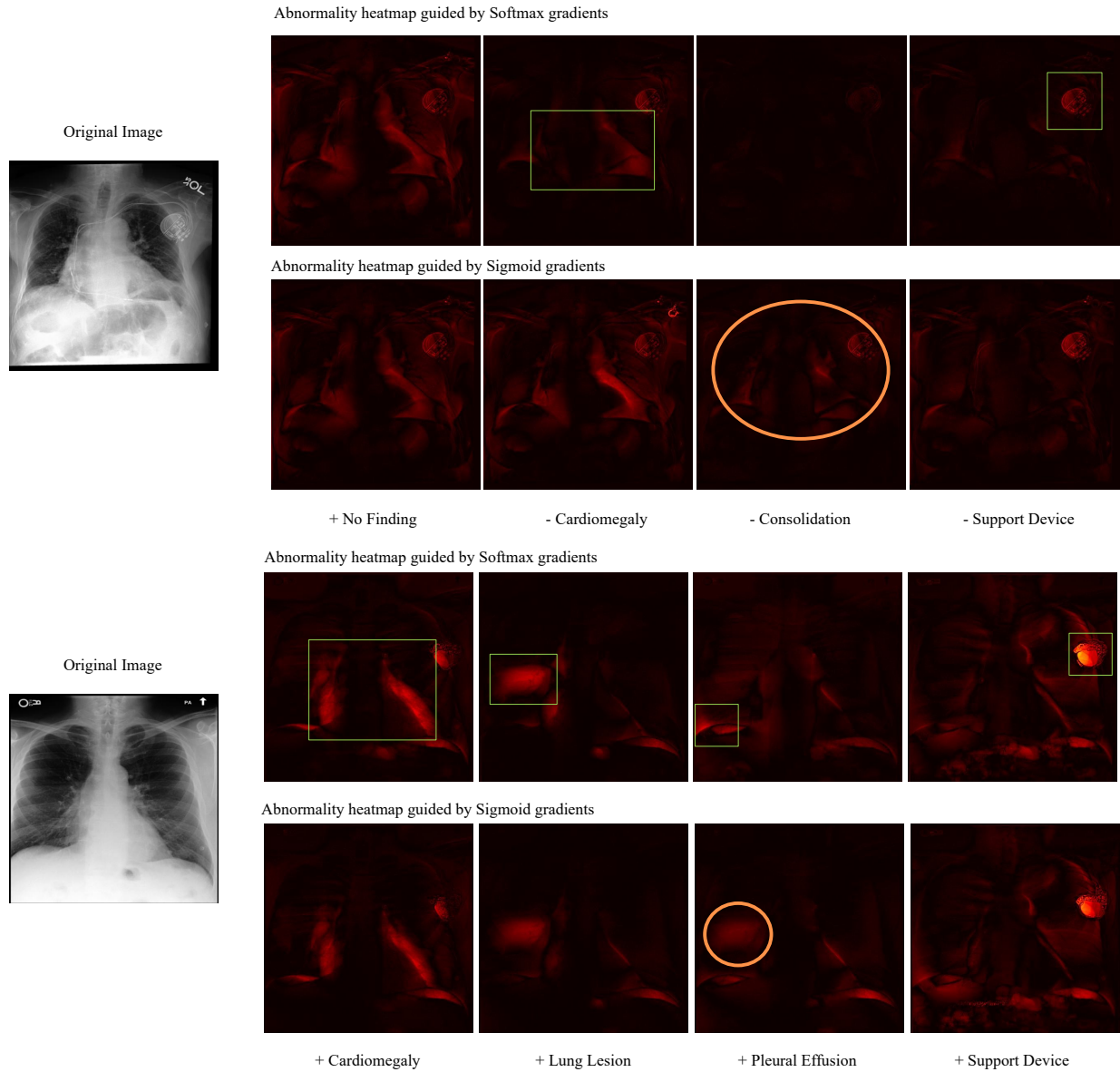


Figure 4.2: Examples of DAug-generated abnormality heatmaps. Each of the two examples covers four abnormalities (*e.g.*, cardiomegaly, consolidation). The **plus** (+) and **minus** (−) signs indicate the direction of the classifier gradient, corresponding to amplifying and reducing the disease, respectively. “+ No Finding” reduces the probability of all potential diseases. For each input, the **first row** shows the output heatmap guided by gradients of **softmax** probabilities (ours), and the **second row** shows the **sigmoid** counterpart used in existing work. The **green bounding boxes** show that our method correctly highlights the disease region. Moreover, softmax gradients provide better guidance than sigmoid gradients because softmax successfully removes false positives (see orange circles). The **orange circle** in the second row highlights false positive areas of consolidation, and the orange circle in the last row highlights incorrect activation of lung lesion when detecting pleural effusion. The corresponding softmax version (green box) makes correct detections.

we transform the task from “generating an image where the probability of cardiomegaly is maximized” to “generating an image where the probability of cardiomegaly, compared to other diseases, is maximized”. We validate the effectiveness with qualitative comparisons, as shown in Figure 4.1 (bottom row) and Figure 4.2.

$$\text{sigmoid}(z_i) = \frac{1}{1 + e^{-z_i}} \quad (4.1a)$$

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} = \frac{1}{\sum_{j=1}^K e^{z_j - z_i}} \quad (4.1b)$$

4.2.3 Channel-wise Feature Augmentation

The generated heatmaps can be integrated into the downstream perception model in various ways. For example, one approach is to pass them through a vision backbone and concatenate their embeddings with the image embedding for downstream classification. This vanilla approach requires modifications to the model architecture. We propose an alternative, channel-wise feature augmentation method that incorporates the heatmaps as additional image channels alongside the medical image.

This channel-wise feature augmentation offers two main advantages. First, it leverages the capabilities of powerful pre-trained vision models designed for RGB three-channel input. As radiology images are typically monochrome, the remaining input channels are under-utilized. Second, channel-wise augmentation does not require any changes to the model architecture, making it easier to use a wide range of pre-trained transformers. This portable approach enhances downstream performance without requiring architectural modifications.

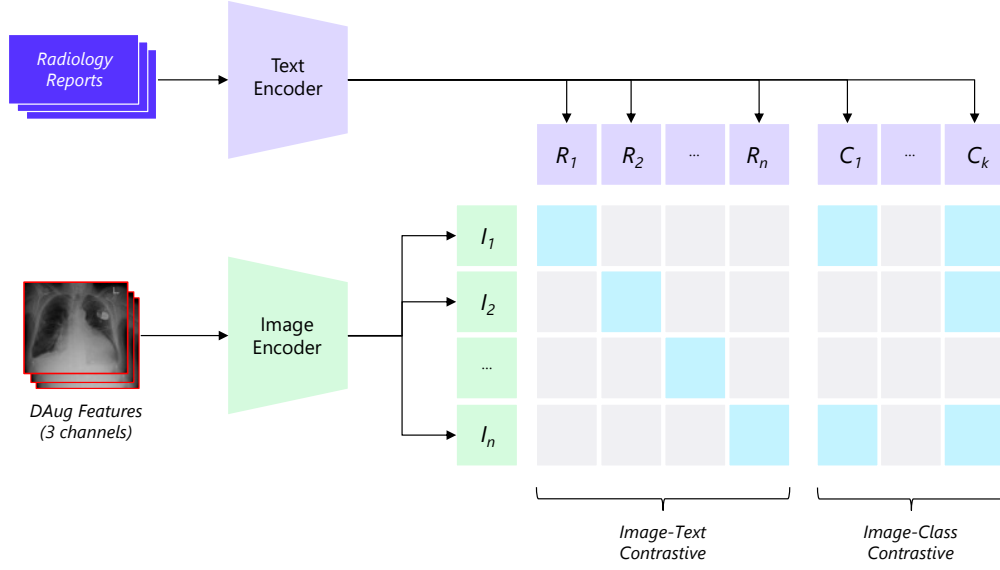


Figure 4.3: Model architecture and Image-Text-Class Hybrid Contrastive Loss. The inputs are pairs of radiology reports and DAug features (3-channel images with the abnormality heatmaps as the third channel). The image and text encoders are pre-trained with CLIP. R and C are text embeddings for the reports and class prompts, respectively. The hybrid contrastive loss includes both the image-text CLIP loss and image-class cross-entropy loss. Blue cells are positive pairs, where the image-class matchings are derived from ground-truth class labels.

4.3 Image-Text-Class Hybrid Contrastive Loss

4.3.1 Model Architecture and Loss Formulation

In this section, we introduce our model architecture and training process. Although image-to-text retrieval and image classification are commonly treated as different tasks, they share the same underlying formulation: comparing an image representation with target representations in an embedding space. In classification, a linear head computes logits as $\mathbf{w}_i \mathbf{x} + b$, where $\mathbf{w}_i$ is the embedding of the i^{th} class, and $\mathbf{w}_i \mathbf{x}$ is their scaled cosine similarity. In image-to-text retrieval, the targets are not fixed classifier weights but text embeddings of candidate reports. From this perspective, classification can be viewed as retrieval over a closed set of discrete classes. This shared formulation motivates us to train retrieval and

classification jointly, with the goal of improving each task through shared supervision. Unlike existing studies that use either image-text or image-class supervision [78, 167], our approach integrates both types of supervision for the same image into a unified loss. We call this objective the Image-Text-Class Hybrid Contrastive Loss.

As illustrated in Figure 4.3, we first convert the classes into a set of text prompts. For instance, the class “cardiomegaly” is converted to “A photo of a Chest X-ray image with cardiomegaly”. During training, as depicted in Figure 4.3, class prompts and radiology reports are both encoded into text embeddings C and R , respectively. We then calculate two sets of losses and adjust their weights using hyper-parameter w :

$$\mathcal{L} = \mathcal{L}_{CLIP} + w \cdot \mathcal{L}_{i2c} \quad (4.2)$$

$$= \mathcal{L}_{i2t} + \mathcal{L}_{t2i} + w \cdot \mathcal{L}_{i2c}, \quad (4.3)$$

where $\mathcal{L}_{CLIP}$ is the CLIP loss composed of an image-to-text term $\mathcal{L}_{i2t}$ and a text-to-image term $\mathcal{L}_{t2i}$ [127], and $\mathcal{L}_{i2c}$ is the image-to-class binary cross-entropy loss. Specifically,

$$\mathcal{L}_{i2c} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\text{sim}(i, j)) + (1 - y_i) \log(1 - \text{sim}(i, j))], \quad (4.4)$$

where $\text{sim}(i, j)$ is the cosine similarity between the image embedding I_i and the text embedding for class prompt C_j . During training, the class prompt embeddings are regenerated in each batch. Several works have discussed the connection between contrastive learning and cross-entropy loss [78, 167]. In our scenario, $\mathcal{L}_{i2c}$ essentially contrasts image I with the class prompts, where multiple positive pairs may exist as determined by the ground-truth class labels. The shared cross-entropy nature of $\mathcal{L}_{CLIP}$ and $\mathcal{L}_{i2c}$ justifies training both tasks jointly.

Table 4.1: **Retrieval performance** for both report-to-image (r-x) and image-to-image (x-x) scenarios, measured with retrieval mAP@K, where K=5. wAvg and Avg are the weighted average over classes and the average over classes, respectively. Our model, DAug-CLIP, surpasses existing methods by a clear margin in the prevalent report-to-image scenario. In image-to-image retrieval, DAug-CLIP achieves state-of-the-art performance by wAvg. The gain is larger on the most common classes because DAug features for these classes are more reliable, as the classifier guiding the diffusion performs better on these classes.

			No Finding	Enl. Cardiomeg.	Cardiomegaly	Lung Opacity	Lung Lesion	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural other	Fracture	Support Devices	wAvg	Avg
Zhang et al. [185]			-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	.485
$r \rightarrow x$	X-TRA CLIP ($\mathcal{L}_{CLIP}$)	.62	.52	.93	.88	.50	.60	.29	.44	.75	.54	.85	.50	.36	.71	.606	.723	
	X-TRA CNN+BERT	.77	.54	.73	.91	.52	.83	.39	.87	.77	.63	.74	.23	.61	.73	.662	.735	
	X-TRA PubmedCLIP	.75	.65	.92	.99	.23	.79	.21	.51	.59	.72	.81	.56	.43	.67	.645	.720	
	X-TRA CLIP	.63	.62	.96	.94	.62	.69	.47	.61	.85	.69	.91	.57	.46	.82	.703	.779	
	DAug CLIP (ours)	.77	.86	.83	.89	.64	.79	.80	.71	.81	.64	.79	.66	.76	.79	.799	.767	
Yu et al. [173]			-	.65	.75	.72	.43	.80	.73	.60	.76	.76	.85	.43	.16	.86	-	.680
$x \rightarrow x$	X-TRA[150] CLIP ($\mathcal{L}_{CLIP}$)	.71	.52	.74	.78	.39	.79	.39	.40	.76	.42	.67	.44	.43	.64	.578	.761	
	X-TRA CNN+BERT	.87	.63	.88	.90	.49	.90	.57	.60	.85	.85	.83	.29	.47	.82	.678	.769	
	X-TRA PubmedCLIP	.90	.63	.82	.83	.39	.86	.45	.63	.87	.53	.90	.48	.51	.79	.685	.795	
	X-TRA CLIP	.84	.62	.89	.89	.56	.91	.55	.59	.89	.60	.86	.49	.57	.84	.713	.840	
	DAug CLIP (ours)	.72	.86	.83	.88	.53	.78	.79	.65	.81	.47	.77	.59	.69	.74	.771	.721	

4.3.2 Unified Inference for Retrieval and Classification

After the model is trained, we use it in different ways for the classification and retrieval tasks. For classification, we use only the image encoder and connect it to a linear classifier. The linear classifier is a single linear layer whose weight vectors are populated with the text embeddings of class prompts generated by the text encoder. The bias term is fixed to zero. Essentially, the output logit per class is equivalent to the unscaled cosine similarity between the image and classes $p_i = \text{sigmoid}(I \cdot C_i)$.

For retrieval tasks, both the image and text encoders are used to convert each sample into embeddings. We use the cosine similarity of the embeddings to rank associations for retrieval.

Table 4.2: **Image classification** performance on the MIMIC-CXR dataset. Results are in AUC-ROC. wAvg is the weighted average by number of samples per class. Avg is the average. The augmentation method X-TRA improves baseline performance, with CLIP achieving the best result. DAug-CLIP, our method, surpasses the best setting of X-TRA by a clear margin across all classes.

		No Finding	Enl. Cardiomeg.	Cardiomegaly	Lung Opacity	Lung Lesion	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural other	Fracture	Support Devices	wAvg	Avg
	Aug																
CNN+BERT	-	.81	.63	.73	.67	.62	.83	.69	.59	.68	.75	.83	.70	.58	.84	.71	.79
	X-TRA[150]	.81	.74	.75	.69	.63	.81	.72	.63	.75	.75	.83	.69	.63	.85	.73	.82
PubmedCLIP	-	.78	.65	.72	.66	.61	.82	.70	.61	.73	.76	.81	.62	.54	.84	.70	.78
	X-TRA[150]	.84	.76	.78	.69	.64	.83	.73	.64	.76	.75	.82	.75	.67	.85	.75	.83
CLIP	-	.77	.65	.71	.67	.62	.85	.73	.61	.72	.75	.80	.59	.51	.83	.70	.80
	X-TRA[150]	.82	.78	.74	.70	.71	.82	.75	.63	.79	.78	.86	.74	.72	.91	.77	.85
	DAug, ours	.93	.90	.90	.91	.85	.89	.89	.85	.89	.90	.90	.87	.84	.89	.89	.89

Table 4.3: **Ablation studies** on image retrieval and classification tasks. Compared to baselines, our feature augmentation method DAug improves performance over no augmentation (CLIP Baseline) and X-TRA augmentation (X-TRA Baseline). Additionally, using the proposed Image-Text-Class Hybrid Contrastive loss outperforms baselines using the original CLIP loss. Baseline results are from [150].

			No Finding	Enl. Cardiomeg.	Cardiomegaly	Lung Opacity	Lung Lesion	Edema	Consolidation	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural other	Fracture	Support Devices	wAvg	Avg
Method	Feature Aug	Criterion																
Retrieval: $r \rightarrow x$																		
X-TRA (Baseline)	X-TRA	CLIP	.62	.52	.93	.88	.50	.60	.29	.44	.75	.54	.85	.50	.36	.71	.606	.723
DAug (Ablation)	DAug	CLIP	.77	.85	.80	.86	.60	.76	.78	.67	.79	.61	.76	.65	.75	.79	.778	.745
DAug (Full)	DAug	Hybrid Contrastive	.77	.86	.83	.89	.64	.79	.80	.71	.81	.64	.79	.66	.76	.79	.799	.767
Retrieval: $x \rightarrow x$																		
X-TRA (Baseline)	X-TRA	CLIP	.71	.52	.74	.78	.39	.79	.39	.40	.76	.42	.67	.44	.43	.64	.578	.761
DAug (Ablation)	DAug	CLIP	.71	.86	.81	.88	.49	.76	.78	.62	.81	.44	.76	.58	.69	.74	.765	.711
DAug (Full)	DAug	Hybrid Contrastive	.72	.86	.83	.88	.53	.78	.79	.65	.81	.47	.77	.59	.69	.74	.771	.721
Classification																		
CLIP (Baseline)	-	CLIP	.77	.65	.71	.67	.62	.85	.73	.61	.72	.75	.80	.59	.51	.83	.70	.80
X-TRA (Baseline)	X-TRA	CLIP	.82	.78	.74	.70	.71	.82	.75	.63	.79	.78	.86	.74	.72	.91	.77	.85
DAug (Ablation)	DAug	CLIP	.77	.86	.83	.89	.64	.79	.80	.71	.81	.64	.79	.66	.76	.79	.80	.77
DAug (Full)	DAug	Hybrid Contrastive	.93	.90	.90	.91	.85	.89	.89	.85	.89	.90	.90	.87	.84	.89	.89	.89

4.4 Experiments

4.4.1 Implementation Details

To benchmark our ideas, we need an image-text-class dataset of medical images. We select MIMIC-CXR, the largest Chest X-ray (CXR) medical report dataset. It contains 227,835 image-text pairs, where the texts are radiology reports that include both normal and abnormal findings. As class labels are unavailable, we generate pseudo-class labels with the VisualCheXbert labeler [66]. It is a text classification model that converts a radiology report into binary labels for 14 disease classes. One of them is “No Finding”, indicating a healthy case.

For DAug feature augmentation, we construct a three-channel image with the first two channels containing the medical image and the last channel filled with the diffusion-generated heatmap. We use the heatmap for “No Finding”, a class that negates all diseases. Our approach requires no change to the model architecture. In real-world applications, heatmaps for other disease classes (*e.g.*, cardiomegaly) can be used for optimal performance gains in specific scenarios.

Following existing work [150], we resize images to 256×256 . We fine-tune a pre-trained ViT-Base/32 CLIP model with a cosine learning rate scheduler and a base learning rate of $2e^{-5}$ for 10 epochs. We use a batch size of 256 over eight V100 GPUs. With all images resized in advance, training and evaluation take around 2 hours. Abnormality heatmaps are pre-generated to save time. We evaluate on 14 disease classes and use a fine-tuned CLIP model without heatmaps as the baseline.

4.4.2 Results

We compare our method, DAug-CLIP, with existing methods on both multi-modal retrieval and image classification on the MIMIC-CXR dataset. Table 4.1 demonstrates that our method outperforms existing state-of-the-art approaches in retrieving radiology images with medical reports, a critical clinical scenario where radiologists refer to previous cases

to confirm diagnoses. Table 4.2 shows the image classification performance, where the same model outperforms existing methods by a clear margin.

4.5 Ablation Studies

We conduct ablation studies in Table 4.3. The results show that performance on both tasks improves incrementally with the addition of DAug feature augmentation and the Hybrid Contrastive criterion. This validates that both methods aid medical image understanding tasks.

4.6 Summary

This chapter presents DAug, a generative feature augmentation framework for both medical image retrieval and classification tasks. While the previous chapter explores the use of text generation models to supervise perception through knowledge distillation, DAug demonstrates an alternative generation-to-perception pathway that enhances discriminative learning without requiring additional labels.

Specifically, DAug augments visual representations with disease-specific localization heatmaps generated by a diffusion model. This auxiliary visual feature is portable to any model architecture. In addition, following the unified perspective introduced in Section 3.3.3, we propose an Image-Text-Class Hybrid Contrastive loss to jointly optimize retrieval and classification tasks within a single model. With both methods combined, DAug achieves state-of-the-art performance in both retrieval and classification tasks.

Chapter 5

QRAD: COMBINING GENERATION AND PERCEPTION IN RADIOLOGY REPORT GENERATION

5.1 Overview

Medical imaging plays a crucial role in healthcare diagnostics. However, the worldwide shortage of radiologists poses significant risks to patient care [40, 121, 10]. Automated radiology report generation using AI has emerged as a promising solution to this challenge, with the potential to handle routine cases and let radiologists focus on the most complex ones.

Despite recent advances in radiology report generation, significant gaps remain toward clinical adoption. First, current approaches, which typically follow an image captioning pipeline, struggle with the inherent linguistic uncertainties [147] in radiology reports. Unlike conventional image captioning, radiology reports are longer documents that require precise factual accuracy while exhibiting considerable sentence-level linguistic variation, such as whether a finding is mentioned or omitted, and the order in which medical findings are presented. As a simplified example, if the ground truth has three sentences [A, B, C], a prediction that reorders these findings (*e.g.*, [C, A, B]) is clinically correct but is unfairly penalized [61] by the language modeling loss because it requires exact token-by-token matches¹. Consequently, models tend to overfit such linguistic variations at the expense of factual accuracy. In report generation datasets such as MIMIC-CXR [73], each training sample contains one or more images and an associated text report. Image captioning datasets like COCO [97] provide multiple reference texts to capture linguistic variation; however, this solution is not feasible for radiology report datasets. Furthermore,

¹Teacher forcing during training may reduce the effect.

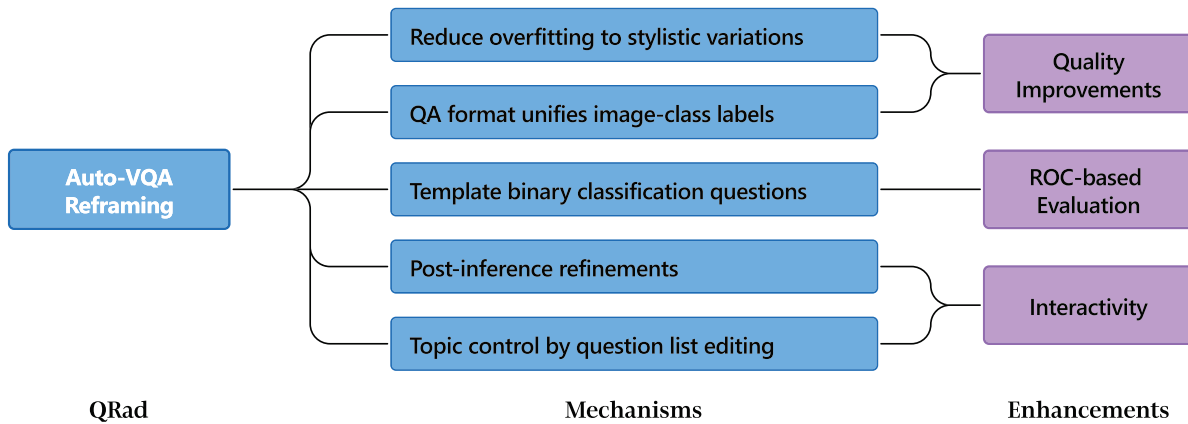


Figure 5.1: Overview of QRad’s Captioning-to-VQA reframing approach. This question-driven framework (the “Q” in QRad) enables five mechanisms that collectively enhance clinical utility across three dimensions: report quality, regulatory evaluation, and interactivity.

conventional approaches that follow a direct image-to-text pipeline [175, 149, 169, 19] offer no interactive mechanisms, preventing physicians from requesting additional information about specific concerns omitted in the initial output [118, 59].

Second, existing generative models do not produce numerical probabilities for individual medical findings. This limitation matters because clinical automation and regulatory clearance both depend on quantitative, per-finding confidence estimates. Downstream applications should be calibrated to operating points with explicit tolerances for false positives and false negatives appropriate to the scenario. For example, cancer screening prioritizes high sensitivity to avoid missed cases, while cohort discovery systems for clinical research require high specificity to accurately identify patients meeting strict inclusion criteria and reduce downstream noise. Without per-class confidence scores, such operating points cannot be selected or audited. FDA device authorization, for instance, requires Receiver Operating Characteristic (ROC) curves, which are obtained by sweeping a threshold over per-class confidence scores to characterize sensitivity and specificity across these clinical applications [37]. A model that produces confidence scores for requested disease classes can therefore facilitate regulatory approval, moving one step closer to real-world

adoption.

To close these gaps, we introduce QRad, a novel approach that reframes radiology report generation as a self-directed Visual Question Answering (Auto-VQA) process. QRad operates in two steps: (1) Question Generation, which produces a chain of relevant clinical questions conditioned on the input radiograph, effectively planning the report’s structure; (2) Answer Generation, which answers those questions by examining visual features. The answers are concatenated to form the final report. To facilitate training, we convert reference reports into QA pairs by segmenting each report into contiguous topical spans (answers), and GPT-5² generates a single question that captures each span’s topic. This design offers two immediate benefits: it operationalizes Chain-of-Thought [162] via explicit planning-and-answering decomposition, and it provides an interactive capability, allowing physicians to request specific information beyond the initial report by editing or issuing follow-up questions, a feature unavailable in previous single-step models.

Formally, traditional approaches model report generation as $Y = f(I)$, where an image I directly maps to a report Y . Due to linguistic variations in radiology reports such as sentence reordering, this formulation suffers from a one-to-many mapping from I to multiple valid Y , causing the learning process to overfit to surface phrasing at the expense of clinical accuracy. This limitation arises from the language modeling loss, which demands exact token matches. When the linguistic variations follow no consistent pattern, the loss from factually critical tokens (*e.g.*, presence and severity of diseases) is diluted by the loss from surface phrasing differences. To address this, QRad reframes the process as $Y = f_A(I, Q)$, where $Q = f_Q(I)$. By providing the Answer Generator f_A with a question Q , we explicitly require the model to state a diagnosis for the given clinical topic, eliminating the uncertainty of sentence ordering. The ground truth for f_A is a single topical span from the report, which leaves less room for linguistic variation. The Question Generator f_Q captures linguistic variability. At test time, even when it produces questions that differ

²We use a private, in-house deployment to satisfy data-usage requirements.

from the training data, they tend to be clinically valid variations that preserve diagnostic utility. In essence, we isolate linguistic variability in f_Q and concentrate factual supervision in f_A . Moreover, the VQA reframing allows us to augment the training data with additional image classification questions; in these cases, the ground truth is a single Yes/No token, which further reduces linguistic variability and concentrates supervision on diagnostic accuracy.

Furthermore, typical regulatory processes (*e.g.*, FDA approval) require ROC-based validation, which depends on class probabilities like those produced by perception models. QRad bridges this gap via *closed-vocabulary* VQA: for each predefined disease class, we pose a binary, template-based query (*e.g.*, “Is this image classified as [CLASS]? (yes/no)”), extract the token logits for predefined answers {Yes, No}, and compute the softmax as class probabilities. In contrast, conventional report-generation models emit free-form sentences that may mention multiple diseases or omit a disease entirely, so token-level probabilities are not class-specific and cannot serve as per-class confidence scores to support ROC analysis. Conversely, image classifiers do not produce open-vocabulary reports that describe medical findings with flexibility. Our VQA reframing unifies both regimes, providing open-vocabulary narratives and closed-vocabulary class probabilities within a single backbone to support ROC/AUC analysis and offering a practical path toward regulatory clearance and real-world adoption.

In summary, we propose QRad, a captioning-to-VQA reframing approach that addresses key limitations in radiology report generation. As illustrated in Figure 5.1, our question-driven framework enables five core mechanisms that collectively enhance clinical utility across three critical dimensions: improving report quality by reducing overfitting to stylistic variations, enabling ROC-based evaluation through quantitative confidence scores, and providing interactivity through post-inference refinements and topic control. Experiments show that QRad outperforms state-of-the-art models [181, 189, 19, 175] while using only 13% of their model size.

5.2 *MedImageInsight: Contrastive Pre-trained Medical Foundation Model*

In this section, we lay the groundwork by introducing *MedImageInsight* [21], an open-source medical foundation model developed as a collaborative project during this Ph.D. study. *MedImageInsight* uses image-text contrastive learning to pre-train its image and text encoders. Our contribution focuses on the generative downstream task, where we connect the pre-trained image encoder to a text decoder and propose an effective training recipe for radiology report generation (Section 5.2.2). This demonstrates how image-text pairs can support both discriminative (contrastive) and generative objectives to achieve state-of-the-art performance on radiology report generation. This work serves as a foundation for our subsequent QRad method, which further enhances report generation by reframing it as a question-answering process and integrating classification capabilities.

5.2.1 *Medical Image-Text Contrastive Pre-training*

MedImageInsight is a medical image-text foundation model trained on a large collection of medical datasets across 12 imaging domains. It is derived from the Florence [174] computer vision foundation model, with a DaViT [30] image encoder and a UniCL [167] objective, and is further trained on medical images with associated text or categorical labels across diverse domains, including X-ray, CT, MRI, clinical photography, dermoscopy, histopathology, ultrasound, and mammography.

The overall architecture of *MedImageInsight* is shown in Figure 5.2. *MedImageInsight* is a two-tower architecture similar to CLIP [127]. The DaViT image encoder used in *MedImageInsight* has 360M parameters and the language encoder has 252M parameters.

Training Setup: We train the foundation model using the UniCL objective with mixed precision (FP16) enabled. Training uses the AdamW [84] optimizer with a learning rate of 0.00001 and weight decay of 0.2. The batch size is set to 1024 and distributed across multiple GPUs, with gradient accumulation steps computed based on the batch size and

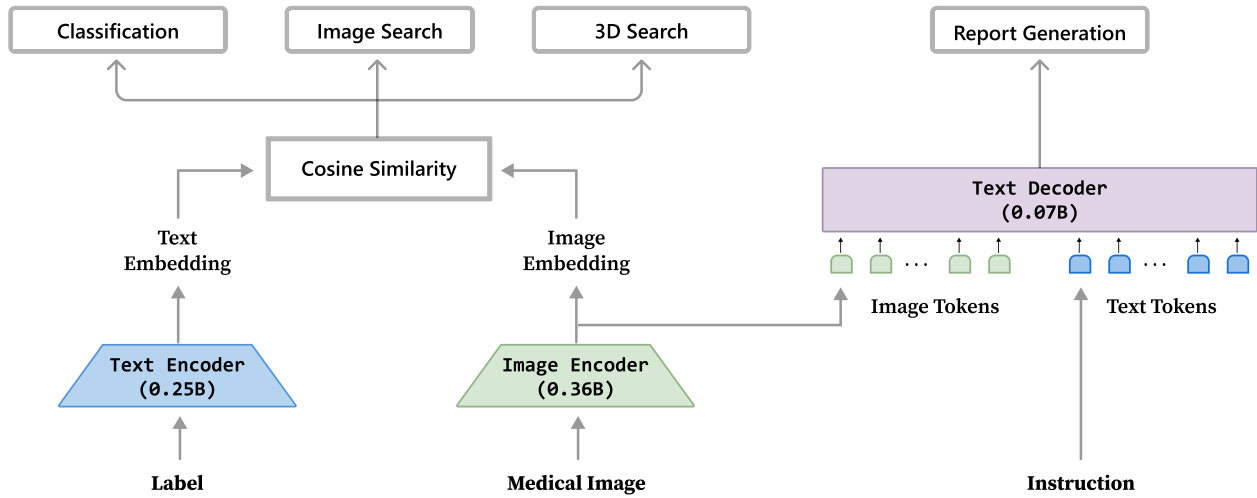


Figure 5.2: Overview of the *MedImageInsight* foundation model architecture. UniCL is used as the pre-training objective function.

number of GPUs. The learning rate follows a cosine decay schedule with warmup steps. Training runs for 40k iterations.

Medical Pre-Training Datasets: Datasets used for model pre-training span 12 medical imaging domains and are summarized in Table 5.1. Datasets include MIMIC-CXR [73], NIH-CXR-LT [56], IRMA 2009 [86], RSNA BoneAge [52], GP BoneAge Atlas [44], UPENN [6], TCGA [132], SD198 [144], ISIC2019 [20, 22, 23], PatchCamelyon [33], RSNA Mammography [12], LIDC-IDRI [104], PAD-UFES-20 [117], ODIR-5K [13], OCT2017 [76], Breast Ultrasound Image (BUSI) [3], HMC-QU [53], Medical Segmentation Decathlon (MSD) [35], ROCO [123], PMC15M [179], and partner and proprietary datasets.

5.2.2 Radiology Report Generation Baselines

As part of *MedImageInsight*, we build a simple baseline for radiology report generation using *MedImageInsight*'s image encoder as the pre-trained backbone. We propose a training recipe that first freezes the vision backbone and pre-trains the text decoder on medical

Table 5.1: *MedImageInsight* Pre-training & Evaluation Datasets

Dataset	Modalities*	Format	Data Used for Pre-training	Data Used for Evaluation
MIMIC-CXR [73]	CXR	Image-Text	203,170	0
NIH-CXR-LT [56]	CXR	Image-Label (20)	68,058	20,279
IRMA 2009 [86]	XR	Image-Label (137)	12,677	1,733
RSNA BoneAge [52]	XR	Image-Age	12,611	1,425
GP BoneAge Atlas [44]	XR	Image-Age	59	0
UPENN [6]	MRI	Image-Text	4,645	0
TCGA [132]	MRI/CT	Image-Text	5,643	0
SD198 [144]	CP	Image-Label (198)	3,253	3,331
ISIC2019 [20, 22, 23]	DM	Image-Label (8)	20,268	5,063
PatchCamelyon [33]	Hist.	Image-Label (2)	262,144	32,768
RSNA Mammography [12]	XR	Image-Label (4)	43,764	10,942
LIDC-IDRI [104]	CT	Image-Label (4)	80,201	4,574
PAD-UFES-20 [117]	CP	Image-Label (6)	2,065	233
ODIR-5K [13]	Fundus	Image-Label (79/30)	5,228	1,267
OCT2017 [76]	OCT	Image-Label (4)	83,484	968
OCT2018 [77]	OCT	Image-Label (4)	0	1000
BUSI [3]	US	Image-Label (3)	623	157
HMC-QU [53]	US	Image-Label (2)	4,714	2,132
Bing Echo View	US	Image-Label (7)	100	54
MSD [35]	CT	Image-Text	116,500	0
Bing Exam Parameter	CT/MRI	Image-Label (21)	1202	401
ROCO [123]	Mixed	Image-Text	70,308	0
MGB CXR	CXR	Image-Label	731	684
MGB BoneAge	XR	Image-Age	0	120
PMC15M [179]	Mixed	Image-Text	952,556	0
Proprietary	Mixed	Image-Text	1,554,068	0
Total	Mixed	Image-Text/Label/Age	3,508,072	86,131

* CXR = Chest X-ray, XR = Other X-ray, MRI = Magnetic Resonance Imaging, CT = Computed Tomography, CP = Clinical Photography (Dermatology), Hist. = Histology, DM = Dermoscopy (Dermatology), Fundus = Fundus Imaging, US = Ultrasound, OCT = Optical Coherence Tomography

image-text pairs, then fine-tunes the full model on report generation. Unlike prior state-of-the-art methods that use large language models as the decoder [175, 88], our decoder is randomly initialized and has only 0.07B parameters, about 1% of their size. Yet it outperforms these much larger counterparts, showing that the training recipe, rather than scaling the model, is the key factor for radiology report generation. We additionally fine-tune GPT-4 [2] as a further baseline and report all results in Section 5.5.2 (Table 5.2).

5.3 Auto-VQA Training Data Generation

Building on the *MedImageInsight* baseline, QRad switches the training objective from image-to-report generation to the proposed Auto-VQA process, where the model generates a sequence of questions and then answers them, with the answers forming the report. QRad requires two types of question-answering (QA) training data: a report generation QA dataset converted from the image-report dataset, and an image classification QA dataset converted from image-class labels. Compared to using the original full reports as supervision, the first dataset reduces linguistic variation at the sentence level, such as the omission and ordering of sentences, while the second dataset, being closed-vocabulary with answers in {Yes, No}, further reduces linguistic variation at the phrase level.

5.3.1 Report Generation Question-Answer Pairs (Open Vocabulary)

To generate these datasets from image captions, we use an LLM³ to split the reports into sentence groups. Consecutive sentences in a report covering the same topic are treated as a cohesive unit. Then, we use each sentence group as an answer and compose a corresponding question with the LLM. As shown in Figure 5.3, the questions should be precise enough to indicate the topics while not being so specific that they leak the answer.

For the MIMIC-CXR [73] dataset, we generate a total of 818,867 question-answer pairs across all radiology studies. There are 110,959 unique questions (based on string matches, not semantic similarity). 91.3% of the reports have no more than 5 sentences, and 99.4% of the reports have no more than 8 sentences. Typical answers contain only one sentence.

5.3.2 Image Classification Question-Answer Pairs (Closed Vocabulary)

One benefit of the Captioning-to-VQA reframing is the possibility of unifying both text and class supervision in a VQA format, allowing our model to learn from both kinds of

³We use a private, in-house deployment of GPT-5 [138] to ensure compliance with the dataset usage requirements. The data processing does not assume a particular LLM.

- [Q1] "What type of view is used in the chest X-ray?"
[A1] "Single AP view of the chest provided."
- [Q2] "Are there any support devices visible?"
[A2] "An endotracheal tube ends 2.0 cm above the Carina. A transesophageal tube courses below the level of the diaphragm, however the tip cannot be visualized."
- [Q3] "What is the condition of the lung volumes and clarity?"
[A3] "Lung volumes are low, however grossly clear."
- [Q4] "Is there any atelectasis?"
[A4] "Bibasilar atelectasis is moderately increased."
- [Q5] "Are there signs of pleural effusion or pneumothorax?"
[A5] "No pleural effusion or pneumothorax."

Figure 5.3: Examples of the converted report generation QA dataset. We show the first five sentences from a radiology report, split into five answers. Q_i and A_i are the i^{th} question and answer, respectively.

annotations for improved performance. Here, we augment image-report data with image-class labels. Specifically, in addition to the report generation QA pairs, we convert image class labels obtained from VisualCheXbert [66] into a closed-vocabulary VQA format. In later sections, we show that this integration not only enhances our model’s image understanding performance but also provides a natural interface for confidence score extraction.

We show examples of the converted dataset in Figure 5.4, where class labels from CheXpert [64] are transformed into closed-vocabulary QA pairs. Questions are constructed by randomly sampling from a template pool. The closed-vocabulary nature of these QA pairs focuses training on the model’s image classification capabilities, similar to an image classifier, in a process free from linguistic variation. When training on such datasets, the model receives no reward for writing a full sentence that has token-wise overlap with the

- [Q1] "Is this image classified as cardiomegaly? (yes/no)"
[A1] "Yes"
- [Q2] "Does this chest X-ray demonstrate edema? (yes/no)"
[A2] "No"
- [Q3] "Is pleural effusion evident in this chest X-ray? (yes/no)"
[A3] "Yes"
- [Q4] "Does this radiograph indicate pneumothorax? (yes/no)"
[A4] "No"
- [Q5] "Does this chest X-ray reveal support devices? (yes/no)"
[A5] "No"

Figure 5.4: Examples of question-answer pairs converted from image classification labels. The questions are formulated using question templates with class names, and a “(yes/no)” suffix that distinguishes them from report generation QA pairs to indicate that a single-token binary answer is expected.

ground-truth sentence but is factually incorrect.

5.4 Auto-VQA Pipeline and QRad Model Architecture

Conventional approaches to long text generation from visual inputs frame the task as direct image-to-text mapping, *i.e.*, image captioning. Because valid linguistic variations are prevalent in radiology reports and amplified by their length, factual accuracy is hindered when the training loss is dominated by unmatched tokens representing linguistic variation rather than factual correctness. We propose a general approach that reframes long text generation into Auto-VQA, a self-directed Visual Question Answering process where the questions are self-generated and serve as an explicit plan akin to chain-of-thought [162] models.

Specifically, QRad decomposes image-to-text generation from $Y = f(I)$ into two distinct components: a Question Generation Module $Q = f_Q(I)$ and an Answer Generation

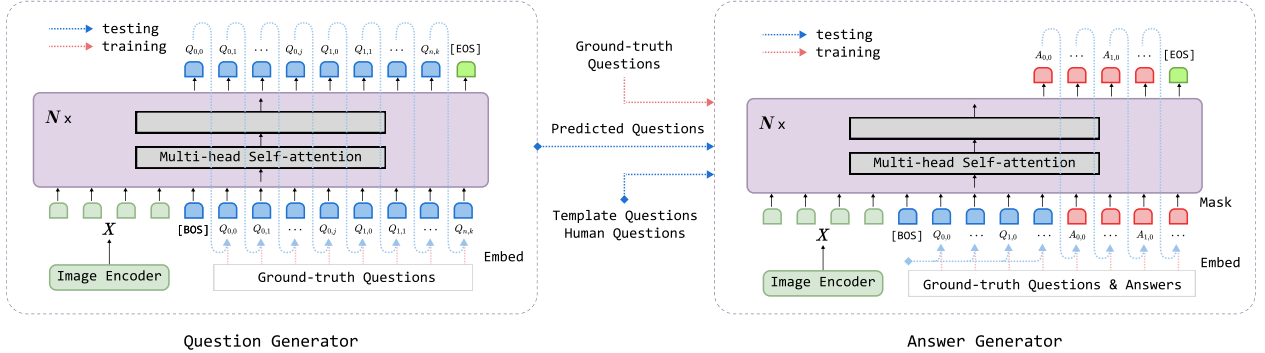


Figure 5.5: QRad’s Auto-VQA pipeline. The Question Generator predicts a sequence of questions $[Q_0, \dots, Q_n]$ given the image features X , where Q_i is the i^{th} question containing m tokens $[Q_{i,0}, \dots, Q_{i,m-1}]$. The Answer Generator predicts corresponding answers $[A_0, \dots, A_n]$ given the image features and the questions. When generating A_i , attention masks are used to control the visibility of prefix questions $[Q_0, \dots, Q_i]$ and previous answers $[A_0, \dots, A_{i-1}]$. During training, the ground-truth questions and answers are used as input (red arrows). During testing, model-predicted questions, optionally modified or extended by humans, are used as input (blue arrows), and their answers are concatenated to form the radiology report. For simplicity, we omit details such as the input instructions in the figure.

Module $Y = f_A(I, Q)$, where I, Q, Y denote the input image, questions, and answers (sentences in the report), respectively. As shown in Figure 5.5, both modules use an identical transformer architecture: an MI2-based [21] visual backbone and a tiny text decoder of six transformer layers, the same architecture as the baseline described in Section 5.2.2.

The proposed Captioning-to-VQA reframing method in QRad is generalizable to different model architectures. In our experiments, it effectively elevates the performance of a small model to match that of models 10 times larger. In this chapter, we benchmark our method on MIMIC-CXR [73], one of the largest public radiology report datasets.

5.4.1 Question Generation Module

The Question Generation Module conducts sequence generation autoregressively with previously generated questions as the context. Concretely, it generates $m + 1$ output tokens for the i^{th} question $q_i = [q_{i,0}, q_{i,1}, \dots, q_{i,m}]$, where $q_{i,m} = [\text{EOS}]$, by modeling Equation (5.1):

$$P(Q | X) = \prod_{i=0}^{k+1} P(q_i | X, q_0, q_1, \dots, q_{i-1}), \quad (5.1)$$

The ground-truth Q is the concatenated questions $[q_0, \dots, q_k]$. Q is predicted as a whole, and we split it by the question mark ("?.") for individual questions.

5.4.2 Answer Generation Module

The Answer Generator learns to generate a sentence of n tokens $Y_i = (y_{i,1}, y_{i,2}, \dots, y_{i,n})$ conditioned on the image and a question Q_i . Mathematically, the module models the following:

$$P(Y_i | X, Q_i) = \prod_{j=0}^{n+1} P(y_{i,j} | X, Q_i, y_{i,0}, y_{i,1}, \dots, y_{i,j-1}), \quad (5.2)$$

where Y_i denotes a sentence from the radiology report, which becomes the i^{th} answer corresponding to question q_i . By iterating q_i through all questions, the Answer Generator generates k sentences $Y = (Y_1, Y_2, \dots, Y_k)$ and composes the whole radiology report.

5.4.3 Training Recipe

Training Stages: The Question and Answer Generators are separate modules using the same encoder-decoder architecture. The vision encoder is MedImageInsight (MI2) [21], a 0.36B model trained on medical images. The decoder is a six-layer, randomly initialized transformer text decoder of 0.07B parameters. The encoder and decoder are connected via a linear projection layer. The total model size is $2 \times (0.36 + 0.07) \approx 0.9\text{B}$, around 13% of current state-of-the-art models that are based on 7B language models. We first pre-train

the model as introduced in Section 5.2.2. Then, we duplicate the model and fine-tune each one for question and answer generation tasks, respectively, where the full model is made trainable.

Mixture of VQA Data: As discussed in Section 5.3, our training data contain both report generation QA data and image classification QA data to improve model performance. The data mixture ratio is discussed in the Appendix, Section 5.8.3.

The Prompt Template: Following existing studies, we use a short instruction that includes the Indication for specifying the goal of the radiology study. We use the ground-truth questions as input when training the Answer Generator.

Attention Masks for Training Efficiency: After converting the training data from image-report to image-QA pairs, the number of training samples increases to the number of sentences in all reports, which significantly increases training cost. To improve training efficiency, we concatenate all QA pairs for the same image and construct attention masks to control context visibility, thereby enabling us to run forward in one pass.

5.5 Experiments and Ablation Studies for Report Generation

5.5.1 Implementation Details and Evaluation Metrics

Training Parameters: We use an image size of 512×512 . We follow the pre-training setup in Section 5.2.2. For the Question Generator, we fine-tune the model to generate questions in the Report Generation VQA dataset. For the Answer Generator, we fine-tune the same pre-trained model on both the Report Generation and Image Classification VQA datasets with a mixing ratio of 6:4. For both models, we use a learning rate of $1e - 5$ for the image encoder and $5e - 5$ for the text decoder. The batch size is 512. The Question Generator and the Answer Generator are trained for 40 and 60 epochs, respectively.

Evaluation Metrics: In Table 5.2, CheXbert [139] is a Clinical Efficacy (CE) metric that classifies generated reports into 14 disease categories and evaluates F1 scores, focusing on factual accuracy rather than textual overlap. As CheXbert produces an *uncertain* class in addition to positive and negative classes, existing methods treat *uncertain* as either positive or negative for evaluation. RadGraph-ER [65] is designed specifically for radiology reports and assesses the correctness of extracted entities and their attributes. BLEU and ROUGE are standard lexical metrics that measure n-gram similarity to evaluate text overlap. For results not reported in the original papers, we use values from [175]. Table 5.3 uses ReXrank [183], a more recent benchmark for radiology report generation. The leaderboard ranks models by the inverse of RadCliQ [172], a composite metric combining BLEU-2 [119], BERTScore [180], SembScore [139], and RadGraph-F1 [172], where BERTScore and SembScore are BERT and CheXbert embedding similarities, respectively. Other individual metrics include RaTEScore [186] and GREEN [115], where RaTEScore is based on embedding similarities of extracted medical entities, and GREEN is an LLM-based metric.

5.5.2 Radiology Report Generation

We conduct the main experiments on MIMIC-CXR [73], one of the largest radiology report generation datasets. It has 227,835 image-report pairs. We use only the frontal-view radiograph from each training sample. Following recent studies [175, 62], we use the IU X-ray dataset [27] as a fully held-out evaluation set. Results are provided in Section 5.8.2.

On the MIMIC-CXR dataset, Table 5.2 and Table 5.3 provide both the conventional benchmark, including lexical and clinical efficacy (CE) metrics, and the newer ReXrank [183] leaderboard. QRad outperforms major state-of-the-art methods across both benchmarks, despite using 13% of their model size. On the ReXrank leaderboard, both QRad and UniRG-CXR are our models: QRad is the SFT model, while UniRG-CXR is a 7B-parameter RFT model initialized from an SFT model. UniRG-CXR achieves superior performance in RadCliQ by optimizing it as the reward. QRad achieves the same or better performance in

Table 5.2: Report Generation Performance on MIMIC-CXR

Model	CheXbert								RadGraph	BLEU		ROUGE
	("uncertain" as <i>negative</i>)				("uncertain" as <i>positive</i>)							
	Micro-avg		Macro-avg		Micro-avg		Macro-avg			ER	(1)	
	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5				
Single Image, Model size $\geq 7B$												
LLaVA-Rad [175] ^F	57.3	57.4	39.5	47.7	57.3	60.2	44.0	53.3	29.4	38.1	15.4	30.6
Med-Gemini [169] ^F	-	-	-	-	-	-	-	-	-	-	20.5	28.3
VILA-M3 40B [109]	-	-	-	-	-	-	-	-	-	-	21.6	32.2
Med-PaLM M [149]	53.6	57.9	39.8	51.6	-	-	-	-	-	32.3	11.3	27.3
MAIRA-1 [62] ^F	55.7	56.0	38.6	47.7	55.3	58.8	42.3	51.7	29.6	39.2	14.2	28.9
GPT-4V	35.5	25.8	20.4	19.6	35.6	33.3	25.3	29.6	13.2	16.4	1.9	13.2
CheXagent [19]	39.3	41.2	24.7	34.5	39.4	42.1	27.3	35.8	20.5	16.9	4.7	21.5
LLaVA-Med [88] ^F	27.2	22.0	15.5	16.6	27.3	24.4	18.7	20.5	6.5	22.2	1.0	13.3
LLaVA [99] ^F	22.9	23.4	15.4	17.5	23.7	26.9	17.0	20.3	4.5	21.0	1.3	13.8
Single Image, Model size = 4B												
Baseline ^a	54.3	55.2	36.9	46.6	54.1	57.4	40.4	50.5	31.1	40.1	17.8	32.7
QRad (ours, 4B) ^a	57.6	59.0	40.8	51.0	57.1	61.4	44.3	54.4	31.1	40.6	17.5	32.5
Single Image, Model size <1B												
PromptMRG [68]	-	-	-	-	-	-	-	-	-	-	11.2	26.8
Flamingo [4]	-	-	-	-	51.9	56.5	-	-	-	-	10.1	29.7
CvT2Dist. [111]	44.2	-	30.7	-	-	-	-	-	-	39.3	12.7	28.6
M ² trans [108]	-	-	-	-	-	56.7	-	-	-	-	11.4	-
RGRG [146]	-	-	-	-	-	54.7	-	-	-	37.3	12.6	26.4
R2Gen [18]	-	-	-	-	22.8	34.6	-	-	-	35.3	10.3	27.7
TieNet [158]	-	-	-	-	-	27.1	-	-	-	-	8.1	-
MI2 (baseline, ours) [21]	56.3	57.9	38.4	49.3	55.7	59.3	43.2	52.1	28.5	37.3	15.3	31.7
QRad^F (ours, 0.9B)	58.4	59.5	41.5	51.8	57.9	62.2	45.1	55.2	31.5	40.0	16.9	32.5

^F The testing set includes only frontal-view images.

^a The 4B models use BiomedCLIP [179] as the vision encoder and Phi-3-mini [1] as the text decoder.

RaTEScore and GREEN, two metrics not included in RadCliQ. UniRG-CXR is collaborative work, and we provide details of its RFT in Section 5.8.1. Qualitative examples of generated questions and answers are included in Section 5.8.4. ROC-based evaluation for each class is provided in Section 5.8.5.

5.5.3 Ablation of Each Component

In Table 5.4, we conduct an ablation study on an MI2-based small model with 0.9B parameters. The table shows that each component brings consistent performance gains. The largest improvements are observed on Clinical Efficacy metrics (CheXbert, RadGraph),

Table 5.3: Performance on the ReXrank Benchmark

Model	1/RadCliQ	BLEU	BERTScore	SembScore	RadGraph	RaTEScore	GREEN
UniRG-CXR [100], ours ³	1.217	<u>0.248</u>	0.493	0.487	<u>0.265</u>	0.596	0.352
QRad-0.9B [71], ours ³	<u>1.143</u>	0.264	<u>0.482</u>	<u>0.479</u>	0.243	0.596	<u>0.362</u>
MedVersa [189]	1.103	0.209	0.448	0.466	0.273	<u>0.550</u>	0.374
Libra [181]	0.898	0.232	0.402	0.403	0.218	0.523	0.356
RadPhi3.5Vision [129]	0.888	0.223	0.386	0.431	0.207	0.534	0.294
CXRMate-ED [113]	0.872	0.208	0.383	0.396	0.223	0.531	0.327
CXRMate-RRG24 [112]	0.870	0.198	0.367	0.423	0.220	0.521	0.338
CheXpertPlus-CheX [14]	0.805	0.142	0.367	0.379	0.181	0.490	0.305
DD-LLava-X*	0.801	0.154	0.348	0.402	0.182	0.505	0.301
RaDialog [125]	0.799	0.127	0.363	0.387	0.172	0.485	0.273
CheXpertPlus-MIMIC [14]	0.788	0.145	0.361	0.375	0.170	0.485	0.311
RGRG [146]	0.755	0.130	0.348	0.344	0.168	0.491	0.273
MedGemma [133]	0.744	0.165	0.346	0.339	0.159	0.549	0.293
CheXagent [19]	0.741	0.113	0.346	0.347	0.148	0.474	0.257
MoERad-MIMIC*	0.726	0.163	0.341	0.334	0.143	0.465	0.240
Cvt2distilgpt2 [111]	0.719	0.126	0.331	0.329	0.149	0.432	0.268

¹ Results shown are for the Findings Generation task on the MIMIC-CXR dataset.

² Models are ranked by 1/RadCliQ-v1 (higher is better for all metrics). **Bold** and underline indicate the best and second-best results, respectively. An introduction to metrics is available in Section 5.5.1.

³ UniRG-CXR is our model that demonstrates reinforcement fine-tuning (RFT, see Section 5.8.1). It uses RadCliQ (a composite metric of BLEU, BERTScore, SembScore, and RadGraph) as the reward function during RFT and achieves the best 1/RadCliQ score. QRad matches UniRG-CXR in RaTEScore and outperforms it in GREEN, two metrics not included in UniRG-CXR’s reward, showing promising generalizability of the QRad approach.

* DD-LLava-X and MoERad from the leaderboard have no associated publications yet.

which reflect factual accuracy in the medical domain. Additional comparisons of training configurations are provided in Section 5.8.3.

Ablations of the Question Generator: Table 5.5 demonstrates the importance of using a learned Question Generator over fixed template questions. The key challenge in medical report generation is the vast and complex space of possible medical conditions that can appear in an image. It is infeasible to enumerate all potential diseases as a predefined set of template questions. Moreover, even if such an exhaustive list existed, requiring the model

Table 5.4: Ablation Studies on MIMIC-CXR

Model	CheXbert								RadGraph	BLEU		ROUGE
	("uncertain" as <i>negative</i>)				("uncertain" as <i>positive</i>)							
	Micro-avg		Macro-avg		Micro-avg		Macro-avg					
	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5				
									ER	(1)	(4)	(L)
Baseline (MI2)												
<i>median</i>	56.2	57.8	38.3	49.2	55.7	59.2	42.1	52.0	31.1	37.3	15.3	31.7
<i>ci_l</i>	55.1	56.2	36.7	47.1	54.7	57.8	40.6	50.6	30.5	36.8	14.9	31.2
<i>ci_h</i>	57.3	59.4	40.0	51.3	56.7	60.7	43.5	53.7	31.8	37.8	15.7	32.1
Baseline + Captioning-to-VQA												
<i>median</i>	57.9	59.8	40.0	50.7	57.6	62.7	44.2	55.5	31.4	39.9	16.5	32.4
<i>ci_l</i>	56.8	58.3	38.1	48.9	56.6	61.3	42.6	53.9	30.8	39.3	16.0	31.8
<i>ci_h</i>	59.0	61.3	41.6	52.5	58.7	64.0	45.8	57.2	32.1	40.6	17.1	32.9
Baseline + Captioning-to-VQA + Classification QA (QRad)												
<i>median</i>	58.3	59.5	41.5	51.8	57.9	62.2	45.1	55.2	31.6	40.2	16.7	32.5
<i>ci_l</i>	57.3	57.9	39.8	49.7	56.9	60.8	43.7	53.6	30.9	39.4	16.2	32.0
<i>ci_h</i>	59.4	61.0	43.0	53.7	59.0	63.5	46.6	57.0	32.2	40.9	17.2	33.1

¹ The baseline ablates Captioning-to-VQA reframing while keeping the model architecture and pre-training the same. It is equivalent to the baseline in Section 5.2.2.

² To demonstrate statistical significance, we report the median and 95% confidence intervals (*ci_l* and *ci_h*) over 500 bootstrap replicates for all metrics.

to answer questions about every possible condition would be computationally prohibitive and inefficient. Our Question Generator addresses this by dynamically predicting relevant questions based on the input image, focusing only on conditions likely to be present.

Quality of Generated Questions: In Table 5.5, the experiment using oracle questions provides a performance upper bound by removing the effect of sentence-level stylistic variations from the evaluation. The superior performance reflects the strength of the Answer Generator, which no longer needs to guess the omission and ordering of findings during training. In practice, we still want the Question Generator to produce meaningful questions related to the input image, even though a perfect match with the ground truth is not possible. We assess the quality of the generated questions from several angles. In terms of length, predicted questions average 5.34 sentences (std 1.57) and 40.85 words

Table 5.5: Ablation of the Question Generator

Model	CheXbert								RadGraph	BLEU		ROUGE
	("uncertain" as <i>negative</i>)				("uncertain" as <i>positive</i>)							
	Micro-avg		Macro-avg		Micro-avg		Macro-avg					
	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5		ER	(1)	(4)
Template Questions ^a	44.7	52.1	28.5	37.5	46.6	56.3	34.2	45.5	23.5	31.1	8.2	20.9
Predicted Questions ^b	58.4	59.5	41.5	51.8	57.9	62.2	45.1	55.2	31.5	40.0	16.9	32.5
Oracle Questions ^c	74.7	78.0	60.5	72.2	76.1	79.8	66.0	74.9	48.0	54.4	30.5	52.8

^a. We use template questions for the 14 widely used CheXbert classes.

^b. We use the Question Generator to learn and predict the questions for each input image (the QRad method).

^c. Oracle questions are GPT-generated from the ground-truth reports. This is the ground truth used to train the Question Generator.

(std 12.21), close to the oracle questions' 5.72 sentences (std 1.75) and 41.81 words (std 14.37). Regarding disease coverage, oracle reports mention 3.10 positive diseases per report on average (std 1.61), and the predicted questions recover them with a recall of 0.79. The missed questions correspond to diseases that are not detected by the image encoder. Improving disease coverage is the biggest opportunity for future work.

5.6 Class Confidence Score Extraction

Unlike conventional report generation models, QRad can produce numerical class probabilities alongside free-text reports. This dual capability allows the same model to be evaluated using both natural-language generation (NLG) metrics and classification metrics (*e.g.*, ROC curves). This section describes how class probabilities are extracted (Section 5.6.1) and calibrated (Section 5.6.2) from QRad, which form the foundation for the threshold-control framework introduced later in Section 5.7.

5.6.1 Image Classification with a Text Generation Model

To extract per-class probabilities from QRad, we leverage its VQA capability: we send a template classification question (*e.g.*, “Is there pleural effusion? (yes/no)”) to request a binary answer. Because the answer is a single token, we read the class probability directly from the model’s output distribution over a vocabulary of two as:

$$P(C_i = 1) = \frac{e^{x_{\text{yes}}}}{e^{x_{\text{yes}}} + e^{x_{\text{no}}}}, \quad (5.3)$$

where x_{yes} and x_{no} are logits for [yes] and [no]. Our approach is related to existing work [75] that uses $P([\text{true}])$ as a self-assessed confidence score for LLM answers. Restricting the softmax to [yes] and [no] rather than the full vocabulary implements classification using the generative model’s weights. In comparison, prior approaches represent binary classifications with free-text sentences that span multiple tokens, making clean probability extraction technically challenging.

5.6.2 Confidence Score Calibration

In theory, classifiers trained with proper scoring rules are naturally calibrated [8, 39]. For example, recent work (*e.g.*, ConfTuner [91]) uses single-token probabilities as confidence without the need for calibration. The theory also applies to QRad, as the binary [yes] / [no] classification is trained with cross-entropy loss, a typical proper scoring rule. In practice, however, post-hoc calibration still helps classes with fewer training samples. Table 5.6 reports the Expected Calibration Error (ECE) before and after temperature scaling, showing improved calibration across classes. An additional ROC evaluation is available in Section 5.8.5.

Table 5.6: Expected Calibration Error (ECE) Before and After Calibration

Classes	wAVG	Enl.	Car.	L.O.	L.L.	Ede.	Con.	Pmn.	Ate.	Pmt.	P.E.	P.O.	Fra.	S.D.
Ratio ¹	-	0.62	0.54	0.63	0.13	0.44	0.38	0.22	0.45	0.07	0.38	0.17	0.31	0.47
ECE (Before) ²	0.18	0.21	0.22	0.17	0.34	0.19	0.15	0.25	0.15	0.39	0.17	0.31	0.10	0.05
ECE (After) ²	0.15	0.22	0.15	0.22	0.24	0.12	0.05	0.17	0.08	0.25	0.08	0.23	0.16	0.08

¹ **Ratio** is the percentage of positive samples, showing class imbalance in the MIMIC-CXR dataset. **wAVG** is the average of all classes weighted by their ratio.

² **Before** and **After** show the ECE improvements from temperature scaling calibration.

³ The disease class shorthands represent Enlarged Cardiomeastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, and Support Devices, respectively.

5.7 Threshold-Controllable Radiology Report Generation

5.7.1 Motivation

Image classifiers produce calibrated probabilities over a set of classes, supporting precise evaluation via ROC curves and sensitivity–specificity trade-offs, as well as integration with rule-based decision systems. Their outputs, however, are inherently constrained by predefined classes: a disease classifier cannot describe the anatomical location, extent, or underlying cause. Generative report models offer the complementary capability of producing expressive, open-vocabulary text that resembles clinical reporting practice, but typically lack calibrated class probabilities, making it difficult to quantitatively assess factual correctness or adjust the diagnostic operating points.

This tension between *structured evaluation* and *expressive flexibility* is particularly acute in radiology. Clinical deployment depends on operating-point selection: emergency triage favors high sensitivity to reduce missed findings, while confirmatory assessment favors high specificity to limit unnecessary follow-up [37]. In addition, regulatory pathways such as FDA device authorization rely on assessing performance at various operating points [37], which is not supported by conventional language models. Recent report generation methods [175, 62, 19, 189, 181, 149] produce a single, fixed report with no mechanism

to adjust the operating point. Related lines of work, including controllable text generation [25, 168], LLM confidence estimation [75, 164], confidence-guided generation [160], and structured reporting [124], address adjacent issues without linking report content to calibrated, class-specific thresholds.

We introduce *threshold-controllable radiology report generation*, a framework that aligns the diagnostic content of generated reports with the ROC behavior of a classification model. The intuition is that, if a report’s diagnostic decisions conform to a calibrated classifier’s operating point, the report inherits that classifier’s well-characterized sensitivity–specificity trade-off while retaining the descriptive richness of generative text. Varying the threshold then produces a family of reports spanning a range of diagnostic operating points.

5.7.2 Method

The framework combines three components: a perception model that produces calibrated per-class confidence scores; a report generator that produces detailed free-text reports and supports targeted follow-up queries (QRad); and an off-the-shelf LLM that rewrites the report so its stated diagnoses are consistent with the classifier’s threshold-adjusted decisions, while remaining grounded in the report’s clinical details. The classifier and report generator play complementary roles: the classifier provides precise, threshold-adjustable decisions, while the report provides details (*e.g.*, location, severity). The LLM acts as a *language-level conformer* that translates structured binary decisions into clinical language without making diagnostic judgments of its own.

Several choices are available for the perception model: (1) a standalone image classifier such as *MedImageInsight*, (2) the report generator’s image encoder with a fitted classification head, or (3) extracting class probabilities directly from QRad as described in Section 5.6. In this work, we use option (1) by fine-tuning *MedImageInsight* with DAUG’s image-text-class Hybrid Contrastive loss. For the LLM rewriter, we use a GPT-5 model deployed in a private, in-house Azure environment.

Given a chest X-ray I and a threshold τ , the inference pipeline proceeds in three steps:

1. **Classification.** The perception model produces confidence scores for K classes $\{P(C_i = 1 \mid I)\}_{i=1}^K$ and uses τ to partition them into a positive set (findings that the report should mention) and a negative set (findings that the report can omit or negate).
2. **Evidence collection.** QRad generates an initial report R . Missing medical conditions in R can be elicited by a follow-up VQA query, producing an evidence pool $\mathcal{E} = \{R\} \cup \{E_i\}_{i=1}^K$.
3. **LLM rewriting.** The LLM receives R , the positive/negative class lists derived from $\{\hat{y}_i\}$, and a structured instruction (provided in Section 5.8.6), and rewrites R to be *consistent* with the positive and negative classes: for each positive class, an existing description is preserved, or, if absent or contradicted, added using $\mathcal{E}$; for each negative class, any present mention is removed or negated; non-class content (*e.g.*, imaging quality) is preserved.

A practical challenge of this rewriting step is that the LLM must recognize the correspondence between structured class names and their varied natural-language manifestations in clinical text. For example, “Pleural Effusion” may surface as “fluid in the costophrenic angles” or “blunting of the costophrenic recesses”. The structured instruction therefore includes example templates for each class to guide this mapping.

5.7.3 Evaluation Protocols and Results

We evaluate on MIMIC-CXR [73] and sweep the threshold τ from 0.0 to 1.0 in steps of 0.1 to yield 11 operating points. The same τ is applied uniformly to all 13 disease classes (“No Finding” is excluded because it semantically overlaps with other classes). At each τ , the classifier’s confidence scores are binarized and the LLM rewrites QRad’s initial report accordingly.

To verify that the rewritten report faithfully reflects the intended classifications, we convert it back to binary class labels using VisualCheXbert [66], a BERT-based model that extracts disease mentions from free-text radiology reports across the same disease categories used by the classifier. It is also the labeler used to generate ground-truth class labels from the reference reports, ensuring a consistent evaluation framework. Comparing the extracted classes against the ground truth yields a sensitivity and specificity at each τ , producing one ROC point per threshold; connecting these points traces the ROC curve of the threshold-controlled reports. This closed-loop evaluation (classes to reports, and then back to classes) tests whether the threshold-controlled report faithfully conforms to the classifier’s ROC curve.

Figures 5.6 and 5.7 compare three conditions on the same axes: the *MedImageInsight* classifier ROC (blue) as the reference; the threshold-controlled report ROC (red); and the original QRad report (green cross) as a single operating point without threshold adjustment. We observe the following:

- **ROC conformance.** In Figure 5.6, across the four representative classes spanning common (Cardiomegaly), threshold-sensitive (Edema), hard (Atelectasis), and rare (Pleural Other, 1.3% prevalence) findings, the threshold-controlled report ROC closely tracks the classifier ROC with comparable AUC, and the concave shape confirms a monotonic sensitivity–specificity trade-off. The close overlap validates that the LLM translates threshold-adjusted decisions into clinical language without introducing diagnostic changes, enabled by QRad’s VQA-grounded evidence pool. The same pattern is observed in the full 13-class evaluation in Figure 5.7.
- **Improved factual correctness.** The ROC curve of the threshold-controlled report lies consistently *above* the original QRad report (the green cross): at matched specificity, sensitivity improves by 6.9%; at matched sensitivity, specificity improves by 20.7%. The classifier detects findings the generative model missed and suppresses hallucinated ones, yielding more accurate reports across the operating range.

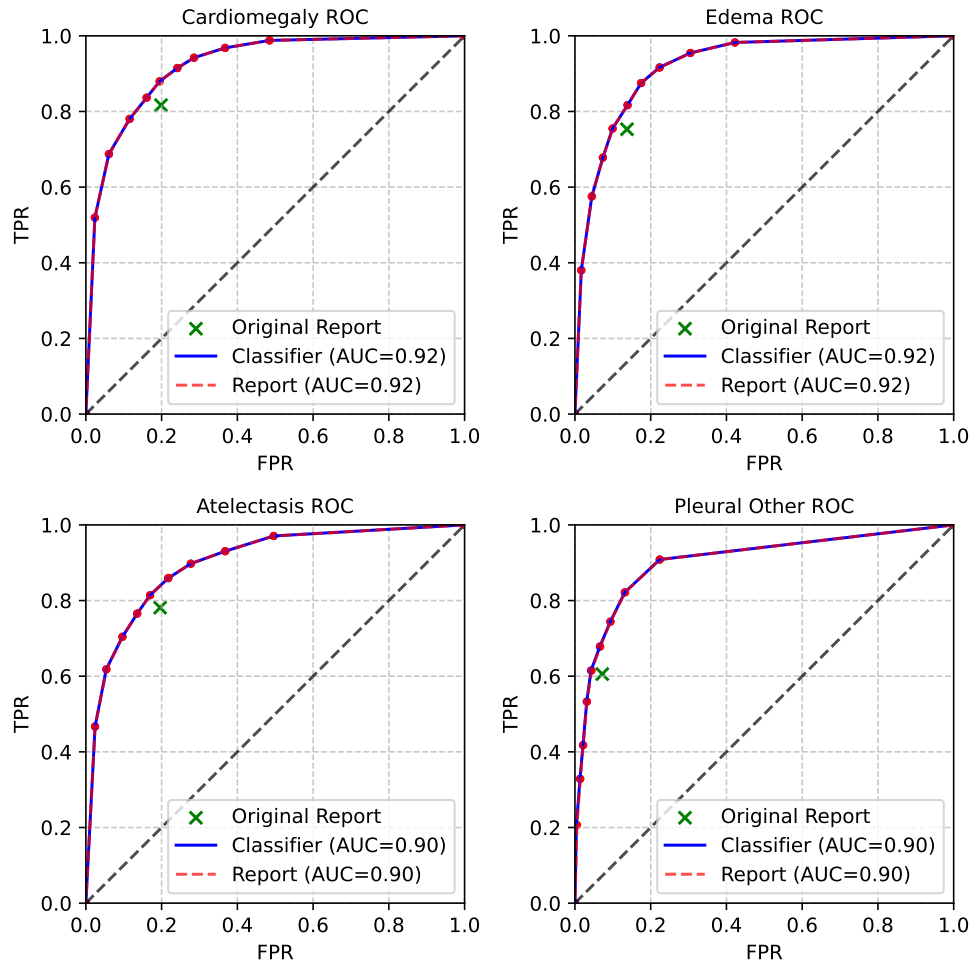


Figure 5.6: Threshold-controlled report generation on four representative classes. **Cardiomegaly**: common, both models perform well. **Edema**: threshold tuning is clinically valuable since mild cases often need no treatment. **Atelectasis**: challenging, requires fine visual details. **Pleural Other**: rare (1.3% prevalence). Across all four, the report ROC (red) closely tracks the classifier (blue), and the original report (green cross) falls below the curves.

- **Clinical utility.** Low thresholds prioritize sensitivity for screening or triage; high thresholds prioritize specificity for confirmatory assessment. The framework makes this trade-off explicit and tunable, replacing the fixed, opaque operating point of conventional report generators and offering a path toward ROC-based regulatory

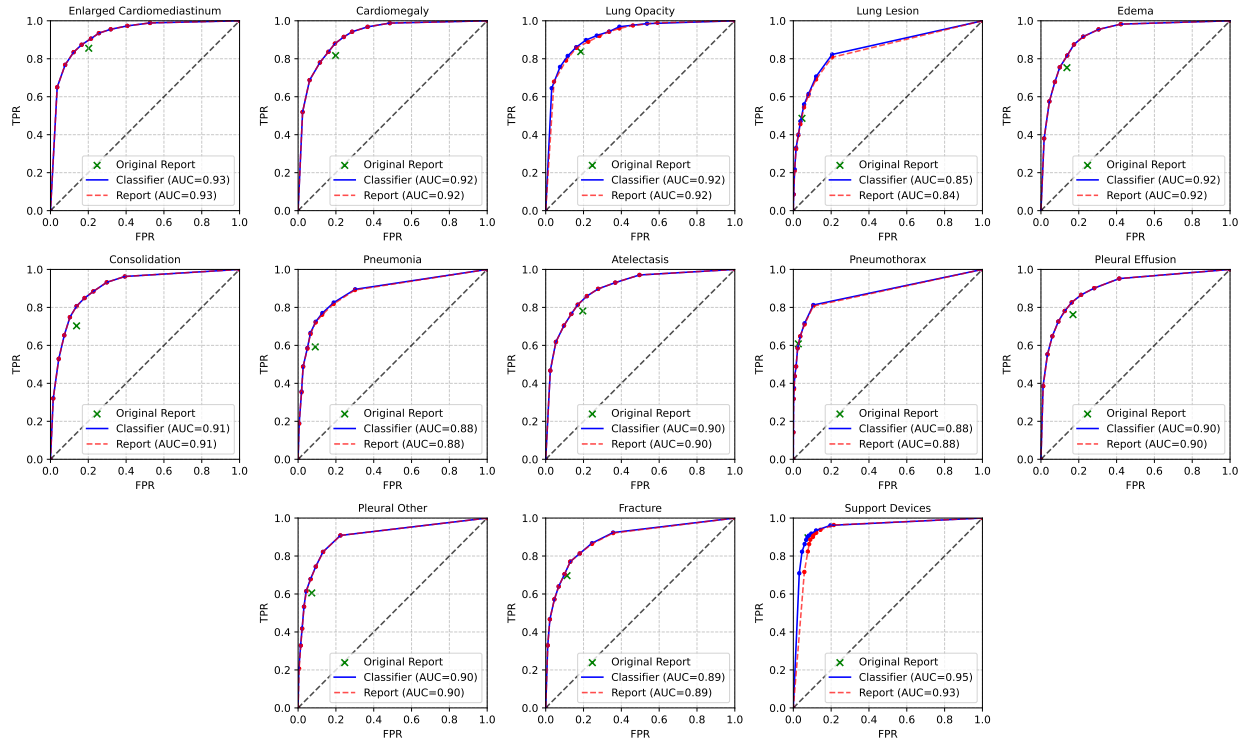


Figure 5.7: Threshold-controlled report generation evaluation on 13 disease classes. High overlap between classifier ROC (blue) and report ROC (red) confirms conformance. The original report (green cross) falls below the curves. VisualCheXbert labels are used as the ground truth.

evaluation of generated reports.

5.8 Additional Results and Discussion

5.8.1 UniRG-CXR: Reinforcement Fine-tuning for Report Generation

While QRad employs supervised fine-tuning (SFT) to optimize next-token prediction, SFT objectives are inherently misaligned with the evaluation metrics used to assess report quality, such as RadCliQ [172] and CheXprompt [175]. This *loss-metric mismatch* implies that minimizing the training loss does not directly maximize the metrics that matter for

clinical deployment. To address this gap, we contribute to UniRG-CXR [100], a reinforcement fine-tuning (RFT) framework that directly optimizes evaluation metrics as reward signals using Group Relative Policy Optimization (GRPO) [135]. UniRG-CXR adopts a two-stage pipeline consisting of SFT followed by RL: the SFT stage establishes a foundation for report generation, where QRad can serve as the SFT model, and the RL stage subsequently refines the model by optimizing a composite reward that aggregates rule-based (BLEU [119]), model-based (BERTScore [180], SemScore [139], RadGraph-F1 [65]), and LLM-based (CheXprompt [175]) metrics. The RL procedure first targets RadCliQ-oriented optimization for overall report quality, then incorporates an error-reduction step that explicitly minimizes clinical errors via CheXprompt. The combined SFT and RL strategy consistently outperforms either approach in isolation.

Trained on approximately 560,000 studies from more than 80 institutions spanning MIMIC-CXR [73], CheXpert Plus [14], ReXGradient [182], and IU [27], UniRG-CXR achieves state-of-the-art performance on the ReXrank [183] benchmark, surpassing prior models including MedVersa, MedGemma, and MAIRA-2 by substantial margins across all datasets and evaluation settings. Notably, UniRG-CXR yields improvements not only on the metrics it explicitly optimizes but also on metrics such as RaTEScore [186] that are excluded from the reward, suggesting that the RL framework promotes genuinely improved clinical reasoning rather than overfitting to specific objectives. The framework also substantially reduces clinical errors in generated reports relative to prior systems, validating the effectiveness of using evaluation metrics as rewards to bridge the loss–evaluation gap. The results of UniRG-CXR are included in Table 5.3.

5.8.2 Results on IU X-ray

Following recent studies [175, 62], we use the IU X-ray dataset [27] as a fully held-out evaluation set. All 3,198 frontal-view X-rays are used as the testing split, unseen during training. Results in Table 5.7 show that QRad generalizes well to unseen data despite using

Table 5.7: Report Generation Performance on IU-XRay

Model	CheXbert								RadGraph	BLEU		ROUGE		
	("uncertain" as <i>negative</i>)				("uncertain" as <i>positive</i>)									
	Micro-avg		Macro-avg		Micro-avg		Macro-avg			ER	(1)		(4)	(L)
	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5						
R2Gen [18]	-	-	13.6	-	-	-	-	-	-	32.5	5.9	25.3		
CvT2Dist. [111]	-	-	16.8	-	-	-	-	-	-	38.3	8.2	27.7		
RGRG [146]	-	-	18.0	-	-	-	-	-	-	26.6	6.3	18.0		
LLaVA-Rad [175]	53.5	-	-	-	-	-	-	-	-	-	-	25.3		
QRad	46.5	36.9	27.0	27.2	44.3	38.8	28.7	31.2	29.4	41.9	10.8	25.3		

a small model.

5.8.3 Comparison of Hyperparameter Configurations

Table 5.8 provides additional comparisons for hyperparameter configurations, including the data mixture ratio, the source of pseudo-labels for the classification QA data, and whether previous QA pairs are provided as input context. By comparing experiments from (a) to (f), we find:

- Performance is robust to data mixture ratios: (b) *vs.* (c).
- Using P+U as the positive label, which aligns with the “uncertain as positive” evaluation setting, leads to consistent performance gains across metrics. This is likely because uncertain labels correspond to diseases mentioned in prior studies but ambiguously stated in current reports: (f) *vs.* (c), (d).
- Providing previous QA pairs as context improves performance: (e) *vs.* (f).

Table 5.8: Comparison of Hyperparameter Configurations on MIMIC-CXR

Model	CheXbert								RadGraph	BLEU		ROUGE
	("uncertain" as <i>negative</i>)				("uncertain" as <i>positive</i>)					(1)	(4)	
	Micro-avg		Macro-avg		Micro-avg		Macro-avg					
	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5	F1-14	F1-5				
									ER			(L)
(a) Classification QA = 0%												
<i>median</i>	57.9	59.8	40.0	50.7	57.6	62.7	44.2	55.5	31.4	39.9	16.5	32.4
<i>ci_l</i>	56.8	58.3	38.1	48.9	56.6	61.3	42.6	53.9	30.8	39.3	16.0	31.8
<i>ci_h</i>	59.0	61.3	41.6	52.5	58.7	64.0	45.8	57.2	32.1	40.6	17.1	32.9
(b) Classification QA = 20%, Label={P}												
<i>median</i>	57.9	59.6	40.4	51.2	57.4	61.8	44.6	54.7	31.4	40.5	16.6	32.4
<i>ci_l</i>	56.8	57.8	38.8	49.1	56.4	60.5	43.1	53.2	30.7	39.8	16.1	31.8
<i>ci_h</i>	59.0	61.1	42.0	53.1	58.5	63.3	46.2	56.4	32.0	41.2	17.2	32.9
(c) Classification QA = 40%, Label={P}												
<i>median</i>	57.8	59.5	40.2	50.9	57.3	61.7	44.5	54.4	31.4	40.0	16.6	32.5
<i>ci_l</i>	56.7	58.0	38.6	48.9	56.4	60.2	43.0	52.9	30.8	39.3	16.1	31.9
<i>ci_h</i>	58.9	61.0	41.9	52.7	58.4	63.1	46.2	56.2	32.1	40.8	17.1	33.1
(d) Classification QA = 40%, Label={P, Random U}												
<i>median</i>	57.9	59.2	40.5	50.8	57.5	61.6	44.5	54.5	31.3	40.1	16.6	32.4
<i>ci_l</i>	57.0	57.7	38.8	48.8	56.5	60.3	42.9	53.0	30.7	39.4	16.1	31.9
<i>ci_h</i>	59.1	60.7	42.0	52.6	58.5	63.1	46.0	56.3	31.9	40.8	17.2	33.0
(e) Classification QA = 40%, Label={P, U}, w/o QA Context												
<i>median</i>	56.6	58.8	39.6	50.7	56.3	61.3	43.4	54.1	28.0	41.8	13.8	27.9
<i>ci_l</i>	55.5	57.4	38.2	48.7	55.4	59.9	42.1	52.6	27.5	41.3	13.5	27.5
<i>ci_h</i>	58.0	60.3	41.1	52.5	57.4	62.7	44.8	55.8	28.6	42.3	14.2	28.3
(f) Classification QA = 40%, Label={P, U}												
<i>median</i>	58.3	59.5	41.5	51.8	57.9	62.2	45.1	55.2	31.6	40.2	16.7	32.5
<i>ci_l</i>	57.3	57.9	39.8	49.7	56.9	60.8	43.7	53.6	30.9	39.4	16.2	32.0
<i>ci_h</i>	59.4	61.0	43.0	53.7	59.0	63.5	46.6	57.0	32.2	40.9	17.2	33.1

¹. The Classification QA ratio (0%, 20%, 40%) indicates their proportion in the training data.

². The Label field defines which CheXbert classes are mapped to positive classes in the Classification QA data: P (positive only), U (positive and uncertain), Random U (uncertain randomly used as positive)

³. QA Context indicates whether previous QA pairs are provided as input context. The w/o QA Context ablates this feature.

5.8.4 Qualitative Results

Qualitative examples illustrating the QRad pipeline are shown in Figure 5.8 and Figure 5.9. The figures highlight two aspects:

- **Intrinsic VQA capability.** The predicted answers are directly relevant to the input questions, demonstrating the model’s ability to perform visual question answering.

(a) QRad Predicted Question-Answer Pairs

- [Q1] "What is the position of the endotracheal tube?"
[A1] "The endotracheal tube is 3.5 cm above the carina."
- [Q2] "Where does the right-sided picc line terminate?"
[A2] "A right-sided picc line terminates in the mid svc."
- [Q3] "What is the course of the enteric tube?"
[A3] "An enteric tube courses below the diaphragm with the tip out of the field of view."
- [Q4] "Are there any changes in the lung opacities compared to previous studies?"
[A4] "Since the prior study, there has been improvement in the bilateral parenchymal opacities."
- [Q5] "What is the status of the lung volumes and the cardiomeastinal silhouette?"
[A5] "Lung volumes are low and the cardiomeastinal silhouette is enlarged, but stable."

(b) QRad Concatenated Report

- [Report] "The endotracheal tube is 3.5 cm above the carina. A right-sided picc line terminates in the mid svc. An enteric tube courses below the diaphragm with the tip out of the field of view. Since the prior study, there has been improvement in the bilateral parenchymal opacities. Lung volumes are low and the cardiomeastinal silhouette is enlarged, but stable."

Figure 5.8: Qualitative examples of QRad's Auto-VQA process. (a) shows the Auto-VQA process, where the model generates both questions and answers sequentially. (b) shows the final report generated by concatenating the predicted answers.

- **Factual correctness.** The model generates factually accurate answers, although there may be stylistic differences such as sentence structure or order.

- [GT-Q1] "Are there any support devices present?"
 - [GT-A1] "An endotracheal tube, NG tube, and right upper extremity PICC with its tip at the cavoatrial junction are unchanged."
 - [PD-A1] "An endotracheal tube is in place with the tip terminating 3.5 cm above the carina. A nasogastric tube is seen coursing below the diaphragm with the tip terminating in the left upper quadrant. A right picc is unchanged in position with the tip terminating in the mid-to-low svc."
- [GT-Q2] "Is there any change in lung opacity?"
 - [GT-A2] "There is no change in left lower lobe opacity."
 - [PD-A2] "There has been interval improvement in the right basilar opacity."
- [GT-Q3] "Are there signs of pleural effusion or pneumothorax?"
 - [GT-A3] "There is no large pleural effusion, or pneumothorax."
 - [PD-A3] "No pleural effusion or pneumothorax is detected."
- [GT-Q4] "What is the condition of the cardiac silhouette and mediastinal contours?"
 - [GT-A4] "The cardiac silhouette remains moderately enlarged, mediastinal contours are notable for calcification of the aortic arch."
 - [PD-A4] "The cardiac silhouette is mildly enlarged but stable, the mediastinal contours are prominent but stable."

Figure 5.9: Qualitative examples of QRad-generated answers using oracle questions. [GT- Q_n]: oracle questions that exactly match the reference report. [GT- A_n]: ground-truth answers from the reference report. [PD- A_n]: QRad’s predicted answers for each question. This example shows the Answer Generator’s reliability. The module produces factually correct answers that are closely aligned with the ground-truth sentences, showing the effectiveness of the Captioning-to-VQA reframing approach.

5.8.5 ROC-Based Evaluation for Regulatory Validation

To the best of our knowledge, QRad is the first report generation model to produce class probability scores for defined diseases directly from its text generation components. Unlike multi-task models that use separate modules for classification and text generation, QRad generates both outputs from the same component. This design enables the evaluation of

confidence scores to directly reflect the report generation model’s intrinsic classification capability.

The class probabilities are used to generate Receiver Operating Characteristic (ROC) curves, which are commonly used in FDA approval studies for diagnostic systems. As shown in Figure 5.10, this confidence-based evaluation provides more granular insights into clinical utility, such as the characteristics of the sensitivity and specificity trade-off. This is especially valuable because clinical applications often have different costs for false positives and false negatives.

From Figure 5.10, we observe that QRad performs reliably on classes such as Enlarged Cardiomedastinum, Cardiomegaly, and Lung Opacity, but is less reliable on Pleural Other, Fracture, and Pneumothorax. We attribute this difference to two main factors. First, conditions such as fracture are challenging because they require detecting subtle details and are rare in the dataset. Second, some classes (*e.g.*, Pleural Other) are umbrella categories that aggregate many rare, long-tail disease names, which cannot be fully covered in our LLM rewriting instruction.

5.8.6 LLM Instructions for Threshold-Controllable Report Generation

Figure 5.11 shows the instruction provided to the LLM rewriter for threshold-controllable radiology report generation. The instruction consists of three parts: (1) a task description specifying the rewriting objective, (2) a template dictionary providing examples of positive and negative text segments for CheXbert disease classes, and (3) examples illustrating the rewriting rules for different cases. At inference, the placeholder {INPUT} is replaced with the original report and the classification results.

System Instruction:

You are given an input Chest X-ray radiology report and the classification result of 14 CheXbert classes, which includes both positive and negative diseases.

Modify the report so that it matches each class in the CheXbert classification results. Add text segments or sentences to the report if the class is not mentioned in the report, or change/remove sentences in the report so that it agrees with the classification.

Your modified report should be a valid Findings section of a Chest X-ray radiology report, with information such as the location and severity of medical findings based on the original report.

When overwriting the report with the positive and negative classes, consider using text segments from the following templates, and fit them into the report as appropriate.

Template Dictionary (abbreviated; full version includes all 14 classes):

```
{
  "Cardiomegaly": {
    "positive": ["The heart is moderately enlarged.",
                "Mild enlargement of the cardiac silhouette.", ...],
    "negative": ["The cardiac silhouette is normal.",
                "The heart is normal in size.", ...]
  },
  "Pneumothorax": {
    "positive": ["small apical pneumothorax", ...],
    "negative": ["no pneumothorax", ...]
  }, ...
}
```

Rewriting Rules:

Example 1: If "Fracture" is a positive class:

- If the original report indicates Fracture is positive, do nothing.
- If the original report indicates Fracture is negative, replace that sentence with a positive text segment.
- If the original report doesn't mention Fracture, choose a positive text segment and add it to the report.

Example 2: If "Support Devices" is a negative class:

```

- If the original report doesn't indicate the existence of any support
device, do nothing.
- If the original report mentions any kind of support device exists,
remove it.

The input and output report are the Findings section of a radiology report.
Make sure all positive classes are identifiable in the modified report.
Return the modified version as JSON with "output" as the key.

Only return the JSON without any explanation. Double check and make sure
you did not forget any classes in the lists, and did not make unnecessary
changes:

{INPUT}

```

Figure 5.11: The LLM rewriting instruction for threshold-controlled report generation. The LLM receives the original report, binary classification results, and a template dictionary of example text segments for CheXbert disease classes.

5.8.7 Ethical Considerations

Medical datasets often contain sensitive patient information. To ensure the ethical use of such data, this study adheres to strict guidelines. All participants who accessed the MIMIC-CXR dataset, including the authors and radiologists involved in this research, completed the required onboarding process through PhysioNet⁴. For the IU X-ray dataset, we complied with the license⁵.

To maintain compliance with PhysioNet’s policy on the use of large language models, we use a private, in-house deployment of GPT-5. This approach ensures that no sensitive information is shared with external parties.

Furthermore, to protect patient privacy, X-ray images presented in this dissertation

⁴MIMIC-CXR on PhysioNet: <https://physionet.org/content/mimic-cxr/2.0.0/>

⁵IU X-ray dataset license: <https://creativecommons.org/licenses/by-nc-nd/4.0/>

were selected from open, compliance-free sources, ensuring that no identifiable patient information is disclosed.

5.9 Summary

This chapter introduces QRad, a novel framework for radiology report generation that reformulates long-form text generation as an Auto-VQA process. By unifying image-class and image-text labels into QA training data, QRad improves factual correctness, enables user interaction, and supports probability-based evaluation such as ROC analysis. The Captioning-to-VQA reframing further supports threshold-based control and regulatory validation, improving the clinical utility of radiology report generation.

In contrast to earlier chapters that explore how generative models enhance perception tasks, this chapter examines the inverse direction by incorporating perception-oriented capabilities into a generative model. By casting image classification as a single-token text generation step, QRad demonstrates how perception mechanisms can be embedded into text generation to improve evaluation, controllability, and clinical utility. Together with the previous chapters, this work completes a bidirectional perspective on bridging perception and generative models.

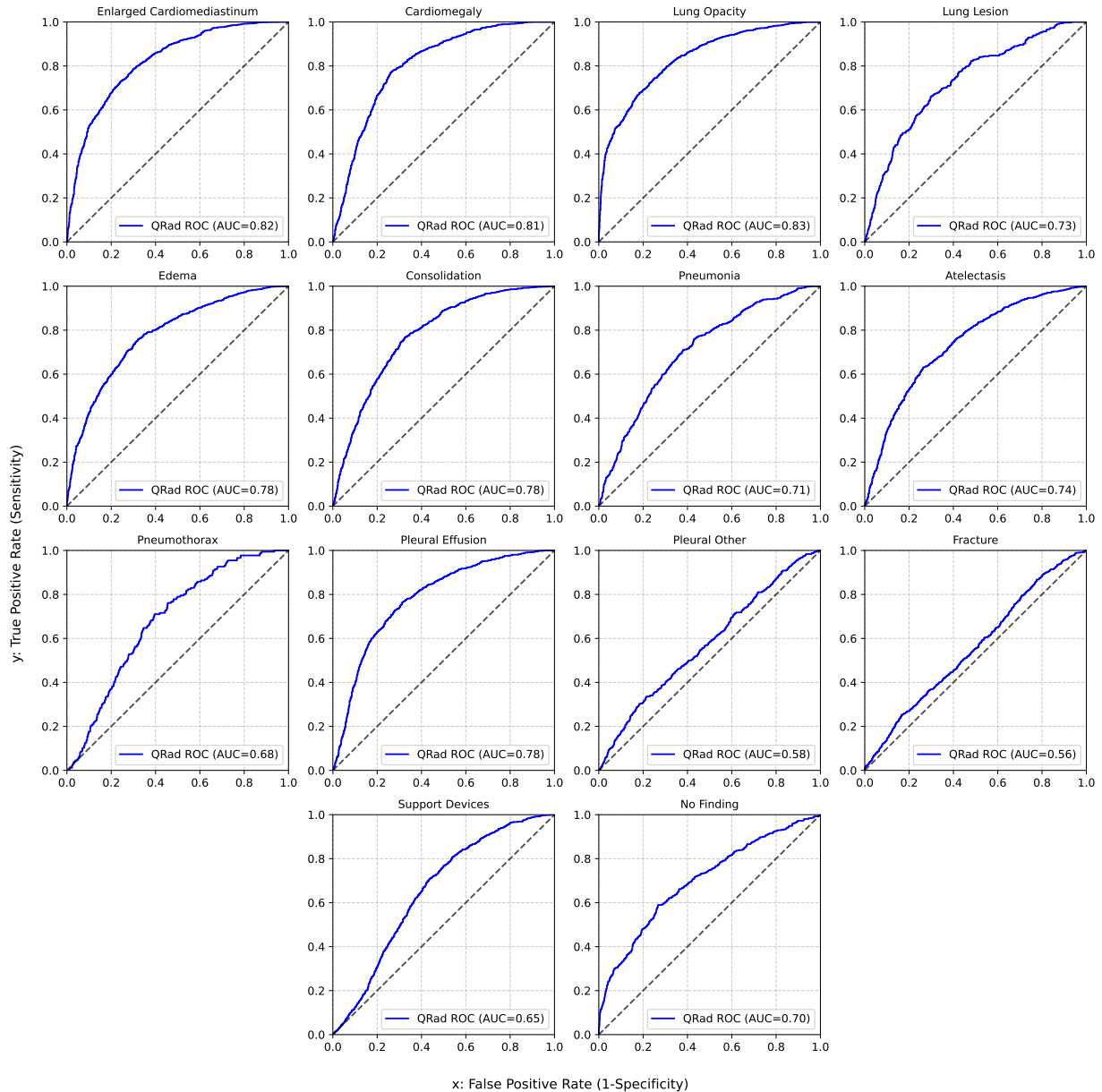


Figure 5.10: Receiver Operating Characteristic (ROC) curves for QRad across multiple disease classes. The x-axis shows the False Positive Rate (1-specificity), and the y-axis shows the True Positive Rate (sensitivity). Each curve illustrates the trade-off between sensitivity and specificity. The ROC analysis factors out the noise introduced by CheXbert, which converts reports into classes for evaluation, and enables a more nuanced assessment of QRad's clinical utility across different operating points and disease categories.

Chapter 6

CONCLUSION

This dissertation investigates how perception and generative models, two traditionally separate paradigms in machine learning, can be combined to improve vision-language tasks in terms of performance, evaluation, and practical utility. The central insight is that the two paradigms can be complementary for vision-language understanding despite differing in the direction of the mapping they learn between images and labels. Across multiple methods and domains, this dissertation shows that bridging these paradigms is both feasible and effective.

First, we show that generative models can improve perception through multiple mechanisms. For Human-Object Interaction recognition, an image captioning model supervises a classifier through pseudo-label distillation, enabling the recognition task to be learned without human annotations. For medical image understanding, a diffusion model generates class-specific abnormality heatmaps that augment the visual input, directing the model’s attention to clinically significant regions and improving both image classification and retrieval. In radiology report generation, we further show that perception-oriented capabilities enhance generation. By reframing the task as an Auto-VQA process, QRad enables a text generation model to support class confidence extraction, threshold control, and ROC-based evaluation. Taken together, these studies show that generation can supply richer supervision for perception, while perception can bring verifiability and controllability to generation. Bridging the two provides a design principle for building vision-language models that are at once more capable and more reliable.

One limitation of this work is that some experiments were conducted before large language models became the dominant paradigm, and therefore do not build on the strongest

models available today. Nevertheless, the underlying principle remains applicable, as bridging perception and generation is independent of a specific model architecture. Looking forward, we envision agentic frameworks as a natural extension of this work, where expert models serve as specialized tools that an agent orchestrates to support clinical decision making. Trained through supervised fine-tuning and reinforcement learning, such an agent could learn from clinical workflows to weigh the reliability of different expert models. This direction moves beyond static model integration toward adaptive systems that more closely reflect how clinicians combine multiple sources of evidence in practice.

BIBLIOGRAPHY

- [1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.
- [2] 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.
- [3] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.
- [4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- [5] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, pages 3686–3693, 2014.
- [6] Spyridon Bakas, Chiharu Sako, Hamed Akbari, Michel Bilello, Aristeidis Sotiras, Gaurav Shukla, Jeffrey D Rudie, Natali Flores Santamaría, Anahita Fathi Kazerooni, Sarthak Pati, et al. The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics. *Scientific data*, 9(1):453, 2022.
- [7] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita

- Sharma, Felix Meissen, et al. MAIRA-2: Grounded Radiology Report Generation. *arXiv preprint arXiv:2406.04449*, 2024.
- [8] Jaroslaw Blasiok, Parikshit Gopalan, Lunjia Hu, and Preetum Nakkiran. When does optimizing a proper loss yield calibration? *Advances in Neural Information Processing Systems*, 36:72071–72095, 2023.
- [9] Brent Burbridge. *Undergraduate diagnostic imaging fundamentals*. Distance Education Unit, University of Saskatchewan, 2017.
- [10] Daniel J Cao, Casey Hurrell, and Michael N Patlas. Current status of burnout in canadian radiology. *Canadian Association of Radiologists Journal*, 74(1):37–43, 2023.
- [11] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.
- [12] Chris Carr, Felipe Kitamura, J Kalpathy-Cramer, J Mongan, K Andriole, M Vazirabad, M Riopel, R Ball, and S Dane. RSNA Screening Mammography Breast Cancer Detection. 2022.
- [13] G Challenge. Peking University International Competition on Ocular Disease Intelligent Recognition (odir-2019), 2019.
- [14] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.
- [15] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1017–1025, 2015.

- [16] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In *ECCV*, 2020.
- [17] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*, 2022.
- [18] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, November 2020.
- [19] Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.
- [20] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the 2017 International symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172, 2018.
- [21] Noel C. F. Codella, Ying Jin, Shrey Jain, Yu Gu, Ho Hin Lee, Asma Ben Abacha, Alberto Santamaria-Pang, Will Guyman, Naiteek Sangani, Sheng Zhang, Hoifung Poon, Stephanie Hyland, Shruthi Bannur, Javier Alvarez-Valle, Xue Li, John Garrett, Alan McMillan, Gaurav Rajguru, Madhu Maddi, Nilesh Vijayrania, Rehaan Bhimai, Nick Mecklenburg, Rupal Jain, Daniel Holstein, Naveen Gaur, Vijay Aski, Jenq-Neng Hwang, Thomas Lin, Ivan Tarapov, Matthew Lungren, and Mu Wei. Medimageinsight: An open-source embedding model for general domain medical imaging, 2024.

- [22] Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen W. Dusza, David A. Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael A. Marchetti, Harald Kittler, and Allan Halpern. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International skin imaging collaboration (ISIC). *CoRR*, abs/1902.03368, 2019.
- [23] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C. Halpern, Susana Puig, and Josep Malvehy. BCN20000: Dermoscopic lesions in the wild, 2019.
- [24] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.
- [25] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.
- [26] Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandiyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, et al. Automated structured radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26813–26829, 2025.
- [27] 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.
- [28] 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.
- [29] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [30] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. DaViT: Dual attention vision transformers. In *European conference on computer vision*, pages 74–92. Springer, 2022.
- [31] Leizhen Dong, Zhimin Li, Kunlun Xu, Zhijun Zhang, Luxin Yan, Sheng Zhong, and Xu Zou. Category-aware transformer network for better human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19538–19547, 2022.
- [32] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. *arXiv preprint arXiv: 2111.02387*, 2021.
- [33] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, , and the CAMELYON16 Consortium. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, 12 2017.
- [34] 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.
- [35] Michela Antonelli et al. The Medical Segmentation Decathlon. *Nature Communications*, 13, 2022.

- [36] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 51–67, 2018.
- [37] US Food. Drug administration. statistical guidance on reporting results from studies evaluating diagnostic tests-guidance for industry and fda staff. *Food and Drug Administration, Center for Devices and Radiological Health Diagnostic Devices Branch, Division of Biostatistics, Office of Surveillance and Biometrics*, 2007.
- [38] Maayan Frid-Adar, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 289–293. IEEE, 2018.
- [39] Christian Fröhlich and Robert C Williamson. Scoring rules and calibration for imprecise probabilities. *arXiv preprint arXiv:2410.23001*, 2024.
- [40] Dhakshinamoorthy Ganeshan, Andrew B Rosenkrantz, Roland L Bassett Jr, Lori Williams, Leon Lenchik, and Wei Yang. Burnout in academic radiologists in the united states. *Academic radiology*, 27(9):1274–1281, 2020.
- [41] Chen Gao, Jiarui Xu, Yuliang Zou, and Jia-Bin Huang. Drg: Dual relation graph for human-object interaction detection. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020.
- [42] Chen Gao, Yuliang Zou, and Jia-Bin Huang. ican: Instance-centric attention network for human-object interaction detection. *arXiv preprint arXiv:1808.10437*, 2018.
- [43] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

- [44] Cree M. Gaskin, S. Lowell Kahn, J. Christopher Bertozzi, and Paul M. Bunch. *Skeletal Development of the Hand and Wrist: A Radiographic Atlas and Digital Bone Age Companion*. Oxford University Press, 11 2011.
- [45] Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 33–44, 2017.
- [46] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8359–8367, 2018.
- [47] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1080–1088, 2015.
- [48] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.
- [49] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [50] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. Stack-captioning: Coarse-to-fine learning for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- [51] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 503–519, 2018.

- [52] Safwan S Halabi, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Artem B Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, et al. The RSNA Pediatric Bone Age Machine Learning Challenge. *Radiology*, 290(2):498–503, 2019.
- [53] Tampere University Hamad Medical Corporation Heart Hospital, Qatar University. HMC-QU Dataset, 2021.
- [54] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.
- [55] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [56] Gregory Holste, Song Wang, Ziyu Jiang, Thomas C. Shen, George Shih, Ronald M. Summers, Yifan Peng, and Zhangyang Wang. *Long-Tailed Classification of Thorax Diseases on Chest X-Ray: A New Benchmark Study*, page 22–32. Springer Nature Switzerland, 2022.
- [57] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *European Conference on Computer Vision*, pages 584–600. Springer, 2020.
- [58] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, 2021.
- [59] Xinyue Hu, Lin Gu, Kazuma Kobayashi, Liangchen Liu, Mengliang Zhang, Tatsuya Harada, Ronald M Summers, and Yingying Zhu. Interpretable medical image visual question answering via multi-modal relationship graph learning. *Medical Image Analysis*, 97:103279, 2024.

- [60] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.
- [61] Chen Huang, Shuangfei Zhai, Walter Talbott, Miguel Bautista Martin, Shih-Yu Sun, Carlos Guestrin, and Josh Susskind. Addressing the loss-metric mismatch with adaptive loss alignment. In *International conference on machine learning*, pages 2891–2900. PMLR, 2019.
- [62] 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.
- [63] ASM Iftekhar, Hao Chen, Kaustav Kundu, Xinyu Li, Joseph Tighe, and Davide Modolo. What to look at and where: Semantic and spatial refined transformer for detecting human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5353–5363, 2022.
- [64] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, 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.
- [65] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021.
- [66] Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh

- Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 105–115, 2021.
- [67] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.
- [68] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2607–2615, 2024.
- [69] Ying Jin, Yinpeng Chen, Jianfeng Wang, Lijuan Wang, Jenq-Neng Hwang, and Zicheng Liu. Zero-shot human-object interaction (hoi) classification by bridging generative and contrastive image-language models. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1970–1974. IEEE, 2023.
- [70] Ying Jin, Yinpeng Chen, Lijuan Wang, Jianfeng Wang, Pei Yu, Lin Liang, Jenq-Neng Hwang, and Zicheng Liu. The overlooked classifier in human-object interaction recognition. *arXiv preprint arXiv:2203.05676*, 2022.
- [71] Ying Jin, Noel C Codella, Yanbo Xu, Yu Gu, Mu Wei, Haoquan Fang, Thomas Lin, Paul Vozila, and Jenq-Neng Hwang. Qrad: Enhancing radiology report generation by captioning-to-vqa reframing. In *The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance*.
- [72] Ying Jin, Zhuoran Zhou, Haoquan Fang, and Jenq-Neng Hwang. Daug: Diffusion-based channel augmentation for radiology image retrieval and classification. In

Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond, 2024.

- [73] 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.
- [74] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.
- [75] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- [76] Daniel Kermany. Labeled Optical Coherence Tomography (OCT) for Classification, 2017.
- [77] Daniel Kermany, Kang Zhang, and Michael Goldbaum. Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images, 2018.
- [78] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.
- [79] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *European Conference on Computer Vision*, pages 498–514. Springer, 2020.

- [80] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. Hotr: End-to-end human-object interaction detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 74–83, June 2021.
- [81] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19578–19587, 2022.
- [82] Dong-Jin Kim, Xiao Sun, Jinsoo Choi, Stephen Lin, and In So Kweon. Detecting human-object interactions with action co-occurrence priors. In *European Conference on Computer Vision*, pages 718–736. Springer, 2020.
- [83] Wonjae Kim, Bokyoung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.
- [84] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [85] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [86] Thomas Martin Lehmann, Henning Schubert, Daniel Keysers, Michael Kohnen, and Berthold B Wein. The IRMA code for unique classification of medical images. In *Medical Imaging 2003: PACS and Integrated Medical Information Systems: Design and Evaluation*, volume 5033, pages 440–451. SPIE, 2003.
- [87] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*, 2020.

- [88] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLava-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *Advances in Neural Information Processing Systems*, 36, 2024.
- [89] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient self-supervised vision transformers for representation learning. *arXiv preprint arXiv:2106.09785*, 2021.
- [90] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.
- [91] Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. Conftuner: Training large language models to express their confidence verbally. *arXiv preprint arXiv:2508.18847*, 2025.
- [92] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Mingyang Chen, Ze Ma, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*, 2019.
- [93] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 382–391, 2020.
- [94] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020.

- [95] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. *arXiv preprint arXiv:2203.13954*, 2022.
- [96] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [97] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [98] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019.
- [99] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- [100] Qianchu Liu, Sheng Zhang, Guanghui Qin, Yu Gu, Ying Jin, Sam Preston, Yanbo Xu, Sid Kiblawi, Wen-wai Yim, Tim Ossowski, et al. Scaling medical imaging report generation with multimodal reinforcement learning. *arXiv preprint arXiv:2601.17151*, 2026.
- [101] Yang Liu, Qingchao Chen, and Andrew Zisserman. Amplifying key cues for human-object-interaction detection. In *European Conference on Computer Vision*, pages 248–265. Springer, 2020.
- [102] Hector Lopez-Cardona. Chest x-ray review: Abcde. *Radiopaedia*, 2023.

- [103] Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar. Relvit: Concept-guided vision transformer for visual relational reasoning. *arXiv preprint arXiv:2204.11167*, 2022.
- [104] K Scott Mader. The Lung Image Database Consortium image collection (LIDC-IDRI). *IEEE Dataport*, 2021.
- [105] Arun Mallya and Svetlana Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *European Conference on Computer Vision*, pages 414–428. Springer, 2016.
- [106] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, pages 570–576, 1998.
- [107] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
- [108] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. *arXiv preprint arXiv:2010.10042*, 2020.
- [109] Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yucheng Tang, Pengfei Guo, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. *arXiv preprint arXiv:2411.12915*, 2024.
- [110] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *European Conference on Computer Vision*, pages 167–184. Springer, 2022.
- [111] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144:102633, 2023.

- [112] Aaron Nicolson, Jinghui Liu, Jason Dowling, Anthony Nguyen, and Bevan Koopman. e-health csiro at rrg24: Entropy-augmented self-critical sequence training for radiology report generation. *arXiv preprint arXiv:2408.03500*, 2024.
- [113] Aaron Nicolson, Shengyao Zhuang, Jason Dowling, and Bevan Koopman. The impact of auxiliary patient data on automated chest x-ray report generation and how to incorporate it. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–203, 2025.
- [114] Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. *arXiv preprint arXiv:2102.09777*, 2021.
- [115] Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 374–390, 2024.
- [116] Övgü Özdemir and Erdem Akagündüz. Enhancing visual question answering through question-driven image captions as prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1562–1571, 2024.
- [117] Andre GC Pacheco, Gustavo R Lima, Amanda S Salomao, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio CR Alves Jr, José GM Esgario, Alana C Simora, Pedro BC Castro, et al. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in brief*, 32:106221, 2020.
- [118] Ankit Pal, Jung-Oh Lee, Xiaoman Zhang, Malaikannan Sankarasubbu, Seunghyeon Roh, Won Jung Kim, Meesun Lee, and Pranav Rajpurkar. Rexvqa: A large-scale

- visual question answering benchmark for generalist chest x-ray understanding. *arXiv preprint arXiv:2506.04353*, 2025.
- [119] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.
- [120] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021.
- [121] Jay R Parikh, Darcy Wolfman, Claire E Bender, and Elizabeth Arleo. Radiologist burnout according to surveyed radiology practice leaders. *Journal of the American College of Radiology*, 17(1):78–81, 2020.
- [122] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*, 2018.
- [123] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology Objects in Context (ROCO): A Multimodal Image Dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018.
- [124] Chantal Pellegrini, Matthias Keicher, Ege Özsoy, and Nassir Navab. Rad-restruct: A novel vqa benchmark and method for structured radiology reporting. In *International*

- Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 409–419. Springer, 2023.
- [125] Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*, 2023.
- [126] Xian Qu, Changxing Ding, Xingao Li, Xubin Zhong, and Dacheng Tao. Distillation using oracle queries for transformer-based human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19558–19567, 2022.
- [127] 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. *arXiv preprint arXiv:2103.00020*, 2021.
- [128] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training.
- [129] Mercy Ranjit, Shaury Srivastav, and Tanuja Ganu. Radphi-3: Small language models for radiology. *arXiv preprint arXiv:2411.13604*, 2024.
- [130] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- [131] Abi Rimmer. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, 359, 2017.
- [132] C Roche, E Bonaccio, and J Filippini. The Cancer Genome Atlas Sarcoma Collection (TCGA-SARC) (Version 3) [Data set]. The Cancer Imaging Archive. *The Cancer Imaging Archive*, 2016.

- [133] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.
- [134] Francesco Dalla Serra, Patrick Schrempf, Chaoyang Wang, Zaiqiao Meng, Fani Deligianni, and Alison Q O’Neil. Grounding chest x-ray visual question answering with generated radiology reports. *arXiv preprint arXiv:2505.16624*, 2025.
- [135] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
- [136] Liyue Shen, Serena Yeung, Judy Hoffman, Greg Mori, and Li Fei-Fei. Scaling human-object interaction recognition through zero-shot learning. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1568–1576. IEEE, 2018.
- [137] Hoo-Chang Shin, Neil A Tenenholtz, Jameson K Rogers, Christopher G Schwarz, Matthew L Senjem, Jeffrey L Gunter, Katherine P Andriole, and Mark Michalski. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 1–11. Springer, 2018.
- [138] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.
- [139] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations

- for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.
- [140] Robin Smithuis and Delden Otto. Chest x-ray - basic interpretation. *Radiology Assistant*, 2022.
- [141] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- [142] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 55(13s):1–40, 2023.
- [143] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*, 2021.
- [144] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. A Benchmark for Automatic Visual Classification of Clinical Skin Disease Images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI 14*, pages 206–222. Springer, 2016.
- [145] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10410–10419, June 2021.
- [146] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023.
- [147] Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, et al.

- Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608, 2025.
- [148] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.
- [149] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards Generalist Biomedical AI. *NEJM AI*, 1(3):AIoa2300138, 2024.
- [150] Tom van Sonsbeek and Marcel Worring. X-tra: Improving chest x-ray tasks with cross-modal retrieval augmentation. In *International Conference on Information Processing in Medical Imaging*, pages 471–482. Springer, 2023.
- [151] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.
- [152] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.
- [153] Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. UFO: A unified transformer for vision-language representation learning. *arXiv preprint arXiv:2111.10023*, 2021.
- [154] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022.
- [155] Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report generation. In *Computer Vision–ECCV 2022: 17th European Conference*,

- Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 563–579. Springer, 2022.
- [156] Suchen Wang, Yueqi Duan, Henghui Ding, Yap-Peng Tan, Kim-Hui Yap, and Junsong Yuan. Learning transferable human-object interaction detector with natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2022.
- [157] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, 2020.
- [158] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9049–9058, 2018.
- [159] Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, et al. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7246–7255, 2024.
- [160] Yixin Wang, Zihao Lin, Jianzong He, Min Zhang, Ling Wang, Shuwen Li, Jingyu Li, S. Kevin Zhou, and Zhiqiang He. Trust it or not: Confidence-guided automatic radiology report generation. *Neurocomputing*, 2024.
- [161] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022.

- [162] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.
- [163] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 35–45. Springer, 2022.
- [164] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations*, 2024.
- [165] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.
- [166] Benjamin Yan, Ruochen Liu, David E Kuo, Subathra Adithan, Eduardo Pontes Reis, Stephen Kwak, Vasantha Kumar Venugopal, Chloe P O’Connell, Agustina Saenz, Pranav Rajpurkar, et al. Style-aware radiology report generation with radgraph and few-shot prompting. *arXiv preprint arXiv:2310.17811*, 2023.
- [167] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022.
- [168] Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of*

- the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, 2021.
- [169] Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024.
- [170] Yanzhi Yi, Hangyu Deng, and Jinglu Hu. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994, 2020.
- [171] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659, 2016.
- [172] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.
- [173] Yang Yu, Peng Hu, Jie Lin, and Pavitra Krishnaswamy. Multimodal multitask deep learning for x-ray image retrieval. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 603–613. Springer, 2021.
- [174] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

- [175] Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, et al. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108, 2025.
- [176] Aixi Zhang, Yue Liao, Si Liu, Miao Lu, Yongliang Wang, Chen Gao, and Xiaobo Li. Mining the benefits of two-stage and one-stage hoi detection. *arXiv preprint arXiv:2108.05077*, 2021.
- [177] Chongsheng Zhang, George Almpanidis, Gaojuan Fan, Binqun Deng, Yanbo Zhang, Ji Liu, Aouaidjia Kamel, Paolo Soda, and João Gama. A systematic review on long-tailed learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.
- [178] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13319–13327, October 2021.
- [179] 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.
- [180] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
- [181] Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond SL Ho. Libra: Leveraging temporal images for biomedical radiology analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17275–17303, 2025.
- [182] Xiaoman Zhang, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports, 2025.

- [183] Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. Rexrank: A public leaderboard for ai-powered radiology report generation. *arXiv preprint arXiv:2411.15122*, 2024.
- [184] Yan Zhang, Jiaqing Lin, Miao Zhang, Kui Xiao, Xiaoju Hou, Yue Zhao, and Zhifei Li. Scra-vqa: Summarized caption-rerank for augmented large language models in visual question answering. *arXiv preprint arXiv:2509.20871*, 2025.
- [185] Yong Zhang, Weihua Ou, Jiacheng Zhang, and Jiaxin Deng. Category supervised cross-modal hashing retrieval for chest x-ray and radiology reports. *Computers & Electrical Engineering*, 98:107673, 2022.
- [186] WeiKe Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. *arXiv preprint arXiv:2406.16845*, 2024.
- [187] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13234–13243, June 2021.
- [188] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19568–19577, 2022.
- [189] Hong-Yu Zhou, Julián Nicolás Acosta, Subathra Adithan, Suvrankar Datta, Eric J Topol, and Pranav Rajpurkar. Medversa: A generalist foundation model for medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.
- [190] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Cheng-guang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object

interaction detection with hoi transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11825–11834, June 2021.