UNIVERSITY
OF TECHNOLOGY
SYDNEY

Exploring Clinical Knowledge to Enhance Deep Learning Models for Medical Report Generation

by Mingjie Li

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Professor Xiaojun Chang

University of Technology Sydney
Faculty of Engineering and Information Technology

February 2023

Certificate of Original Ownership

I, *Mingjie Li* declare that this thesis, is submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the *School of Computer Science*, *Faculty of Engineering and Information Technology* at the University of Technology Sydney, Australia.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Production Note:

SIGNATURE: Signature removed prior to publication.

DATE: 18th February, 2023

PLACE: Sydney, Australia

ABSTRACT

Automatic generation of long and coherent medical reports regarding the given medical images (*e.g.* Chest X-ray and Fundus Fluorescein Angiography (FFA)) has great potential to support clinical practice. Researchers have explored advanced methods, especially deep learning, from computer vision and natural language processing for the generation of readable medical reports. However, when writing a report, experts make inferences with prior clinical knowledge. Not surprisingly, existing methods with insufficient medical knowledge find it hard to achieve comparable promising performances in generic image caption fields since even researchers without a medical background cannot understand those images thoroughly, either. Thus, this thesis mainly investigates how to explore clinical knowledge to enhance deep learning models for automatic report generation. The thesis first explores knowledge by mimicking radiologists' working patterns and utilizes such knowledge to guide an encoder-decoder framework to generate accurate reports. Since medical decisions may lead to life-or-death consequences, a reliable rationale for interpretation is also excepted, along with accurate prediction. However, existing medical report generation (MRG) benchmarks lack both explainable annotations and reliable evaluation tools, also hindering the current research advances. This thesis then proposes an explainable and reliable MRG benchmark based on FFA Images and Reports (FFA-IR). Based on the FFA-IR, the thesis extracts structural information from clinical recorded reports and explores such clinical knowledge to enhance a cross-modal Transformer for ophthalmic report generation along with corresponding disease diagnosis. In the last, to stimulate the potential of backbone networks, the thesis explores clinical knowledge to enhance the pretraining progress to improve the quality of predicted reports. To validate proposed approaches and components, extensive experiments are also conducted in various downstream tasks, such as disease classification, medical VQA and medical image-text retrieval.

Dedication

To my beloved parents, wife and my new born baby.

Acknowledgments

First and foremost, I would like to thank my main supervisor Professor Xiaojun Chang, without whom the thesis could not be completed, and I would not have been able to start my academic journal. He is also the one who led me into the realm of deep learning and medical image analysis. I incredibly appreciate him for his kind, selfless and tremendous support on both research and my life. He guided me on vision-and-language and medical image analysis which are my favourite research topics. Under his kind supervision, I have found the right way to do research which affects me profoundly. Thanks to Professor Yi Yang, my associate supervisor, I learned a rigorous research attitude and was inspired a lot by his passion for exploring new techniques. I would also like to thank Associate Professor Xiaodan Liang, who is not on my supervisors' list but guided me throughout the research with useful suggestions and tremendous assistance with her theoretic knowledge and research skills. Dr. Wenjia Cai is an expert ophthalmologist. Thanks to her generous help and medical knowledge, I had a better understanding of those tasks. Thanks to Dr. Po-Yao Huang, my senior, from whom I learned many multi-modal methodologies and shared many experiences when I was a junior. Thanks to Professor Karin Verspoor, from whom I learned the academic writing skills and published my first paper.

I would also like to thank many mates when I was at Monash. Thanks to Professor Shirui Pan, my supervisor in Monash, who provided useful suggestions on adopting knowledge graph techniques. Thanks to Associate Professor Winston Chong for inviting me to his project, and it was the first time I found the chance to apply my research in clinical practice. Thanks to Professor Jianfei Cai, from whom I rethought the relationship between the project and research. Thanks to Professor Dana Kulić and Associate Professor Ben Beck, my mentors, when I worked at Monash, from whom I systematically learned the academic system, critical thinking and what is research impact. Dr. Tharindu Rathnayake, my colleague, helped me so much when I started working in the team. It was really nice to work with Associate Professor. Flora Wong, my mentor when I worked at Hudson Institute, from whom I realized how important our research could be.

I want to thank Mingfei Han, Siyi Hu, Wanqiang Zhang, Shiyu Ning, Wenhao Yu, Zhihang Xu, Sihao Lin, Changlin Li, Yuetian Weng, Haoran Li, Xiaoyun Zhao, Qi Duan, Dr. Di Yuan, Zutao Jiang, Dr. Fenglei Xu, Qi Hao, Xiangtan Lin, Guangsi Shi, Yicheng Wu, Yi Zhang, Fengda Zhu, Bingqian Lin, Zicong Chen, Haokun Lin and Yi Zhang. Their accompany makes my research and life much easier.

In the last, I would like to thank my father, Hongqiang Li, my mother, Jun Wan, my mother-in-law XiuZhen Zhao for their support and love throughout the years. I sincerely appreciate my beloved wife, Rui Liu, for her continuous accompany and encouragement and bringing up our newborn son Yichen Li. All the best to them all.

List of Publications

Related to the Thesis :

1. **Li, M.**, Liu, R., Wang, F., Chang, X., & Liang, X. (2022). Auxiliary signal-guided knowledge encoder-decoder for medical report generation. World Wide Web, 1-18.

2. **Li, M.**, Cai, W., Liu, R., Weng, Y., Zhao, X., Wang, C., ... & Chang, X. (2021, August). Ffa-ir: Towards an explainable and reliable medical report generation benchmark. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2).

3. **Li, M.**, Cai, W., Verspoor, K., Pan, S., Liang, X., & Chang, X. (2022). Cross-modal Clinical Graph Transformer for Ophthalmic Report Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 20656-20665).

Others :

4. **Li, M.**, Huang, P. Y., Chang, X., Hu, J., Yang, Y., & Hauptmann, A. (2022). Video Pivoting Unsupervised Multi-Modal Machine Translation. IEEE Transactions on Pattern Analysis and Machine Intelligence.

5. **Li, M.**, Zhao, X., Liu, R., Li, C., Wang, X., & Chang, X. (2022). Generalizable Memory-driven Transformer for Multivariate Long Sequence Time-series Forecasting. arXiv preprint arXiv:2207.07827.

6. Li, C., Li, Z., Ge, Z., & **Li, M.** (2019). Knowledge driven temporal activity localization. Journal of Visual Communication and Image Representation, 64, 102628.

Table of Contents

List of Figures

List of Tables

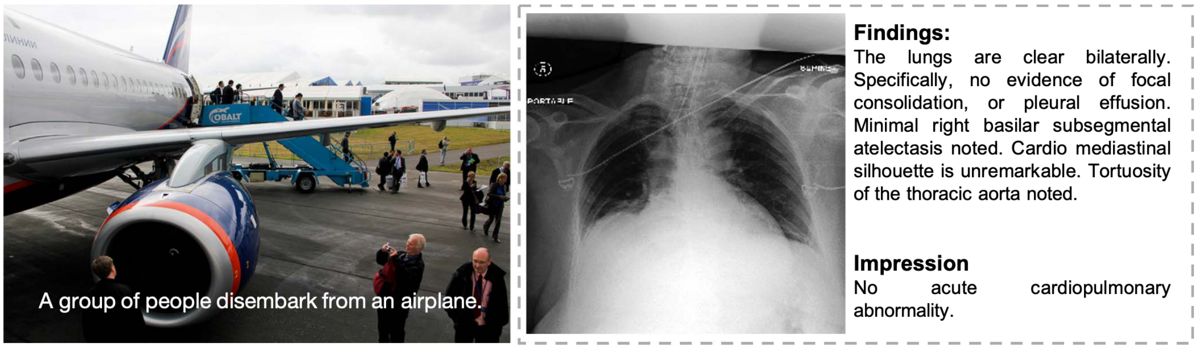
Chapter 1

Introduction

Radiology medical images (*e.g.* Chest X-Ray, Lung CT-Scan, or Fundus Fluorescein Angiography (FFA)) are essential examinations in clinical practice. Such images assist radiologists with observing inside symptoms and then making medical decisions. However, reading those images are laborious and costly, especially during pandemic period, like the rapidly increasing amounts of Chest X-Ray examinations since the novel COVID-19 outbreak [35]. Thus, the enormous demands from radiologists in clinical practice attract researchers from both automatic medicine and machine learning fields for driving medical imaging analysis's evolution and developments [91].

Within the medical imaging analysis (MIA) field, medical report generation (MRG) is a fundamental but challenging research topic. MRG tasks ask the computer to a free-text description, or *report*, summarising observations and findings of lesions or abnormalities regarding a given medical image. In clinical practice, this is done primarily to provide an interpretation of the images that supports making medical decisions. This writing process is error-prone and time-consuming, especially for those junior radiologists. Thus, given the complexity of image interpretation and to lighten the workload for radiologists considerably, automatic report generation techniques have great potential in clinical practice. One of the most similar tasks is generic image captioning [29], describing generic image instead. Thus, several successful concepts from generic image captioning have also been employed in MRG systems. However, the modality of source image is not the only difference between generic image caption and MRG tasks.

Most MRG datasets exists sever textual and visual deviation. Figure 1.1 presents two samples including images and their related descriptions from the COCO [61] and MIMIC-CXR [40], respectively. On the one hand, it is observed that medical reports, usually the lengthier one, is comprised of two sections, namely FINDINGS and IMPRESSION, while generic image caption only requires one and concise sentence. The FINDINGS usually section describes the imaging characteristics of a body structure of function that have a clinical impact. Meanwhile, the IMPRESSION section, which is usually the shorter one, summarizes the most important findings and interprets their clinical value, giving the referring physician a direction for the management of the disease or a final diagnosis [78]. Most existing MRG works only use the FINDINGS section as the target for prediction. Obviously, MRG systems call for more powerful and effective long sequence processing capabilities. Although medical reports are longer, the paradigms of these sentences are highly similar [62], especially for those that describe normal parts. Such repeated textual data is prone to degenerate MRG models, in other words, underfitting MRG models will keep repeating those sentences ignoring the image encoding. Moreover, compared with generic sentences, medical reports are hard to read and understand for researchers without medical background, due to the existences of medical terminologies and lack of medical domain knowledge; On the other hand, medical images are even harder. Due to the human tissues themselves and imaging methods, global features of the same kind of medical images are highly similar. To tell the abnormalities from normal images, shape and texture features play a critical role. Unfortunately, such features are likely to be drown by other patch or object features when the training data is insufficient.



A group of people disembark from an airplane.

Findings:
The lungs are clear bilaterally. Specifically, no evidence of focal consolidation, or pleural effusion. Minimal right basilar subsegmental atelectasis noted. Cardio mediastinal silhouette is unremarkable. Tortuosity of the thoracic aorta noted.

Impression
No acute cardiopulmonary abnormality.

Figure 1.1: The left one is a image caption sample from COCO dataset [61], the right one is a medical report generation sample from MIMIC-CXR dataset [40]

For the past decade, the *Artificial Intelligence* (AI), especially deep learning (DL) techniques, has achieved incredible achievements and evolved rapidly. By recognizing visual or textual characteristics like human beings, DL techniques have achieved state-of-the-art (SOTA) performances in vision-and-language tasks [74], including image caption. Therefore, researchers have employed various deep neural networks (DNNs) for automatic report generation [14, 49, 56]. Most often, the encoder-decoder frameworks with supervised learning or reinforcement learning are employed in MRG systems. Systems traditionally adopt a convolutional neural network (CNN) [25, 88] as the encoder first to encode medical images as dense vectors. They then employ a recurrent neural network (RNN) [27, 96] as the decoder to generate diagnostic texts from the image encoding. A cross-modal module or mechanism is utilized to attend the visual vectors to the textual representations. Since current DNNs are data greedy and sensitive, eliminating the data deviation that was previously discussed [49, 58, 65, 66] motivates the evolution and development of MRG research. Existing works have well investigated the architecture of MRG systems, and thus this thesis focuses on enhancing deep learning models with clinical knowledge to improve MRG systems. To endow MRG systems with clinical knowledge, this thesis explores various kinds of clinical medical knowledge; what is more, we also investigate when and how to inject such knowledge to MRG systems.

In Chapter 3, inspired by the radiologists' working patterns, we explore auxiliary signals' power to facilitate generating medical reports. Generally, when a radiologist describes a medical image, he/she will carefully inspect the suspicious regions after quickly browsing the global image. Then, he/she will write a report that draws on the knowledge he/she learned from the external medical domain and his/her working experience. Accordingly, to mimic the behavior of medical experts, we propose an Auxiliary Signal-Guided Knowledge (**ASGK**) approach including two kinds of auxiliary signals to improve a Transformer to generate medical reports. Firstly, we automatically find a suspicious region where the pre-trained neural visual extractor paid the most attention. After resizing and cutting, the auxiliary patches are concatenated to the original patch features before being fed to the encoder. These patches ensure that the Transformer will learn better visual hidden representations. Then, we collect a medical corpus to pre-train the decoder, in which all the sentences that record related medical knowledge are easily accessed online. It was the first time we find that pre-training steps can improve the model robustness to alleviate the training corpus deviation and decrease the sensitivity to similar linguistic patterns.

In Chapter 4, we propose a Cross-modal clinical Graph Transformer (**CGT**) for ophthalmic report generation (ORG). In particular, we first invoke an information extraction scheme based on a natural language processing pipeline, including named entity recognition and entity linking, to obtain a clinical knowledge graph. More details will be

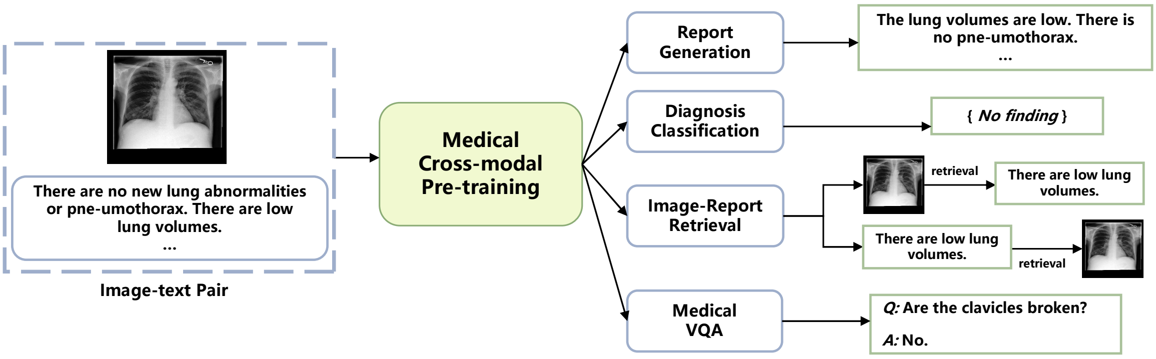

There are no new lung abnormalities or pne-umothorax. There are low lung volumes.
...
Image-text Pair
Medical Cross-modal Pre-training
Report Generation
Diagnosis Classification
Image-Report Retrieval
Medical VQA
The lung volumes are low. There is no pne-umothorax.
...
{ No finding }
retrieval
There are low lung volumes.
There are low lung volumes.
retrieval
Q: Are the clavicles broken?
A: No.

Figure 1.2: Our proposed MONITOR, which is a multi-task benchmark for enabling the comprehensive evaluation of unified medical vision-language models.

introduced in Chapter 3. As discussed in [36], the structured clinical information behind the free-text reports can enhance the diagnostic methods. In addition, the entities and relations in our clinical graph are in the homogeneous embedding space with the training corpus. Given a set of ophthalmic images, the extracted visual features are transformed to a compressed visual token and a sub-graph with relevant restored triples. Since the sub-graph is not guaranteed to be a completely accurate representation of the given images and natural noise exists in the clinical graph, we adopt a cross-modal encoder to encode the universal feature token and sub-graph information. To avoid influence from unrelated entities, a visible matrix is introduced during the cross-modal encoding process. Finally, reports are generated via a Transformer[96] decoder.

In Chapter 5, inspired by the recent progress in vision language pretraining (VLP), we propose a multi-task benchmark dubbed Medical cross-mOdal uNderstandIng and generaTion with knOwledge-enhanced pRetraining (**MONITOR**). As shown in Fig. 1.2, in addition to *MRG tasks*, MONITOR also covers a set of fundamental medical cross-modal tasks, including *diagnosis classification*, *image-report retrieval*, and *medical visual question answering*. Through MONITOR, the comprehensive evaluation of unified medical cross-modal models can be fulfilled. To establish a baseline model on MONITOR for encouraging the future research, we develop Med-KEP, which is a unified model pretrained on large-scale medical data and finetuned on both understanding and generation downstream medical tasks. The expert knowledge has been demonstrated to be crucial in enhancing the performance of medical vision-language models as well as improving their explainability [65, 70, 110, 116]. To study the impact of the expert medical knowledge on the unified pretrained model in different downstream tasks, we further introduce three kinds of medical knowledge construction and injection strategies during the pretraining process of Med-KEP: 1) Triplet Concatenation (TC) concatenates multiple knowledge triplets (each is formed as <*head entity, relation, tail entity*>) into one single sentence to obtain the knowledge encoding, 2) Triplet Insertion (TI) replaces the entities in the text by the knowledge triplets, and 3) Symbolic Knowledge Graph (SKG) represents different relations as edge weights and encodes the knowledge graph through the self-attention mechanism.

In addition to the above frameworks, this thesis also contributes to medical report generation dataset and benchmark to benefit this community. In Chapter 3, we introduce a new COVID-19 CT Report (**COV-CTR**) dataset for use in validating the robustness and generalization ability of ASGK. Along with the accurate predicted reports and diagnosis, the reliable rationale for interpretation is strongly encouraged by specialists and patients to trust those predictions. To improve MRG systems' explainability, researchers have explored text-image attention mappings [14, 39] to explain the automatic generation procedure. However, the accuracy of these explanations is unclear. Since existing MRG datasets including COV-CTR fail to provide explainable annotations, development of interpretable MRG methods to improve trustworthiness is a great challenge. Therefore, in Chapter 4, we further present a new benchmark, **FFA-IR**, towards an explainable and reliable MRG benchmark based on FFA Images and Reports. Specifically, FFA-IR is *large*, with 10,790 reports along with 1,048,584 FFA images from clinical practice; it includes *explainable annotations*, based on a schema of 46 categories of lesions; and it is *bilingual*, providing both English and Chinese reports for each case. Besides using the widely used natural language generation (NLG) metrics, we also propose a set of nine human evaluation criteria to evaluate the generated reports. Due to the different modality, auxiliary signals explored in the Chapter 3 are hard to transfer for facilitating ophthalmic report generation.

This thesis is organised as follows. After this Chapter, we review the related literature on medical report generation methods, medical report generation benchmarks, medical knowledge enhanced models, and knowledge enhanced pretraining in Chapter 2. In Chapter 3, an auxiliary signal-guided knowledge encoder-decoder framework is proposed for automatic report generation. We publish a COVID-19 CT Report and Image dataset, and evaluate our proposed model with it for MRG. In Chapter 4, we first propose an explainable and reliable MRG benchmark based on FFA Images and Reports. And then it is for use in validating the proposed cross-modal clinical graph Transformer. In Chapter 5, we propose benchmarking the medical cross-modal understanding and generation with knowledge-enhanced pretraining (MONITOR), providing a multi-task benchmark for enabling the comprehensive evaluation of unified medical vision-language models.

In Chapter 6, we summarize this thesis and imagine our future directions for MRG research.

In this thesis, we make the following contributions:

1. We are the first to release a COVID report generation dataset regarding Lung CT-Scan examinations.

2. We are the first to present an explainable and reliable MRG benchmark, before us, there are rare explainable annotations and reliable evaluation tools.

3. We explore medical knowledge by mimicking radiologists' working patterns and utilize those knowledge to guide a proposed encoder-decoder MRG system.

4. We construct a clinical graph from medical reports automatically and propose a cross-modal clinical graph Transformer for ophthalmic report generation. Based on our knowledge extraction scheme, we are the first work that can provide sliver sub-graph to supervise the knowledge restoration process.

5. We propose a medical enhanced pretraining benchmark that introduces three kinds of medical knowledge construction and injection strategies.

CHAPTER 2

Literature Review

2.1 Medical Report Generation Methods

2.1.1 Early Approaches

At the beginning, medical reports were not predicted by an End-to-end manner. Varges et al. [95] first encoded given medical images into triplets that represented cardiological findings. And then such triplets were extended to a readable free-text reports by an ontology-based natural language generation approach. In contrast, Schlegl et al.[86] utilized the gold reports as input instead of treating them as targets and combined reports with images for 3D pixels classification. Then they employed a structured Support Vector Machines[94] to generate semantic tags for each lesion, with the concepts of a radiology lexicon.

2.1.2 Medical Report Generation Frameworks

Traditionally, MRG works [39, 68] employ an encoder-decoder (ED) architecture as the backbone, visual encoding is attended to textual representations with or without attention mechanism with the supervised [66], unsupervised [67] and reinforcement learning [49]. Most often, such ED MRG systems typically use CNN-based image encoders [25, 88] to first encode the medical images as dense vectors. They then typically use a RNN-based natural language decoders [27, 96] to generate diagnostic reports from those vectors. During the decoding procedure, a cross-modal attention mechanism

Table 2.1: Comparing existing deep learning based medical report generation approaches.

Methods	Modality	Backbone Network	Adopted Knowledge	Learning Type
CoAtt[39]	CXR	CNN+LSTM	Medical Tags	Supervised Learning
HRGR[58]	CXR	CNN+LSTM	Template Database	Reinforcement Learning
KERP[49]	CXR	CNN+Graph Transformer	Template Database+Terminology Graph	Supervised Learning
R2Gen[14]	CXR	CNN+Transformer	Memory Driven	Supervised Learning
PPKED[65]	CXR	CNN+Transformer	Posterior and Prior Knowledge	Supervised Learning
CA[66]	CXR	CNN+LSTM	None	Contrastive Learning
CMCL[64]	CXR	CNN+LSTM	Working Patterns	Curriculum Learning
MKG[116]	CXR	CNN+GCN+LSTM	Terminology Graph	Supervised Learning
MGSK[110]	CXR	CNN+GCN+Transformer	General and Specific Knowledge	Supervised Learning
ASKG	CXR+LCT	CNN+Transformer	Working Patterns+Terminology	Finetuning
CGT	FFA	I3D+Transformer	Clinical Graph	Supervised Learning
MONITOR	CXR	ViT+Transformer	General and Specific Knowledge	Supervised Learning

may be employed to let the decoder focus on particular visual vectors when predicting each word. Such mechanisms can also be used to highlight the specific regions where the findings described in the report, are widely applied in recent MRG systems. In the beginning, Jing *et al.* [39] presented an encoder-decoder framework and employed a co-attention mechanism over both visual and textual features to predict medical tags and generate a single sentence simultaneously. To generate multi-sentences, Xue *et al.* [109] adopted a hierarchical RNN, consisting of a topic-level LSTM and a word-level LSTM as the decoder. The sentence-level LSTM produces a sequence of sentence embeddings, each intuitively specifying the information to be expressed by a sentence of the image descriptions. This concepts have been adapted by the following works[49, 116]. Zhang et al.[117] proposed the MDNet model, which was the first to utilize visual attention mechanism in MRG tasks. With the success of Transformer [96] in NLP tasks, Chen *et al.*[14] proposed a memory-driven Transformer to enhance the decoding procedure's memory. Before encoding the visual vectors, researchers usually first adopted a CNN like VGG, or ResNet to extract equally sized visual patch features. Then those patch features are treated as visual tokens for encoding. Despite progress in developing models, the lack of accurate explanation and reliable evaluation undermines the trustworthiness of these methods.

2.1.3 Types of Learning

Besides the supervised learning, reinforcement learning is another learning type which dominates the MRG research. The key concept of reinforcement learning (RL) in MRG is to treat the MRG systems like an agent. In this type, researchers can propose vary rewards to encourage the system in understanding clinical correctness. Most importantly, non-differentiable evaluation measures can be used directly during training in RL, so that systems are not required to be optimized by loss functions like cross-entropy during training, while being assessed with metrics such as BLEU [76], CIDER [97], or clinical F1[116] at testing procedure. Rennie *et al.* [85] are the first to employ the RL algorithm with a reward based on CIDER value. Li *et al.* [58] utilized RL to decide if a sentence will be generated from scratch, or if it will be retrieved from a database with frequently occurring sentences. Their experimental results on CX-CHR and Open-IU [17] datasets were close to baseline performances with supervised learning. For other perspectives, Liu *et al.* [68] used RL to optimise reports' readability. Liu *et al.* also introduced a reward based on comparing labels that extracted by CheXpert[34] from the system-generated text and the human-authored report, in order to optimize clinical accuracy.

Recently, unsupervised MRG systems also attract increasing attention in the community. Those works argue that collecting large amounts of paired gold reports and medical images are prohibitively expensive, and most deep learning based MRG systems are severely data greedy, especially for Transformer-based systems. Thus, exploring the effectiveness of MRG systems with limited data should become a fundamental research topic in this field. Liu *et al.* [67] made the first attempt to train a medical report generation model without using any coupled image-report training pairs. To relax the dependency on paired data, they proposed an unsupervised model (KAGE) which accepts independent sets of images and reports in training. The KGAE consists of a pre-constructed knowledge graph, a knowledge-driven encoder and a knowledge-driven decoder. They converted the image features to the prior knowledge representations in the latent space to bridge the visual and textual domains. The knowledge-driven encoder projects medical images and reports to the corresponding coordinates in this latent space and the knowledge-driven decoder generates a medical report given a coordinate in this space. Since the knowledge-driven encoder and decoder can be trained with independent sets of images and reports, KGAE is unsupervised. But the experimental results on Open-IU and MIMIC-CXR [40] demonstrated that the gap was quite large between the unsupervised and supervised models.

2.1.4 Explainability

In clinical practices, both patients and doctors expect the accurate predictions among with a reliable rationale to explain their decisions. Therefore, in this section, we discuss how the existing MRG works present their explainability.

On the one hand, the multi-modal attention mechanism is a great manner to represent the expalinability and make the diagnosis more easily interpretable. For example, Jing

Table 2.2: Comparison of existing widely used MRG datasets, where * means the average number. Report length and number of lesions are marked as – for data sets that do not provide this figure.

Dataset	Image			Report			Lesions
	Number	Modality	View*	Length*	Language	Cases	
Open-IU[17]	7,470	X-Ray	2	32.5	En	2,955	–
MIMIC-CXR[40]	377,110	X-Ray	1	53.2	EN	276,778	–
PadChest[10]	160,868	X-Ray	2	–	Es	22,710	–
CX-CHR[58]	45,598	X-Ray	2	66.9	Zh	40,410	√
COV-CTR[56]	728	CT-Scans	1	77.3	En/Zh	728	√
DEN[31]	15,709	CFP+FFA	1	7	En	–	–
STARE[28]	397	CFP+FFA	5	–	En	397	–
DIARETDB1[42]	89	CFP	1	–	En	89	–
MESSIDOR[16]	1,200	CFP	2	–	Fr	587	–
FFA-IR [55]	1,048,584	FFA	87	91.2	En/Zh	10,790	√

et al. [39] and Chen *et al.* [14] visualized the text-image attention maps to highlight the specific regions where the predicted word is based on. Thanks to this, doctors can easily find the lesion or abnormal regions when reading the disease keywords. On the other hand, the medical knowledge graph can be an internal output to assist researchers or doctors understanding the relations between different entities. Zhang *et al.* [116] proposed a unified knowledge graph consisting organs and diseases. In each case, a subgraph is restored from the unified graph and used as the feature to attend image features. This graph is also used in [65]. To represent more information, Li *et al.* [49] used the same medical terminology and disease graphs with us to bridge the visual and textual modalities. To evaluate the explainability quantitatively, our FFA-IR provides the location information for most lesions and release the first explainability comparison benchmark by calculating the Inter-over-union between the gold lesion location and the text-image attention maps.

2.2 Medical Report Generation Benchmarks

2.2.1 Medical Report Generation Datasets

In Table.2.2, we compare eight widely used and publicly available MRG benchmarks with our proposed COV-CTR and FFA-IR, in terms of their statistic, including image modality, report language, report length and others. Firstly, among all the MRG benchmarks, Open-IU [17] and MIMIC-CXR [40] are the two most widely-used medical report benchmarks. Both of them provide chest X-Ray images along with related English written reports. Their reports usually contain 'FINDINGS', 'COMPARISON', 'INDICATION' and 'IMPRESSION' sections, among which 'FINDINGS' and 'IMPRESSION' are of primary interest. The 'FINDINGS' section summarizes the symptoms or clinical impacts from all the observed imaging characteristics of a body structure of function. The 'IMPRESSION' usually summarizes the most important findings and interprets their clinical value, giving the referring physician a direction for the management of the disease or a final diagnosis. However, sometimes the 'IMPRESSION' (or 'FINDINGS') includes a conclusion that does not follow from the previous sections and the images of the current exam. Some previous work used only the 'IMPRESSION' section as the target text to be generated[87], but most previous work either uses the 'FINDINGS' as the target[58, 68] or aims to generate the concatenation of the two sections[39].

PadChest [10] is another large-scale chest X-ray report dataset, which comprises 160,868 images and multi-label annotated reports. However, these diagnosis texts are not complete and are written in Spanish. Furthermore, MIMIC-CXR provides extra related disease impressions, which can be used for disease classification. Due to the characteristic of Chinese words, CX-CHR [58] and our COV-CTR have a more considerable average report length than English medical report datasets. Compared with our FFA-IR, DEN mainly contains CFP images (13,898 CFP and 1,811 FFA). Table.2.2 presents that our FFA-IR has the most significant number of medical images and the average length of reports among all these datasets. Unlike all existing medical report datasets, FFA-IR provides explainable annotation by label 46 kinds of lesions in a total of 12,166 regions along with FFA images and reports which play an essential role in identifying disease and writing reports. There are also three more retinal datasets comprising retinal images and text. STARE [28] was conceived and initiated in 1975 and released in 2004 with 397 images including CFP and FFA. Their diagnosis texts are almost the retinal diseases that are unsuitable for training a medical report generation model. DIARETDB1 [42] is well annotated with lesion location and size yet has a limited number of CFP images. MESSIDOR [16] comprises 1,200 CFP images and 600 fine-gained French reports.

These are also several datasets which are not publicly available or lack of related annotations. Such datasets are BCIDR[117], consisting of 1,000 pathological bladder cancer images, each with five reports; Frontal Pelvic X-Rays[22], which comprises 50,363 images, each accompanied by a radiology report simplified to follow a standard template; and Chest X-Ray 14[102], which is publicly available, but does not include any medical reports in its public version.

Notably, all the MRG benchmarks share the same issue, the inevitable data biases. Those biases comes from two aspects. On the one hand, among each benchmark, medical images are highly similar due to the nature of human tissues, angle of imaging examinations and imaging methods. On the other hand, normal tissues are described in the same manner which brings out large amounts of repeated sentences. As a results, investigating how to estimate those two biases is one of the key motivations in MRG research.

2.2.2 Medical Report Generation Metrics

2.2.2.1 Natural Language Generation Metrics

One essential way to evaluate the performances of proposed MRG methods is to judge the quality of predicted reports. To this end, researchers employ the proposed automatic evaluation metrics from natural language generation (NLG) tasks (*e.g.* machine translation and text summarizing) to assess the predicted reports. The key concept of such NLG metrics (*e.g.* BLEU[76], ROUGE[60], and METEOR[6]) is to calculate the similarity between the generated texts and ground truth texts. In MRG tasks, ground truth texts are annotated by radiologists in daily practices. In more details, the similarity can be quantified by the shared occurrences of n-grams, phrases of n consecutive words. Such metrics have been verified to perform well in ranking systems, and the rank lists correlate well with human judgments of information content in machine translation and text summarizing tasks. However, recent studies argue that those metrics do not capture promising clinical correctness. Due the reason that, once the models keep repeating sentences which describe normal findings, they still achieve considerable measurements. Therefore, the occurrences of abnormal terminologies and positive disease keywords should acquire more attention than any other single words. In generic image caption tasks, CIDER measure[97] becomes the main metric. Due to the reason that, CIDER assigns different importance to different n-grams and performances more like human annotators. The more recent SPICE [2] extracts knowledge or keywords from the reference and generated texts to measure both the gram-level and semantic-level similarities. Subsistent MRG works utilize either BLEU-4 (4-gram) or CIDER as the main metric to compare their methods with others.

2.2.2.2 Clinical Correctness Measures

As we discussed, the above NLG metrics do not always capture clinical correctness. To evaluate the MRG systems in a reliable way, researchers employ other measures to assess the predicted reports. The first one is human study. During a human study, researchers randomly selected plenty of examinations, including the images, gold standard reports and reports produced by baseline MRG models and the proposed elaborate system. Then available annotators are invited to choose the best system-generated report with consulting the gold reports and examination. At the beginning, Li *et al.* [49] employed crowd-sourcing annotators. However, consulting the gold reports and examination requires medical background annotators. Hiring such annotators are expensive and sometimes it may have been inadequate. Thus, Zhang *et al.* [116] invited experienced radiologists to judge the information content. However, it may be too subjective to select the best reports. In this thesis, we propose nine criteria to judge the reports, including the clinical correctness of the reported abnormalities, fluency, and content coverage compared to the ground truth report. Since human study may be prohibitively expensive, recent work measure the clinical correctness through medical terms.

To this end, the concept of Tags in MRG tasks have been involved to facilitate the evaluation. Xue *et al.*[109] were the first to use an evaluation measure that considers medical tags extracted from system-generated and human-authored reports. The authors called the measure Keyword Accuracy, but it should not be confused with the conventional classification Accuracy, since it only measures Recall. Huang *et al.* [32] followed the same approach, but they used only the MTI tags as their ground truth. In both of these works, however, where gold tags were compared with predicted tags, it is unclear how the predicted tags were extracted from the system-generated reports. Liu *et al.* [68] used the CheXpert medical abnormality mention detection system, which generates one out of 4 labels (presence, absence, negative, not sure) for each one of 14 thoracic diseases. Recently, Zhang *et al.* [116] proposed a new metric based on the Recall, however the code of this evaluation is still not available.

2.3 Knowledge Enhanced Models

2.3.1 Medical Knowledge Enhanced Models

In this section, we will introduce medical knowledge enhanced models for medical report generation and other medical domain tasks, medical QA, or memorization. The incorporated medical knowledge can be divided into three groups.

The first kind is from radiologists' working patterns [65]. In clinical practice, radiologists read images and write reports in a specific pattern to remind them of not missing any part of the images. After browsing the whole image, radiologists will focus on the suspicious regions. To make the model achieve this capability, we introduce two kinds of auxiliary signals to guide the MRG model. Similarly, Liu *et al.* adopted both posterior and prior knowledge to imitate the pattern with retrieved reports and a universal knowledge graph. Secondly, researchers explored the clinical knowledge behind the free-text reports to drive MRG models. Li *et al.* [49] extracted normal and abnormal terminologies from corpus as nodes and automatically predict weights between these findings as edges to construct a knowledge graph. This graph worked as prior knowledge to drive the decoding procedure and restore a unique sub-graph for each case. In contrast, Zhang *et al.* [116] and Liu *et al.* [65] adopted an universal graph covering 20 findings to enhance the MRG models. In the last, the existing biomedical knowledge base is adopted to incorporate medical knowledge. The unified medical language system (UMLS) [9] is the largest biomedical knowledge base and is adopted in [70] and [24] to enhance pretrained medical models for medical QA tasks. However, utilizing the existing knowledge base will bring in inconsistencies due to the heterogeneous embedding space arising from vocabulary and context mismatch. Since the entities and relations in UMLS are derived independently of the training corpus, when embedding node information, the embedded token vectors are inconsistent. Additionally, utilizing the full UMLS in MRG tasks will place a burden on the computation resources since it has $13,555,037$ triples, and most of them are irrelevant to our task.

2.3.2 Knowledge Enhanced Pretraining

Vision-language pretraining (VLP) aims to improve performance of the downstream unimodal or cross-modal tasks via pretraining the model on a large amount of image-text pairs. Typically, the inputs of VLP models are from different features, such as object-based region features [54, 71], CNN-based grid features [33], ViT-based patch features [46, 100] and word-level embeddings, which are fed into one single transformer encoder [54, 89] or two transformer encoders [93, 112] interacted with the cross-attention mechanism for multi-modal fusion. Motivated by the success of BERT [20] in the natural language processing (NLP) field, researchers pre-train VLP models by using different unsupervised learning objectives, including masked language modeling (MLM) [20], masked vision modeling (MVM) [81], image-text matching (ITM) [53], image-text contrastive learning

(ITC) [52], *etc*.

Most existing VLP methods focus on understanding tasks and do not possess the ability to generate. To tackle this problem, some works attempt to develop unified VLP models for addressing both understanding and generation tasks [52, 69, 119]. In this paper, we utilize BLIP [52], a multi-modal mixture of encoder-decoder model, as the backbone architecture to handle both discriminative and generative downstream tasks on our proposed MONITOR.

Although VLP models have the capability to store certain knowledge or facts from training data, their knowledge awareness is still far from satisfactory. For understanding capabilities, VLP models are easily fooled by negated or misprimed probes [41] and fail in reasoning tasks [92]. For generation capabilities, the predicted sentences could be grammatically correct but not logical [59]. Thus, recent works have explored knowledge graphs from linguistic [20], encyclopedia [99] or domain specific [9] knowledge bases to enhance VLP models. For example, Sun *et al.* [90] proposed a knowledge masking strategy for MLM to enhance language representations by encyclopedia knowledge. Zhang *et al.* [116] utilized both linguistic knowledge and relations between entities in knowledge graphs to train an enhanced language-pretrained model. To further align vision and language, Yu *et al.* [112] used a semantic scene graph parsed from the text as the bridge. In this work, we introduce three kinds of knowledge construction and injection strategies in the pretraining phase to sufficiently study how the expert medical knowledge impacts the unified pretrained model in different downstream medical tasks.

CHAPTER 3

AUXILIARY SIGNAL-GUIDED KNOWLEDGE ENCODER-DECODER

3.1 Introduction

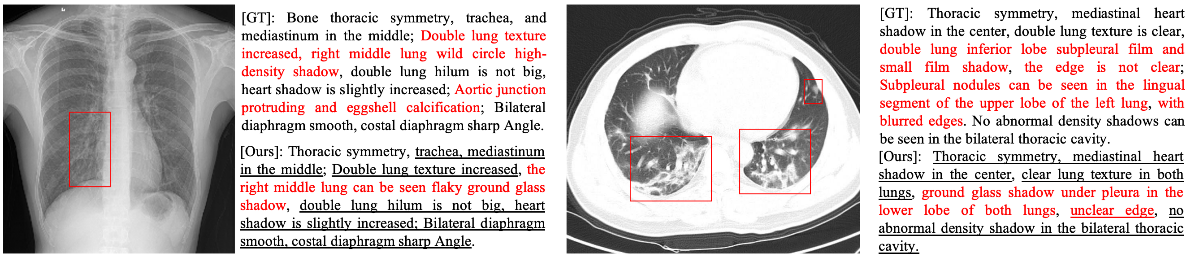
When you take a medical image in any hospital, you will receive a medical report. This medical report describes both normal and abnormal terminologies, and can assist radiologists and specialists in diagnosing and reviewing. However, writing medical reports is error-prone and time-consuming, especially during a pandemic like COVID-19, because radiologists may have to diagnose hundreds of images per day. Therefore, the topic of automatically generating medical reports has attracted research attention from both artificial intelligence and clinical medicine fields.

As discussed, the most similar task to medical report generation in the computer vision field is image captioning. Beyond the common difficulties in natural image captioning, there are three more bottlenecks for medical report generation. Firstly, the amount of image-report pairs in existing datasets are considered small compared to the captioning datasets, which are insufficient to learn visual representations; Secondly, it is hard to acquire the object features which are widely used in the natural image captioning tasks [3] from medical images. Only a few medical images can provide the well-annotated segmentation or location information of lesions; Thirdly, there are severe data deviation exists in these datasets. Some diseases are rare in nature, and their positive samples are hard to collect. Moreover, there are many similar sentences used in each report to describe the routine observation, which leads to the overfitting problem and limits the generalization of neural approaches [58].

Recently, many approaches have been designed to address these problems and achieved promising performance on automatically generating medical reports [14, 49, 58]. Inspired by the radiologists' working patterns, in this section, we explore auxiliary signals' power to facilitate generating medical reports. Generally, when a radiologist describes a medical image, he/she will carefully inspect the suspicious regions after quickly browsing the global image. Then, he/she will write a report that draws on the knowledge he/she learned from the external medical domain and his/her working experience. As shown in Fig 3.1, the suspicious region takes up only a tiny portion of the global image but has been treated equally to other regions in previous works. Therefore, other regions could be considered irrelevant noise that distracts the model. Although these regions may get more attention based on the self-attention mechanism in Transformer, Dosovitskiy *et al.*[21] pointed out that Transformer can learn a better visual representation when fed with original image patches instead of the encoded visual features. Using large extra corpora to pre-train the Transformer is an effective way to alleviate the corpus deviation in the training datasets [19, 83]. However, there is a considerable textual semantic gap between the medical and common domains.

Accordingly, to mimic the behavior of medical experts and address the above mentioned learning difficulties, we propose an Auxiliary Signal-Guided Knowledge (ASGK) approach including two kinds of auxiliary signals to improve a Transformer to generate medical reports. Firstly, we automatically find a suspicious region where the pre-trained neural visual extractor paid the most attention. After resizing and cutting, the auxiliary patches are concatenated to the original patch features before being fed to the encoder. These patches ensure that the Transformer will learn better visual hidden representations. Then, we collect a medical corpus to pre-train the decoder, in which all the sentences that record related medical knowledge are easily accessed online. The pre-training steps can improve the model robustness to alleviate the training corpus deviation and decrease the sensitivity to similar linguistic patterns.

We further introduce a new COVID-19 CT Report (COV-CTR) dataset for use in validating the robustness and generalization ability of ASGK. Since December 2019, the novel COVID-19 virus has caused a global pandemic and infected millions of people across 200 countries. A key step in controlling the infection is that of identifying infected people. In addition to the Reverse Transcription Polymerase Chain Reaction (RT-PCR) tests, lung CT scan analysis has emerged as another essential testing method. Therefore,



[GT]: Bone thoracic symmetry, trachea, and mediastinum in the middle; Double lung texture increased, right middle lung wild circle high-density shadow, double lung hilum is not big, heart shadow is slightly increased; Aortic junction protruding and eggshell calcification; Bilateral diaphragm smooth, costal diaphragm sharp Angle.

[Ours]: Thoracic symmetry, trachea, mediastinum in the middle; Double lung texture increased, the right middle lung can be seen flaky ground glass shadow, double lung hilum is not big, heart shadow is slightly increased; Bilateral diaphragm smooth, costal diaphragm sharp Angle.

[GT]: Thoracic symmetry, mediastinal heart shadow in the center, double lung texture is clear, double lung inferior lobe subpleural film and small film shadow, the edge is not clear; Subpleural nodules can be seen in the lingual segment of the upper lobe of the left lung, with blurred edges. No abnormal density shadows can be seen in the bilateral thoracic cavity.

[Ours]: Thoracic symmetry, mediastinal heart shadow in the center, clear lung texture in both lungs, ground glass shadow under pleura in the lower lobe of both lungs, unclear edge, no abnormal density shadow in the bilateral thoracic cavity.

Figure 3.1: Two samples from CX-CHR and our COV-CTR datasets. Red bounding boxes annotated by a radiologist indicate the regions that he pays more attention to describing this image. The red text describes the abnormalities. Underlined text indicates alignment between ground truth reports and generated reports.

an accurately written report could assist patients and doctors to understand their health condition. We invited three radiologists with more than five years of working experience to apply their diagnostic skills to the public COVID-CT dataset[118] and use this information to construct the COV-CTR dataset. The main contributions of this section are three-fold as follows:

- We identify and produce two kinds of auxiliary signals, namely the internal fusion visual features and the external medical linguistic information to facilitate graph encoding and medical knowledge learning respectively.
- We design a medical tag graph encoder to transfer input features into higher-level information and adopt Generative Pre-Training (GPT) [83] as our natural language decoder to generate accurate and robust medical reports.
- We invite three radiologists with more than five years of experience to apply their diagnostic skills to the COVID-19 CT images [118] and use this information to construct a new medical report dataset, COVID-19 CT Report which will be available.

3.2 Approach

3.2.1 Problem Setup

Similar to the previous studies [39, 49, 58, 116], the task of medical report generation involves asking a model to generate a topic related paragraph consisting of a series of sentences to describe a medical image of a patient case. We represent the image as

I and the report as $S = \{w_1, w_2, ..., w_l | w_i \in \mathbf{V}\}$, where w_i presents the index of word in $\mathbf{V}$ the vocabulary of all words contained in the datasets. To generate fine-grained and semantically coherent sentences, we propose a graph encoder-decoder framework that first encodes inputs feature vectors to a medical tag graph and then decodes them to a medical report. We represent the medical tag as $G = (V, E)$, where $V = \{v_i\}_{i=1:N_t}$ and $E = \{e_{i,j}\}_{i,j=1:N_t}$ is a set of edges. In our task, we represent each node feature v_i by its detected tag classification probability, then encode the correlation between each of the two tags as edge weights. N_t represents the total number of medical tags composes abnormal terminologies, such as "pneumothorax" and "colon shadow", and normal terminologies such as "normal spine", "normal intercostal space" and so on.

Generally, when a radiologist describes a image, he will inspect the abnormal region carefully after quickly browsing the global image, then write a report that reflects both his inspection and the knowledge obtained from external medical domain information and his working experience. To mimic this pattern, we firstly pretrain the framework with the external medical signals collected from an appropriate website in order to correctly phrase and learn medical knowledge. Subsequently, the internal visual fusion signals facilitate graph encoding and bridge the gap between linguistic and visual domain. More details regarding these internal visual fusion signals are described in Section 3.2.3.

3.2.2 The structure of ASGK

An overview of our approach is shown in Figure 3.2. The main structure of ASGK comprises a medical graph encoder and natural language decoder.

3.2.2.1 Medical Graph Encoder.

This component is built to encode the input features into higher level information, *i.e.* a medical tag graph. In the medical graph, each node denotes one detected medical tag, the features of which are the classification probabilities and can be written as Equation 3.1.

$$V = \text{Sigmoid}(W_v f_{input}) \tag{3.1}$$

where W_v is a projection matrix of size $N \times d$; here, d represent the dimension of the input features, and N is the number of total tags. Given that the truth edge information is not available in our case, we conduct an attention operation to learn edge weights automatically, which can be written as follows:

$$e_{i,j} = \text{Norm}(\text{Attention}(W_v v_i, W_v v_j)) \tag{3.2}$$

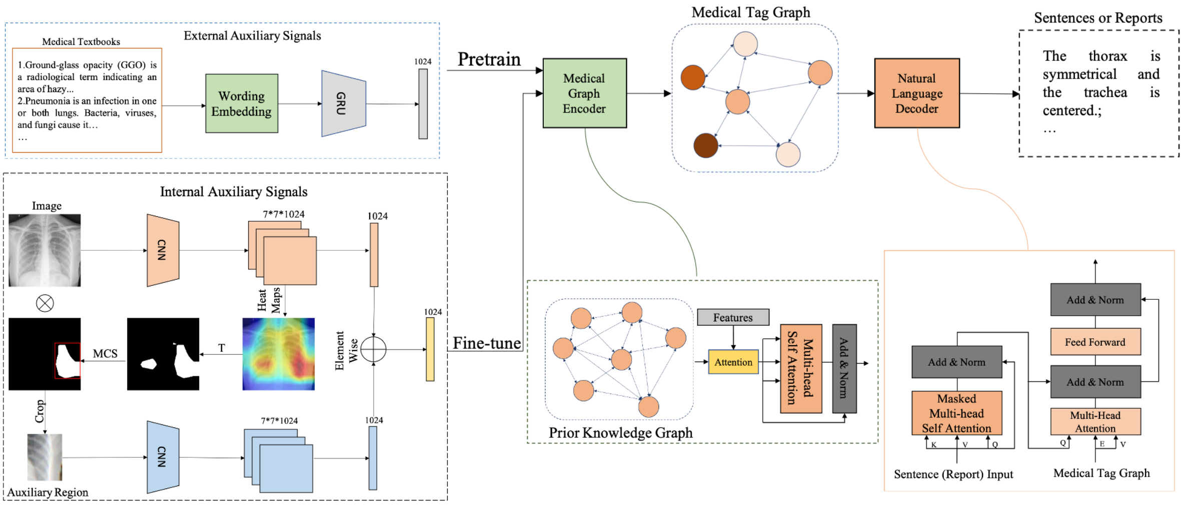


Figure 3.2: An overview of our ASGK approach. The ASGK model consists of a medical graph encoder and a natural language decoder. The medical graph encoder encodes input features into the corresponding medical tag graph, while the natural language decoder transfers high-level information to sentences or reports. The external signals guide the pretraining procedure, while the internal signals guide the model to bridge linguistic and visual information. T and MCS represent threshold and max connection select operation respectively.

where *Norm* is the normalization operation, while *Attention* is executed as a scaled dot-product operation. Then the medical tag graph is incorporated with the prior medical knowledge which is represented as a set of nodes of size N with initialized features and edges via attention mechanism following by [49], which can be written as follows:

$$G = \text{att}(G_{prior}, V, E) \tag{3.3}$$

To enhance the correlation between each of the nodes, we employ a multi-head self attention operation on G to get the final graph. We further treat medical tag detection as a multi-label classification task and adopt BCE loss to maximize the prediction scores

$$L_{tagcls} = -\sum_{i=0}^{N-1} y_i \log v_i + (1 - y_i)\log(1 - v_i) \tag{3.4}$$

where W_v is a projection matrix of size $N \times d$; here, d represent the dimension of the input features, y_i is the ground truth label, and v_i is the final graph tag features.

3.2.2.2 Natural Language Decoder.

Inspired by GPT [83], we design a natural language decoder consisting of $N = 3$ blocks, similar to the Transformer decoder, to interpret the medical tag graph and enable semantic alignment in the visual and linguistic domain. The structure of the block is presented in Figure 3.2. This block applies a masked, multi-head self-attention operation to the medical report or sentences tokens $T = \{t_1, t_2, ..., t_l\}$ embedded from Glove vectors pretrained on our datasets. We use [83] to maximize the likelihood in the following formulation:

$$L_t(T) = -\sum_i \log P(t_i | t_1, ..., t_{i-1}; \Theta) \tag{3.5}$$

where P is the conditional probability of the next token prediction, modeled using a neural network with parameters Θ and history sentences. Then, followed by position-wise feed forward layers, the natural language decoder aims to produce an output distribution over all token vocabulary.

$$h_0 = I_W W_e + I_P W_p, \tag{3.6}$$

$$H_l = \mathbf{block}(h_{l-1}, V, E) \forall l \in [1, N], \tag{3.7}$$

$$P_i = \text{Softmax}(h_N W_e^T) \tag{3.8}$$

where I_W is the index of input tokens in the vocabulary, I_P is the index of the token's position, W_e is the pretrained wording embedding matrix, and W_p is the position embedding matrix.

3.2.3 Auxiliary Signal-Guide Learning

3.2.3.1 Pretraining with External Auxiliary Signals.

The direct application of general pretrained language models to medical domain tasks leads to unsatisfactory results, since the word distributions differ from those of those of general and medical corpora. To resolve this problem, we collect medical textual information from an appropriate website to construct a large-scale medical textbook. This textbook provides sufficient information about medical knowledge, including the symptoms, manifestations and other information about COVID-19 and thoracic diseases. Before feeding it into the medical graph encoder, we divide the medical textbook into sentences and embed the word tokens with embedding vectors, which are trained in our datasets using Glove [79]. After embedding, sentences are encoded using a single-layer GRU with 1024 hidden units to produce the external medical auxiliary signals.

3.2.3.2 Training with Internal Auxiliary Signals.

Evidently, the quality of the encoded medical graph will significantly affect the accuracy of the generated reports. Therefore, we produce internal fusion visual signals to facilitate medical graph encoding and bridge the gap between linguistic and visual information. As shown in Figure 3.2, we first classify the global image using DenseNet-121 and obtain the feature maps $f_c \in R^{7*7*1024}$ before the final pooling layers and output from last pooling layers $f_g \in R^{1*1024}$. To produce the mask, we perform a threshold operation on a heat map acquired by Equation 3.9 and select the max connected area:

$$H = \max_k(|f_c^k|), k \in 1:0124 \tag{3.9}$$

We adopt another DenseNet to extract the attended region features $f_l \in R^{1*1024}$ from the final pooling layers, then perform the element-wise operation on f_g and f_l to produce the fusion signals f_f. To balance the deviation in medical tags, we optimize the parameters of three branch via focal loss, as follows:

$$p_i^* = \begin{cases} p_i, & if \;\; y_i = 1 \\ 1 - p_i, & otherwise \end{cases} \tag{3.10}$$

$$L_{focal} = -\sum_{i=0}^{N-1} \alpha(1-p_i^*)^\gamma \log p_i^* \tag{3.11}$$

where y_i represents the label, p_i represents the prediction probability, α is a hyper-parameter set according to diverse datasets, and $(1-p_i^*)^\gamma$ is treated as a modulating factor with a tunable focusing parameter $\gamma \geq 0$. We set α to 0.25 and γ to 2 in our task.

3.3 Experiments and Results

3.3.1 Datasets

We invited three Chinese radiologists with more than five years of working experience to apply their diagnostic skills to the public COVID-CT [118] and use these image-report pairs to construct the COV-CTR. All the images are lung CT-scans and collected from the published papers. The references to these papers are listed in [118]. Notably, the quality of these images are degraded in following aspects: the Hounsfield unit (HU) values are lost; the number of bits per pixel is reduced; the resolution of images is reduced. However, as explained in [118], experienced radiologists are able to make an accurate diagnosis from low quality CT images. For example, given a photo taken by smart phone of the original CT image, experienced radiologists can make an accurate diagnosis by just looking at the photo, though the CT image in the photo has much lower quality than the original CT image. Likewise, the quality gap between CT images in papers and original CT images will not largely hurt the accuracy of diagnosis.

For each image in COV-CTR, we present the related reports and the impression which indicates the patient is COVID or not. There are 349 and 379 images for COVID and Non-COVID, respectively. More details and comparisons with other datasets are reported in Table. 3.1 Medical report generation tasks aim to describe all the visual grounding in the image with medical terminologies. Therefore, one CT scan is enough for neural models to diagnose.

Table 3.1: Statistics of COV-CTR, CX-CHR and Open-IU.

Statistics	COV-CTR	CX-CHR	IU X-Ray
Patients	-	35,609	3867
Images	728	45,598	7470
Normalities	-	18	-
Abnormalities	-	155	-
Vocabulary Size	235	27683	2791
Max. Sen. Num.	14	24	18
Max. Sen. Len.	37	38	42
Max. Rep. Len.	127	216	173
Avg. Sen. Len.	8.197	7.111	6.997
Avg. Rep. Len.	77.274	64.858	32.450

We conduct experiments on both Chinese annotated CX-CHR, COV-CTR dataset and English described Open-IU dataset in order to validate the robustness and generalization ability of ASGK. CX-CHR is a large-scale chest X-ray dataset, constructed by a professional medical institution, that consists of 35,609 patients and 45,598 images paired with their corresponding Chinese diagnostic reports. We collect 173 medical tags comprising 155 abnormal terminologies and 28 normal terminologies from the ’findings’ section and annotate paired images with these tags. Moreover, the COV-CTR datasets consist of 728 images (349 for COVID-19 and 379 for Non-COVID) collected from published papers and their corresponding paired Chinese reports. We perform the same operation described above and collect 68 tags (50 abnormalities and 18 normalities). We adopt the same

Chinese textbook when conducting experiments on two Chinese datasets. We tokenize all reports and the medical textbook and filter tokens with a minimum frequency of three, which results in 27683 unique Chinese tokens covering over 98.7% of words in the corpus including four special tokens *pad*, *eos*, *sep* and *unk*. On both Chinese datasets, we randomly split the data into training, validation, and testing sets using a ratio of 7 : 1 : 2; there is no overlap between these branches.

We perform the same operations on the Open-IU dataset to clarify the performance of our ASGK to generate English medical reports, we collected medical papers‚Äô abstracts from Pubmed to construct the English Medical Textbook and provide the external signals with 2791 unique English tokens. Then we included 20 finding keywords as disease categories the same as [116] to extract the internal signals.

3.3.2 Evaluation Metrics

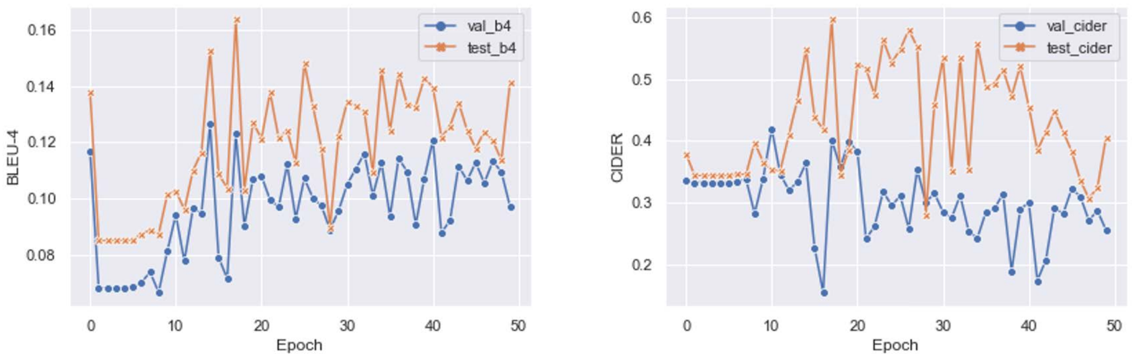


Figure 3.3: We evaluate our model each epoch and report BLEU-4 and CIDER values on validation and testing sets.

Following [49], we adopt three kinds of metrics to evaluate our approach. Firstly, we use area under the curve (AUC) to evaluate the performance of all medical tag classifications. We compare our approach with existing approaches, including conventional natural image captioning models and typical medical report generation pipelines on the metrics including CIDER-D [97], ROUGE-L [60], BLEU [76] and clinical efficacy. Most existing medical report generation approaches adopt the BLEU-4 as the primary metric. However, as shown in Fig. 3.3, the model achieves a high BLEU value in the first epoch, where all outputs of models are the same. Obviously, BLEU has limits on evaluating medical reports. Compared with BLEU, CIDER pays more attention to the different words between each sentence, and most of the words describe abnormal terminologies in this task. Therefore, we adopt the CIDER as our primary metric. We also conduct human evaluation, inviting senior radiologists to judge the quality of generated reports. Specifically, we randomly select 200 samples from the testing set and generate corresponding medical reports using CoAtt [39] and our approach. Then we invite senior radiologist to find which predicted reports are described the given images more accurately.

3.3.3 Training Details

The whole network is implemented using a PyTorch framework based on Python 3.6 and trained on two GeForce RTX 2080Ti GPUs. We adopt DenseNet-121 with no pretraining as the backbone to extract visual features. There are three steps in our training process: external auxiliary signal-guide pretraining, DenseNet pretraining, and internal auxiliary guide training. In the first step, the maximum length of the sentence is 300 (padded with 0s), and the word embedding dimension is 300. We train ASGK for 30 epochs until convergence. The natural language decoder consists of three blocks. We adopt ADAM for optimizing and the training rate is 5e-4. For the second step, we resize the image to 224×224 for both global and region images. The batch size is 32. We jointly train two DenseNets for 50 epochs until convergence. The learning rate starts from 1e-2 and delays by 0.1 every 10 epochs until 1e-5. We threshold the heat map by 0.7 to acquire region images. We adopt the model that achieves the best performance on test datasets as a visual extractor in the third step. In the final step, we resize the images to 224×224 and train the entire network for 30 epochs until convergence. The learning rates for the visual extractor and ASGK are 1e-5 and 5e-4, respectively. We also adopt the ADAM optimizer to minimize the loss function. Among the multi-tasks, we set all loss weights to 1.

3.4 Results and Analysis

3.4.1 Automatic Evaluation

Table 3.2 summarizes the performances on the automatic evaluation metrics of different models. The results on both datasets indicate that ASGK outperforms all existing state-of-the-art models through its exploitation of auxiliary signals to guide the framework in knowledge pretraining and knowledge transfer procedures. The results demonstrate the robustness and superior generalization ability of ASGK. We also combine our medical graph encoder with V-Bert [19] and V-GPT[83] in order to validate the capability of the

Table 3.2: Evaluation metrics on CH-CHR and COV-CTR datasets comparing ASGK with other methods. C and R are short for CIDER-D and ROUGE-L. B-n denotes that the BLEU score uses up to n-grams. Hit represents the human evaluation results.

Dataset	Model	C	R	B@1	B@2	B@3	B@4	Hit(%)
CX-CHR	CoAtt	273.5	**64.5**	64.7	57.5	52.5	48.7	8.0
	HRGR	289.5	61.2	67.3	58.7	53.0	48.6	-
	KERP	285.0	61.8	67.3	58.8	53.2	47.3	-
	V-BERT	302.4	63.7	**68.6**	60.1	54.1	50.3	19.0
	V-GPT	301.8	63.0	67.9	59.6	54.0	48.7	-
	SAT	311.2	63.3	62.3	55.2	53.9	48.1	-
	R2Gen	310.2	63.3	68.1	60.2	54.3	50.1	-
	Ours	**324.5**	64.1	**68.6**	**60.8**	**55.8**	**52.3**	**20.0**
COV-CTR	CoAtt	67.2	**74.8**	70.9	64.5	60.3	55.2	25.0
	SAT	65.9	72.3	69.7	62.1	56.8	51.5	-
	AdaAtt	68.2	72.6	67.6	63.3	59.6	51.4	-
	V-BERT	**68.4**	74.7	71.0	65.3	60.6	55.8	26.0
	V-GPT	68.0	74.6	70.8	64.5	60.0	54.9	-
	R2Gen	67.2	73.2	69.3	61.1	55.9	51.8	-
	TopDown	63.1	72.1	70.5	65.3	60.9	56.1	-
	Ours	**68.4**	74.6	**71.2**	**65.9**	**61.1**	**57.0**	**27.0**

language-to-vision transfer. We adopt CIDER-D as the main metric to validate our model. On the large-scale CX-CHR dataset, ASGK significantly boosts performance compared with other baselines, it increases the CIDER score by 51.0, 35.0, 39.5, 22.1 and 22.7 respectively. It is observed that ASGK achieves a slightly low ROUGE-L score than the CoAtt[39] method. We speculate the reason is that ROUGE-L is proposed to calculate the longest common subsequence among the ground truth and predicted reports. In the meantime, it fails to evaluate sentences' frequency. Our ASGk can generate fine-gained sentences covering more abnormal terminologies. But, there are some subtle differences between the way they are expressed and the ground truth report. ASGK also outperforms other baselines in COV-CTR dataset.

Compared with the results present in Table 3.3, ASGK performed better than TieNet [103], CARG [68], SentSAT [113] and SentSAT+KG [116]. The most Cider score indicates that our generated reports have the least redundancy as there are many similar sentences used in each medical report to describe the normal terminology in which patients care less.

Table 3.3: Comparison of report generation models on three metrics on the Open-IU dataset. As some of their works are outsourced, we directly use the results reported in their papers.

Model	Bleu-4	Cider-D	Rouge-L
CARG [68]	11.3	-	35.4
KERP [49]	**16.2**	28.0	33.9
TieNet [103]	8.1	-	31.1
SentSAT [113]	14.3	26.8	35.9
SentSAT+KG [116]	14.7	30.4	**36.7**
Ours	12.5	**30.6**	27.9

3.4.2 Medical Tags Classification

The AUCs of medical tag classification, which contains both normal and abnormal terminologies on both datasets, are presented in Table 3.4. Our framework, which is guided by two auxiliary signals, outperforms the baseline on both datasets. Baseline outputs are predicted by a DenseNet-121 without pretraining. We attempt to boost the performance through the use of internal auxiliary signals and the adaptation of focal loss to balance the deviation. This demonstrates that internal auxiliary signals effectively promote the medical graph encoder and facilitate the medical tag classification.

3.4.3 Human Evaluation

Given 200 random images from these two datasets equally, we invited three radiologists to evaluate the corresponding outputs of our methods, CoAtt[39] and Vison-Bert[19]. They are encouraged to select a more accurate result from each pair. The human evaluation results are presented in Table 3.2. It shows that in the CX-CHR and COV-CTR datasets, radiologists thought 20%, and 27% portions of our reports are more accurate than others' respectively, and while they thought 53%, and 22% portions of results are same. The human evaluation demonstrates that our method is capable of generating accurate and semantic-coherent reports.

3.4.4 Visualization

An illustration of heat maps, suspicious regions, is presented in Figure 3.4. It is clear from the results that suspicious regions suggest the region on which the model should focus. For example, in the first row, the auxiliary region focuses on the inferior lobe of the left lung which presents a shadow. In the fourth row, moreover, the auxiliary region

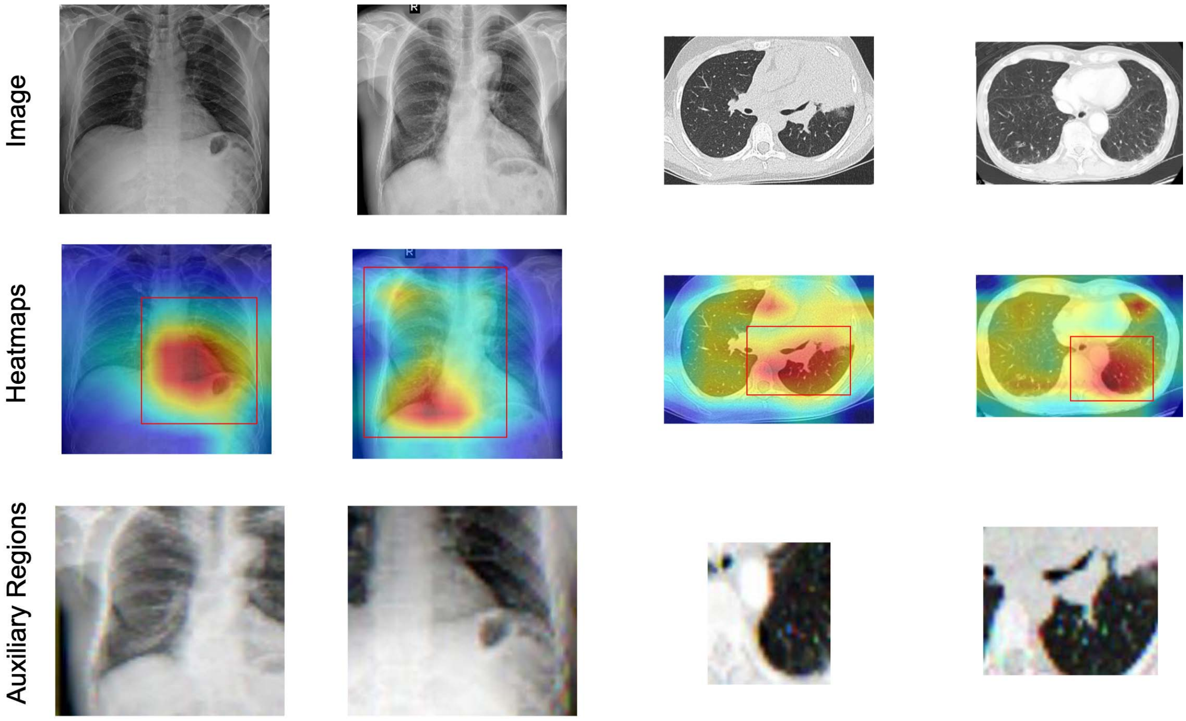


Figure 3.4: Sample output of our approach on both CX-CHR and COV-CTR datasets. We use the outputs before the last pooling layer in DenseNet-121 to generate heat maps, then threshold them by $\tau = 0.7$ to produce the suspicious regions

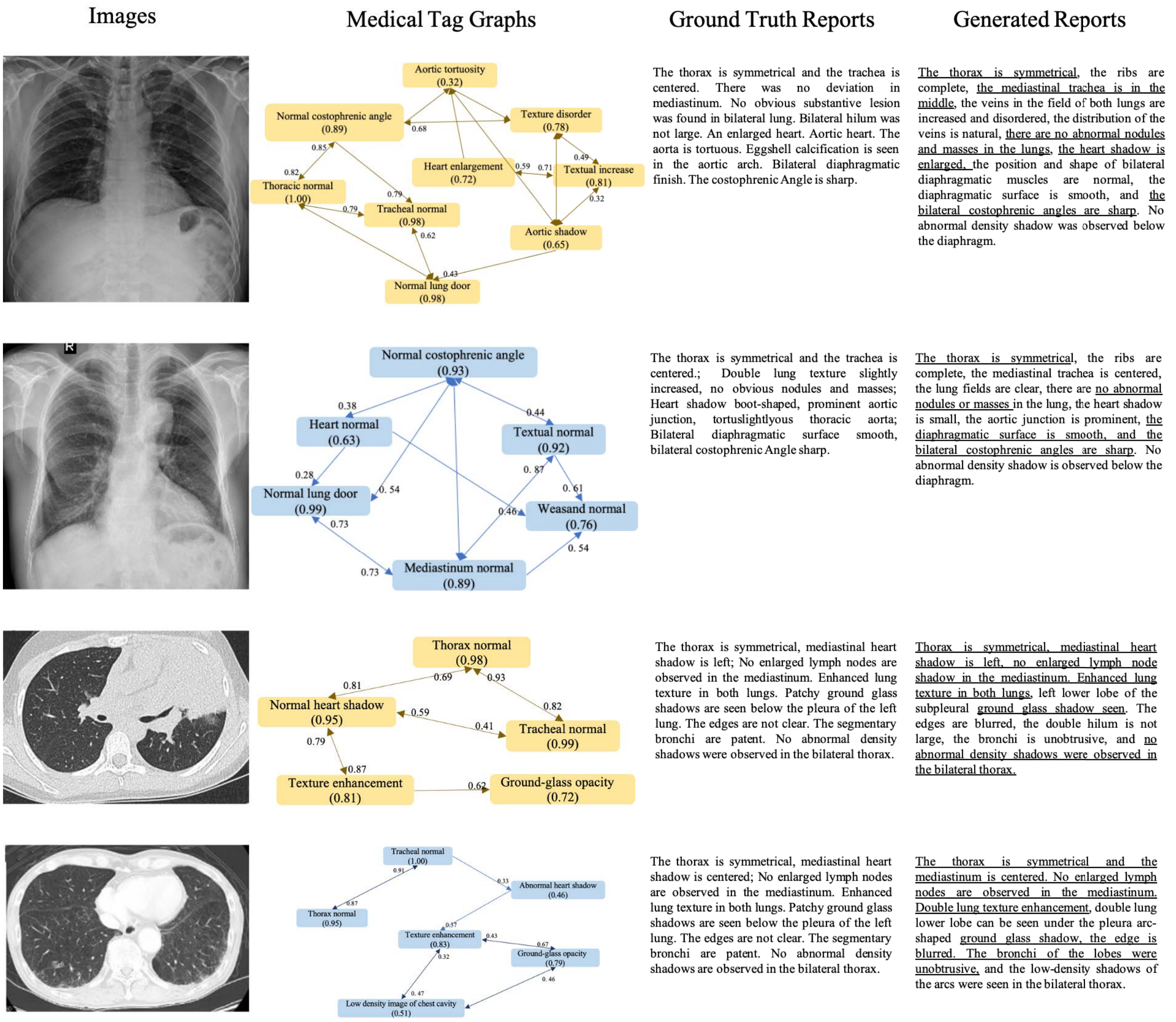
.

focuses the inferior pleural of the left lung, which covers ground-glass opacity, one of the symptoms of COVID-19.

Figure 3.5 shows the illustration of medical tag graphs, and paragraphs of medical reports. The medical tag graph demonstrates that ASGK is capable of encoding input features into a high-level knowledge graph; as we lack the ground truth of the corresponding graph, we train in an end-to-end way to encode the graph. The generated reports demonstrate the high quality and provide significant alignment with the ground truth.

3.4.5 Ablation Studies

We conduct ablation experiments to compare the performance of the two auxiliary signals. Table 3.4 presents the results of automatic evaluation metrics and tag classification. The baseline represents the direct training of the ASGK model without any auxiliary signals. In addition to extra notes, we adopt focal loss as our training strategy.



The thorax is symmetrical and the trachea is centered. There was no deviation in mediastinum. No obvious substantive lesion was found in bilateral lung. Bilateral hilum was not large. An enlarged heart. Aortic heart. The aorta is tortuous. Eggshell calcification is seen in the aortic arch. Bilateral diaphragmatic finish. The costophrenic Angle is sharp.

The thorax is symmetrical, the ribs are complete, the mediastinal trachea is in the middle, the veins in the field of both lungs are increased and disordered, the distribution of the veins is natural, there are no abnormal nodules and masses in the lungs, the heart shadow is enlarged, the position and shape of bilateral diaphragmatic muscles are normal, the diaphragmatic surface is smooth, and the bilateral costophrenic angles are sharp. No abnormal density shadow was observed below the diaphragm.

The thorax is symmetrical and the trachea is centered.; Double lung texture slightly increased, no obvious nodules and masses; Heart shadow boot-shaped, prominent aortic junction, tortuslightlyous thoracic aorta; Bilateral diaphragmatic surface smooth, bilateral costophrenic Angle sharp.

The thorax is symmetrical, the ribs are complete, the mediastinal trachea is centered, the lung fields are clear, there are no abnormal nodules or masses in the lung, the heart shadow is small, the aortic junction is prominent, the diaphragmatic surface is smooth, and the bilateral costophrenic angles are sharp. No abnormal density shadow is observed below the diaphragm.

The thorax is symmetrical, mediastinal heart shadow is left; No enlarged lymph nodes are observed in the mediastinum. Enhanced lung texture in both lungs. Patchy ground glass shadows are seen below the pleura of the left lung. The edges are not clear. The segmentary bronchi are patent. No abnormal density shadows were observed in the bilateral thorax.

Thorax is symmetrical, mediastinal heart shadow is left, no enlarged lymph node shadow in the mediastinum. Enhanced lung texture in both lungs, left lower lobe of the subpleural ground glass shadow seen. The edges are blurred, the double hilum is not large, the bronchi is unobtrusive, and no abnormal density shadows were observed in the bilateral thorax.

The thorax is symmetrical, mediastinal heart shadow is centered; No enlarged lymph nodes are observed in the mediastinum. Enhanced lung texture in both lungs. Patchy ground glass shadows are seen below the pleura of the left lung. The edges are not clear. The segmentary bronchi are patent. No abnormal density shadows are observed in the bilateral thorax.

The thorax is symmetrical and the mediastinum is centered. No enlarged lymph nodes are observed in the mediastinum. Double lung texture enhancement, double lung lower lobe can be seen under the pleura arc-shaped ground glass shadow, the edge is blurred. The bronchi of the lobes were unobtrusive, and the low-density shadows of the arcs were seen in the bilateral thorax.

Figure 3.5: Sample output of our approach on both CX-CHR and COV-CTR datasets. In the medical tag graphs, we show the nodes whose value (which is equal to the classification probability) exceeds 0.5 and edges whose weights are more than 0.3. To read the image clearly, we show the values of some edges in the appropriate places. The underlined text indicates alignment between ground truth reports and generated reports.

3.4.5.1 Do internal auxiliary signals help?

From Table 3.4, we can determine that auxiliary signals significantly boost the tag classification performance and improve the quality of generated reports. The internal auxiliary signal-guided learning outperforms the automatic metrics 15.6%, 1.4% and 0.6% respectively, and also performs 4.5% better than the baseline in terms of classification accuracy on the CX-CHR dataset. The quality of the medical tag graphs significantly

Table 3.4: Ablation studies for different auxiliary signals. IA, EA and CE are short for “internal auxiliary signals”, “external auxiliary signals’ and “cross entropy”. Four metrics are adopted to evaluate our model on two datasets.

Dataset	Model	CIDER-D	ROUGE-L	BLEU-4	AUC
CX-CHR	baseline	289.7	61.3	48.3	78.7
	baseline+IA+CE	304.6	62.5	48.9	82.1
	baseline+IA	305.3	62.7	49.1	83.2
	baseline+EA	317.2	63.8	52.0	79.3
	baseline+IA+EA	**324.5**	**64.1**	**52.3**	**85.9**
COV-CRT	baseline	59.1	68.3	52.5	72.7
	baseline+IA+CE	61.3	70.2	54.1	79.0
	baseline+IA	62.8	70.5	54.2	79.7
	baseline+EA	66.9	72.0	55.6	74.5
	baseline+IA+EA	**68.4**	**74.6**	**57.0**	**80.4**

impacts the natural language decoder. We produce internal auxiliary signals to mimic radiologists’ working patterns, since abnormal regions provide richer visual features. These experiments demonstrate that focusing on abnormal regions benefits the detection of medical tags and the generation of medical reports.

3.4.5.2 What is the use of focal loss?

Radiologists are asked to describe all of their observations on one medical image, which leads to serious data deviation on medical tag labels and reports. Typically, each image contains three to five normal tags and a few abnormal terminologies. To alleviate the deviation in multi-label classification tasks, we adopt focal loss in order to optimize the parameters in DenseNet and the medical tag decoder. When the second and third rows are compared, the performance shows its capability to balance deviation and improve AUC metrics. Without focal loss, the performances on AUC metrics decrease by 0.9% and 0.7% respectively on the two datasets.

3.4.5.3 Are external auxiliary signals useful?

The external auxiliary signals guide the pretraining procedure to assist the model in memorizing and phrasing medical knowledge. As expected, ASGK benefits a lot from the pretraining procedure. The performance on automatic metrics are boosted substantially from 289.7% to 317.2% and 59.1% to 66.9% on the two datasets respectively, which indicates that external auxiliary signal-guided training is capable of generating accurate and semantically coherent sentence. However, it improves the classification accuracy slightly, by 0.6%, and 1.8% respectively on the two datasets, which demonstrates that exploiting medical domain knowledge primarily promotes the natural language decoder. Furthermore, our findings show that without external auxiliary signals, the model fails to alleviate the data bias and is therefore prone to repeating several specific words and sentences in one report.

Overall, the internal signals mainly facilitate the medical tag encoder's effectiveness in generating fine-grained sentences and describing more medical tags. The external signals enable the natural language decoder to generate more semantically coherent sentences.

3.5 Broader Impacts

This work practically analyzes a meaningful task combined with the computer vision and natural language processing task, medical report generation. Especially when pandemic happens like COVID-19, robust and accurate medical report generation technology is of great clinical value, which can reduce the burden on doctors and enable people to more accurately grasp their health status. We propose an anthropomorphic model, mimicking radiologists' working patterns, to promote the medical report generation task via acquiring easily-accessed auxiliary signals. This approach may inspire those researchers who have limited access to medical image resources to dig deeper into adopting unsupervised learning methods to acquire more auxiliary signals to supervised this task and achieve state-of-the-art performances. However, it still needs more effort to provide theoretical interpretation for these auxiliary signals. And our algorithm should be utilized carefully in clinical practice since medical decisions may lead to live-or-death consequences.

3.6 Conclusions

In this section, we proposed an Auxiliary Signal-Guided Knowledge Encoder-Decoder approach that mimics radiologists' working patterns to generate fine-grained and semantically coherent medical reports. We investigated how to best crop the auxiliary region from the global medical image, how to exploit medical domain knowledge from medical textbook, and how these auxiliary signals work. Experiments demonstrate that ASGK outperforms existing methods and boosts the performance of medical report generation tasks on report generation and tag classification on two medical datasets. Moreover, we have constructed and released a new medical report dataset, COV-CTR, to contribute to the community.

However, our COV-CTR is a relatively small image and report dataset, which poses a more challenging task to the current data-greedy deep learning based networks. We solved this problem by adopting external signals from online medical textbooks to pretrain the knowledge encoder and natural language decoder. However, it will be complicate and time-consuming to collect different modality textbooks.

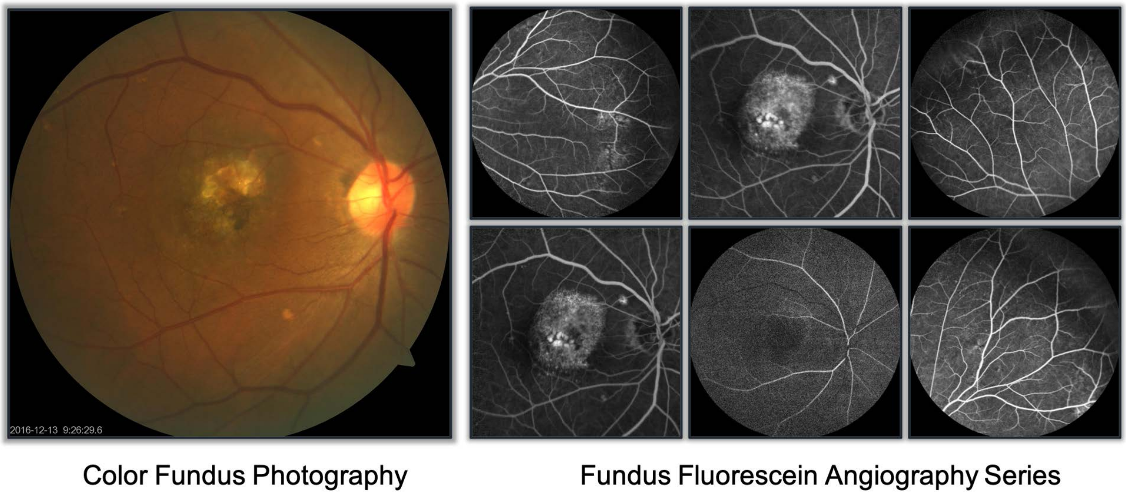
CHAPTER 4

CROSS-MODAL CLINICAL GRAPH TRANSFORMER TOWARDS EXPLAINABLE AND RELIABLE OPHTHALMIC REPORT GENERATION

4.1 Introduction

4.1.1 Fundus Fluorescein Angiography Examination

The World Health Organization (WHO) estimates that 2.2 billion people have visual impairments, and 500 million of them are caused by specific retinal diseases such as age-related macular degeneration (AMD) and diabetic retinopathy (DR)[80]. FFA is one of the most common and essential examination methods in the differentiation, diagnosis, treatment, and prognosis of fundus ophthalmic diseases. FFA is a kind of dynamic imaging procedure, and as shown in Fig. 4.1, with sodium fluorescein flowing through the blood into the fundus vessels, the whole procedure can be divided into five parts: Preaterial, Arterial, Arteriovenous, Venous, and Late period. At different periods, ophthalmologists determine different diseases based on the morphology of different lesions. For example, the nature of new blood vessels in different areas from the fluorescein leakage pattern, the scope and size of the non-perfusion area of the retina. After browsing all the FFA images, ophthalmologists will select several typical FFA images according to their observations and write a report summarizing their findings. This process of reading and interpreting dozens of FFA images is laborious.



Color Fundus Photography

Fundus Fluorescein Angiography Series

Figure 4.1: Patient can receive one CFP image or series of FA images from one diagnosis. Compared with CFP, FA images can present more details about each part of retinal and blood vessel.

Compared with Color Fundus Photography (CFP) imaging, FFA is a high-cost, invasive and complex imaging method but has a high confirmation rate. As some patients may be allergic to fluorescein, FFA is also not suitable for large-scale screening. Therefore, it is challenging and costly to collect large-scale data set FFA images and reports, making the FFA-IR collection highly valuable. A practical, interpretable, and reliable MRG model derived using FFA-IR can assist ophthalmologists in understanding these images and improve the conventional retinal disease diagnosis procedure.

4.1.2 Problem Statement

In this section, we aim at proposing an explainable and reliable ophthalmic report generation (ORG) model to assist ophthalmologists in improving diagnosis efficiency and accuracy. To achieve this goal, we should first construct a reliable and large-scale ORG benchmark, FFA-IR, and then design a data-driven neural network to automatically predict reports.

4.1.2.1 Medical Report Generation Benchmarks

Although DNNs-based methods have made some promising progress in the field of MRG, the black-box characteristics of DNNs discourage specialists and patients from trusting the predicted reports in clinical practice since medical decisions may have life-or-death consequences. To address this limitation, researchers have explored text-image attention mappings [14, 39] to explain the automatic generation procedure. However, the accuracy of these explanations is unclear. Since existing MRG datasets fail to provide explainable annotations, development of interpretable MRG methods to improve trustworthiness is a great challenge.

Besides explainable annotations, the lack of reliable evaluation tools hinders research advances. Natural-language generation (NLG) metrics, including BLEU [76], CIDER [97], Meteor [6] and Rouge [60], have been widely used to evaluate the quality of the predicted reports. These methods focus on the linguistic similarity of target and source sentences and are based on counting the occurrences of overlapping N-grams, in which they treat each word in the sentences equally. They ignore the fact that certain words carry more weight in specific contexts. For image reports, identified lesions and their corresponding attribute descriptors are most important in diagnosis. Thus, these terms should carry larger weights in report quality evaluation than other words [116]. In addition, serious data bias commonly exists in medical reports. For example, a majority of the sentences in reports are descriptions of normal findings. In this context, overall performance in terms of standard NLG metrics will appear to be promising, although models are underfitting, particularly sentences describing abnormal findings and prone to repeat common sentences.

4.1.2.2 Clinical Knowledge Driven Medical Report Generation Approaches

Despite significant progress in generic image captioning models[3, 12], when transferring them into medical knowledge-driven tasks, they fail to achieve promising and competitive performance due to a lack of prior medical knowledge. When describing ophthalmic images, ordinary people can only recognize the common visual information, such as the shape and color, while ophthalmologists make inferences with their prior clinical knowledge. For models to achieve this capability, recent work explores the incorporation of medical knowledge to enhance diagnostic models [50, 65, 116].

On the one hand, researchers[50] have explored graph structure weights as posterior knowledge to alleviate the textual bias. In each graph, the nodes are observed abnormalities selected from prior knowledge, such as external medical corpus, and the nodes are the predicted weights correlating each pair abnormalities. However, the weight graph limits the effectiveness of the knowledge graph from two aspects. Firstly, some entities are extracted from the external medical corpus or knowledge graph database separated from the training corpus. These entities will bring in a heterogeneous embedding space[70] which makes the embedding vectors inconsistent. Secondly, there are no ground truth weights to supervise the message passing procedure, and the model is still prone to be distracted by the visual bias in medical images[65]. On the other hand, a universal graph is proposed with prior knowledge on 20 chest findings[116] to enhance models. Since these findings are not always depicted in one report, incorporating all this knowledge may divert the visual features from their original meaning.

4.1.3 Summary of Achievements

4.1.3.1 FFA-IR Benchmark

On the one hand, we first present a new benchmark, FFA-IR, towards an explainable and reliable MRG benchmark based on **FFA I**mages and **R**eports. There are two main motivations for building and releasing FFA-IR. In terms of clinical application, FFA is one of the most commonly used imaging methods for the diagnosis of retinal diseases [5]. Compared with other imaging methods, FFA can significantly improve the positive diagnosis rate. Thus, there is an urgent need to collect large-scale FFA datasets with images and reports. In terms of scientific research, FFA-IR also provides a new challenge to MRG researchers. Compared against existing MRG datasets [17, 40], which provide only one or two views for each case, FFA-IR provides dozens of medical images for each case. Among these given medical images, only a few may capture lesions. In addition, the lesions are usually localized in a small area of the global image. Thus, we cannot simply concatenate the visual features from different views, as in traditional MRG methods [14, 49, 116] do, because other features will inundate lesion features in the same channel after concatenation.

The unique features of FFA-IR include:

- A large-scale medical dataset. Our FFA-IR contains $10,790$ reports describing $1,048,584$ FFA images in total, representing the most significant number of medical images among the existing medical report datasets. All these data are collected from real-world clinical practice and accurately represent the practical writing patterns of ophthalmologists.

- Explainable annotations. Compared with other datasets, our FFA-IR includes annotations of 46 categories of lesions with a total of $12,166$ regions along with FFA images and reports to make the diagnosis process more explainable.

- Bilingual reports. The original reports obtained in the dataset are in Chinese. To make the dataset more broadly accessible, we also provide translations of these reports in English. The translations were derived from automatic translation followed by expert humans correction.

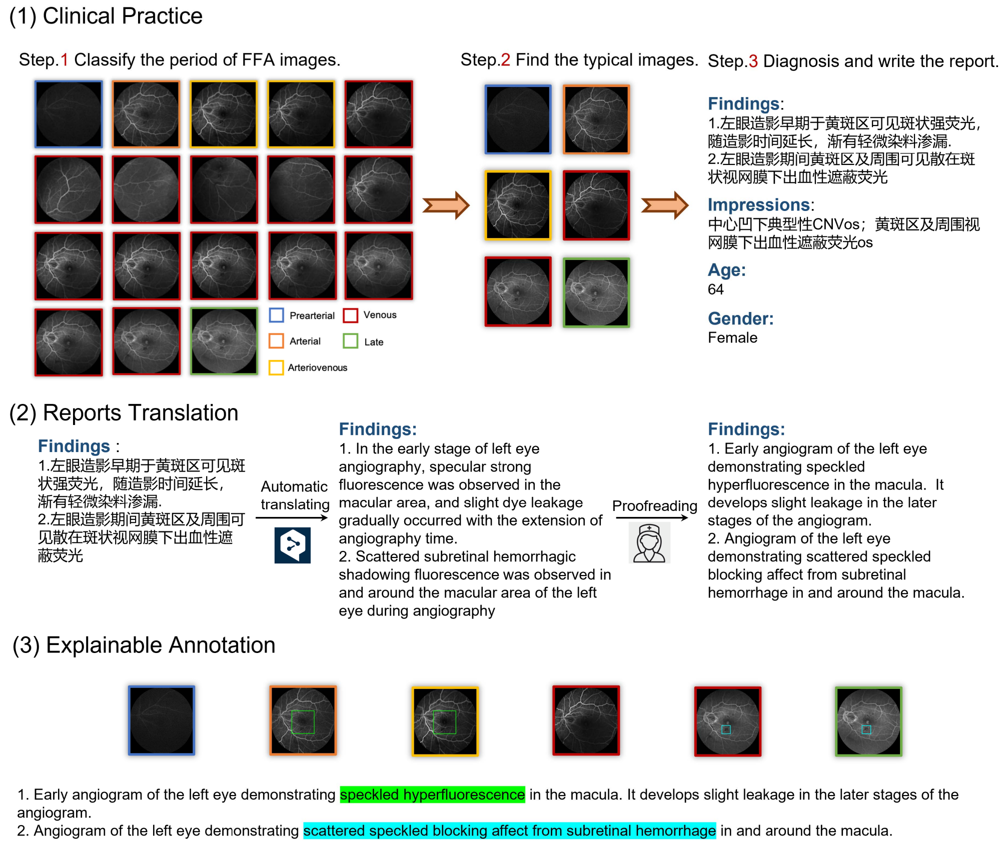
4.1.3.2 Cross-modal Clinical Graph Transformer

On the other hand, we propose a **C**ross-model clinical **G**raph **T**ransformer (CGT) for ophthalmic report generation (ORG). In particular, we first invoke an information extraction scheme based on a natural language processing pipeline, including named entity recognition and entity linking, to obtain a clinical knowledge graph. As discussed in [36], the structured clinical information behind the free-text reports can enhance the diagnostic methods. In addition, the entities and relations in our clinical graph are in the homogeneous embedding space with the training corpus. Given a set of ophthalmic images, the extracted visual features are transformed to a compressed visual token and a subgraph with relevant restored triples. Since the sub-graph is not guaranteed to be a completely accurate representation of the given images and natural noise exists in the clinical graph, we adopt a cross-modal encoder to encode the universal feature token and sub-graph information. To avoid influence from unrelated entities, a visible matrix is introduced during the cross-modal encoding process. Finally, reports are generated via a Transformer[96] decoder.

We also conduct extensive experiments on our FFA-IR benchmark[55]. Experiments show that our CGT achieves the state-of-the-art performance of predicted reports under four automatic evaluation metrics and high AUC scores for the restored triples, providing a solid rationale for the explanation.

4.2 FFA-IR Benchmark

In this section, we introduce the process to build FFA-IR, a dataset focusing on diagnosing FFA images. In general, for each case, we provide: 1) the clinically annotated Chinese reports and the translated English reports; 2) annotated lesion information, including lesion category and regions on FFA images, to explain the diagnostic procedure. We summarize our process for creating FFA-IR in Figure 4.2.



Findings:
1.左眼造影早期于黄斑区可见斑状强荧光，随造影时间延长，渐有轻微染料渗漏.
2.左眼造影期间黄斑区及周围可见散在斑状视网膜下出血性遮蔽荧光

Impressions:
中心凹下典型性CNVos；黄斑区及周围视网膜下出血性遮蔽荧光os

Findings :
1.左眼造影早期于黄斑区可见斑状强荧光，随造影时间延长，渐有轻微染料渗漏.
2.左眼造影期间黄斑区及周围可见散在斑状视网膜下出血性遮蔽荧光

Findings:
1. In the early stage of left eye angiography, specular strong fluorescence was observed in the macular area, and slight dye leakage gradually occurred with the extension of angiography time.
2. Scattered subretinal hemorrhagic shadowing fluorescence was observed in and around the macular area of the left eye during angiography

Findings:
1. Early angiogram of the left eye demonstrating speckled hyperfluorescence in the macula. It develops slight leakage in the later stages of the angiogram.
2. Angiogram of the left eye demonstrating scattered speckled blocking affect from subretinal hemorrhage in and around the macula.

1. Early angiogram of the left eye demonstrating speckled hyperfluorescence in the macula. It develops slight leakage in the later stages of the angiogram.
2. Angiogram of the left eye demonstrating scattered speckled blocking affect from subretinal hemorrhage in and around the macula.

Figure 4.2: Process for creating FFA-IR. Firstly, we collect FFA images and reports from the clinical practice. To translate the reports, we invited bilingual ophthalmologists to proofread the automatically translated documents. They also labeled the described lesions along with FFA images and reports to provide explainable annotations.

4.2.1 Data collection and Annotation

The data were collected from patients at the Zhongshan Ophthalmic Center of Sun Yat-Sen University in Guangzhou, China, during the period between 11/2016 and 12/2019. Institutional review board (No.2021KYPJ039) and ethics committee approval were obtained in Zhongshan Ophthalmic Center, Sun Yat-Sen University. This study followed the tenets of the Declaration of Helsinki [4]. All angiography images and reports were anonymized and de-identified before the analysis.

During the data collection period, our system captured 15, 232 reports, containing findings, impressions, and clinical information, along with 1, 716, 825 DICOMs in which clinical information and pixel values of FFA images are stored. However, we removed some reports and FFA images due to data quality issues. First, there were some reports that could not be matched to FFA images with the same case ID number; Second, the pixel values were missing for some images when we converted the DICOMs to JPG pictures; Third, some reports were incomplete, with key information like findings or impressions missing. After processing the raw data, we finally obtained $10,790$ reports with $1,048,584$ FFA images for our FFA-IR data set.

4.2.1.1 Annotator Information

The original medical reports were generated by about 12 ophthalmologists of the fundus department. Around five ophthalmologists with 1-3 years of experience in fundus diseases generated reports under the supervision of residents or attending physicians in fundus specialty. About 3-4 residents or attending physicians in total, each with over five years of experience in the clinical retina, all of whom created reports independently. Finally, 3-4 senior retinal specialists with the title of professor or associate professor had been in the field of the retina for more than 15 years. They either wrote reports independently or helped make final decisions on complicated cases.

For image labeling and annotation, three ophthalmologists with about 2-5 years of experience in ophthalmology labeled lesion regions on FFA images. 2 residents with more than seven years in ophthalmology verified all lesion labels. One professor and one associate professor in ophthalmology checked the accuracy of the random sample of the labels and helped make final decisions on complicated cases. Images with problematical labels were discussed until all specialists agreed on the grading.

4.2.1.2 Explainable Annotations

Our FFA-IR annotation schema includes 46 categories of retinal lesions, such as Cystoid Macular Edema (CME) and Diabetic Macular Edema (DME). The schema was developed by the ophthalmologists based on their expert knowledge, and covers most typical retinal lesions. The schema can be viewed as defining the set of ”explanations” that are relevant to the interpretation of the FFA images.

The ophthalmologists annotated each lesion with its minimum enclosing rectangle and providing the lesion category. All the lesions in one FFA are recorded in a dict format, and the key name is the combination of the case ID and the image name while the value is a list data, and each element contains the category and positional information.

The medical reports aim to describe the size, location, and period of detected lesions on the corresponding images. Therefore, any lesions annotated on the images should also be described in the report. The terms corresponding to each of the 46 categories of retinal lesions can be identified in the reports, and used to evaluate the accuracy of explanations generated by the models. Effectively, the schema serves as prior medical knowledge that enables connecting the visual features on the images and the linguistic information describing those features.

4.2.1.3 Bilingual Reports

To make the dataset more broadly accessible, we translate these reports to English and provide bilingual reports for each case. As it is laborious to translate tens of thousands of reports, we firstly uses DeepL Translator[105] to automatically translate all the reports and invited the bilingual ophthalmologists to proofread these reports. Due to the particularity of the Chinese language, we also provided a vocabulary containing medical nomenclature to help researchers tokenize the Chinese reports. Along with the bilingual reports, FFA-IR is the first benchmark to evaluate qualitative and quantitative influences of different languages on MRG methods. Thanks to these bilingual reports, FFA-IR can also facilitate the development of multi-modal machine translation models.

4.2.2 Dataset Statistics

We report the statistics of our FFA-IR in Table 4.1. In total, our FFA-IR dataset contains $10,790$ cases describing $1,048,584$ FFA images. For each case, FFA-IR provides FFA images, free-text reports, and explainable annotations.

Five percent of the cases in FFA-IR are entirely healthy and are negative training samples.[1] Consistent with most large-scale datasets for deep learning research, we created standard splits, separating the whole dataset into 75%, 10%, 15%, *i.e.*, $8,016$ (train), $1,069$ (val), and $1,604$ (test) cases, respectively. The vocabulary sizes of English and Chinese reports are 918 and $6,181$, respectively. The training corpus covers most of the words, with words appearing less than 3 times in the corpus replaced by <unk> during the training process. Training Chinese models requires a larger wording embedding space which may influence the efficiency. Furthermore, there is no obvious data bias in the Gender and Age distributions. There are slightly more reports describing the left eye than the right in FFA-IR. The resolutions of FFA images in FFA-IR range from 384×384 to 3216×2696.

[1] We note that this data set may therefore differ from data sets derived from diagnostic screening applications (such as breast cancer screening), where the positive samples would be expected to be in the minority.

Table 4.1: The FFA-IR dataset statistics, where * represents the average number.

	Attribute	Train		Val		Test	
		En	Zh	En	Zh	En	Zh
Report	Length*	63.4	91.3	63.6	91.1	63.5	91.0
	Vocabulary(%)	89.1	95.4	39.0	68.1	46.1	73.6
Case	Number		8,016		1,069		1,604
	Healthy(%)		5.6		6.1		5.5
	Unhealthy(%)		94.4		93.9		94.5
FFA	Image*		87.2		87.3		86.0
Gender	Male(%)		55.6		54.4		57.8
	Female(%)		44.4		45.6		42.2
Eyes	Right(%)		29.0		30.1		29.7
	Left(%)		39.6		38.2		40.1
	Both(%)		31.4		31.7		30.2
Age	Average		47.7		47.6		47.8
	Range		3~92		3~87		4~91
Lesion	Number		9,336		1,220		1,610
	Category		46		46		46

4.2.3 Data and Code Availability

Our dataset with all images and documentation, including bilingual reports, findings, explainable annotations, and lesion code dictionary, is hosted and maintained on PhysioNet under the following license: PhysioNet Credentialed Health Data License 1.5.0. It can be accessed at the following link: https://physionet.org/content/ffa-ir-medical-report/1.0.0/. Our start up codes can be accessed at https://github.com/mlii0117/FFA-IR, under the MIT licences.

4.2.4 More Data Usage Discussion

Our FFA-IR data set can be used in various medical image analysis domains along with explainable annotations and bilingual reports. We highly recommend three cases. The first one is to develop an explainable and reliable MRG model to describe lesions relating to retinal diseases identified in FFA images, in Chinese or English language reports. Secondly, due to the dynamic imaging procedure, exploring the use of temporal information or interactions between related images to improve lesion detection or disease classification should be encouraged. The last case is to develop a multi-modal machine translation model. It would be interesting to investigate whether medical images can facilitate aligning the source and target sentences in the latent space. There are limited resources available for machine translation in biomedicine, with existing resources focusing primarily on scientific literature [7, 37].

Prior errors may exist due to the unbalanced distributions across attributes, such as gender and age. As recommended by Saahil *et al.*[36], researchers should audit performance disparities across these attributes when developing clinical models.

4.3 Cross-modal Clinical Graph Transformer

In this section, we introduce the clinical graph extraction scheme, and the process is shown in Figure 4.3. Then we detail the implementation of CGT, and the overall framework for ophthalmic report generation (ORG) is presented in Figure 4.4.

4.3.1 Notation

In ORG task, given a set of FFA images which represented by $I = \{x_1, x_2, \ldots, x_{Ni}\}$, where x_j and Ni refer to the j-th FFA image and the number of total images, model is asked to generate a descriptive report encoded as $R = \{y_1, y_2, \ldots, y_{Nr}\}$. While we denote the ground truth report by $\hat{R} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{N\hat{r}}\}$. We extract entities and relations from all the training $\hat{R}$ to construct a clinical graph (CG), denoted as $\mathscr{G}$, which is a collection of triples $\epsilon = (e_s, r, e_o)$, where e_s and e_o denote the names of subjective and objective entities, and r is the relation between them. All the triples are in CG, i.e., $\epsilon \in \mathscr{G}$. In this work, English tokens are taken at the word-level and each token y_i, e_i and r_i are in the same vocabulary $\mathscr{V}$ whose size is d_V to make all the embedding vectors consistent.

4.3.2 Clinical Graph Extraction Scheme

Recently, extracting clinical information from medical reports has received increasing attention[36, 107]. The structured clinical information within the free-text reports is valuable for clinical reasoning and a variety of critical healthcare applications. We believe that ORG is one such application. However, due to the huge domain discrepancy between different medical models, transferring information from existing biomedical knowledge databases is unlikely to be effective. In this subsection, we will introduce our information extraction scheme to detail how we construct a clinical graph $\mathscr{G}$ from ophthalmic reports. This scheme is implemented by a SpaCy[73] natural language parser in an AI accelerating human-in-the-loop manner[108]. Notably, the ophthalmic reports used in this scheme are all derived from the training set to avoid target leakage.

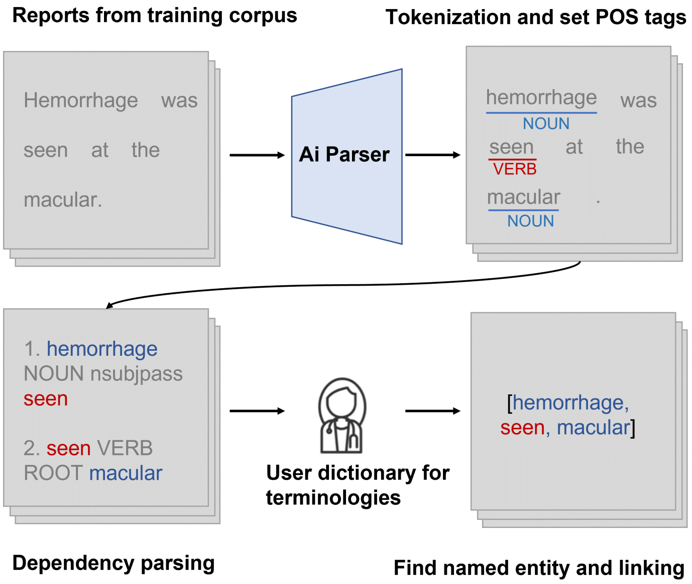

Reports from training corpus
Tokenization and set POS tags
Hemorrhage was seen at the macular.
Ai Parser
hemorrhage NOUN was seen VERB at the macular NOUN .
1. hemorrhage NOUN nsubjpass seen
2. seen VERB ROOT macular
User dictionary for terminologies
[hemorrhage, seen, macular]
Dependency parsing
Find named entity and linking

Figure 4.3: Process for extracting entities and relations from ophthalmic reports.

Table 4.2: Statistics of our clinical graph.

# Entities	# Relations	# Triples
1,811	29	4,823

To save the writing space, we take one sentence, "*Spotted obscured fluorescence (hemorrhage?) was seen at the inferior edge of the macular arch ring during left eye imaging.*" from an ophthalmic report as an example, and the whole process is shown in Figure 4.3. Our scheme contains seven steps by following: **Tokenization**, taking the sentence into word-level and segmenting tokens into words, punctuation marks etc; **Part-of-speech tagging**, before automatically recognizing the relations between each pair tokens, we assign work types to each token, such as verb or nun; **Dependency parsing**, assigning syntactic dependency labels to describe the relations between individual tokens, such as '*spotted*' is the attributive of subjective '*fluorescence*'; **Lemmatization**, digging the base form of tokens. For example, the lemma of 'was' is 'is'; **Sentence boundary detection**, finding individual sentences to prevent the calculation across sentences; **Named entity recognition**, we create a user-dictionary to assist the machine in recognizing rare ophthalmic terminologies, such as 'macular'; **Entity linking**, linking entities with their relation to creating triples. Triples extracted from the sample are "*fluorescence, seen, macular*" and "*hemorrhage, seen, macular*", respectively. Then we collect all the unique triples to construct the whole clinical graph $\mathscr{G}$. In total, our clinical

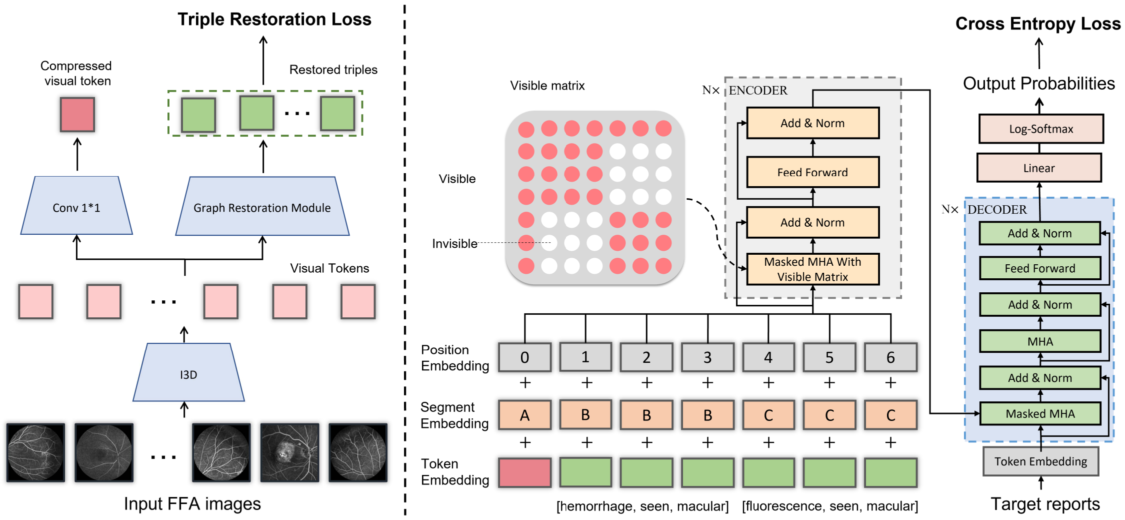


Figure 4.4: Illustration of our proposed cross-modal clinical graph transformer. Visual features extracted by an I3D are utilized to restore sub-graph information and compressed to one visual token. Then the cross-modal information encoded with visible matrix masked multi-head attention is used for report generation.

graph contains 4,823 triples, and more details are presented in Table 4.2.

4.3.3 CGT Framework

The traditional report generation models are based on an encoder-decoder architecture. Among all the encoder-decoder frameworks, Transformer[96] has achieved great success in various tasks. Therefore, we adopt a Transformer, the backbone of our proposed CGT, to describe ophthalmic images from the FFA-IR benchmark. As shown in Figure 4.4, our CGT is composed of a visual extractor, a graph construction module, a cross-modal encoder, and a language decoder.

4.3.3.1 Visual Extractor

Different from describing radiology images, the average number of input images for each case is 97 in the FFA-IR. Despite the benchmark proposed by [55] is adopting lesion features via a Faster-RCNN[84], we utilize an I3D[2] model pretrained on Kinetics[11] to extract visual features from given images. Due to the reason that the entities in our CG contain both abnormalities and normal tissues, while the lesion information provided

[2] https://github.com/piergiaj/pytorch-i3d

by the FFA-IR is all about the lesions or abnormalities. This data bias may mislead the message passing inter the CG.

Since the image numbers are different among each case, we first transform the given images and set a fixed length of 96 for all the input images. For those whose length is more than 96, we randomly down-sample some images. In contrast, we repeat the whole sequence until its length is 96, when its length is below the threshold. The I3D model extracts one feature from every eight images, and the final visual features can be denoted as $f_V = \{f_1, f_2, \ldots, f_{12}\}$, where $f_i \in \mathbb{R}^{12\times1024}$.

4.3.3.2 Graph Restoration Module

The graph construction module is proposed to restore a sub-graph according to the visual features generated by the visual extractor. The sub-graph encoded as $\mathcal{G}_s = \epsilon_1, \epsilon_2, \ldots, \epsilon_{Ngs}$ is a combination of triples. The whole process can be written as follows:

$$\mathcal{G}_s = max(0; BN(conv_{3\times3}(f_v)))W_f + b_f \tag{4.1}$$

where $max(0;*)$ and BN represent the ReLU activation function and batch normalization operation, respectively; $W_f \in \mathbb{R}^{1024\times d_V}$ denotes learnable matrix for linear transformation, while b_f refers to the bias terms. Firstly, we adopt a convolution layer with a 3 × 3 kernel followed by an operation sequence of batch normalization and ReLU activation to fuse the temporal information inside the f_v. Then the output has been projected by a linear transformation layer to the dimension of $d = d_V$. As mentioned, all the tokens in CG are in the same vocabulary with the training corpus; then, each vector is used to restore the index of entity or relation in $\mathcal{V}$.

4.3.3.3 Cross-modal Encoder

In this module, the visual features, and the graph information are encoded by self-attention mechanism[96]. The input of the cross-modal encoder comes from the visual extractor and the graph restoration module. As mentioned in [65, 104], serve visual bias exists in most medical datasets for two reasons: the abnormal regions only take a small portion of the whole image, and the human tissues are highly similar. To alleviate the impact of visual bias, we compress the f_v into one compressed visual token, encoded as $T_v \in \mathbb{R}^d$, and concatenate it with a sub-graph before fed to the embedding layer. The compressed visual token has two more advantages. Firstly, it promises that the sub-graph information is dominant to the input features. More importantly, it can be used to resist the inevitable noise inside the clinical graph adaptively since the knowledge graph can not be completely accurate.

We utilize an ‘argmax’ function on $\mathcal{G}_s$ and transform it into the one-hot format to represent the sub-graph, represented as $T_g = \{t_1, t_2, \ldots, t_{Nt} | t_i \in \mathbb{R}^{d_v}\}$. After concatenation, we feed the cross-modal tokens, encoded as $T = \{T_v, T_g\}$, to the embedding layer. The function of the embedding layer is to convert the cross-modal tokens into embedding representations. Similar to the BERT[19], the embedding representation of CGT is the sum of three parts. Firstly, each token in T_g is converted to an embedding vector of dimension $d = 512$ via a trainable lookup table. Different from BERT, the classification tag $[CLS]$ is replaced by T_v. Secondly, position embedding is added to the token embedding, and the formulation is written as follows:

$$PE_{pos,2i} = sin(pos/1000^{2i/d}) \tag{4.2}$$

$$PE_{pos,2i+1} = cos(pos/1000^{2i/d}) \tag{4.3}$$

where *pos* is the position of each token, i is the index of embedding dimension, and d is the dimension of the hidden states. Lastly, segment embedding is employed to identify each sentence. Notably, we find that most sentences in the training corpus contain two triples. Therefore, we consider every six tokens as a sentence. The T is marked with a sequence of segment tags, $\{A, B, \ldots, B, C, \ldots, C\}$, where A represents the compressed visual token.

Then the embedded tokens are encoded by a cross-modal encoder, the whole process of an encoder layer can be written as:

$$f_e(t) = BN(FFN(e_{attn}) + e_{attn}) \tag{4.4}$$

$$e_{attn} = BN(MMHA(t) + t) \tag{4.5}$$

Where *FFN* represents the feed forward layer, and *MMHA* represents the mask multi-head attention. The feed forward layer contains two linear layers with ReLU activation. It makes sure the dimensions of transformer input and output are the same. Another difference between our CGT and Transformer is that we adopt *MMHA* instead of *MHA* during the encoding process and introduce a visible matrix, M_v, to limit the impact of unrelated triples. The computation between unrelated triples is useless and untrue, which may also lead the changes to the original meanings. The visible matrix is presented in Figure 4.4, and it can limit the message passing inter the sentence or between the universal token. The *MMHA* can be written as:

$$\mathbf{h}_i^t = \text{softmax}(\frac{\mathbf{Q}_i(\mathbf{K}^t)' M_v}{\sqrt{d}})\mathbf{V}^t \tag{4.6}$$

where $\{\mathbf{Q},\mathbf{K}^*,\mathbf{V}^*\}$ are the packed d-dimensional *Query, Key, Value* vectors.

4.3.3.4 Language Decoder

We adopt the vanilla Transformer decoder as our language decoder. The whole process of a decoder layer can be written as:

$$f_d(\mathbf{y}) = BN(FFN(e_{c_attn}) + e_{c_attn}) \tag{4.7}$$

$$e_{c_attn} = BN(MHA(e_{attn}, f_e(\mathbf{t})) + e_{attn}) \tag{4.8}$$

$$e_{attn} = BN(MMHA(\mathbf{y}) + \mathbf{y})) \tag{4.9}$$

where *MMHA* represents the original masked multi-head self-attention, $\mathbf{y}$ is the input of decoder and y_t is the $t-$th input token in time step t. Cross-attention sublayer receives the output of encoder $f_e(\mathbf{t})$ and previous sublayer e_{attn}. In where, for each head, $\{\mathbf{Q},\mathbf{K}^*,\mathbf{V}^*\}$ comes from $\mathbf{Q} = W_q * e_{attn}$, $\mathbf{K} = W_k * f_e(\mathbf{x})$, and $\mathbf{V} = W_v * f_e(\mathbf{x})$, where W_* is the weight of a Linear layer. The $f_d(\mathbf{y})$ will be sent to a Linear & Log-Softmax layer to get the output of target sentences. Notably, only token embedding is adopted during the decoding procedure. The entire recursive generation process can be written as follows:

$$p(\hat{R}|I) = \prod_{t=1} p(\hat{y}_t|\hat{y}_1,\ldots,\hat{y}_{t-1},I) \tag{4.10}$$

4.3.3.5 Objectives

Similar to the image captioning tasks, existing medical report generation approaches adopt cross-entropy loss to evaluate the differences between the predicted and the ground truth reports at the word level. Although many works attempt to explore auxiliary signals to drive the report generation, these signals can not supervise the learning process. For example, Li *et al.*[56] introduced an internal visual signal to locate the abnormal regions. However, there is no ground truth for the abnormal region bounding. Similarly, the accurate weights correlated paired findings in [49, 116] are also unavailable. Therefore, the effect of auxiliary signals has been limited.

In this work, we additionally introduce a triple restoration loss [24] to supervise the sub-graph restoration process since our clinical graph extraction scheme provides the ground truth structured information. It guarantees that the accurate graph information will be encoded with the visual features for report generation and is also what makes this method so effective. The cross-entropy loss that is widely used in classification tasks can also achieve the similar goals. However, the considerable amounts of target triples lead to extremely large computation time when operating the Softmax function.

In sum, the total loss function used in our CGT can be written as follows:

$$\mathscr{L}_{RG} = \lambda_{CE}\mathscr{L}_{CE} + \lambda_{TR}\mathscr{L}_{TR} \tag{4.11}$$

where λ_{CE} and λ_{TR} are hyper-parameters balancing two terms. The first loss term $\mathscr{L}_{CE}$ is the cross-entropy loss. The second loss term is the triples restoration loss function to measure the energy of a knowledge triple. The specific process can be written as follows:

$$\mathscr{L}_{TR} = \sum_{\epsilon \in \mathscr{G}} \max(d(\epsilon) - d(f(\epsilon)) + \gamma, 0) \tag{4.12}$$

where $\epsilon = (e_s, r, e_o)$, $d(\epsilon) = |e_s + r - e_o|$, $\gamma > 0$ is a margin hyper-parameter, $f(\epsilon)$ is an entity replacement operation that the subjective or objective entity in a triple is replaced and the replaced triple is an invalid triple in $\mathscr{G}$. Here, e_s,e and e_o are the indexes of the subjective, relation and objective tokens in $\mathscr{V}$.

4.3.4 Evaluation Metrics

4.3.4.1 Report Evaluation

We employ the automatic metrics and nine kinds of human evaluation results to evaluate the quality of the generated reports. The automatic metrics including BLEU [76], Cider [97], Meteor[6] and Rouge [60] aim to calculate the similarity between source and target sentence based on the occurrences of N-gram or word matching. However, these metrics cannot give reliable evaluations for medical fields as the detection of positive disease keywords should largely determine the quality of whole reports. Therefore, we propose a human evaluation of the reports, making use of four experienced ophthalmologists to answer a series of questions about the generated reports:

H1: Are the left and right eyes identified accurately?

H2: Is the imaging period accurately described?

H3: Does this report describe any lesion?

H4: Is the category of the described lesion accurate?

H5: Is the location of the described lesion accurate?

H6: Is the imaging period of the described lesion accurate?

H7: Fluency of the text, on a scale of [1-5] with 5 most fluent.

H8: Intelligibility of the text, on a scale of [1-5] with 5 most intelligible.

H9: The time savings (in seconds) achieved with the help of this report.

For binary questions H1-H6, we set "yes" for 1 and "no" for 0.

For H9, ophthalmologists first record the average time to diagnose one case with the first half samples. Then they will record the average time they used to diagnose one case with the help of generated reports with the remaining samples.

4.3.4.2 Intersection-Over-Union

The existing methods visualize text-image attention mappings to explain the generation process. However, few of them justify the accuracy of their explanations due to a lack of ground truth regions. In FFA-IR, we quantify the accuracy of models' explanation by calculating the Intersection-Over-Union (IOU) (or Jaccard similarity) between the lesion-image attention mapping regions and ground truth regions. Then we draw the minimum rectangles to cover each maximum connection region. The IOU between generated rectangles and annotated regions is calculated. However, to capture the semantics of the annotation, each word in the relevant lesion label must correspond to a word that the model attends to in order to be counted.

4.3.4.3 Mean Average Precision

Medical reports aim to describe lesions from the given medical images and can be considered as the interpretable foundations for disease diagnosis. Therefore, we conduct disease classification experiments and report the mean average precision (mAP) to compare the accuracy of each model.

4.3.4.4 Area Under Curve

Besides the lesion categories, we have to evaluate the accuracy of restored sub-graph. Therefore, we calculate the area under micro-average of receiver operating characteristic curve to verify if our CGT could restore an accurate sub-graph.

4.4 Achievements And Analysis

4.4.1 Experiments Details

Our CGT and other models are all implemented by Pytorch [77] based on Python 3.7 and trained on four GeForce RTX 2080Ti GPUs. The images are resized to 224 before being fed into the I3D. The maximum length of T is 90, padded with tag $[PAD]$. The embedding space for both visual and graph tokens is 512, and the dimension of the hidden states in the Transformer is also 512. Both encoder and decoder consist of six blocks and 8 heads. The ADAM [47] is utilized for optimizing all the parameters in our CGT, and the learning rate is $1e-4$. The whole network is trained for 50 epochs. We adopt greedy decoding when testing. We use greedy decoding for inference models.

4.4.2 FFA-IR Benchmark

In FFA-IR, we pose a new medical report generation task focusing on describing retinal diseases on FFA images. We present benchmark results over FFA-IR using baseline and existing MRG methods.

4.4.2.1 Baseline model

MRG models usually contain two modules, visual extractor, and natural language decoder. In this section, we develop three simple, transformer-based [96] baseline approaches, namely CNN [25]+Transformer [106], I3D [11]+Transformer and Faster-RCNN [84]+Transformer. Firstly, we use ResNet [25], I3D, and Faster-RCNN as the visual extractor to extract spatial, temporal, and object features, respectively. For CNN+T, we employ ResNet [25] to extract the spatial features of each FFA image and then fuse them and feed 49 visual tokens to a transformer. For I3D+T, we first employ I3D pretrained on Kinetics [11] to extract temporal features and pad these features to 49 tokens and then feed to a transformer. For F-R+T, we first use Faster-RCNN pretrained with our lesion regions to extract object features. The object features have also been fused before being sent to a transformer. The batch size of training I3D+T is 32, while others are 2.

4.4.2.2 Existing Approaches

CoAtt[39], Show-Tell[98] and AdaAtt[72] propose similar CNN-LSTM neural network with different attention methods. Top-Down[3] and Gounded[120] extract object fea-

Table 4.3: The results of automatic and human evaluations, where B*N represents the N-gram of Bleu value, H*N represents the index of human evaluation, and T is the short for Transformer[96].

	B1	B2	B3	B4	Meteor	Rouge	Cider	H1	H2	H3	H4	H5	H6	H7	H8	H9
CoAtt[39]	0.313	0.200	0.144	0.111	0.197	0.247	0.254	0.615	0.515	0.430	0.04	0.269	0.315	**4.96**	**4.93**	**20.9**
Show-Tell[98]	0.306	0.197	0.142	0.109	0.191	0.247	0.232	0.646	0.523	0.415	0.02	0.276	0.353	**4.96**	**4.93**	19.7
Top-Down[3]	0.320	0.217	0.162	0.127	0.207	0.289	0.363	**0.684**	**0.584**	0.430	0.01	0.292	**0.376**	4.83	4.70	**20.9**
Gounded[120]	0.396	0.319	0.261	0.218	**0.229**	**0.353**	0.576	0.538	0.361	0.423	0.03	0.307	0.292	4.82	4.84	**20.9**
AdaAtt[72]	0.292	0.181	0.127	0.095	0.205	0.236	0.234	0.553	0.338	0.515	0.06	0.384	0.284	**4.96**	4.78	18.6
R2Gen[14]	0.330	0.225	0.167	0.132	0.210	0.296	0.367	0.423	0.230	0.507	**0.1**	0.361	0.176	4.85	4.82	19.8
CNN[25]+T	0.321	0.211	0.154	0.122	0.198	0.268	0.283	0.423	0.238	0.523	0.079	0.369	0.176	4.77	4.76	18.9
I3D[11]+T	0.428	0.341	0.276	0.229	0.213	0.334	0.561	0.530	0.3	0.461	0.092	0.330	0.223	4.86	4.83	20.7
F-R[84]+T	**0.443**	**0.355**	**0.288**	**0.240**	0.205	0.341	**0.590**	0.590	0.3	**0.576**	0.084	**0.392**	0.215	4.83	**4.93**	18.4

Table 4.4: The results of IOU between lesion-image mapping regions and ground truth.

	B4	C	IoU
CoAtt[39]	0.111	0.254	0.163
R2Gen[14]	0.132	0.367	0.203
CNN[25]+T	0.122	0.283	0.185
F-R[84]+T	**0.240**	**0.590**	**0.312**

Table 4.5: Classification results with using visual features and generated reports, where GT refers to the results using ground truth reports.

	Vision	Report	GT
CoAtt[39]	0.733	0.513	0.728
Top-Down[3]	0.811	**0.531**	
R2Gen[14]	0.734	0.494	
I3d[11]+T	0.762	0.494	
F-R[84]+T	**0.821**	0.527	

tures as visual grounding for captions. R2Gen[14] integrate relational memory into Transformer to describe medical images with spatial features.

4.4.2.3 Benchmark Results and Analysis

FFA-IR benchmark model In Table 4.3, we report values of the automatic metrics and human evaluation to compare various models. Firstly, our FFA-IR benchmark model is F-R+T which achieves almost all the highest numbers of automatic metrics. It outperforms Gounded by 0.022, 0.014 in Bleu-4 and Cider, respectively.

Although Gounded achieves stronger performance on the Meteor and Rouge metrics, we should note that both F-R+T and Gounded generate medical reports based on object features. Secondly, we find that in FFA-IR, the performance of models exploring object

Table 4.6: Comparison of different language reports from the same models, where Zh, Zh-T and En represent generating reports by Chinese words, Chinese tokens, and English.

	Bleu-4			Cider			Hit
	Zh	Zh-T	En	Zh	Zh-T	En	
CoAtt[39]	0.223	0.111	0.113	0.577	0.254	0.250	En
R2Gen[14]	0.231	0.132	0.131	0.623	0.367	0.367	En
Gounded[120]	0.303	0.218	0.220	0.854	0.576	0.569	Zh-T
I3D[11]+T	0.297	0.229	**0.231**	0.849	0.561	0.559	En
F-R[84]+T	**0.365**	**0.240**	**0.231**	**0.882**	**0.590**	**0.590**	Zh-T

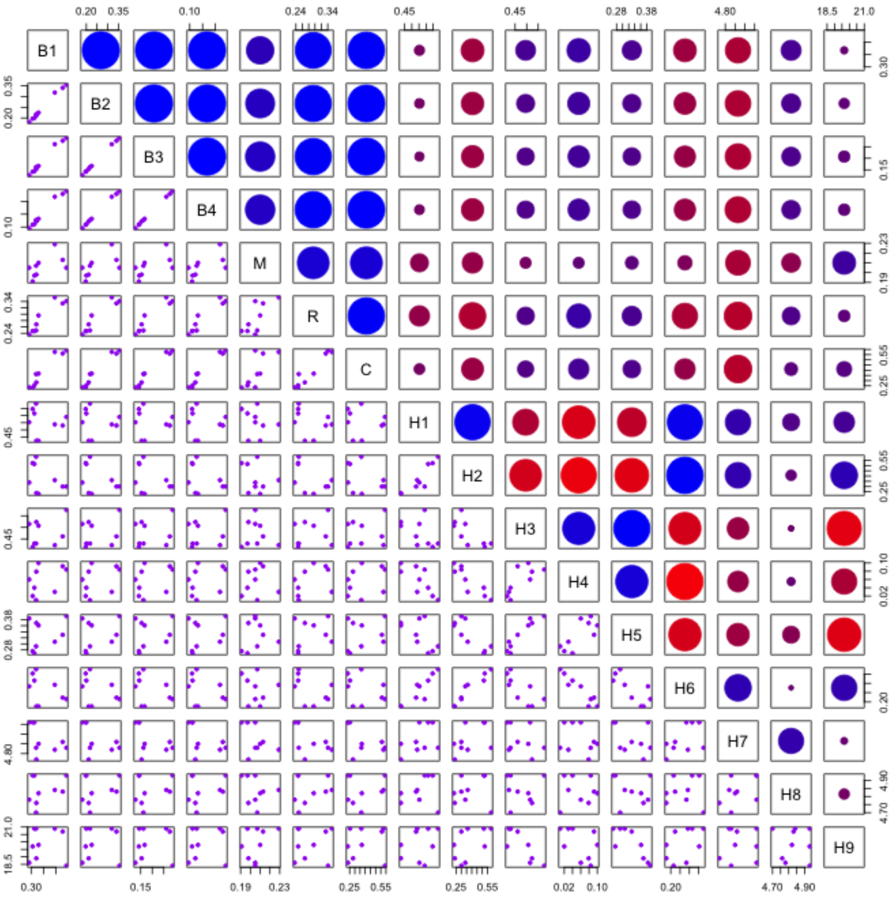


Figure 4.5: The Pearson correlations between each pair of metrics, where the blue and red refer to positive and negative correlation, respectively.

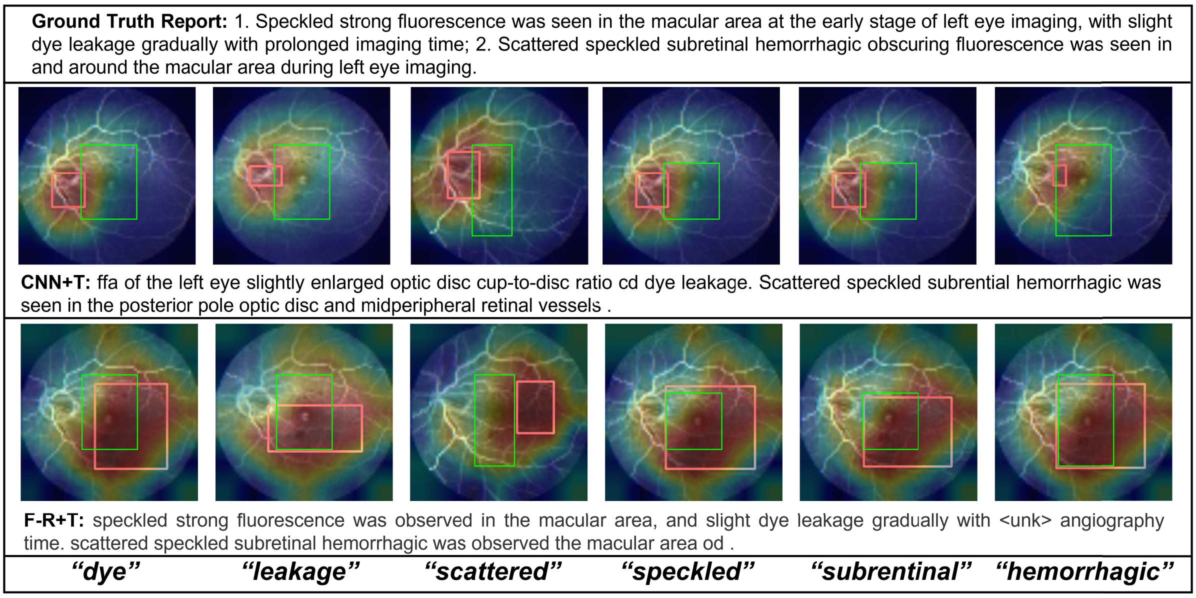
features is significantly higher than other models. These results demonstrate that lesion features are essential in MRG models, but they can be easily inundated by global features without supervised signals. Thirdly, in FFA-IR, Transformer is more efficient in generating long sequences than LSTM[27]. Fourthly, although CNN+LSTM models[39, 72, 98] perform poorly under the automatic metrics, ophthalmologists find that these models generate the most fluent and intelligible reports.

As mentioned, serious data bias exists in medical reports. CNN+LSTM models are prone to underfitting, generating repetitive and non-essential sentences. From the other human evaluation questions, we find that CNN+LSTM models have difficulty with accurate and detailed lesion information. Fifthly, based on H4 results, MRG models struggle to describe the correct category of lesions. Finally, to our knowledge, we are the first to verify the value of medical report generation for clinical diagnosis. Based on H9 results, the automatically generated reports can significantly save ophthalmologists time in the image interpretation and diagnosis procedure. Notably, the average time required for our ophthalmologists to diagnose cases in FFA-IR is 38.2 seconds.

Correlations between automatic metrics and human evaluation Based on Figure 4.5, NLG metrics are correlated with each human evaluation criteria at various degrees. For instance, the Bleu values and H1-6 criteria are highly correlated with themselves and Meteor. Rouge is also negatively correlated with Bleu to a large degree. In addition, all B measures are correlated with H2-H8 with an absolute correlation of between $0.2-0.5$. Meteor, Rouge, and Cider are also correlated with human evaluation criteria to a certain degree, especially with H6-H8, where the absolute correlation value range from 0.12 to 0.55. Another critical measure is the H7, as it is correlated with all other variables with an absolute correlation measure of bigger than 0.3. In sum, the existing NLG metrics are not the most reliable and appropriate evaluation tools in medical fields.

Explanation accuracy In Table 4.4, we calculate the IOU between the lesion-image attention mapping regions and annotated lesion regions to evaluate the accuracy of the explanation. We can find that F-R+T significantly outnumbers CNN+T, R2Gen, and CoAtt by 0.127, 0.109, and 0.149, respectively, proving that the FFA-IR benchmark model also has excellent explanation accuracy. We also visualize the explanation accuracy evaluation process in Figure 4.6. First, F-R+T can generate more fine-grained, coherent-semantic, and accurate reports than CNN+T. On the other hand, we can find that lesion-image attention mapping rectangles are closer to the ground truth regions.

Disease classification results Since the medical reports are used for facilitating



Ground Truth Report: 1. Speckled strong fluorescence was seen in the macular area at the early stage of left eye imaging, with slight dye leakage gradually with prolonged imaging time; 2. Scattered speckled subretinal hemorrhagic obscuring fluorescence was seen in and around the macular area during left eye imaging.

CNN+T: ffa of the left eye slightly enlarged optic disc cup-to-disc ratio cd dye leakage. Scattered speckled subrential hemorrhagic was seen in the posterior pole optic disc and midperipheral retinal vessels .

F-R+T: speckled strong fluorescence was observed in the macular area, and slight dye leakage gradually with <unk> angiography time. scattered speckled subretinal hemorrhagic was observed the macular area od .

Figure 4.6: The visualization of lesion-image attention mapping regions and ground truth among samples from CNN-T and F-R+T, respectively, where the green boxes are the annotated region for each lesion word, and the red boxes are the lesion-image attention mapping regions.

the disease diagnosis procedure by ophthalmologists, we also conduct experiments to investigate whether the generated reports can be used for disease classification. Based on the results presented in Table 4.5, for each model, the results from using visual features alone are significantly higher than using the generated reports. Using ground truth reports, in contrast, can achieve comparable classification results. The results suggest that the generated reports are not yet strong enough to support disease classification.

Does language affect the model? In Table 4.6, we compare the quality of different language reports predicted from the same model. We find that different languages do not affect the model performance. However, the tokenization strategy does. Chinese sentences can be tokenized by words or tokens, as Chinese sometimes requires several words to describe a concept. The vocabulary sizes of Chinese words, Chinese tokens, and English words in FFAIR are 918, 2581, and 3241, respectively. Therefore, two reasons lead to that generating reports by Chinese words achieves higher automatic metric values. One reason is that generating reports by Chinese words has more matching words once the model recognizes a terminology; Another reason is that the word embedding space of Chines words is smaller than the other two's, decreasing the task difficulty. However, the human evaluation shows that these reports are dispreferred.

Table 4.7: The results of NLG metrics of our proposed CGT and other state-of-the-art methods on the FFA-IR dataset. Bold numbers denote the best performance in their columns.

Methods	Year	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
CoAtt[39]	2018	0.313	0.200	0.144	0.111	0.197	0.247	0.254
Show-Tell[98]	2015	0.306	0.197	0.142	0.109	0.191	0.247	0.232
Top-Down[3]	2018	0.320	0.217	0.162	0.127	0.207	0.289	0.363
Gounded[120]	2020	0.396	0.319	0.261	0.218	**0.229**	**0.353**	0.576
AdaAtt[72]	2017	0.292	0.181	0.127	0.095	0.205	0.236	0.234
R2Gen[14]	2020	0.330	0.225	0.167	0.132	0.210	0.296	0.367
I3D+T[55]	2021	0.428	0.341	0.276	0.229	0.213	0.334	0.561
Faster+T[55]	2021	0.443	0.355	0.288	0.240	0.205	0.341	0.590
CGT	Ours	**0.456**	**0.363**	**0.295**	**0.243**	0.227	0.345	**0.599**

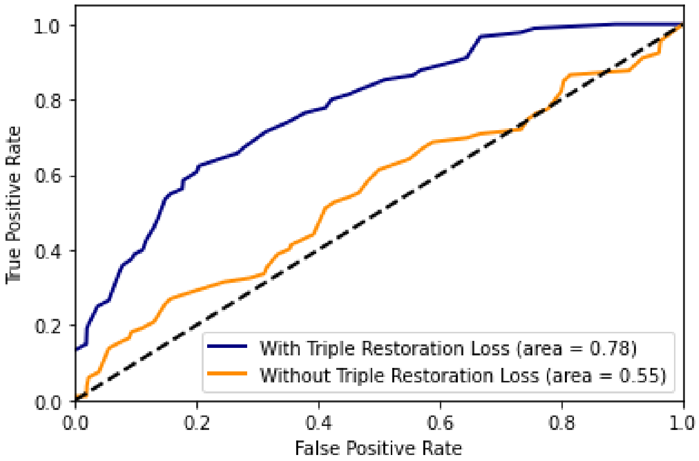


Figure 4.7: Micro-average of receiver operating characteristic curve for sub-graph restoration.

4.4.3 CGT Analysis

4.4.3.1 Main Results

Report generation In Table 4.7, we compare our CGT with a wide range of existing models. I3D+T [55] and Faster+T [55] are the two benchmark models achieving the state-of-the-art performance on FFA-IR dataset. R2Gen [14] and CoAtt [39] are the state-of-the-art radiology report generation models. The remaining presented works are from image captioning approaches. As shown in Table 4.7, our CGT outperforms the state-of-the-art method across all NLG metrics. The improved performance of CGT demonstrates the validity of our practice in incorporating prior medical into ophthalmic report generation.

Table 4.8: Quantitative analysis and human study of proposed method, where CVT, VM and TRL are the short for compressed visual token, visible matrix and triple restoration loss, respectively.

Settings	I3D	Triples	CVT	VM	TRL	CIDEr	BLEU-4	ROUGE	METEOR	Hit(%)
Baseline	✓					0.561	0.229	0.334	0.213	21.6
(a)		✓				0.223	0.087	0.218	0.200	-
(b)		Random				0.223	0.085	0.220	0.204	-
(c)		✓			✓	0.561	0.226	0.287	0.209	-
(d)		✓		✓	✓	0.569	0.231	0.309	**0.228**	-
(e)		✓	✓		✓	0.586	0.240	0.332	0.225	-
(f)	✓	✓			✓	0.573	0.242	0.324	0.226	-
CGT		✓	✓	✓	✓	**0.599**	**0.243**	**0.345**	0.227	**44.7**

Sub-graph restoration In Figure 4.7, we show the micro-average of ROC for sub-graph restoration and present the AUC scores when the proposed model is trained with triple loss restoration loss or not. With the triple restoration loss, the AUC score increased from 0.55 to 0.78 significantly. This improvement demonstrates the effectiveness of triple restoration loss and the accuracy of our restored sub-graph. Without the triple restoration loss, the restored sub-graph is similar to a sequence of random triples. It also verifies the importance of our clinical graph extraction scheme.

4.4.3.2 Quantitative Analysis

In Table 4.8, we present the results of quantitative analysis to investigate the contribution of each component in our CGT. The baseline model is a combination of I3D and Transformer proposed by [55].

Effect of clinical graph In this section, we evaluate the effectiveness of the proposed clinical graph, including triples and triples restoration loss.

Comparing the results in baseline and (a) in Table 4.8, we can find that without the triple restoration loss, the automatically restored sub-graph fails to drive the model to generate an accurate report. In (b), we randomly restore a sub-graph instead of based on the input visual features. Along with the AUC scores in Figure 4.7, these demonstrate that only the relevant and accurate prior knowledge can improve the effectiveness of diagnostic models. Encouragingly, Table 4.8 Baseline and (c) illustrates that the results of utilizing the clinical graph only are competitive to the baseline. These results verify that the triples restoration loss can supervise the sub-graph restoration process and guarantee the accuracy of the incorporated prior knowledge.

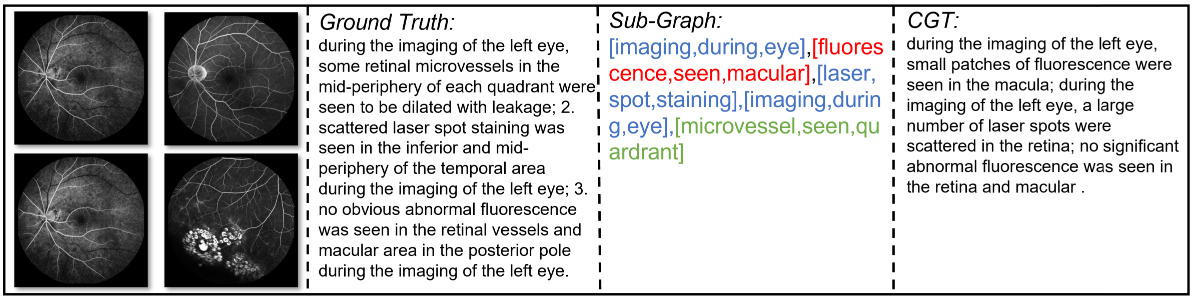
Effect of visible matrix Visible matrix is another essential component in our CGT.

This concept is widely used in knowledge-enhanced pretraining works [24, 70] with various formulations. In this section, the visible matrix is adopted during the cross-modal encoding process for two purposes. On the one hand, we hope it can limit the impact of unrelated triples; On the other hand, we want the message can pass between the visual features and each triple.

The results between (c) and (d), (e) and CGT in Table 4.8 demonstrate the effectiveness of the visible matrix. We can see that the performances increase substantially when integrating visible matrix with (c) and (e), e.g., 0.561 → 0.569 and 0.586 → 0.599 in CIDEr score. Firstly, by comparing the results of (c) and (d), the visible matrix limits the impact from unrelated triples and greatly enhances the information interaction between related triples. Therefore, we speculate that the entity and relation representations can be well trained and improve the quality of predicted reports. When working in CGT, the visible matrix additionally facilitates the message passing between the visual features and each triple. There is inevitable noise among the knowledge graph since the relation is not a 'hard' label. Although triple representations can be well learned, the triple may not be relevant to the input case. Therefore, the visual features play a role in de-noise adaptively. Furthermore, the visible matrix makes sure that the cross-modal signals can interact with each other.

Effect of compressed visual token The effectiveness of the compressed visual token is verified when comparing the results of (c), (e), and (f) in Table 4.8. As discussed, there are always noises existing in a knowledge graph. Therefore, one of the purposes for proposing a compressed visual token is to keep the accurate signals from original meanings when the sub-graph is inaccurate. When integrating the compressed visual token, the quality of predicted reports improves significantly comparing (c) and (e) and outperforming the baseline method. It demonstrates the importance of visual signals in the T. We also conducted an experiment to compare the performances of injecting prior knowledge into the compressed visual token and temporal features ((e) and (f)). We can find that the performances decrease slightly when using all the temporal features, e.g., 0.586 → 0.573 in the CIDEr score. We speculate the reason is that too many visual tokens will impair the effectiveness of prior knowledge. Therefore, using the compressed visual token can make the prior knowledge dominant. Notably, the visible matrix is modified when using all temporal features.

Human study In this section, we invited three senior ophthalmologists to evaluate the quality of predicted reports by the baseline model and our CGT. As shown in Table 4.8, ophthalmologists regarded that 44.7% of predicted reports by CGT can describe the given



Ground Truth:	*Sub-Graph:*	*CGT:*
during the imaging of the left eye, some retinal microvessels in the mid-periphery of each quadrant were seen to be dilated with leakage; 2. scattered laser spot staining was seen in the inferior and mid-periphery of the temporal area during the imaging of the left eye; 3. no obvious abnormal fluorescence was seen in the retinal vessels and macular area in the posterior pole during the imaging of the left eye.	[imaging,during,eye],[fluorescence,seen,macular],[laser,spot,staining],[imaging,during,eye],[microvessel,seen,quardrant]	during the imaging of the left eye, small patches of fluorescence were seen in the macula; during the imaging of the left eye, a large number of laser spots were scattered in the retina; no significant abnormal fluorescence was seen in the retina and macular .

Figure 4.8: Illustrations of reports from the ground truth and CGT, and the restored sub-graph. The blue, red, and greed triples represent the true positive, false positive, and false negative.

FFA images more accurately. The human study results demonstrate that our CGT outperforms the baseline model in both NLG metrics and clinical practice. Ophthalmologists also mentioned that there were 33.7% of predicted reports by both methods that failed to describe any key finding.

4.4.3.3 Qualitative Analysis

In this section, we conduct qualitative analysis for better understanding our approach via an intuitive example. Given a set of input FFA images, our CGT first restores a sub-graph which is further incorporated with visual features to generate a report.

As shown in Figure 4.8, one restored sub-graph consists of four triples, and each triple describes a relation between the subjective and objective entity, e.g., *[fluorescence,seen,macular]* represents that based on the prior clinical knowledge, *'fluorescence'* can be seen in the *'macular*. The number of triples is depended on the length of the input FFA images. Among the restored triples, *[fluorescence,seen,macular]* is the false positive triple which leads to the incorrect sentence *during the imaging of the left eye, small patches of fluorescence were seen in the macular.* This phenomenon shows that our CGT is capable of extending triples to a relevant sentence. Notably, due to the serve textual bias among the training corpus, the sub-graph restoration also suffers since the clinical graph is constructed from the training corpus. *[fluorescence,seen,macular]* is one of the bias triples and exists in 92% training samples. Accurately restored the triple *[laser,spot,staining]* verifies the effectiveness of our CGT to detect abnormalities among the input images and translate them into sentences. It also demonstrates that our CGT is highly capable in sub-graph restoration owing to the triple restoration loss. The last predicted sentence is not relevant to any triple in the restored sub-graph. However, this information can be provided by the compressed visual token.

4.4.4 Limitations and Discussion

4.4.4.1 FFA-IR Benchmark

Our FFA-IR still has the following limitations: First, all these data are only collected from a single medical center. Second, as the original reports are collected from clinical practice, various writing patterns belonging to different report authors can be observed in FFA-IR, affecting the automatic metrics. Third, there are still several rare lesions that are not captured in FFA-IR. Fourth, FFA-IR suffers data bias due to the naturally unbalanced distribution of pathological statistics. Prior errors may also exist due to the unbalanced distributions across attributes, such as gender and age. Fifth, training models in FFA-IR require considerable GPU memories. The models have to read 87 images for each case on average.

4.4.4.2 Cross-modal Clinical Graph Transformer Limitation

Our clinical graph is constructed in an automatic manner from a training corpus; therefore, we cannot guarantee the complete accuracy of our graph. We are inviting more experienced ophthalmologists to verify this graph. In addition, our method is not sufficiently general to support other report generation tasks. For each task, we will need to update the information extraction methods and construct a new clinical graph.

4.4.4.3 Negative Societal Impact

As with other automatic diagnostic methods, our algorithm should be utilized carefully in clinical practice since medical decisions may lead to significant consequences, including death. Therefore, while our AI diagnostic method can provide a strong rationale for judgment along with satisfactory performances, it should only be used as an auxiliary resource.

4.5 Conclusion

In this report, we introduce and discuss our progress since confirmation in details. Our major achievements can be summarized as follows:

On the one hand, we contribute a Fundus fluorescein Angiography Images and Reports (FFA-IR) dataset towards an explainable and reliable benchmark. The FFA-IR dataset has the following characteristics: 1) FFA-IR is a large-scale MRG dataset containing $10,790$ reports along with $1,048,584$ FFA images collected from clinical practice. 2) In FFA-IR, we label $12,166$ lesion regions and collected reports and images to make the diagnosis procedure more explainable. 3) For each case, FFA-IR provides both English and Chinese reports that can facilitate medical multi-modal models. To the best of our knowledge, our work with FFA-IR is the first attempt to quantify the explanation of challenging medical report generation models, propose targeted human evaluation to judge the quality of predicted reports, and investigate the reliability of natural language generation metrics in the medical field. By releasing FFA-IR, we hope this task can be extensively explored in the future to advance research from both vision-and-language and medicine fields significantly and further improve the conventional retinal disease diagnosis procedures.

On the other hand, we present an effective cross-modal clinical graph transformer for ophthalmic report generation. To obtain prior medical knowledge, we propose an information extraction scheme to construct a clinical graph from ophthalmic reports. The prior knowledge inside this graph is further restored to a sub-graph which is injected into the visual features for report generation. The experiments and analyses on our FFA-IR dataset support our arguments and verify the effectiveness of our approach. Along with achieving state-of-the-art performances, the restored sub-graph also improves the explainability of our approach.

CHAPTER 5

MEDICAL KNOWLEDGE ENHANCED CROSS-MODAL PRETRAINING

5.1 Introduction

In addition to medical report generation, there are also several medical vision-language tasks, such as medical visual question answering [26, 43, 63, 75, 114] and medical image text retrieval [18] have received increasing research interests in recent years. Existing works typically design task-specific models for different medical cross-modal tasks, which, however, is inefficient in real-world applications. Furthermore, the reasoning skills and the expert medical knowledge required by different medical cross-modal tasks overlap significantly. Consider the example in Fig. 1.2. Both generating the report and conducting the diagnosis classification regarding the given image require the model capable of distinguishing whether a lung is healthy or abnormal from its visual appearance. Therefore, developing a unified model for handling different medical cross-modal tasks is promising and significant, but it is rarely explored.

In the past few years, vision-language pretraining (VLP) has achieved remarkable success on many vision-language tasks with natural images [13, 51, 57, 71]. Through making only minor additions to the base model architecture and conducting simple finetuning, VLP models show great superiority over conventional task-specific models in various downstream tasks. Most existing VLP models cannot simultaneously address both vision-language understanding (e.g., image-text retrieval) and generation (e.g., image captioning) tasks. To address this issue, some works propose unified VLP frameworks which can handle both understanding-based tasks and generation-based tasks [52, 69, 119] in non-medical domains.

In this section, inspired by the recent progress in VLP, we propose a multi-task benchmark dubbed medical cross-modal understanding and generation with knowledge-enhanced pretraining (MONITOR). As shown in Fig. 1.2, MONITOR covers a set of fundamental medical cross-modal tasks, including *m*edical report generation, *d*iagnosis classification, *i*mage-report retrieval, and *m*edical visual question answering. Through MONITOR, the comprehensive evaluation of unified medical cross-modal models can be fulfilled. To establish a baseline model on MONITOR for encouraging the future research, we develop Med-KEP, which is a unified model pretrained on large-scale medical data and finetuned on both understanding and generation downstream medical tasks. The expert knowledge has been demonstrated to be crucial in enhancing the performance of medical vision-language models as well as improving their explainability [65, 70, 110, 116]. To study the impact of the expert medical knowledge on the unified pretrained model in different downstream tasks, we further introduce three kinds of medical knowledge construction and injection strategies during the pretraining process of Med-KEP: 1) Triplet Concatenation (TC) concatenates multiple knowledge triplets (each is formed as <head entity, relation, tail entity>) into one single sentence to obtain the knowledge encoding, 2) Triplet Insertion (TI) replaces the entities in the text by the knowledge triplets, and 3) Symbolic Knowledge Graph (SKG) represents different relations as edge weights and encodes the knowledge graph through the self-attention mechanism.

We conduct extensive experiments on our MONITOR to evaluate the proposed Med-KEP and knowledge enhancing strategies. Experimental results show the great potential in developing unified medical vision-language models for addressing different downstream tasks. Moreover, the proposed knowledge enhancing approaches consistently improve the performance of Med-KEP and provide better explainablity on different downstream tasks, showing the importance of injecting the expert medical knowledge for assisting the medical cross-modal pretraining.

5.2 MONITOR

MONITOR is a multi-task medical vision-language benchmark aiming to enable the comprehensive evaluation of unified medical cross-modal models. In this section, we first describe the tasks and the datasets contained in the MONITOR benchmark in

Section 5.2.1. Then we introduce the architecture and the pretraining objectives of the proposed Med-KEP in Section 5.2.2. Finally, we present the knowledge-enhanced pretraining strategies in Section 5.2.3.

5.2.1 Tasks & Datasets.

Medical Report Generation. Medical report generation is our main task in this thesis, as mentioned, writing diagnostic reports manually for radiology images is a time-consuming and error-prone process for radiologists. Therefore, MRG becomes an important medical cross-modal generation task, which requires the system to automatically generate a free-text report given a radiology image. In MONITOR, we use the popular chest x-ray dataset IU-Xray [18] for the medical report generation task. The widely-used BLEU [76], METEOR [6], ROUGE-L [60], and CIDEr [97] for MRG are adopted as the evaluation metrics.

Image-Report Retrieval. Similar to most existing works of VLP [52, 53, 115], we develop an image-report retrieval (IRR) task in MONITOR, which is a simple and effective way to evaluate the cross-modal understanding ability of VLP models. Two subtasks are included in IRR: report retrieval (RR), where images and reports are queries and targets, respectively; and image retrieval (IR), where reports are queries and images are targets. The R@K (recall with top k predictions) metric is used for the performance evaluation for both subtasks. We use the IU-Xray [18] and MIMIC-CXR [40] datasets for evaluating the image-report retrieval performance. IU-Xray contains 7,470 chest Xray images with 3,955 radiology reports. Following [14, 38, 50], the training-validation-testing split is 7:1:2. MIMIC-CXR is a larger dataset includes 377,110 chest X-ray images and 227,835 reports.

Medical Visual Question Answering. Medical visual question answering (Med-VQA) models take a medical image and a clinical question about the image as input and output an answer in natural language. Two publicly available Med-VQA datasets, VQA-RAD [48] and SLAKE [63], are adopted in MONITOR. VQA-RAD contains 315 radiology images and 3515 question-answer pairs generated by clinicians. SLAKE includes 642 images and 14,028 questions. According to the answer form, the questions in both VQA-RAD and SLAKE can be categorized into two types. The answers of "closed-ended" questions are "yes/no" or other limited choices, and the answers of "open-ended" questions are free-form texts. Compared with the MRG datasets that only contain the chest X-ray images, the Med-VQA datasets contain images of different organs (e.g., head and abdomen) or modalities (e.g., CT and MRI). Following [63, 114], we use accuracy as the evaluation metric.

Diagnosis Classification. We also include diagnosis classification into MONITOR, which is an important medical task. Following [113], it is formulated as a multi-label image classification task including 14 common radiographic observations: *e*nlarged cardiom, cardiomegaly, lung opacity, lung lesion, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture, support devices, and *n*o finding. We adopt two datasets for diagnosis classification, MIMIC-CXR [40] and ChestX-ray14 [101]. The evaluation metrics are F1, macro-F1, micro-F1, and AUROC.

5.2.2 Med-KEP

We develop a unified medical vision-language model to establish a baseline on our proposed MONITOR. Our constructed baseline, named Med-KEP, follows the architecture and the pretext tasks of the recently proposed BLIP [52], which is a large-scale cross-modal pretrained model capable of addressing with both multi-modal understanding and generation tasks. As shown in Fig. 5.1, Med-KEP contains two unimodal encoders, a cross-modal encoder, and a cross-modal decoder. It is pretrained with three popular pretext tasks, i.e., Image-Text Contrastive (ITC), Image-Text Matching (ITM), and Language Modeling (LM). In the following, we introduce the architecture and the pretext tasks of Med-KEP in detail.

Unimodal Encoder. Med-KEP contains two unimodal encoders, i.e., an image encoder for encoding the image, and a text encoder for encoding the text. The image encoder is a ViT [21] model and the text encoder is the same as BERT [20].

Cross-modal Encoder. The cross-modal encoder contains multiple transformer blocks which are composed of the self-attention layer, the cross-attention layer, and the feed forward network. For the text fed to the cross-modal encoder, a [Encode] token is appended in the beginning. And the output embedding of [Encode] is viewed as the multimodal representation of the image-text pair.

Cross-modal Decoder. Different from the cross modal encoder that use bi-directional masks in the self attention layers, the cross modal decoder adopts causal masks in its self-attention layers. A [Decode] token and an end-of-sequence token are inserted in the beginning and the end of the text sequence to serve as the indicators, respectively.

Pretext Tasks. Following [52], we use three pretext tasks in the pretraining phases of Med-KEP. The Image-Text Contrastive (ITC) task is adopted for improving unimodal encoders by enforcing the alignment of positive image-text pairs in the feature space. The

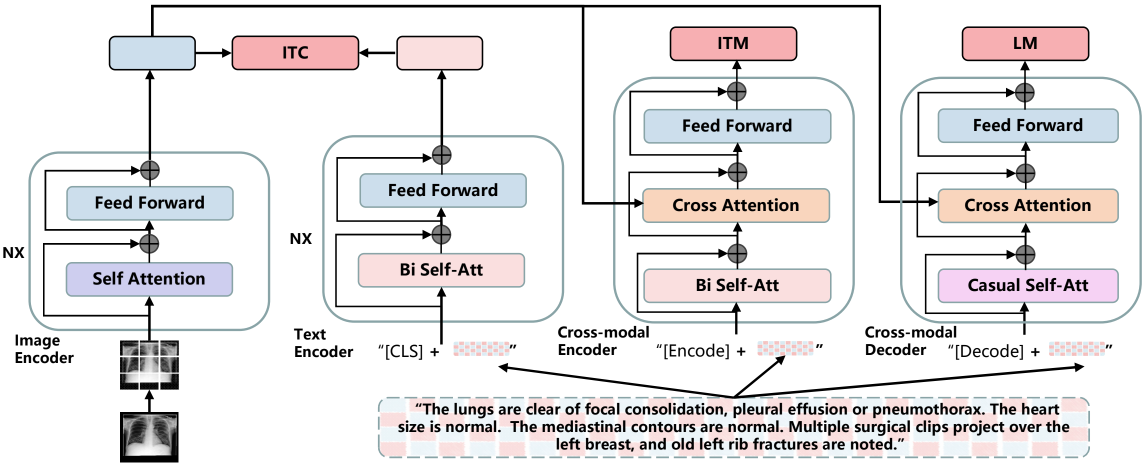

ITC
ITM
LM
Feed Forward
Self Attention
NX
Image Encoder
Bi Self-Att
Text Encoder
"[CLS] + "
Cross Attention
Cross-modal Encoder
"[Encode] + "
Casual Self-Att
Cross-modal Decoder
"[Decode] + "
"The lungs are clear of focal consolidation, pleural effusion or pneumothorax. The heart size is normal. The mediastinal contours are normal. Multiple surgical clips project over the left breast, and old left rib fractures are noted."

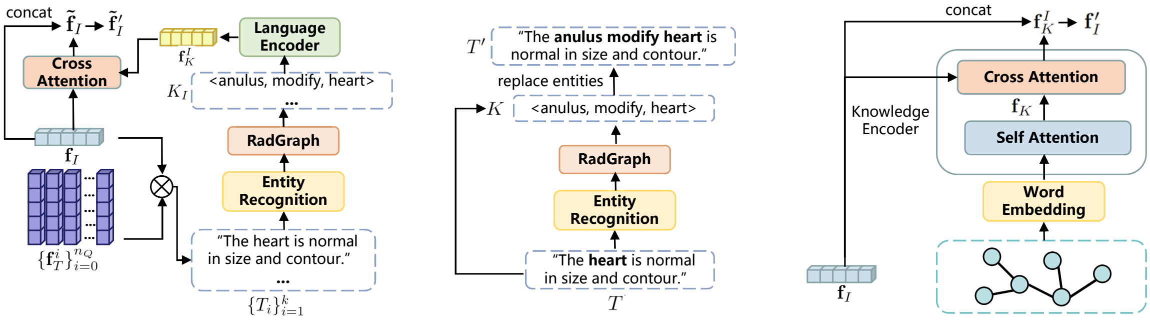
Figure 5.1: The model architecture of Med-KEP. Med-KEP contains two unimodal encoders, a cross-modal encoder, and a cross-modal decoder. Three pretext tasks, i.e., Image-Text Contrastive (ITC), Image-Text Matching (ITM), and Language Modeling (LM) are adopted for pretraining Med-KEP.

Image-Text Matching (ITM) task is designed for the cross-modal encoder by requiring the model to predict whether an image-text pair is positive or negative given the related multi-modal features. The Language Modeling (LM) task is conducted for activating the cross-modal decoder by asking the model to generate textual descriptions given an image through optimizing a cross-entropy loss.

5.2.3 Knowledge-enhanced Pretraining

The expert medical knowledge has been shown to have great potential in improving the performance and enhancing the explainability in many medical cross-modal tasks [24, 49, 56, 70, 110]. Moreover, some recent works have shown that injecting the knowledge in the pretraining phase of VLP models can effectively enable them to learn better cross-modal alignments and therefore benefiting the downstream tasks [15, 112]. To sufficiently analyze how the medical knowledge impact the medical cross-modal pretrained model, we introduce three kinds of knowledge construction and injection strategies during the pretraining of Med-KEP, which are described below.

Triplet Concatenation (TC). Inspired by [110], for each input image I, we first retrieve top k similar texts $\{T_i\}_{i=1}^{k}$ from the text queue Q through calculating the cosine similarity between the image feature $\mathbf{f}_I$ and text features $\{\mathbf{f}_T^i\}_{i=0}^{n_Q}$. n_Q is the size of Q. $\mathbf{f}_I$ and $\{\mathbf{f}_T^i\}_{i=1}^{n_Q}$ are obtained through the image encoder and the text encoder, respectively.



(a) Triplet Concatenation (b) Triplet Insertion (c) Symbolic Knowledge Graph

Figure 5.2: The illustration of Triplet Concatenation (TC), Triplet Insertion (TC), and Symbolic Knowledge Graph (SKG).

Then we extract the named entities $E_i = \{e_i^1, ..., e_i^{n_i}\}$ using a named entity recognizer provided by [82] from each retrieved text T_i. n_i is the number of the named entities. We query specific knowledge using the extracted entities $\{E_i\}_{i=1}^k$ from the public available knowledge graph, e.g., RadGraph [36]. Each queried knowledge is a triplet containing the source entity, target entity, and the relation.

Denote the queried knowledge set regarding to the image I as $K_I = \{k_I^1, ..., k_I^{n_K}\}$. n_K is the size of K_I. We concatenate the knowledge in K_I into one single sentence and feed the sentence to a BERT-like language encoder $E^K(\cdot)$ to obtain the knowledge feature $\mathbf{f}_K^I$:

$$\mathbf{f}_K^I = E^K([k_I^1; ...; k_I^{n_K}]). \tag{5.1}$$

Then we use $\mathbf{f}_K^I$ to update the image feature $\mathbf{f}_I$ through a cross-attention module $E^c(\cdot)$:

$$\tilde{\mathbf{f}}_I = E^c(\mathbf{f}_I, \mathbf{f}_K^I, \mathbf{f}_K^I), \tag{5.2}$$

where $\mathbf{f}_I$ is the query and $\mathbf{f}_K^I$ is the key/value. We concatenate $\tilde{\mathbf{f}}_I$ and $\mathbf{f}_I$ to obtain the final image feature $\tilde{\mathbf{f}}'_I$ and feed $\tilde{\mathbf{f}}'_I$ instead of $\mathbf{f}_I$ into the following modules in Med-KEP.

Triplet Insertion (TI). Inspired by [70], for each image-text pair (I, T), we first extract the named entities $E = \{e^n\}_{n=1}^{n_E}$ using [82] for the text T. n_E is the entity number. Then we query specific knowledge $K = \{\{k_i^n\}_{i=1}^{i_n}\}_{n=1}^{n_E}$ using the extracted entities E similar to that in the TC strategy from the public available knowledge graph. For each entity e^n in E, we randomly choose one knowledge triplet k_i^n from $\{k_i^n\}_{i=1}^{i_n}$ and use it to replace e^n in the original text T. The obtained text T' is used instead of T as the input of Med-KEP in the pretraining phase.

Symbolic Knowledge Graph (SKG). Following [65, 116], we choose 20 keywords (categories) which cover the most common abnormalities or findings in chest X-rays to construct the symbolic knowledge graph. Concretely, we use the category word embedding

$\{\mathbf{w}_i\}_{i=1}^{20}$ to initialize the node feature. Then we resort to a Transformer-like knowledge encoder, which is composed of the self-attention block $E^{\mathrm{SA}}(q,k,v)$ and the cross-attention block $E^{\mathrm{CA}}(q,k,v)$ to obtain the knowledge feature. q, k, v represent query, key, and value, respectively. For each image I, we obtain the image-specific knowledge feature $\mathbf{f}_K^I$ through:

$$\begin{aligned} \mathbf{f}_K &= E^{\mathrm{SA}}(\{\mathbf{w}_i\},\{\mathbf{w}_i\},\{\mathbf{w}_i\}), \\ \mathbf{f}_K^I &= E^{\mathrm{CA}}(\mathbf{f}_K,\mathbf{f}_I,\mathbf{f}_I), \end{aligned} \tag{5.3}$$

where $\mathbf{f}_I$ is the image feature. Then we concatenate $\mathbf{f}_I$ and $\mathbf{f}_K^I$ to obtained the updated image feature $\mathbf{f}_I'$. Like that in the TC strategy, we use $\mathbf{f}_I'$ as the image input of Med-KEP during pretraining.

5.3 Experiments

5.3.1 Pretraining Details

We use MIMIC-CXR [40] dataset for pretraining Med-KEP, which is a large-scale dataset including 377,110 chest X-ray images and 227,835 reports. The dataset contains both the frontal and the lateral view images, and we use the same image encoder for different views following [14]. Considering the large gap between medical texts and general texts, we use SciBERT [8] instead of BERT to serve as the text encoder. And we present the comparison of different text encoders in Table 5.1. We pretrain Med-KEP on 8 NVIDIA V100 GPUs with the batch size 32 and 30 epochs. The learning rate is set as 1e-4 and the optimizer is AdamW. In TC, we retrieve top 3 similar texts from the text queue Q. And the size of Q is set as 65536. The max length of the knowledge sequence in TC is set as 90. We use the tokenizer of SciBERT [8] for tokenizing the knowledge texts in both TC and SKG.

5.3.2 Finetuning Details

Like that during pretraining, we also use SciBERT to serve as the text encoder. Since the additional model architecture, i.e., the knowledge encoder is introduced into Med-KEP during pretraining in TC and SKG, we also inject the knowledge during downstream finetuning when using TC and SKG for enhancing MED-KEP. The finetuning details on MONITOR are described below.

Medical Report Generation. The language modeling loss in the pretraining phase is used for finetuning on the medical report generation task. Following [14], we train the model for 50 epochs with the AdamW optimizer. The batch size is 32 and the learning rate is 1e-5. Considering that the MIMIC-CXR dataset [40] is used for pretraining using the same objective with the medical report generation task, we do not compare Med-KEP with existing methods on MIMIC-CXR for fairness.

Image-Report Retrieval. We use the image-text contrastive loss used in the pretraining phase for finetuning the model on the image-report retrieval task. The model is trained with the batch size 32 and 30 epochs. The learning rate is set as 5e-5 and the optimizer is AdamW.

Medical Visual Question Answering. Following [114], we use the pretrained task type classifier for distinguishing the answers of different types, i.e., open-ended or closed-ended at first. Then we use two cross-modal encoders whose architectures are the same for separately addressing the questions whose answers are open-ended and closed-ended. We train the model with 200 epochs and 300 epochs on VQA-RAD [48] and SLAKE [63], respectively. The learning rate is set as 5e-3 and the optimizer is AdamW. To mitigate the gap among the knowledge regarding to different organs or diseases, we also resort to the knowledge graph released by SLAKE [63] dataset besides the RadGraph [36] when using TC and SKG during finetuning.

Diagnosis Classification. We add 14 linear classifier heads on top of the [CLS] feature of the image feature encoded by the image encoder for diagnosis classification. The model is optimized through the binary cross-entropy loss. The learning rate, the batch size, and the training epochs on both MIMIC-CXR [40] and ChestX-ray14 [101] are set as 1e-5, 32, and 50, respectively.

5.3.3 Downstream Task Results

Medical Report Generation. As shown in Table 5.1, Med-KEP outperforms the previous approaches significantly in the CIDEr metric, which is an important metric used for evaluating the image captioning systems. By comparing Med-KEP w/o knowledge and the proposed TC, TI, and SKG, we can find that the introduction of medical knowledge improve the generation performance in all metrics, showing the effectiveness of the three kinds knowledge enhancing methods.

Image-Report Retrieval. Since there are no available reported results of existing methods on the image-report retrieval task on the proposed MONITOR, we present the results of Med-KEP without/with knowledge on IU-Xray dataset in Table 5.2 to

Table 5.1: Medical report generation performance of different methods on IU-Xray dataset. * represents the main metric.

Methods	CIDEr*	ROUGE_L	METEOR	BLEU_4
CoAtt [39]	0.277	0.369	-	0.154
R2Gen [14]	0.398	0.322	0.165	0.124
Con-Trans [1]	0.257	0.289	0.164	0.111
SentSAT+KG [116]	0.304	0.367	-	0.147
PPKED [65]	0.351	0.376	0.190	0.168
Med-KEP w/o knowledge (BERT)	0.355	0.275	0.152	0.100
Med-KEP w/o knowledge	0.386	0.278	0.164	0.113
Med-KEP w TC	0.516	0.314	0.179	0.139
Med-KEP w TI	0.480	0.297	0.176	0.136
Med-KEP w SKG	0.507	0.304	0.178	0.140

Table 5.2: Image-Report Retrieval performance of different methods on IU-Xray dataset.

Method	TR			IR		
	Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
Med-KEP w/o knowledge	1.17	5.85	6.69	0.71	2.59	4.31
Med-KEP w TC	4.24	12.54	17.63	4.07	10.68	16.44
Med-KEP w TI	1.18	4.92	7.97	0.83	2.63	4.39
Med-KEP w SKG	4.91	10.34	13.56	3.39	11.53	16.27

study the impact of the proposed knowledge enhancing strategies. From Table 5.2, we can find that TC, and SKG significantly improve the Med-KEP, showing that the proposed two kinds knowledge-enhancing methods can effectively encourage the model to learn better cross-modal alignment. In contrast, TI can only consistently outperform the baseline. Due to the reason that, TI knowledge is linguistic information and can only work on enhancing textual modality representations and contribute nothing to the visual vectors. Moreover, we can also observe that there is a large gap between the image-report retrieval performance and the image-text retrieval performance in the general domain [52], showing that learning the cross-modal alignment for medical data is more challenging.

Diagnosis Classification. The results on ChestX-ray14 [101] and MIMIC-CXR [40] are given in Table 5.4 and Table 5.3, respectively. From Table 5.4, we can find that Med-KEP w TC and Med-KEP w SKG are comparable to the state-of-the-art method [44], showing the effectiveness of the knowledge-enhanced pretraining. Moreover, we can find in Table 5.4 that our three kinds of knowledge enhancing methods cover the best classification result regarding most categories. The comparison results among Med-KEP

Table 5.3: Diagnosis classification performance of different methods on MIMIC-CXR dataset. * represents the re-implementation.

Methods	Macro-F1 (%)	Micro-F1 (%)	F1 (%)	AUROC (%)
ResNet101 [25]*	23.1	40.8	37.1	75.2
DenseNet121 [30]*	25.0	40.5	36.8	76.4
Med-KEP w/o knowledge	25.6	42.5	39.4	76.6
Med-KEP w TC	**27.7**	**45.7**	**42.5**	**77.4**
Med-KEP w TI	26.2	44.3	40.9	77.4
Med-KEP w SKG	26.5	45.4	42.2	77.0

Table 5.4: Diagnosis classification performance of different methods on ChestX-ray14 dataset. $^\diamondsuit$ and $^\heartsuit$ represents using ResNet50 [25] and DenseNet-121 [30] as the backbone, respectively.

Methods	Atel	Card	Effu	Infi	Mass	Nodu	Pne1	Pne2	Cons	Edem	Emph	Fibr	P.T.	Hern	Mean
Wang et al.$^\diamondsuit$ [101]	0.700	0.810	0.759	0.661	0.693	0.669	0.658	0.799	0.703	0.805	0.833	0.786	0.684	0.872	0.745
Guan et al.$^\diamondsuit$ [23]	0.779	0.879	0.824	0.694	0.831	0.766	0.726	0.858	**0.758**	0.850	0.909	0.832	0.778	0.906	0.814
Guan et al.$^\heartsuit$ [23]	0.781	0.883	0.831	0.697	0.830	0.764	0.725	0.866	0.758	0.853	0.911	0.826	0.780	0.918	0.816
Kim et al.$^\heartsuit$ [44]	0.780	0.887	0.835	0.710	0.831	**0.804**	0.734	**0.871**	0.747	0.840	**0.941**	0.815	**0.799**	0.909	**0.822**
Med-KEP w/o knowledge	0.722	0.859	0.761	0.659	0.749	0.681	0.631	0.821	0.660	0.775	0.862	0.782	0.741	0.911	0.756
Med-KEP w TC	**0.787**	0.941	**0.875**	**0.720**	0.833	0.754	**0.745**	0.828	0.726	**0.882**	0.893	0.801	0.763	0.939	0.821
Med-KEP w TI	0.755	0.901	0.808	0.701	0.796	0.737	0.683	0.869	0.708	0.835	0.895	**0.834**	0.761	**0.950**	0.802
Med-KEP w SKG	0.780	**0.943**	0.861	0.720	**0.838**	0.752	0.732	0.842	0.726	0.865	0.890	0.806	0.760	0.923	0.817

w/o knowledge, Med-KEP w TC, Med-KEP w TI, and Med-KEP w SKG in both Table 5.4 and Table 5.3 show the obvious advantage of the proposed knowledge enhancing methods.

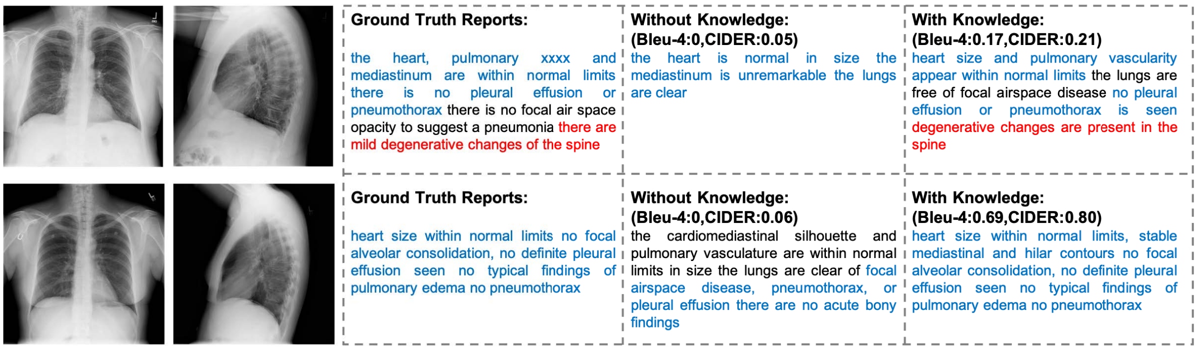
Medical Visual Question Answering. From the results in Table 5.5, we can find that although the gap between the pretraining dataset (MIMIC-CXR) and two downstream VQA datasets is large, Med-KEP w TC is still comparable or even superior than previous methods (71.43 *vs.* 71.60 on VQA-RAD, 80.96 *vs.* 78.60 on SLAKE). This demonstrates the good generalization ability of the proposed Med-KEP with the help of the knowledge enhancement. We can also find that the overall accuracy of TC, TI, and SKG is consistently superior than Med-KEP w/o knowledge on both VQA-RAD and SLAKE. Moreover, Med-KEP w TC effectively improves the open-ended accuracy. These results show that injecting knowledge is also helpful in the Med-VQA task.

5.3.4 Visualization

Medical Report Generation In Figure 5.3, we present two Chest X-ray examinations along with reports from gold annotation, Med-KEP (our baseline model) and TC knowledge enhanced Med-KEP. First, we can observe that enhanced by the medical pretraining, the both Bleu-4 and CIDEr measures increase significantly. In the upper

Table 5.5: Medical visual question answering accuracy (%) of different methods on VQA-RAD and SLAKE datasets.

Method	VQA-RAD			SLAKE		
	Open-ended	Closed-ended	Overall	Open-ended	Closed-ended	Overall
SAN [111]	31.30	69.50	54.30	74.00	79.10	76.00
BAN [45]	37.40	72.10	58.30	74.60	79.10	76.30
MEVF-SAN [75]	49.20	73.90	64.10	75.30	78.40	76.50
MEVF-BAN [75]	49.20	77.20	66.10	77.80	79.80	78.60
QCR+TCR [114]	60.00	**79.30**	**71.60**	-	-	-
Med-KEP w/o knowledge	62.57	73.90	69.41	78.14	83.65	80.30
Med-KEP w TC	**64.02**	76.10	71.43	**79.07**	83.89	**80.96**
Med-KEP w TI	60.89	77.21	70.73	77.98	84.62	80.58
Med-KEP w SKG	62.57	76.94	70.98	77.52	**85.10**	80.49



Ground Truth Reports:

the heart, pulmonary xxxx and mediastinum are within normal limits there is no pleural effusion or pneumothorax there is no focal air space opacity to suggest a pneumonia there are mild degenerative changes of the spine

Without Knowledge:
(Bleu-4:0,CIDER:0.05)

the heart is normal in size the mediastinum is unremarkable the lungs are clear

With Knowledge:
(Bleu-4:0.17,CIDER:0.21)

heart size and pulmonary vascularity appear within normal limits the lungs are free of focal airspace disease no pleural effusion or pneumothorax is seen degenerative changes are present in the spine

Ground Truth Reports:

heart size within normal limits no focal alveolar consolidation, no definite pleural effusion seen no typical findings of pulmonary edema no pneumothorax

Without Knowledge:
(Bleu-4:0,CIDER:0.06)

the cardiomediastinal silhouette and pulmonary vasculature are within normal limits in size the lungs are clear of focal airspace disease, pneumothorax, or pleural effusion there are no acute bony findings

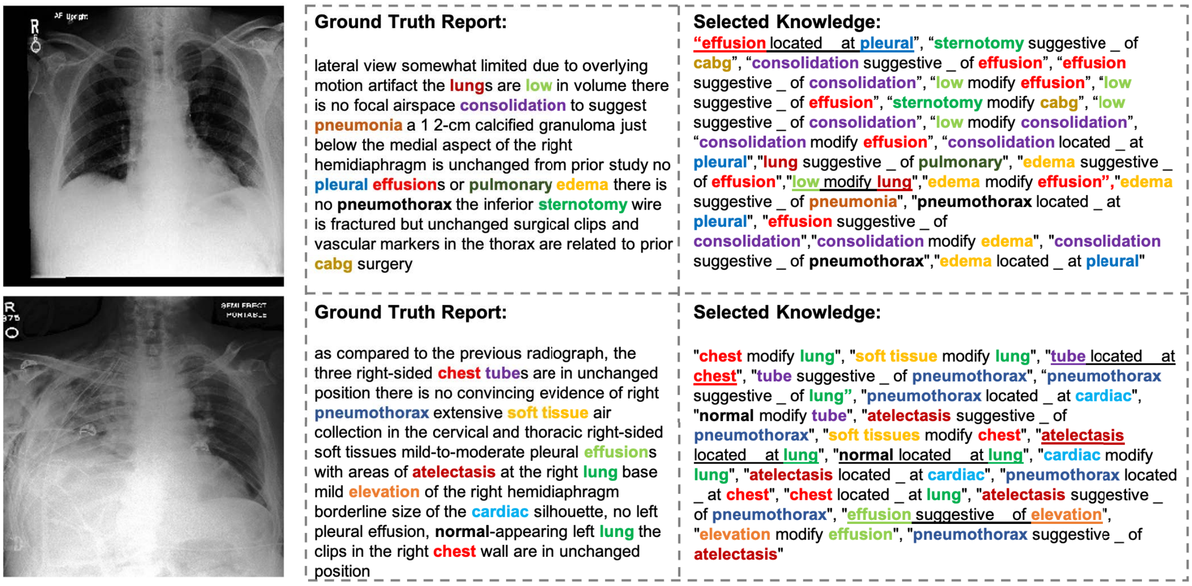
With Knowledge:
(Bleu-4:0.69,CIDER:0.80)

heart size within normal limits, stable mediastinal and hilar contours no focal alveolar consolidation, no definite pleural effusion seen no typical findings of pulmonary edema no pneumothorax

Figure 5.3: Illustrations of reports from ground-truth, baseline without knowledge and baseline with TC knowledge for two X-ray chest examinations. To better distinguish the content in the reports, different colors highlight normal and abnormal medical terms, respectively.

case, the improvements are from abnormal terms description in red texts. The enhanced model is capable to detect it degenerative changes are present in the spine. In the bottom case, report from knowledge enhanced Med-KEP is almost completely similar to the ground truth. The large amounts of normal term descriptions in training corpus may lead to this.

Selected TC Knowledge To validate the effectiveness of our selected knowledge triplets, we present two Chest X-ray examinations with ground truth reports and our selected knowledge triplets from the retrieved TC knowledge in Figure 5.4. To better observe the knowledge triplets' effectiveness, we employ different colors to highlight the same medical entities. We can find that almost all the observed entities are selected in our knowledge triplets, and then their relations between other entities are also presented. Notably, those triplets are selected from the retrieved reports regarding the given images, demonstrating the effectiveness of TC enhanced method.



Ground Truth Report:

lateral view somewhat limited due to overlying motion artifact the lungs are low in volume there is no focal airspace consolidation to suggest pneumonia a 1 2-cm calcified granuloma just below the medial aspect of the right hemidiaphragm is unchanged from prior study no pleural effusions or pulmonary edema there is no pneumothorax the inferior sternotomy wire is fractured but unchanged surgical clips and vascular markers in the thorax are related to prior cabg surgery

Selected Knowledge:

“effusion located _ at pleural”, “sternotomy suggestive _ of cabg”, “consolidation suggestive _ of effusion”, “effusion suggestive _ of consolidation”, “low modify effusion”, “low suggestive _ of effusion”, “sternotomy modify cabg”, “low suggestive _ of consolidation”, “low modify consolidation”, “consolidation modify effusion”, “consolidation located _ at pleural","lung suggestive _ of pulmonary", "edema suggestive _ of effusion","low modify lung","edema modify effusion”,"edema suggestive _ of pneumonia", "pneumothorax located _ at pleural", "effusion suggestive _ of consolidation","consolidation modify edema", "consolidation suggestive _ of pneumothorax","edema located _ at pleural"

Ground Truth Report:

as compared to the previous radiograph, the three right-sided chest tubes are in unchanged position there is no convincing evidence of right pneumothorax extensive soft tissue air collection in the cervical and thoracic right-sided soft tissues mild-to-moderate pleural effusions with areas of atelectasis at the right lung base mild elevation of the right hemidiaphragm borderline size of the cardiac silhouette, no left pleural effusion, normal-appearing left lung the clips in the right chest wall are in unchanged position

Selected Knowledge:

"chest modify lung", "soft tissue modify lung", "tube located _ at chest", "tube suggestive _ of pneumothorax", “pneumothorax suggestive _ of lung”, "pneumothorax located _ at cardiac", "normal modify tube", "atelectasis suggestive _ of pneumothorax", "soft tissues modify chest", "atelectasis located _ at lung", "normal located _ at lung", "cardiac modify lung", "atelectasis located _ at cardiac", "pneumothorax located _ at chest", "chest located _ at lung", "atelectasis suggestive _ of pneumothorax", "effusion suggestive _ of elevation", "elevation modify effusion", "pneumothorax suggestive _ of atelectasis"

Figure 5.4: Illustrations of our selected knowledge in TC. To better distinguish the content in the reports, different colors highlight medical entities.

Selected TC Knowledge To validate the effectiveness of our selected knowledge triplets, we present two Chest X-ray examinations with ground truth reports and our selected knowledge triplets from the retrieved TC knowledge in Figure 5.4. To better observe the knowledge triplets' effectiveness, we employ different colors to highlight the same medical entities. We can find that almost all the observed entities are selected in our knowledge triplets, and then their relations between other entities are also presented. Notably, those triplets are selected from the retrieved reports regarding the given images, demonstrating the effectiveness of TC enhanced method.

Knowledge-Image Attention Mapping To demonstrate the explainability, we first visualize all the attention weights in the knowledge-image attention layer. We can find that the organ-level weights are different from their corresponding diseases, since all the organs can affect each other and directly are attended with the global node. *Heart, spine* and *pleural* acquire more importance than the other organs, like *bone or lung*. It may be the reason that those are the primary inspected organs during the Chest X-ray examinations. In Figure 5.6, we visualize the SKG graph and a specific Chest X-ray attention mappings to highlight the suspicious regions. Notably, the weights from six attention heads are visualized, individually. It is observed that the heart and right-bottom lung regions acquire more importance, which is also described in its diagnostic reports.

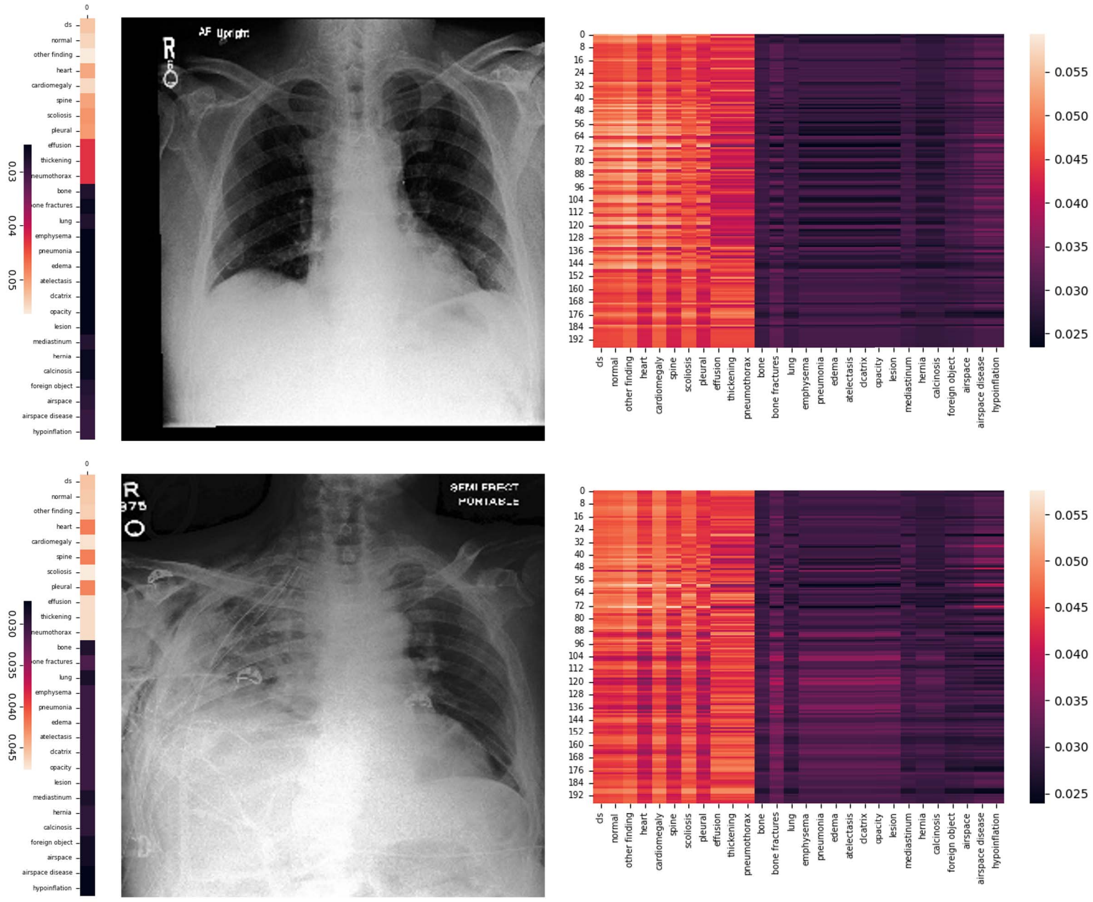


Figure 5.5: Heatmap visualizations among chest X-ray examinations and our proposed SKG graph. In the right, we present the global attention weights among proposed SKG when attended with the global visual information. In the left, we visualize all the attention weights when enhancing medical knowledge.

5.4 Conclusion

This chapter proposes a multi-task benchmark, named MONITOR, for facilitating the comprehensive evaluation of the unified medical vision-language model. Four popular medical downstream tasks are included in MONITOR. A unified pretraining model, Med-KEP, is developed to establish a strong baseline on MONITOR. Three kinds of knowledge enhancing strategies are also introduced into Med-KEP to investigate the impact of the medical knowledge on the unified pretraining model. Experimental results show that the proposed knowledge enhancing approaches consistently improve the performance on different downstream tasks as well as provide better interpretability. We hope that this

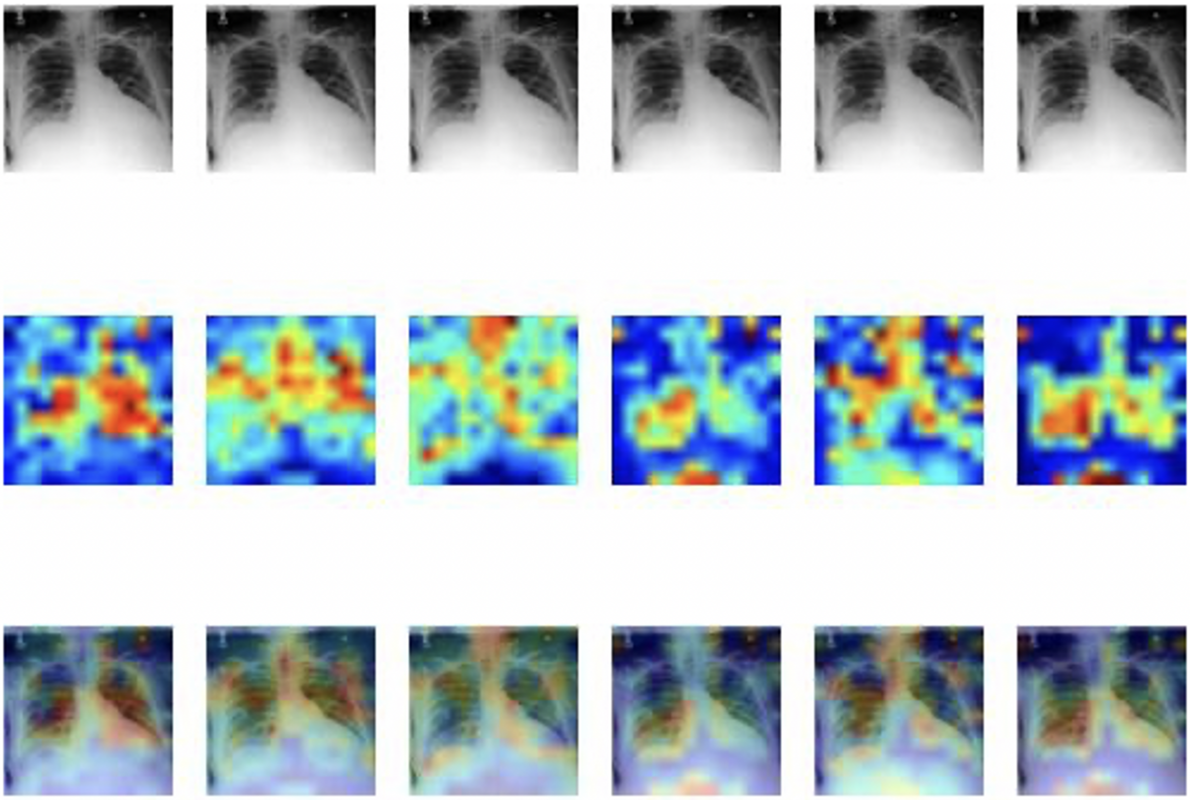


Figure 5.6: Visualization of knowledge and image mappings between a specific chest X-ray examination and our proposed SKG graph.

work motivates the unified medical vision-language research in the future.

Right now, our MONITOR are basically built upon radiology datasets consisting of radiology images, reports and knowledge. There is still a huge gap to integrate other modality data, such as Lung CT and FFA, into our model. The knowledge we utilized in this chapter is all pre-constructed. Therefore, we plan to propose a unified way to construct appropriate knowledge from online data. Then the whole system can be trained in an end-to-end manner.

Chapter 6

Conclusion

In this thesis, we investigated clinical knowledge enhanced deep learning models for medical report generation. We first explore two kinds of clinical knowledge by mimicking radiologists' working patterns to drive an encoder-decoder framework for automatic report generation. Then we construct a clinical knowledge graph by extracting structural clinical information from diagnostic reports, and then propose a cross-modal clinical graph Transformer for ophthalmic report generation. We further introduce an explainable and reliable medical report generation benchmark, and envision them as a testbed for explainable and reliable medical report generation. We also hope that it can broadly accelerate medical imaging research and facilitate interaction between the fields of medical imaging, computer vision, and natural language processing. In the end, we explore three kinds of medical knowledge construction and injection strategies during the pretraining process of an unified medical knowledge enhanced cross-modal model. We demonstrated the importance of enhancing medical report generation systems with clinical knowledge to improve the explainability and generalization ability. In addition to medical report generation task, we also proved the effectiveness of clinical knowledge in other medical imaging analysis, such as medical VQA, medical image-text retrieval and disease diagnosis.

In the future, we attempt to explore an unified medical knowledge and propose an unified knowledge driven medical report generation framework to handle all kinds of radiology examinations. Due to the modality gap, the current medical report generation systems have poor generalization ability. It is encouraging to investigate how to transfer the inherent knowledge among different systems. Inspired by the recent progress of big data and multi-modal pretraining in vision-and-language, we plan to collect vary kinds of data from the existing medical datasets and collaborated institutes to propose a medical multi-modal pretraining dataset. Based on this dataset, the unified or general knowledge will be introduced either. In theory, the concept of models' explainability should be well formulated and also be promising for other medical imaging analysis tasks. Visualize the attention mechanism is the only way to explore model's explainability. However, it is still hard to propose an interpretable medical report generation systems. One of the reasons is lacking of corresponding theory. We plan to formulate the explainability, such as the concept of entropy for informative representation. Then researchers may find the right way to improve models' interpretability and trustworthiness.

BIBLIOGRAPHY

[1] O. ALFARGHALY, R. KHALED, A. ELKORANY, M. HELAL, AND A. FAHMY, *Automated radiology report generation using conditioned transformers*, Informatics in Medicine Unlocked, 24 (2021), p. 100557.

[2] P. ANDERSON, B. FERNANDO, M. JOHNSON, AND S. GOULD, *Spice: Semantic propositional image caption evaluation*, in European conference on computer vision, Springer, 2016, pp. 382–398.

[3] P. ANDERSON, X. HE, C. BUEHLER, D. TENEY, M. JOHNSON, S. GOULD, AND L. ZHANG, *Bottom-up and top-down attention for image captioning and visual question answering*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6077–6086.

[4] W. M. ASSOCIATION ET AL., *World medical association declaration of helsinki. ethical principles for medical research involving human subjects.*, Bulletin of the World Health Organization, 79 (2001), p. 373.

[5] M. BADAR, M. HARIS, AND A. FATIMA, *Application of deep learning for retinal image analysis: A review*, Comput. Sci. Rev., 35 (2020), p. 100203.

[6] S. BANERJEE AND A. LAVIE, *Meteor: An automatic metric for mt evaluation with improved correlation with human judgments*, in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[7] R. BAWDEN, G. M. DI NUNZIO, C. GROZEA, I. J. UNANUE, A. J. YEPES, N. MAH, D. MARTINEZ, A. NÉVÉOL, M. NEVES, M. ORONOZ, O. PEREZ DE VIÑASPRE, M. PICCARDI, R. ROLLER, A. SIU, P. THOMAS, F. VEZZANI, M. V. NAVARRO, D. WIEMANN, AND L. YEGANOVA, *Findings of the WMT 2020 Biomedical Translation Shared Task: Basque, Italian and Russian as New Additional*

Languages, in 5th Conference on Machine Translation, Online, Unknown Region, 2020.

[8] I. BELTAGY, K. LO, AND A. COHAN, *Scibert: A pretrained language model for scientific text*, in EMNLP, 2019.

[9] O. BODENREIDER, *The unified medical language system (umls): integrating biomedical terminology*, Nucleic acids research, 32 (2004), pp. D267–D270.

[10] A. BUSTOS, A. PERTUSA, J.-M. SALINAS, AND M. DE LA IGLESIA-VAYÁ, *Padchest: A large chest x-ray image dataset with multi-label annotated reports*, Medical image analysis, 66 (2020), p. 101797.

[11] J. CARREIRA AND A. ZISSERMAN, *Quo vadis, action recognition? a new model and the kinetics dataset*, in proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[12] S. CHEN, Q. JIN, P. WANG, AND Q. WU, *Say as you wish: Fine-grained control of image caption generation with abstract scene graphs*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9962–9971.

[13] Y.-C. CHEN, L. LI, L. YU, A. E. KHOLY, F. AHMED, Z. GAN, Y. CHENG, AND J. LIU, *Uniter: Universal image-text representation learning*, in Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 104–120.

[14] Z. CHEN, Y. SONG, T. CHANG, AND X. WAN, *Generating radiology reports via memory-driven transformer*, in EMNLP, B. Webber, T. Cohn, Y. He, and Y. Liu, eds., 2020.

[15] Y. CUI, Z. YU, C. WANG, Z. ZHAO, J. ZHANG, M. WANG, AND J. YU, *Rosita: Enhancing vision-and-language semantic alignments via cross- and intra-modal knowledge integration*, Proceedings of the 29th ACM International Conference on Multimedia, (2021).

[16] E. DECENCIÈRE, X. ZHANG, G. CAZUGUEL, B. LAY, B. COCHENER, C. TRONE, P. GAIN, R. ORDONEZ, P. MASSIN, A. ERGINAY, ET AL., *Feedback on a publicly distributed image database: the messidor database*, Image Analysis & Stereology, 33 (2014), pp. 231–234.

[17] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, *Preparing a collection of radiology examinations for distribution and retrieval*, Journal of the American Medical Informatics Association, 23 (2016), pp. 304–310.

[18] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. M. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, *Preparing a collection of radiology examinations for distribution and retrieval*, Journal of the American Medical Informatics Association : JAMIA, 23 2 (2016), pp. 304–10.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805, (2018).

[20] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, ArXiv, abs/1810.04805 (2019).

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, *An image is worth 16x16 words: Transformers for image recognition at scale*, CoRR, abs/2010.11929 (2020).

[22] W. Gale, L. Oakden-Rayner, G. Carneiro, A. P. Bradley, and L. J. Palmer, *Producing radiologist-quality reports for interpretable artificial intelligence*, arXiv preprint arXiv:1806.00340, (2018).

[23] Q. Guan and Y. Huang, *Multi-label chest x-ray image classification via category-wise residual attention learning*, Pattern Recognition Letters, 130 (2020), pp. 259–266.

[24] B. He, D. Zhou, J. Xiao, Q. Liu, N. J. Yuan, T. Xu, et al., *Integrating graph contextualized knowledge into pre-trained language models*, arXiv preprint arXiv:1912.00147, (2019).

[25] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[26] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, *Pathvqa: 30000+ questions for medical visual question answering*, arXiv preprint arXiv:2003.10286, (2020).

[27] S. Hochreiter and J. Schmidhuber, *Long short-term memory*, Neural computation, 9 (1997), pp. 1735–1780.

[28] A. Hoover and M. Goldbaum, *Locating the optic nerve in a retinal image using the fuzzy convergence of the blood vessels*, IEEE transactions on medical imaging, 22 (2003), pp. 951–958.

[29] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, *A comprehensive survey of deep learning for image captioning*, ACM Computing Surveys (CsUR), 51 (2019), pp. 1–36.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, *Densely connected convolutional networks*, in CVPR, 2017.

[31] J.-H. Huang, C.-H. H. Yang, F. Liu, M. Tian, Y.-C. Liu, T.-W. Wu, I. Lin, K. Wang, H. Morikawa, H. Chang, et al., *Deepopht: medical report generation for retinal images via deep models and visual explanation*, in Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 2442–2452.

[32] X. Huang, F. Yan, W. Xu, and M. Li, *Multi-attention and incorporating background information model for chest x-ray image report generation*, IEEE Access, 7 (2019), pp. 154808–154817.

[33] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, *Pixel-bert: Aligning image pixels with text by deep multi-modal transformers*, CoRR, abs/2004.00849 (2020).

[34] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al., *Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison*, in Proceedings of the AAAI conference on artificial intelligence, vol. 33, 2019, pp. 590–597.

[35] A. Jacobi, M. Chung, A. Bernheim, and C. Eber, *Portable chest x-ray in coronavirus disease-19 (covid-19): A pictorial review*, Clinical imaging, 64 (2020), pp. 35–42.

[36] S. Jain, A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng, C. Langlotz, and P. Rajpurkar, *Radgraph: Extracting clinical entities and relations from radiology reports*, in Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.

[37] A. Jimeno Yepes, A. Névéol, M. Neves, K. Verspoor, O. Bojar, A. Boyer, C. Grozea, B. Haddow, M. Kittner, Y. Lichtblau, et al., *Findings of the WMT 2017 biomedical translation shared task*, in Proceedings of the Second Conference on Machine Translation, 2017, pp. 234–247.

[38] B. Jing, Z. Wang, and E. P. Xing, *Show, describe and conclude: On exploiting the structure information of chest x-ray reports*, in ACL, 2019.

[39] B. Jing, P. Xie, and E. P. Xing, *On the automatic generation of medical imaging reports*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers.

[40] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng, *MIMIC-CXR: A large publicly available database of labeled chest radiographs*, CoRR, abs/1901.07042 (2019).

[41] N. Kassner and H. Schütze, *Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly*, arXiv preprint arXiv:1911.03343, (2019).

[42] T. Kauppi, V. Kalesnykiene, J.-K. Kamarainen, L. Lensu, I. Sorri, A. Raninen, R. Voutilainen, H. Uusitalo, H. Kälviäinen, and J. Pietilä, *The diaretdb1 diabetic retinopathy database and evaluation protocol.*, in BMVC, vol. 1, 2007, pp. 1–10.

[43] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar, *Mmbert: Multimodal bert pretraining for improved medical vqa*, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), (2021), pp. 1033–1036.

[44] E. KIM, S. KIM, M. SEO, AND S. YOON, *Xprotonet: diagnosis in chest radiography with global and local explanations*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15719–15728.

[45] J.-H. KIM, J. JUN, AND B.-T. ZHANG, *Bilinear attention networks*, in NeurIPS, 2018.

[46] W. KIM, B. SON, AND I. KIM, *Vilt: Vision-and-language transformer without convolution or region supervision*, in Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, M. Meila and T. Zhang, eds., vol. 139 of Proceedings of Machine Learning Research, PMLR, 2021, pp. 5583–5594.

[47] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, (2014).

[48] J. J. LAU, S. GAYEN, A. B. ABACHA, AND D. DEMNER-FUSHMAN, *A dataset of clinically generated visual questions and answers about radiology images*, Scientific Data, 5 (2018).

[49] C. LI, Z. LI, Z. GE, AND M. LI, *Knowledge driven temporal activity localization*, Journal of Visual Communication and Image Representation, 64 (2019), p. 102628.

[50] C. Y. LI, X. LIANG, Z. HU, AND E. P. XING, *Knowledge-driven encode, retrieve, paraphrase for medical image report generation*, in AAAI, 2019.

[51] G. LI, N. DUAN, Y. FANG, M. GONG, AND D. JIANG, *Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training.*, in Proceedings of the Association for the Advance of Artificial Intelligence (AAAI), vol. 34, 2020, pp. 11336–11344.

[52] J. LI, D. LI, C. XIONG, AND S. C. H. HOI, *Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation*, ArXiv, abs/2201.12086 (2022).

[53] J. LI, R. R. SELVARAJU, A. D. GOTMARE, S. R. JOTY, C. XIONG, AND S. C. H. HOI, *Align before fuse: Vision and language representation learning with momentum distillation*, in NeurIPS, 2021.

[54] L. H. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang, *Visualbert: A simple and performant baseline for vision and language*, CoRR, abs/1908.03557 (2019).

[55] M. Li, W. Cai, R. Liu, Y. Weng, X. Zhao, C. Wang, X. Chen, Z. Liu, C. Pan, M. Li, et al., *Ffa-ir: Towards an explainable and reliable medical report generation benchmark*, in Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.

[56] M. Li, F. Wang, X. Chang, and X. Liang, *Auxiliary signal-guided knowledge encoder-decoder for medical report generation*, arXiv preprint arXiv:2006.03744, (2020).

[57] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, *Oscar: Object-semantics aligned pre-training for vision-language tasks*, in Proceedings of the European Conference on Computer Vision (ECCV), 2020, pp. 121–137.

[58] Y. Li, X. Liang, Z. Hu, and E. P. Xing, *Hybrid retrieval-generation reinforced agent for medical image report generation*, in NeurIPS, 2018.

[59] B. Y. Lin, W. Zhou, M. Shen, P. Zhou, C. Bhagavatula, Y. Choi, and X. Ren, *Commongen: A constrained text generation challenge for generative commonsense reasoning*, arXiv preprint arXiv:1911.03705, (2019).

[60] C.-Y. Lin, *ROUGE: A package for automatic evaluation of summaries*, in Text Summarization Branches Out, Association for Computational Linguistics, July 2004.

[61] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft coco: Common objects in context*, in European conference on computer vision, Springer, 2014, pp. 740–755.

[62] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, *A survey on deep learning in medical image analysis*, Medical image analysis, 42 (2017), pp. 60–88.

[63] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. F. Yang, and X.-M. Wu, *Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question*

answering, 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), (2021), pp. 1650–1654.

[64] F. Liu, S. Ge, and X. Wu, *Competence-based multimodal curriculum learning for medical report generation*, arXiv preprint arXiv:2206.14579, (2022).

[65] F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou, *Exploring and distilling posterior and prior knowledge for radiology report generation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13753–13762.

[66] F. Liu, C. Yin, X. Wu, S. Ge, P. Zhang, and X. Sun, *Contrastive attention for automatic chest x-ray report generation*, arXiv preprint arXiv:2106.06965, (2021).

[67] F. Liu, C. You, X. Wu, S. Ge, X. Sun, et al., *Auto-encoding knowledge graph for unsupervised medical report generation*, Advances in Neural Information Processing Systems, 34 (2021), pp. 16266–16279.

[68] G. Liu, T.-M. H. Hsu, M. McDermott, W. Boag, W.-H. Weng, P. Szolovits, and M. Ghassemi, *Clinically accurate chest x-ray report generation*, in Machine Learning for Healthcare Conference, PMLR, 2019, pp. 249–269.

[69] T. Liu, Z. Wu, W. Xiong, J. Chen, and Y.-G. Jiang, *Unified multimodal pretraining and prompt-based tuning for vision-language understanding and generation*, ArXiv, abs/2112.05587 (2021).

[70] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng, and P. Wang, *K-bert: Enabling language representation with knowledge graph*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 2901–2908.

[71] J. Lu, D. Batra, D. Parikh, and S. Lee, *Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks*, in Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), vol. 32, 2019, pp. 13–23.

[72] J. Lu, C. Xiong, D. Parikh, and R. Socher, *Knowing when to look: Adaptive attention via a visual sentinel for image captioning*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 375–383.

[73] S. V. L. Matthew Honnibal, Ines Montani and A. Boyd, *spacy: Industrial-strength natural language processing in python.*, (2020).

[74] A. Mogadala, M. Kalimuthu, and D. Klakow, *Trends in integration of vision and language research: A survey of tasks, datasets, and methods*, Journal of Artificial Intelligence Research, 71 (2021), pp. 1183–1317.

[75] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. K. Do, E. Tjiputra, and Q. D. Tran, *Overcoming data limitation in medical visual question answering*, in MICCAI, 2019.

[76] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, *Bleu: a method for automatic evaluation of machine translation*, in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 2002.

[77] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., *Pytorch: An imperative style, high-performance deep learning library*, in NeurIPS, 2019.

[78] J. Pavlopoulos, V. Kougia, I. Androutsopoulos, and D. Papamichail, *Diagnostic captioning: a survey*, arXiv preprint arXiv:2101.07299, (2021).

[79] J. Pennington, R. Socher, and C. D. Manning, *Glove: Global vectors for word representation*, in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[80] L. Pizzarello, A. Abiose, T. Ffytche, R. Duerksen, R. Thulasiraj, H. Taylor, H. Faal, G. Rao, I. Kocur, and S. Resnikoff, *Vision 2020: The right to sight: a global initiative to eliminate avoidable blindness*, Archives of ophthalmology, 122 (2004), pp. 615–620.

[81] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, *Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data*, arXiv preprint arXiv:2001.07966, (2020).

[82] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, *Stanza: A python natural language processing toolkit for many human languages*, in ACL, 2020.

[83] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, *Improving language understanding by generative pre-training*, URL https://s3-us-west-2. ama-

zonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf, (2018).

[84] S. Ren, K. He, R. Girshick, and J. Sun, *Faster r-cnn: Towards real-time object detection with region proposal networks*, Advances in neural information processing systems, 28 (2015), pp. 91–99.

[85] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, *Self-critical sequence training for image captioning*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7008–7024.

[86] T. Schlegl, S. M. Waldstein, W.-D. Vogl, U. Schmidt-Erfurth, and G. Langs, *Predicting semantic descriptions from medical images with convolutional neural networks*, in International Conference on Information Processing in Medical Imaging, Springer, 2015, pp. 437–448.

[87] H.-C. Shin, L. Lu, L. Kim, A. Seff, J. Yao, and R. Summers, *Interleaved text/image deep mining on a large-scale radiology image database*, in Deep Learning and Convolutional Neural Networks for Medical Image Computing, Springer, 2017, pp. 305–321.

[88] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, Y. Bengio and Y. LeCun, eds., 2015.

[89] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, *Vl-bert: Pre-training of generic visual-linguistic representations*, arXiv preprint arXiv:1908.08530, (2019).

[90] Y. Sun, S. Wang, Y. Li, S. Feng, X. Chen, H. Zhang, X. Tian, D. Zhu, H. Tian, and H. Wu, *Ernie: Enhanced representation through knowledge integration*, arXiv preprint arXiv:1904.09223, (2019).

[91] K. Suzuki, *Overview of deep learning in medical imaging*, Radiological physics and technology, 10 (2017), pp. 257–273.

[92] A. Talmor, Y. Elazar, Y. Goldberg, and J. Berant, *olmpics-on what language model pre-training captures*, Transactions of the Association for Computational Linguistics, 8 (2020), pp. 743–758.

[93] H. TAN AND M. BANSAL, *Lxmert: Learning cross-modality encoder representations from transformers*, in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019, pp. 5099–5110.

[94] I. TSOCHANTARIDIS, T. HOFMANN, T. JOACHIMS, AND Y. ALTUN, *Support vector machine learning for interdependent and structured output spaces*, in Proceedings of the twenty-first international conference on Machine learning, 2004, p. 104.

[95] S. VARGES, H. BIELER, M. STEDE, L. C. FAULSTICH, K. IRSIG, AND M. ATALLA, *Semscribe: Natural language generation for medical reports.*, in LREC, 2012, pp. 2674–2681.

[96] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A. N. GOMEZ, Ł. KAISER, AND I. POLOSUKHIN, *Attention is all you need*, in NIPS, 2017.

[97] R. VEDANTAM, C. LAWRENCE ZITNICK, AND D. PARIKH, *Cider: Consensus-based image description evaluation*, in CVPR, 2015.

[98] O. VINYALS, A. TOSHEV, S. BENGIO, AND D. ERHAN, *Show and tell: A neural image caption generator*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.

[99] D. VRANDEČIĆ AND M. KRÖTZSCH, *Wikidata: a free collaborative knowledgebase*, Communications of the ACM, 57 (2014), pp. 78–85.

[100] W. WANG, H. BAO, L. DONG, AND F. WEI, *Vlmo: Unified vision-language pre-training with mixture-of-modality-experts*, arXiv preprint arXiv:2111.02358, (2021).

[101] X. WANG, Y. PENG, L. LU, Z. LU, M. BAGHERI, AND R. M. SUMMERS, *Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2097–2106.

[102] X. WANG, Y. PENG, L. LU, Z. LU, AND R. M. SUMMERS, *Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays*, in CVPR, 2018.

[103] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, *Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays*, in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, 2018.

[104] Z. Wang, L. Zhou, L. Wang, and X. Li, *A self-boosting framework for automated radiographic report generation*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2433–2442.

[105] Wikipedia contributors, *Deepl translator — Wikipedia, the free encyclopedia*, 2021.
[Online; accessed 15-August-2021].

[106] J. Wu, Y. Zhang, J. Wang, J. Zhao, D. Ding, N. Chen, L. Wang, X. Chen, C. Jiang, X. Zou, et al., *Attennet: deep attention based retinal disease classification in oct images*, in International Conference on Multimedia Modeling, Springer, 2020, pp. 565–576.

[107] J. T. Wu, N. N. Agu, I. Lourentzou, A. Sharma, J. A. Paguio, J. S. Yao, E. C. Dee, W. G. Mitchell, S. Kashyap, A. Giovannini, et al., *Chest imagenome dataset for clinical reasoning*, (2021).

[108] J. T. Wu, A. Syed, H. Ahmad, A. Pillai, Y. Gur, A. Jadhav, D. Gruhl, L. Kato, M. Moradi, and T. Syeda-Mahmood, *Ai accelerated human-in-the-loop structuring of radiology reports*, in AMIA Annual Symposium Proceedings, vol. 2020, American Medical Informatics Association, 2020, p. 1305.

[109] Y. Xue, T. Xu, L. R. Long, Z. Xue, S. Antani, G. R. Thoma, and X. Huang, *Multimodal recurrent model with attention for automated radiology report generation*, in MICCAI, 2018.

[110] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, *Knowledge matters: Chest radiology report generation with general and specific knowledge*, Medical Image Analysis, (2022), p. 102510.

[111] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, *Stacked attention networks for image question answering*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016), pp. 21–29.

[112] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang, *Ernie-vil: Knowledge enhanced vision-language representations through scene graphs*, in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 3208–3216.

[113] J. Yuan, H. Liao, R. Luo, and J. Luo, *Automatic radiology report generation based on multi-view image fusion and medical concept enrichment*, in International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2019, pp. 721–729.

[114] L.-M. Zhan, B. Liu, L. Fan, J. Chen, and X.-M. Wu, *Medical visual question answering via conditional reasoning*, Proceedings of the 28th ACM International Conference on Multimedia, (2020).

[115] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, *Vinvl: Revisiting visual representations in vision-language models*, 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2021), pp. 5575–5584.

[116] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. L. Yuille, and D. Xu, *When radiology report generation meets knowledge graph*, in The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020.

[117] Z. Zhang, Y. Xie, F. Xing, M. McGough, and L. Yang, *Mdnet: A semantically and visually interpretable medical image diagnosis network*, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6428–6436.

[118] J. Zhao, Y. Zhang, X. He, and P. Xie, *Covid-ct-dataset: a ct scan dataset about covid-19*, arXiv preprint arXiv:2003.13865, (2020).

[119] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, *Unified vision-language pre-training for image captioning and vqa*, ArXiv, abs/1909.11059 (2020).

[120] Y. Zhou, M. Wang, D. Liu, Z. Hu, and H. Zhang, *More grounded image captioning by distilling image-text matching model*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4777–4786.