Multimodal Dual-Path Large-Model Decoding for Radiology Report Generation

Anonymous ACL submission

Abstract

Radiology report generation requires precise alignment between medical imaging findings and clinically coherent textual descriptions. While current methods predominantly rely on either large vision-language models (LVLMs) for visual grounding or large language models (LLMs) for medical narrative generation, they often fail to effectively integrate multimodal clinical evidence with domain-specific knowledge. This paper proposes a novel multimodal dual-path framework that synergistically combines LVLMs and LLMs to address these limitations. Our approach establishes a dynamic fusion between LVLMs' visual-semantic grounding capabilities and LLMs' clinical knowledge reasoning. Specifically, we employ a structured prompting strategy that models the report generation task into three clinically meaningful sections and introduces fine-grained multi-label classification prompts to guide the models, enabling more accurate and comprehensive clinical report generation. Experiments on the public MIMIC-CXR and IU-Xray benchmarks demonstrate our framework's superiority over state-of-the-art methods.

1 Introduction

Radiology report generation (RRG) aims to automatically analyze complex medical images and generate clinically meaningful textual reports. Accurate and efficient report generation not only alleviates the workload of radiologists but also helps reduce diagnostic errors and ensures consistent documentation, ultimately improving patient care and clinical decision-making (Tanno et al., 2025).

Traditional approaches to RRG (Chen et al., 2020; Nooralahzadeh et al., 2021; Wang et al., 2023b) primarily employ an encoder-decoder based framework. While achieving notable progress, the performance of encoder-decoder based approaches heavily relies on the volume and quality of labeled data. However, the RRG datasets are particularly labor-intensive and expensive to obtain. As a result, the scales of existing widely-used datasets for RRG, *e.g.*, MIMIC-CXR (0.22M samples) (Johnson et al., 2019b) is relatively small compared to image captioning datasets, *e.g.*, Conceptual Captions (3.3M samples) (Sharma et al., 2018).

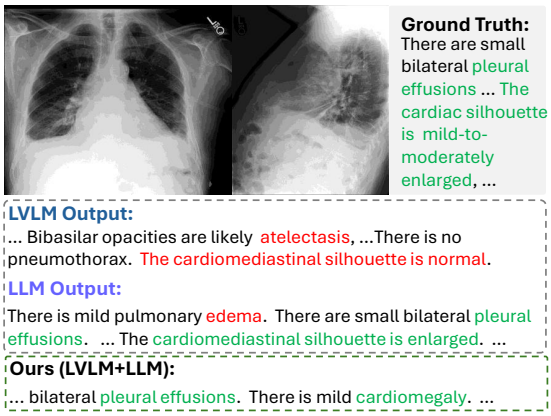

Figure 1: Motivation of our proposed dual-path decoding framework. The text in red indicates errors made by individual models, whereas the text in green denotes correct output. Our framework can correct the errors of the LVLM and the LLM by dual-path decoding.

Recent advances in large-scale models have demonstrated their strong capability in zero-shot/few-shot learning (Brown et al., 2020) which may alleviate the data dependency of RRG task. Existing efforts of applying large scale models to RRG can be categorized into two dominant strategies: First, Large Vision Language Models (LVLMs) (Thawkar et al., 2023; Chen et al.; Wang et al., 2023c) can ground textual descriptions in visual content, enabling more accurate extraction of image-based evidence. However, despite their strong visual grounding abilities, they often struggle to encode prior medical knowledge and generate fine-grained details. On the other hand, Large Language Models (LLMs) have demonstrated remarkable proficiency in understanding and generating natural language, as well as in encoding

extensive prior medical knowledge. These methods (Liu et al., 2025) generate initial reports by Transformer-based models and refine or correct them using LLMs. LLMs can produce contextually rich texts, but typically lack direct access to visual information, limiting their ability to reflect image-based findings in the generated text accurately.

Since LVLMs and LLMs have exhibited complementary strengths and weaknesses for RRG, a natural thought is: **Is it possible and beneficial to ensemble LVLMs and LLMs for radiology report generation?**

Recent research has begun to explore ensemble methods (Jiang et al., 2023; Wang et al., 2023a; Yadav et al., 2023; Yu et al., 2024) that combine multiple LLMs to enhance overall performance. However, most existing ensemble approaches focus on combining multiple *language* models. In contrast, we propose a novel framework that, for the first time, explicitly integrates an LVLM and an LLM during the decoding step of report generation. In our approach, the LVLM focuses on accurately identifying visual information grounded in the image, while the LLM injects additional clinically relevant information to ensure comprehensive and nuanced report generation. As illustrated in Figure 1, our method is able to correct the respective errors of both the LVLM and the LLM after ensemble.

This work proposes a novel multimodal dual-path framework that integrates both LVLMs and LLMs for RRG. The framework harnesses the visual grounding capabilities of LVLMs to extract clinically relevant evidence from medical images, and simultaneously utilizes the language skills of LLMs—prompted with multi-label classification results—to generate fine-grained and clinically accurate reports. By effectively combining the strengths of both types of models, our framework delivers more precise, informative, and clinically useful radiology reports than existing ones.

In summary, our contributions are as follows:

- We propose a novel multimodal dual-path framework that integrates LVLMs and LLMs for RRG, effectively leveraging their complementary strengths to enhance report quality.
- We design a structured prompting strategy that decomposes the RRG task into three clinically meaningful sections: disease categories, overall impression, and imaging findings.
- We introduce fine-grained multi-label classification prompts to guide the LLM, enabling more accurate and comprehensive clinical report generation.
- Extensive experiments on the MIMIC-CXR public benchmark demonstrate that our method performs better on clinical efficacy metrics than state-of-the-art approaches.

2 Related Works

2.1 Radiology Report Generation

Radiology report generation (RRG) aims to automatically report the findings and summarize the impressions from medical images. Initial attempts in RRG predominantly employed encoder-decoder architectures inspired by the successes in natural language processing (NLP) and image captioning domains (Chen et al., 2020; Nooralahzadeh et al., 2021; Yan and Pei, 2022; Long et al., 2025). These models typically leveraged convolutional neural networks (CNNs) or other vision encoders to extract image features, which were then fed into recurrent neural networks (RNNs) or Transformers to generate the textual report. The optimization objective primarily focused on natural language generation (NLG) metrics such as BLEU and ROUGE, which measure lexical and syntactic similarities between generated and reference reports. To address this, subsequent works incorporated fine-grained classification tasks (Wang et al., 2023b; Jin et al., 2024) to enhance the ability to generate clinically relevant and accurate reports. These approaches typically follow a two-stage pipeline: first, extracting image features using a pretrained image encoder (e.g., ResNet (He et al., 2016)), and then concatenating these features with textual representations as input to the report generator. This multimodal fusion strategy allows the model to leverage visual cues and textual context simultaneously. However, this segmentation of encoding and decoding steps may lead to information loss or insufficient semantic alignment between the image and text modalities.

With the advent of large-scale pretrained models, recent research has explored leveraging Large Vision-Language Models (LVLMs) for RRG (Thawkar et al., 2023; Chen et al.; Wang et al., 2023c). These models, often built upon architectures such as multimodal Transformers or vision-language encoders like CLIP, jointly process visual and textual modalities in an end-to-end manner. Extensive pretraining on massive and diverse

datasets—including both natural images, text corpora, and specialized medical data—enables these models to develop strong generalization abilities, rich clinical knowledge, and powerful cross-modal reasoning skills. However, regarding clinical efficiency (i.e., diagnostic accuracy), some LVLM-based methods (Li et al., 2023; Chen et al.) lag behind traditional Transformer-based approaches, highlighting the gap between general language ability and clinically meaningful report generation. Therefore, we believe it is essential to further explore and harness the capabilities of large models, particularly their medical knowledge and reasoning abilities, to advance the quality and clinical relevance of RRG.

In this work, we address these limitations by proposing a multimodal multi-path inference decoding strategy that dynamically integrates the strengths of both LVLMs and Large Language Models (LLMs).

2.2 Large Model Ensemble

Ensembling has proven an effective strategy for addressing the limitations of individual large models and improving overall performance and robustness(Jiang et al., 2023; Yadav et al., 2023; Wei et al., 2025). Existing ensemble methods can be broadly categorized into three types: output ensemble, weight ensemble, and training ensemble. Output ensemble methods (Jiang et al., 2023; Wang et al., 2023a) combine the predictions of multiple models, typically through majority voting, averaging, or more sophisticated aggregation strategies. This approach leverages the diversity among models to improve overall accuracy and reliability. However, output ensembles can also be computationally demanding during inference since multiple models must be run independently, and simple aggregation may not fully exploit shared patterns across models. Weight ensemble techniques (Yadav et al., 2023; Yu et al., 2024), such as model averaging or parameter interpolation, merge the weights of different models to create a single, potentially more powerful model. These methods aim to capture complementary knowledge encoded in the parameters of individual models. A key challenge of weight ensembling is ensuring the compatibility and alignment of weights across models, as differences in initialization, architecture variants, or training regimes can cause naive averaging to fail. Training ensemble involves jointly training multiple models or using techniques like knowledge distillation (Wan et al., 2024) to encourage collaboration and knowledge sharing among models.

While most prior works focus on ensembling multiple LLMs, our approach explores the ensemble of an LVLM and an LLM. Specifically, we leverage the grounding capability of LVLMs to extract visual evidence from medical images, and further enhance clinical guidance by prompting the LLM with multi-label classification results (i.e., positive, negative, uncertain, and not mentioned). Unlike simple output ensembles, our method performs token-level ensembling during the inference process, enabling dynamic collaboration between the LVLM and LLM. This design allows our model to capture more fine-grained and clinically relevant information, effectively combining the strengths of both LVLMs and LLMs for RRG.

3 Method

3.1 Problem Setting

The training dataset consists of fully annotated samples, where each sample is represented as a pair $\{\mathbf{x}, R\}$: $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ denotes a set of chest X-ray (CXR) images from a patient—potentially acquired from multiple views (e.g., posteroanterior and lateral), with $n \leq 3$ images typically—and R is the associated clinical radiology report, composed of words r from a vocabulary $\mathcal{V}$. Each report comprises two sections: (diagnostic) Impression and (imaging) Findings. Notably, most existing report generation methods (Chen et al., 2020; Tanida et al., 2023) only utilize the Findings section. This work aims to develop a framework that, given a set of CXR images $\mathbf{x}$ of a patient, can generate a comprehensive radiology report R covering both Findings and Impression sections.

3.2 Method Overview

The pipeline of our proposed method is illustrated in Figure 2. Our method consists of two stages: 1) model-specific training: We first fine-tune the LVLM (e.g., Qwen2-VL-7B) and LLM (e.g., Qwen2-7B) separately. The LVLM is trained to generate disease categories, Impression, and Findings given the CXR images, whereas the LLM is trained to generate the same three sections following a fine-grained multi-label prompt; 2) multimodal multi-path inference: We then integrate the two models to generate a report. Concretely, our method generates each token in a dual-path manner,

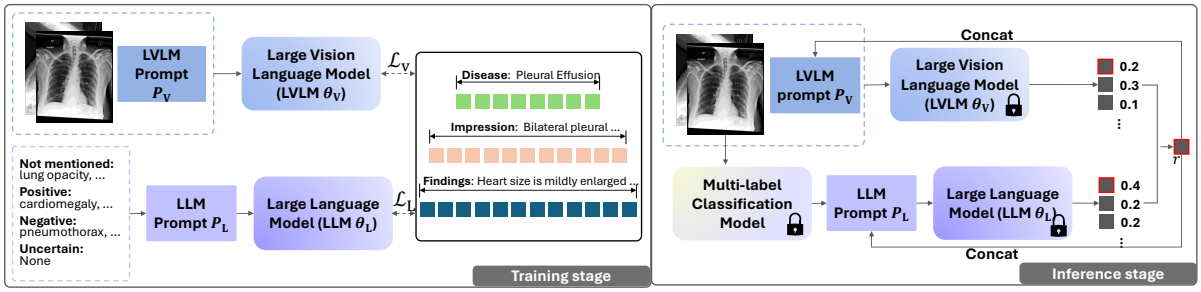

Figure 2: Overview of the proposed method.

Overall Impression: Focal left upper lobe opacity represent atelectasis, however an early focus of infection cannot be excluded.
Disease: Lung Opacity, Pneumonia, Enlarged cardiomediastinum
Findings: Lung volumes are low and the patient is significantly rotated. The endotracheal tube has been removed. A right chest wall port catheter tip terminates at the cavoatrial junction. A focal opacity at the left upper lobe may represent atelectasis, however early infection is also possible. There is no pleural effusion. or pneumothorax. Cardiomediastinal silhouette is mildly enlarged. The imaged upper abdomen is unremarkable.

Figure 3: The proposed three-part training generation targets (i.e., ground truth) incorporating rich information on (diagnostic) Impression, Disease, and (imaging) Findings.

integrating the prediction of both the LVLM and the LLM to produce a comprehensive radiology report collaboratively.

3.3 Disease-Aware Comprehensive Generation Target Construction

Most existing RRG methods (Chen et al., 2020; Shen et al., 2024; Jin et al., 2024) focus solely on generating the Findings section, yet overlook the Impression. We argue that as an indispensable part of a clinical radiology report, the Impression contains important information helpful for RRG. In addition, our preliminary experiments indicate that LVLMs yield limited recall for the generated reports even when trained to produce both the Findings and Impression. To address these issues, we propose to train the model to not only generate the complete Findings and Impression sections, but also a list of positive diseases to boost the recall.

Concretely, our training generation targets are illustrated in Figure 3, structured into three parts: Impression, Disease, and Findings. The Impression and Findings sections are directly copied from the original report written by the radiologist. For the Disease section, we leverage the 14-class multi-label annotations provided by (Johnson et al., 2019b). Then, we enumerate the categories labeled as "positive" to compose the Disease part (e.g., "pleural effusion" and "edema" in Figure 3). The three-part formulation of our generation targets not only aligns with the clinical workflow but also benefits the generation of findings through richer, more structured training signals that explicitly model the diagnostic reasoning process.

3.4 Decoding Path 1: LVLM Training

In accordance with the generation targets, our prompt for the LVLM is designed to instruct a structured output. As shown in Figure 4, the placeholder `<image>` represents the image tokens corresponding to the input radiographs. The following part of the prompt imposes both format and semantic constraints: it requires the model to generate the report in a predefined order: Impression, Disease, and Findings (we study the order's impact empirically in the Experiments section). This prompt enforces a clinically relevant structure, guiding the model to generate comprehensive, logically organized, and interpretable radiology reports.

For training, we employ the instruction tuning (Wei et al., 2021) to teach the model to understand our devised prompt and generate the structured contents. Specifically, given a pretrained LVLM model parameterised by θ_V, we optimize the model using the standard cross-entropy loss commonly adopted in autoregressive language modeling:

$$\mathcal{L}_V = -\sum_{m=1}^{M} \log p_{\theta_V}(r_m \mid \mathbf{x}, P_V, r_{<m}; \theta_V), \quad (1)$$

where r_m denotes the m-th token in the target output sequence, P_V is the prompt devised for the LVLM (Figure 4), and $r_{<m}$ refers to all tokens prior to position m.

<image>\n<image>\n
Please analyze the chest X-ray images and provide a structured report in the EXACT following format:
Overall Impression: Provide a concise 1-2 sentence summary of key observations.
Disease: List ONLY the detected disease categories from: fracture, atelectasis, consolidation, edema, lung lesion, lung opacity, pneumonia, pneumothorax, cardiomegaly, enlarged cardiomediastinum, pleural effusion, pleural other, support devices. If no diseases are detected, output No Finding.
Findings: Write a SINGLE continuous paragraph describing abnormalities. Connect all findings logically to the diseases listed.
Important rules:1. Disease section must ONLY contain detected category words separated by commas. 2. Findings section must be a single paragraph without segmentation.

Figure 4: The proposed prompt for the LVLM.

3.5 Decoding Path 2: Multi-Label Prompted LLM

LVLMs excel at grounding textual descriptions in visual content, enabling more accurate extraction of image-based evidence. However, despite their strong visual grounding abilities, they often struggle to encode prior medical knowledge and generate fine-grained details. To address this limitation, we incorporate an LLM into our framework. By leveraging the LLM's strong language capabilities in integrating fine-grained multi-label classification information, our approach enables the generation of more comprehensive and clinically accurate reports that better reflect radiologists' reporting practices. Following Jin et al. (2024), we assign one of the four possible statuses—"not mentioned", "positive", "negative", or "uncertain"—to each disease category. The multi-label classification results are organized in the format of a dictionary: {"not mentioned": $C_1, ..., C_i$; "positive": $C_{i+1}, ..., C_j$; "negative": $C_{j+1}, ..., C_k$; "uncertain": $C_{k+1}, ..., C_K$}, where C_i represents a specific disease category, and K is the total number of categories.

During training, we utilize the annotations published by PromptMRG (Jin et al., 2024), which provide $K = 18$ multi-label classification results for all training samples. These annotations are used to construct a comprehensive prompt P_{L} for the LLM, as illustrated in Figure 5. The LLM (parameterized by θ_{L}) is trained to minimize the following loss:

$$\mathcal{L}_{\mathrm{L}} = -\sum_{m=1}^{M} \log p_{\theta_{\mathrm{L}}}(r_m \mid \mathbf{x}, P_{\mathrm{L}}, r_{<m}; \theta_{\mathrm{L}}). \quad (2)$$

Note that the training generation targets are the same as the LVLM, as described in Section 3.3. By including this multi-label classification information in the prompt, we provide explicit and structured guidance for the LLM, enabling it to better capture each disease's presence, absence, or uncertainty.

Please generate a chest X-ray report according to the following criteria: \n{**'not mentioned'**: ['cardiomegaly', 'lung lesion', 'edema', 'consolidation', 'pleural other', 'fracture', 'no finding', 'aorta abnormal'], **'positive'**: ['enlarged cardiomediastinum', 'lung opacity', 'support devices', 'bone/spine abnormal', 'hemidiaphragm abnormal', 'lung volume abnormal'], **'negative'**: ['pneumothorax', 'pleural effusion'], **'uncertain'**: ['pneumonia', 'atelectasis']}
The report must be in the EXACT following format:
Overall Impression: Provide a concise 1-2 sentence summary of key observations.
Disease: List ONLY the positive disease categories from: fracture, atelectasis, consolidation, edema, lung lesion, lung opacity, pneumonia, pneumothorax, cardiomegaly, enlarged cardiomediastinum, pleural effusion, pleural other, support devices. If no diseases are detected, output No Finding.
Findings: Write a SINGLE continuous paragraph describing abnormalities. Connect all findings logically to the diseases listed.
Important rules:1. Disease section must ONLY contain positive category words separated by commas.2. Findings section must be a single paragraph without segmentation.3. Do not mention diseases that are not detected, including those in the 'not mentioned' and 'uncertain' categories.4. For negative findings, explicitly state that the negative diseases are not present in the report.

Figure 5: The proposed prompt for the LLM.

For testing, i.e., generating the report for a (set of) new radiograph(s), we apply PromptMRG to the input radiograph to obtain the multi-label classification results, which are then used to compose the prompt P_{L}.

3.6 Multimodal Dual-Path Inference Decoding

For RRG, relying on a single model often fails to simultaneously capture both the precise understanding of visual information and the rich, domain-specific language required for clinical reporting. Specifically, LVLMs excel at extracting fine-grained visual features directly from medical images, enabling intuitive recognition of abnormalities. However, their ability to organize complex clinical narratives and perform sophisticated reasoning is often limited. In contrast, LLMs demonstrate strong capabilities in medical knowledge, clinical reasoning, and structured text generation, producing coherent reports that adhere to medical conventions. Nevertheless, LLMs primarily depend on external prompts for image content and lack direct visual grounding (Zhao et al., 2024b). As a result, single-path decoding approaches relying on either LVLMs or LLMs are subject to the inherent limitations of each model, potentially

leading to omissions, inaccurate descriptions, or a lack of visual evidence in the generated reports. To tackle this problem, during inference, we employ both the LLM and the LVLM to generate the radiology report jointly. Specifically, at each decoding step m, both models independently compute the probability distribution over the vocabulary $\mathcal{V}$ for the next token, conditioned on the input image $\mathbf{x}$, the structured prompt P_{V} or P_{L}, and prior tokens $r_{<m}$. Denoting the two probability distributions by $p_{\theta_{\text{V}}}(\mathcal{V} \mid \mathbf{X}, P_{\text{V}}, r_{<m}; \theta_{\text{V}})$ and $p_{\theta_{\text{L}}}(\mathcal{V} \mid \mathbf{X}, P_{\text{L}}, r_{<m}; \theta_{\text{L}})$, to integrate the predictions of both models, we compute a weighted average of the probability distributions, controlled by a hyperparameter $\alpha \in [0, 1]$:

$$p_{\text{fusion}} = \alpha * p_{\theta_{\text{V}}}(\mathcal{V} \mid \mathbf{X}, P_{\text{V}}, r_{<m}; \theta_{\text{V}}) + (1 - \alpha) * p_{\theta_{\text{L}}}(\mathcal{V} \mid \mathbf{X}, P_{\text{L}}, r_{<m}; \theta_{\text{L}}). \quad (3)$$

The next token r_m^* is then selected by taking the token with the highest probability in the fused distribution: $r_m^* = \arg\max\ p_{\text{fusion}}$. Then, we append it to the prior token sequence, i.e., $r_{<m} \leftarrow [r_{<m}, r_m^*]$, and proceed to the next decoding step. This process is repeated until the end-of-sequence token is generated.

This dual-path decoding approach allows the model to benefit from both the strong language modeling and clinical reasoning capabilities of the LLM, as well as the direct visual grounding provided by the LVLM. By fusing their predictions at each step, we achieve more accurate, comprehensive, and clinically faithful report generation.

4 Experiments

4.1 Datasets and Evaluation Metrics

We conduct extensive experiments on the MIMIC-CXR dataset (Johnson et al., 2019a,b) and **IU-Xray** dataset (Demner-Fushman et al., 2015). MIMIC-CXR is a large, publicly available collection of chest X-rays paired with free-text radiology reports. Following the commonly adopted data split proposed by Chen et al. (2020), we use 270,790 samples for training, 2,130 for validation, and 3,858 for testing. IU-Xray contains 7,470 chest X-ray images and 3,955 reports. Each study strictly pairs posteroanterior (PA) and lateral views with detailed radiology reports. Following (Chen et al., 2020), the dataset is partitioned into 5,226 training samples, 748 validation samples, and 1,496 testing samples.

Four commonly used natural language generation (NLG) metrics are employed to evaluate the quality of generated reports: BLEU (1- and 4-gram) (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and ROUGE (Lin, 2004). Following Nicolson et al. (2023), we evaluate the clinical efficiency (CE) metrics—precision, recall, and F1 score by converting the reports into 14 disease classification labels using CheXbert (Smit et al., 2020) (for fair comparison and alignment with other methods, we only evaluate on these 14 categories that are commonly considered in previous works). Furthermore, we incorporate two clinical semantic integrity (CSI) metrics, RaTEScore (Zhao et al., 2024a) and RadGraph (Jain et al.), which are specifically designed for evaluating report generation and better capture the clinical performance of models by assessing both entities and relations within the generated reports. Unless otherwise specified, we restrict our evaluation to the Findings section, as most previous works only considered the Findings section.

4.2 Comparison with State-of-the-Art (SOTA) Report Generation Methods

Table 1 compares our method with SOTA approaches of three categories: a) Transformer-based RRG methods, including R2Gen (Chen et al., 2020), M2TR (Nooralahzadeh et al., 2021), CliBert (Yan and Pei, 2022), METrans (Wang et al., 2023b), RGRG (Tanida et al., 2023), MAN (Shen et al., 2024), and PromptMRG (Jin et al., 2024); b) LVLMs (without finetuning), such as Qwen2-VL-7B (Wang et al., 2024) and Deepseek-Janus-Pro-7b (Chen et al., 2025); c) medical LVLMs, including LLaVA-Med (Li et al., 2023), XrayGPT (Thawkar et al., 2023), CheXagent (Chen et al.), R2GenGPT (Wang et al., 2023c), and MLRG (Liu et al., 2025).

For the CE Metrics, our model achieves the highest precision (0.591) and F1 Score (0.527), as well as the second-highest recall (0.476), outperforming all other methods. Specifically, compared to the best-performing SOTA Transformer-based method (PromptMRG), our model improves precision by 0.09 and F1 Score by 0.051. Compared with the best-performing medical LVLM method (MLRG), our model demonstrates improved precision by 0.042, recall by 0.008, and F1 score by 0.022. For the CSI Metrics, our model also achieves the highest RaTEScore (0.557) and RadGraph score (0.249), further validating its superiority in gener-

Method	Year	CE Metrics			CSI Metrics		NLG Metrics			
		Precision	Recall	F1 Score	RaTEScore	RadGraph	BLEU-1	BLEU-4	METEOR	ROUGE
R2Gen (Chen et al., 2020)	2020	0.333	0.273	0.276	-	-	0.353	0.103	0.142	0.277
M2TR (Nooralahzadeh et al., 2021)	2021	0.240	0.428	0.308	-	-	0.378	0.107	0.145	0.272
CliBert (Yan and Pei, 2022)	2022	0.397	0.435	0.415	-	-	0.383	0.106	0.144	0.275
METrans (Wang et al., 2023b)	2023	0.364	0.309	0.311	-	-	0.386	0.124	0.152	0.291
RGRG (Tanida et al., 2023)	2023	0.461	0.475	0.447	0.491	-	0.373	0.126	0.168	0.264
MAN (Shen et al., 2024)	2024	0.411	0.398	0.389	-	-	0.396	0.115	0.151	0.274
PromptMRG (Jin et al., 2024)	2024	0.501	**0.509**	0.476	-	-	0.398	0.112	0.157	0.268
Qwen2-VL-7B (Wang et al., 2024)	2024	0.366	0.205	0.213	0.434	0.081	0.137	0.001		0.147
Deepseek-Janus-Pro-7b (Chen et al., 2025)	2025	0.193	0.064	0.096	0.359	0.056	0.053	0.005		0.138
LLaVA-Med (Li et al., 2023)	2023	-	-	0.107	-	-	-	0.110	-	0.151
Xray-GPT (Thawkar et al., 2023)	2023	-	-	0.193	-	-	-	0.054	-	0.220
CheXagent (Chen et al.)	2024	-	-	0.403	0.474	0.148	-	0.073	-	0.259
R2GenGPT (Wang et al., 2023c)	2024	-	-	0.247	-	-	-	0.101	-	0.276
MLRG (Liu et al., 2025)	2025	0.549	0.468	0.505	-	-	**0.411**	**0.158**	**0.176**	**0.320**
Qwen2-VL-7B-FT	-	0.502	0.369	0.404	0.455	0.113	0.230	0.062	0.148	0.293
Qwen2-VL-7B-IDF	-	0.535	0.464	0.497	0.486	0.210	0.246	0.064	0.144	0.288
Ours	-	**0.591**	0.476	**0.527**	**0.557**	**0.249**	0.280	0.070	0.144	0.286

Table 1: Comparison with SOTA methods on the MIMIC-CXR dataset. The best results are in bold, and the second-best results are underlined. The results for Transformer-based methods and medical LVLMs are from Jin et al. (2024) and Pellegrini et al. (2023), respectively.

ating clinically accurate and semantically precise reports.

Although our method does not achieve the highest scores on standard NLG metrics, we argue that clinical efficacy metrics are more critical in the context of medical diagnosis, as they directly reflect model ability to accurately identify and classify clinical conditions—an essential aspect for supporting effective medical decision-making. Moreover, some works have shown that BLEU exhibits weak correlation with human judgment, while F1 demonstrates the strongest (Turian et al., 2003; Callison-Burch et al., 2006). Other studies (Novikova et al., 2017) have indicated that widely used metrics such as BLEU, ROUGE, and METEOR do not consistently align with human evaluations in NLG tasks. Therefore, we report NLG metrics for reference purposes and emphasize more on CE metrics when assessing clinical report generation performance.

Furthermore, Table 1 shows that Qwen2-VL-7B-IDF (fine-tuned with our proposed three-part training targets), outperforms the original Qwen2-VL-7B and Qwen2-VL-7B-FT (fine-tuned with the Findings section of the reports), underscoring the effectiveness of our structured report generation. On top of that, our proposed dual-path multimodal inference—integrating both LVLM and LLM—achieves even better results than both Qwen2-VL-7B-FT and Qwen2-VL-7B-IDF. As illustrated in Figure 6, the LVLM incorrectly predicts two disease categories, atelectasis and lung opacity. However, in the LLM prompt, these categories are marked as "not mentioned", leading to their effective removal in the final report. Additionally, the LVLM misses the category cardiomegaly, which is successfully recovered in the final output. Similarly, the LLM generates an incorrect category, edema, which is also corrected in the final report. These examples demonstrate how the two models can complement each other, collaboratively reducing errors and enhancing the overall clinical accuracy of the generated reports.

Method	Precision	Recall	F1
R2Gen (Chen et al., 2020)	0.141	0.136	0.136
RGRG (Tanida et al., 2023)	0.183	0.187	0.180
PromptMRG (Jin et al., 2024)	0.213	0.229	0.211
Qwen2-VL-7B-FT	0.193	0.196	0.194
Qwen2VL-7B-DF (baseline)	0.202	0.218	0.207
Ours	**0.216**	**0.247**	**0.235**

Table 2: Comparison of clinical efficiency (CE) metrics across different methods on the IU-Xray dataset.

To further verify our model's clinical effectiveness, we assessed Clinical Efficiency (CE) on the IU-Xray dataset against existing SOTA methods in Table 2. Compared to the previous SOTA transformer-based model, PromptMRG, our approach shows an absolute improvement of 0.024 in the overall F1 score. When compared directly to our baseline (Qwen2VL-7B-DF), the enhancements are even more pronounced, with gains of 0.028 in the critical F1 metric. This superior performance further corroborates the clinical utility and effectiveness of our proposed method.

4.3 Ablation Studies

4.3.1 Complementary Strengths of LLM and LVLM

We conducted an ablation study on MIMIC-CXR

LLM Prompt:	Please generate a chest X-ray report according to the following criteria: {**"not mentioned"**: ["enlarged cardiomediastinum", "lung lesion", "lung opacity", "consolidation", "pneumonia", "atelectasis", "pleural other", "fracture", "support devices", "no finding"], **"positive"**: ["cardiomegaly", "edema", "pleural effusion", "aorta abnormal", "bone/spine abnormal", "hemidiaphragm abnormal", "lung volume abnormal"], **"negative"**: ["pneumothorax"], **"uncertain"**: []} ...
LVLM Output:	"Overall Impression": "1. Small bilateral pleural effusions. 2. Bibasilar opacities, likely atelectasis, however, infection cannot be excluded." "Disease":"Atelectasis, Lung Opacity, Pleural Effusion" "Findings": "There are small bilateral pleural effusions. Bibasilar opacities are likely atelectasis, however, infection cannot be excluded. There is no pneumothorax. The cardiomediastinal silhouette is normal. Atherosclerotic calcifications are noted in the aortic arch. There is no free air under the diaphragm."""
LLM Output:	"Overall Impression": "1. Mild pulmonary edema. 2. Small bilateral pleural effusions. 3. Mild cardiomegaly. " "Disease":" Cardiomegaly, Edema, Pleural Effusion" "Findings": "There is mild pulmonary edema. There are small bilateral pleural effusions. There is no pneumothorax. The cardiomediastinal silhouette is enlarged. The aorta is calcified."
Ours (LVLM+ LLM):	"Overall Impression": "1. No pneumothorax. 2. Small bilateral pleural effusions. 3. Mild cardiomegaly." "Disease":"Cardiomegaly, Pleural Effusion" "Findings": "The lungs are well expanded. There is no pneumothorax. There are small bilateral pleural effusions. There is mild cardiomegaly. The aorta is tortuous and calcified. There is no focal consolidation concerning for pneumonia.""
Ground Truth:	"Overall Impression": "1. Small bilateral pleural effusions. 2. Right upper lobe densities, for which followup chest CT could be considered on a non-urgent basis." "Disease":" Cardiomegaly, Pleural Effusion" "Findings": "There are small bilateral pleural effusions with fluid extending into the major and minor fissures bilaterally. There is no focal consolidation. Rounded densities projecting over the peripheral right upper lung zone on the AP view may represent pulmonary nodules. There is mild pulmonary vascular congestion/interstitial edema. The cardiac silhouette is mild-to-moderately enlarged, but stable. The mediastinal and hilar contours are within normal limits. Partial calcification of the aortic knob is noted."

Figure 6: Example of generated radiology reports. Text highlighted with a red background indicates disease categories corrected by our method (previously misclassified by either LVLM or LLM).

	CE Metrics			NLG Metrics			
	Precision	Recall	F1 Score	BLEU-1	BLEU-4	METEOR	ROUGE
LLM-only	0.483	**0.517**	0.485	0.238	0.059	0.141	0.279
LVLM-only	0.535	0.464	0.497	0.246	0.064	0.144	0.288
Ours	**0.591**	0.476	**0.527**	0.280	0.070	0.144	0.286

Table 3: Performance Comparison of Individual and Fused Models on the MIMIC-CXR dataset.

to investigate the individual contributions of the LLM and LVLM in Table 3. In the "LLM-only" setup, we first use the multi-label classifier to process the image and obtain disease classification results. We then input these results, along with the LLM's prompt, to the LLM to generate the final report. The results clearly demonstrate that relying solely on either a Large Language Model or a Large Vision Language Model does not lead to ideal report generation performance. As shown in Table 3, the "LLM-only" model exhibits a respectable recall score (0.517). However, its flat precision and overall F1 score suggest that the report's accuracy is limited without direct visual evidence. In contrast, the "LVLM-only" model achieves a better precision (0.535), proving its advantage in accurately extracting visual cues from the image. Its relatively lower score, however, reflects its limitations in integrating broader clinical knowledge. By dynamically fusing the strengths of both models, our method achieved the best results for both precision (0.591) and F1 score (0.527). This demonstrates that our framework can effectively combine the powerful visual information extraction capabilities of the LVLM with the deep clinical knowledge reasoning abilities of the LLM, resulting in more accurate and comprehensive radiology reports that are significantly superior to those produced by either single model.

5 Conclusion

In this paper, we introduced a novel multimodal dual-path framework that synergistically integrates large vision-language models and large language models for radiology report generation. By establishing a dynamic fusion between visual-semantic understanding and clinical knowledge injection, with a structured prompting strategy employed, our approach effectively enhances the clinical accuracy of generated reports, making a big step towards automatic report generation that are not only fluent but also clinically reliable.

6 Limitations

Despite the promising improvement over existing approaches, our method has several limitations that warrant further investigation. First, the current framework relies on the quality of both the LVLM and LLM base models; improvements in either backbone could further enhance overall performance. Second, it requires the vocabulary of the LVLM and LLM components to be aligned, which may limit the choice of models. In future work, we plan to explore more advanced fusion strategies and investigate the use of other large models to further improve the performance of our framework.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Eval. Measures for Mach. Transl. and/or Summarization*, pages 65–72.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256. Association for Computational Linguistics.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. 2025. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *EMNLP*, pages 1439–1449.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, and 1 others. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, and 1 others. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2607–2615.

Alistair Johnson, Matt Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. 2019a. MIMIC-CXR-JPG-chest radiographs with structured labels. *PhysioNet*.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019b. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arxiv Preprint arxiv:1901.07042*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. 2025. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. *arXiv preprint arXiv:2502.20056*.

Jieting Long, Zhiyuan Li, Jianan Fan, Zhuonan Liang, Ao Ma, Henning Müller, and Weidong Cai. 2025. Diversity-augmented diffusion network with llm assistance for radiology report generation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2288–2296.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144:102633.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive transformer-based generation of radiology reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proc. 40th Annu. Meeting Assoc. for Comput. Linguistics*, pages 311–318.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. PyTorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8026–8037.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. 2023. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv e-prints*, pages arXiv–2311.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. 2024. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4776–4783.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442.

Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, and 1 others. 2025. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. 2023. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*.

Joseph Turian, Luke Shen, and I Dan Melamed. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of machine translation summit IX: papers*.

Fanqi Wan, Xinting Huang, Deng Cai, Xiaojun Quan, Wei Bi, and Shuming Shi. 2024. Knowledge fusion of large language models. *arXiv preprint arXiv:2401.10491*.

Hongyi Wang, Felipe Maia Polo, Yuekai Sun, Souvik Kundu, Eric Xing, and Mikhail Yurochkin. 2023a. Fusing models with complementary expertise. In *Annual Conference on Neural Information Processing Systems*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023b. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023c. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Yongfu Wei, Yan Lin, Hongfan Gao, Ronghui Xu, Sean Bin Yang, and Jilin Hu. 2025. Path-llm: A multi-modal path representation learning by aligning and fusing with large language models. In *Proceedings of the ACM on Web Conference 2025*, pages 2289–2298.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

Bin Yan and Mingtao Pei. 2022. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2982–2990.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024a. Ratescore: A metric for radiology report generation. In *EMNLP*, pages 15004–15019.

Zihao Zhao, Sheng Wang, Jinchen Gu, Yitao Zhu, Lanzhuju Mei, Zixu Zhuang, Zhiming Cui, Qian Wang, and Dinggang Shen. 2024b. ChatCAD+: Toward a universal and reliable interactive CAD using LLMs. *IEEE Transactions on Medical Imaging*.

A Appendix

A.1 Implementation

The PyTorch (Paszke et al., 2019) framework (2.4.0) is used for experiments. Images are resized to 512×512 pixels. We use the Qwen2-VL-7B model (Wang et al., 2024) as our LVLM and the Qwen2-7B model (Yang et al., 2024) as our LLM. We use the Low-Rank Adaptation (LoRA) (Hu et al., 2022) to fine-tune the LVLM and LLM. For the LVLM, we adopt LoRA with a rank and alpha of 64, and a dropout of 0.05. The learning rate is set to 1×10^{-4}. We use a weight decay of 0.1 and a warmup ratio of 0.03. For the LLM, we set the learning rate to 5×10^{-6}, with a weight decay of 0.1. We adopt LoRA with a rank of 8 and an alpha of 16. The batch size is four, with the gradient accumulation set to four steps. We use a warmup of 100 steps. Training is performed with the bf16 precision for one epoch.

A.2 Effect of Fusion Weight α

α	CE Metrics			NLG Metrics			
	Precision	Recall	F1 Score	BLEU-1	BLEU-4	METEOR	ROUGE
0	0.400	0.198	0.265	0.069	0.015	0.092	0.101
0.2	0.438	0.231	0.302	0.111	0.010	0.105	0.153
0.4	0.572	0.471	0.520	**0.280**	**0.070**	0.127	0.275
0.6	**0.591**	**0.476**	**0.527**	0.263	0.063	0.140	0.286
0.8	0.523	0.471	0.510	0.216	0.059	**0.145**	0.286
1.0	0.535	0.464	0.497	0.246	0.064	0.144	**0.288**

Table A1: Ablation study on the fusion coefficient α for combining LVLM and LLM predictions during inference. The best results are in bold, the second best are underlined.

Table A1 presents the performance of our framework under different values of the fusion coefficient α, which controls the relative contribution of the LVLM and LLM in the report generation process. As we can see, when $\alpha = 0$ (i.e., decoding using only the LLM), both classification and generation metrics are significantly lower than other settings, indicating that the LLM alone is insufficient for accurate report generation due to the lack of visual grounding. When $\alpha = 1.0$ (i.e., decoding using only the LVLM), the performance improves substantially, demonstrating the importance of visual information. However, the best results are achieved at $\alpha = 0.6$, where CE metrics reach their highest values. These results indicate that a balanced integration of the LLM and LVLM effectively leverages their complementary strengths, leading to superior report generation performance.

A.3 Effects of Training Generation Target Structure and Section Order

	CE Metrics			NLG Metrics			
	Precision	Recall	F1 Score	BLEU-1	BLEU-4	METEOR	ROUGE
I-D-F	0.535	**0.464**	**0.497**	**0.246**	**0.064**	0.144	0.288
D-I-F	**0.551**	0.440	0.489	0.236	0.061	0.145	0.292
I-F	0.525	0.368	0.433	0.237	0.064	0.137	0.284
D-F	0.534	0.441	0.483	0.232	0.061	0.140	0.285
F	0.502	0.369	0.404	0.230	0.062	**0.148**	**0.293**

Table A2: Ablation study on the effect of different section orders and combinations of Impression (I), Disease (D), and Findings (F) in the composed training targets on the MIMIC-CXR dataset. The best results are in bold, whereas the second-best results are underlined.

Table A2 presents an ablation study on the MIMIC-CXR dataset, analyzing the effect of different section orders and combinations in the structure of our proposed training generation targets. All results are obtained using only the LVLM with the visual probability P_{V}. The results show that using all three sections (I-D-F and D-I-F) generally leads to better performance across both CE and NLG metrics. Specifically, the I-D-F structure achieves the best F1 score and BLEU scores, while D-I-F yields the highest precision and competitive results on other metrics. It is worth noting that these performance variations, though consistent, are of a much smaller magnitude than the degradation caused by removing entire sections, underscoring that the presence of content is paramount while its order offers secondary refinement. Notably, removing the Disease categories (D) section (i.e., comparing DIF/IDF and IF/F) leads to a substantial decrease in classification performance, with the F1 score dropping by up to 9.3%, indicating that the

Disease section provides crucial information for accurate classification. Overall, these results suggest that a comprehensive, information-rich, and well-ordered training generation target is crucial for optimal model performance.