Beyond N-grams: A Hierarchical Reward Learning Framework for Clinically-Aware Medical Report Generation

Yuan Wang[1], Shujian Gao[2], Jiaxiang Liu[1,3], Songtao Jiang[1], Haoxiang Xia[1], Xiaotian Zhang[1], Zhaolu Kang[4], Yemin Wang[1], Zuozhu Liu[1*]

[1]Zhejiang University, Hangzhou, China
[2]Fudan University, Shanghai, China
[3]Guangdong Institute of Intelligence Science and Technology, Zhuhai, China
[4]Peking University, Beijing, China

Abstract

Automatic medical report generation can greatly reduce the workload of doctors, but it is often unreliable for real-world deployment. Current methods can write formally fluent sentences but may be factually flawed, introducing serious medical errors known as clinical hallucinations, which make them untrustworthy for diagnosis. To bridge this gap, we introduce **HiMed-RL**, a Hierarchical Medical Reward Learning Framework designed to explicitly prioritize clinical quality. HiMed-RL moves beyond simple text matching by deconstructing reward learning into three synergistic levels: it first ensures linguistic fluency at the token-level, then enforces factual grounding at the concept-level by aligning key medical terms with expert knowledge, and finally assesses high-level diagnostic consistency at the semantic-level using a specialized LLM verifier. This hierarchical reward is implemented via a **Human-inspired Dynamic Reward Adjustment**, a strategy which first teaches the model to learn basic facts before progressing to more complex diagnostic reasoning. Experimentally, **HiMed-3B** achieves state-of-the-art performance on both in-domain and out-of-domain benchmarks, particularly on the latter, with an improvement of **12.1%** over the second-best baseline. Our work provides a robust paradigm for generating reports that not only improve fluency but clinical fine-grained quality.

Code — https://github.com/Venn2336/HiMed-RL

Introduction

Automatic Medical Report Generation (MRG), which aims to generate textual descriptions from medical images, is a promising solution to alleviate the documentation burden in clinical practice (Kyung et al. 2025; Guo et al. 2024). More than just writing fluently, reports must be factually accurate with the visual information and show a deep understanding of medical knowledge (Zheng et al. 2024). However, current methods often struggle to meet these clinical standards.

Conventional MRG-specific methods, such as R2Gen (Chen et al. 2020) and Att2in (Xu et al. 2016), tend to overfit the training data; while achieving high keyword overlap, they often underperform on clinical metrics that assess semantic accuracy and medical relevance (Li et al. 2025). Besides, another line of research has focused on refining the language modeling process within the Supervised Fine-Tuning (SFT) paradigm for Multi-modal Large Language Models (MLLMs) (Liu et al. 2024; Jiang et al. 2025). These efforts include advanced prompt engineering with medical entities (Jin et al. 2024) or question-driven cues (Zhang et al. 2025a), and enhanced fine-tuning strategies to improve visual comprehension (Zheng et al. 2024). However, such methods are fundamentally bottlenecked by the SFT objective itself. The standard goal of maximizing token-level likelihood (Zeng et al. 2024) does not inherently enforce the factual integrity and logical consistency that are paramount for deployment in clinical practice.

Reinforcement Learning (RL) has demonstrated exceptional capabilities in generalization, complex reasoning, and trustworthy alignment, making it a highly suitable paradigm for MRG. However, the potential of RL in this domain is currently hindered by simplistic reward designs. Current methods rely on rule-based rewards that operate at lower linguistic levels, mainly focusing on token and concept matching. As we illustrate in ***Figure 1(a)***, linguistic quality in MRG can be assessed at three levels: token, concept, and semantic. For instance, token-level approaches often depend on superficial metrics like BLEU (Zou et al. 2025). At the concept-level, a significant body of work incentivizes matching medical keywords to improve clinical utility (Zhang et al. 2025c; Dai et al. 2025a; Fan et al. 2025; Zhang et al. 2025b).

However, we argue that the complexity of MRG necessitates a semantically coherent and procedurally trustworthy reasoning process, which cannot be adequately addressed by the aforementioned reward types alone. Previous reward designs fall short, particularly in generating clinically consistent reports (Hou et al. 2024), mitigating factual hallucinations (Wu et al. 2024; Xu et al. 2023), resolving semantic contradictions (Zhou et al. 2021), and perceiving underlying pathological structures (Bu et al. 2024), since rule-based mechanisms struggle to evaluate these challenges directly.

Hence, we introduce **HiMed-RL**, a hierarchical reward learning paradigm for MRG. Specifically, as shown in ***Figure 1***, at the token-level, we ensure linguistic fluency and syntactic correctness to establish report readability. Next, at the concept-level, our reward encourages alignment with key medical terminology and domain knowledge, reinforcing entity consistency and factual grounding to mitigate clin-

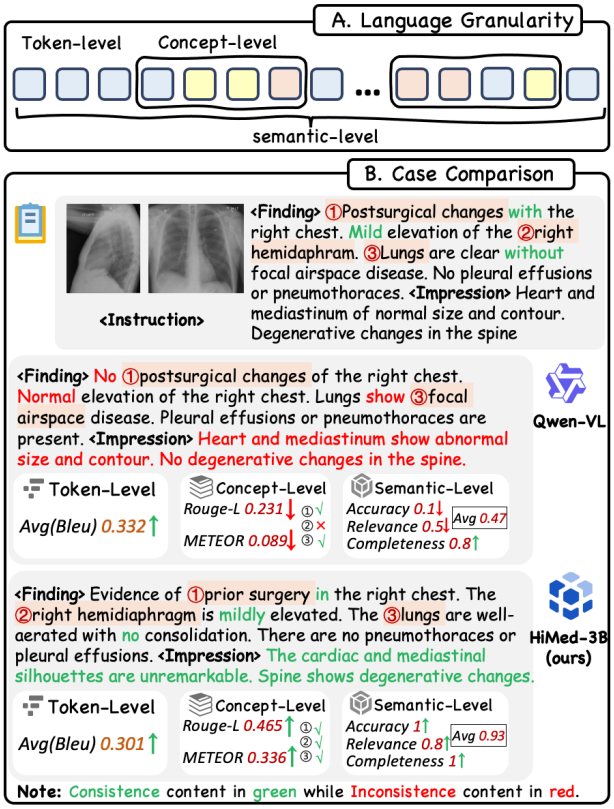

Figure 1: (a) *Token-level*, the report text is decomposed into individual words or sub-word tokens for assessment. *Concept-level*, relevant tokens are aggregated into clinically significant medical concepts, e.g., postsurgical changes. *Semantic-level* aims to comprehend the core meaning and diagnostic conclusion of the entire report. (b) Our HiMed-3B model outperforms Qwen2.5-VL-7B across three evaluation levels: at the token-level, with a comparable average BLEU score; at the concept-level, by correctly identifying three medical entity types and achieving superior ROUGE-L and METEOR scores; and at the semantic-level, demonstrating enhanced accuracy, relevance, and completeness.

ical hallucinations. In contrast to previous works (Fan et al. 2025), our method introduces an LLM verifier to conduct a high-level evaluation of the semantic integrity and diagnostic consistency of the entire report, ensuring the rationality of the reasoning logic and the credibility of the medical judgment. In terms of training strategy, inspired by the real-scenario radiologist-decision process, we use dynamic reward weight adjustment, starting with learning basic facts and gradually transitioning to training more challenging diagnostic reasoning abilities. This phased optimization method not only improves learning stability but also more closely follows the physician's thought process of *from factual statement to clinical judgment* when writing reports. Our contributions are three-fold:

- We reconceptualize MRG as a complex long sequence modeling task. Unlike the previous textual-similarity–driven optimization paradigm, we introduce **HiMed-RL**, a framework that jointly optimizes reward signals at the token, concept, and semantic levels.
- We introduce **Human-inspired Dynamic Reward Adjustment** policy during training. This mechanism guides the model through a progressive optimization process, transitioning from foundational linguistic expression to complex medical reasoning, thereby mirroring the cognitive workflow of clinicians in real-world scenarios.
- We conduct extensive experiments to demonstrate that our proposed **HiMed-3B** achieves state-of-the-art (SOTA) performance on multimodal medical imaging datasets. Our detailed evaluations validate the effectiveness and superiority of our approach.

Related Work

Medical Report Generation

MRG, a specialized sub-task of vision-language learning, aims to automatically generate coherent and accurate diagnostic reports from medical images. Prior efforts have sought to enhance model performance by integrating external structures and knowledge. These strategies include embedding cognitive modules into the network architecture (Huang, Zhang, and Zhang 2023; Huang et al. 2025), leveraging medical knowledge graphs (Li et al. 2024), and employing multi-task learning to foster more robust and clinically-aware representations (Jin et al. 2024). However, those methods are inherently limited by their reliance on training from scratch with small-scale medical datasets, which restricts generalization. The advent of LLMs marks a paradigm shift (Ma et al. 2025); the extensive pre-training on vast corpora provides robust reasoning and fluency that can overcome the data scarcity issues to the medical domain, fundamentally reshaping the technical landscape for MRG.

Reward Learning

Reward learning has significantly advanced the reasoning and generation capabilities of LLMs, enhancing the out-of-domain generalization. Nevertheless, designing an appropriate reward learning framework tailored to the unique demands of different tasks remains a challenge. This often requires task-specific reward engineering. For instance, text summarization may employ rewards for information overlap alongside penalties for excessive length (Aggarwal and Welleck 2025); Question-Answering tasks frequently rely on accuracy-based rewards (Wang et al. 2025). In this work, we address this challenge within the context of MRG.

Methodology

In this section, we present our approach, which employs RL with a Hierarchical Medical Reward Learning Framework and Human-inspired Dynamic Reward Adjustment policy, as shown in ***Figure 2***.

Hierarchical Medical Reward Scaling

As discussed in the introduction, current reward designs for MRG cannot sufficiently address issues such as clinical hallucinations, semantic contradictions, and overall coherence.

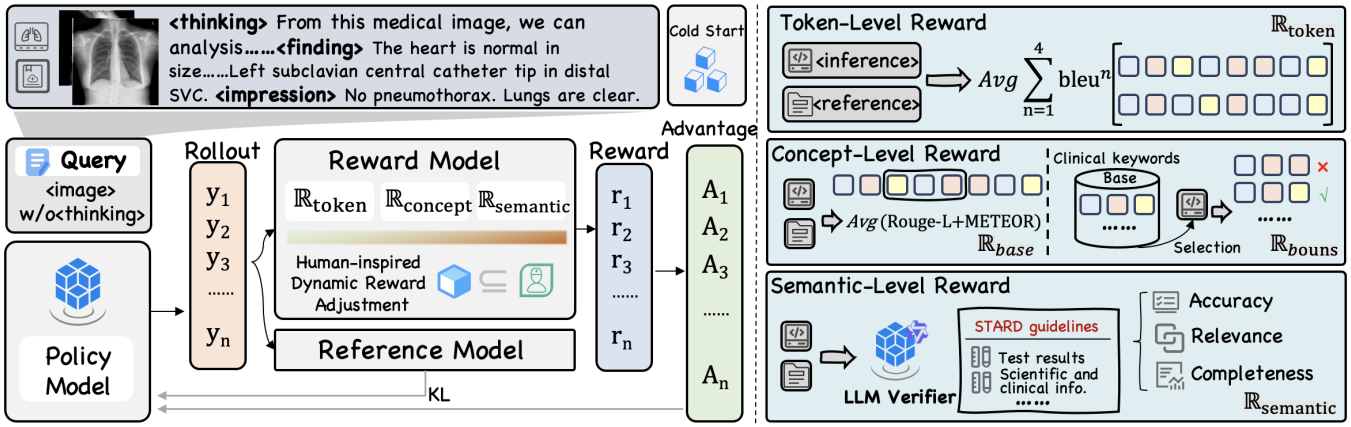

Figure 2: An overview of the HiMed-RL pipeline. MRG task is trained by a Hierarchical Reward Learning Framework that integrates token-level, concept-level, and semantic-level rewards. Human-inspired Dynamic Reward Adjustment strategy guides the model's learning process, transitioning from foundational fluency to complex medical reasoning, to optimize report quality.

This is primarily because a comprehensive reward learning framework for MRG is still absent. To address this limitation, we propose Hierarchical Reward Learning strategy. (Details of prompt design are given in ***Appendix A.2***)

Preliminaries and Formal Definitions In our setting, given a medical image I, the goal of our policy model π is to generate a piece of text description, that is, a medical report Y. We formalize both the generated report Y and the gold standard reference report $\hat{Y}$ as a sequence of tokens:

- **Generated Report:** $Y = (y_1, y_2, \ldots, y_L)$, with a sequence length of L.
- **Reference Report:** $\hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{\hat{L}})$, with a sequence length of $\hat{L}$.

For the precise definition of subsequent reward functions, we first introduce the concept of n-gram. An n-gram is a continuous subsequence of length n extracted from a token sequence. We use $\text{grams}_n(S)$ to denote the set of all n-grams extracted from an arbitrary sequence S.

Token-Level Reward ($\mathbb{R}_{\text{token}}$) It is designed to measure the linguistic similarity between a generated sequence Y and a reference sequence $\hat{Y}$. Its formulation is inspired by the BLEU score and integrates two key components: modified n-gram precision and a brevity penalty.

A simple calculation of n-gram precision measures the fraction of n-grams in the candidate sequence Y that appear in the reference sequence $\hat{Y}$. However, this approach improperly handles repeated words, as a model could be rewarded for the over-generation of high-frequency terms. Therefore, we employ a modified n-gram precision, denoted as p_n, which incorporates a clipping mechanism.

$$p_n(Y, \hat{Y}) = \frac{\sum_{g \in \text{grams}_n(Y)} \min(C(g, Y), C(g, \hat{Y}))}{\sum_{g' \in \text{grams}_n(Y)} C(g', Y)}, \quad (1)$$

where $\text{grams}_n(S)$ represents the multiset of n-grams in a sequence S, and $C(g, S)$ denotes the count of a specific n-gram g within S.

In addition, to penalize generated sequences that are unduly shorter than their references, we introduce a brevity penalty (BP). This factor is defined as:

$$\text{BP} = \min\left(1, \exp\left(1 - \frac{\hat{L}}{L}\right)\right), \quad (2)$$

where L and $\hat{L}$ are the respective lengths of the generated sequence Y and the reference sequence $\hat{Y}$.

The final reward function $\mathbb{R}_{\text{token}}$ combines these elements:

$$\mathbb{R}_{\text{token}}(Y, \hat{Y}) = \text{BP} \cdot \sum_{n=1}^{4} \lambda_n p_n(Y, \hat{Y}), \quad (3)$$

in our implementation, we use the standard BLEU-4 formulation, where λ_n are weights for each n-gram order.

Concept-Level Reward ($\mathbb{R}_{\text{concept}}$) Above the form, $\mathbb{R}_{\text{concept}}$ aims to assess the consistency of the generated sequence Y with $\hat{Y}$ in clinical factual content. It is composed of a base alignment reward R_{base} and a constrained entity reward bonus R_{bonus}. The former is measured through ROUGE-L and METEOR to evaluate the coverage and accuracy of basic clinical concepts,

$$\mathbb{R}_{\text{base}}(Y, \hat{Y}) = F_{\text{LCS}}(Y, \hat{Y}) + \text{METEOR}(Y, \hat{Y}), \quad (4)$$

where F_{LCS} is the F1-score component of ROUGE-L, calculated using the Longest Common Subsequence (LCS) between the candidate sequence Y and reference $\hat{Y}$.

To encourage the generation of key medical entities, we introduce a bonus reward, $\mathbb{R}_{\text{bonus}}$. This reward is granted based on a predefined set of critical keywords, $\mathcal{K}$, which includes terms from USMLE [1] and RSNA (Langlotz 2006) standardized medical vocabularies. A reward is given for each keyword $k \in \mathcal{K}$ (represented as a token sequence) found as a subsequence in the output Y. To prevent the

[1] United States Medical Licensing Examination

model from exploiting this reward through simple repetition, the total bonus is capped at a maximum value of τ_{limit}.

Finally, the total concept-level reward combines the base and bonus components as follows:

$$\mathbb{R}_{\text{concept}}(Y, \hat{Y}) = \mathbb{R}_{\text{base}}(Y, \hat{Y}) + \min\left(\sum_{k \in \mathcal{K}} \beta \cdot \mathbb{I}(k \subseteq Y), \tau_{\text{limit}}\right), \quad (5)$$

where $\mathbb{I}(k \subseteq Y)$ is an indicator function that returns 1 if keyword k is found in the output Y, and β is the reward value for a single match.

Semantic-Level Reward ($\mathbb{R}_{\text{semantic}}$) While token-level and concept-level rewards ensure linguistic fluency and factual accuracy of individual clinical entities, they are insufficient for evaluating the overall clinical coherence and diagnostic integrity of the entire report. To bridge this gap, we introduce a semantic-level reward, R_{semantic}, which leverages a powerful clinical verifier as an impartial judge to perform a holistic evaluation of the generated report. This LLM verifier assesses the report's quality along three critical axes of clinical utility: Accuracy, Relevance, and Completeness, grounded in the Standards for Reporting of Diagnostic Accuracy (STARD) 2015 guidelines (Cohen et al. 2016).

We formalize the scoring function as $\mathcal{J}$, which takes the generated report Y and the ground-truth reference report $\hat{Y}$ as inputs. It then outputs a vector of scores, where each score is normalized to the range $[0, 1]$.

$$\mathcal{J}(Y, \hat{Y}) = s_{\text{acc}} + s_{\text{rel}} + s_{\text{com}}, \quad s_{\text{acc}},\ s_{\text{rel}},\ s_{\text{com}} \in [0, 1]. \quad (6)$$

The evaluation dimensions are defined as follows:

- **Accuracy (s_{acc}):** Assesses factual correctness by penalizing clinical *hallucinations* (unfounded findings) and *contradictions* against the reference report. This metric is guided by STARD's principles of reporting diagnostic accuracy (Items 1, 24) and systematically comparing results against a reference standard (Item 23).
- **Relevance (s_{rel}):** Measures clinical pertinence by evaluating whether the report prioritizes critical pathological findings and avoids verbose, non-contributory descriptions. This aligns with STARD's emphasis on defining the test's clinical role (Item 3) and its implications for practice (Item 27).
- **Completeness (s_{com}):** Quantifies the coverage of all essential clinical observations from the reference, penalizing the omission of significant findings. Inspired by STARD's requirement for comprehensive data reporting (Items 19-21, 23).

For semantic-level reward evaluation, our prompt template follow the STARD guidelines, which includes both the generated report (Y) and the reference report ($\hat{Y}$). (Details of LLM Verifier Case Study will be seen in ***Appendix A.3***).

Format Reward We use regular expression extraction to enforce a structured response format. The model is required to place its reasoning process within `<think></think>` tags and provide the medical report inside `<finding></finding>` and `<impression></impression>` tags. The format reward score ($\mathbb{R}_{\text{format}}$) is computed as:

$$\mathbb{R}_{\text{format}} = \begin{cases} 1, & \text{if format is correct} \\ \text{-1}, & \text{if format is incorrect} \end{cases} \quad (7)$$

Total Reward Function To synergistically optimize for both foundational correctness and high-level clinical reasoning, we construct a composite total reward, $\mathbb{R}_{\text{total}}$. First, Low-Level Reward $\mathbb{R}_{\text{low-level}}$, which assesses structural and factual integrity, is formulated below:

$$\mathbb{R}_{\text{low-level}} = w_t \mathbb{R}_{\text{token}} + w_c \mathbb{R}_{\text{concept}} + w_f \mathbb{R}_{\text{format}}, \quad (8)$$

where w_t, w_c, and w_f are scalar weights.

The total reward, $\mathbb{R}_{\text{total}}$, is then defined as a dynamically weighted combination of this low-level reward and the high-level semantic reward, $\mathbb{R}_{\text{semantic}}$. At any given training step t, the function is:

$$\mathbb{R}_{\text{total}}(t) = \alpha_1(t) \cdot \mathbb{R}_{\text{low-level}} + \alpha_2(t) \cdot \mathbb{R}_{\text{semantic}} \quad (9)$$

The time-varying hyperparameters, $\alpha_1(t)$ and $\alpha_2(t)$, are crucial for our training strategy, enabling a strategic shift in optimization focus as training progresses.

Human-inspired Dynamic Reward Adjustment

A critical question arises within a multi-level reward learning framework: how can we orchestrate its various components to foster a coherent learning progression? Hence, we propose the Human-inspired Dynamic Reward Adjustment policy, which guides the model through a progressive learning process, transitioning from mastering fundamental concepts to performing complex reasoning. The adjustment mechanism is scheduled as follows:

- Initial Phase: At the beginning of the training, we set a high value for $\alpha_1(t)$ and a low value for $\alpha_2(t)$. This encourages the model to focus on generating linguistically fluent and factually accurate text, mastering the building blocks of a valid report.
- Transition Phase: As training progresses, we gradually decrease $\alpha_1(t)$ while simultaneously increasing $\alpha_2(t)$ We employ a linear scheduling function to ensure a smooth transition:

$$\alpha_1(t) = \max\left(1 - \frac{t}{T}, \alpha_{\text{min}}\right), \quad \alpha_2(t) = 1 - \alpha_1(t) \quad (10)$$

RL Algorithm We employ the Group Reward Policy Optimization (GRPO) algorithm (Shao et al. 2024) for training. For each input q, we sample a group of candidate outputs $\{o_i\}_{i=1}^{G}$, compute their advantages A_i based on our rule-metric mixed rewards, and then optimize the policy π_θ by maximizing the GRPO objective function:

$$\begin{aligned} J_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim P(Q),\, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)} \\ \Bigg[\frac{1}{G} \sum_{i=1}^{G} \min\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)} A_i, \\ \text{clip}\Big(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, 1-\varepsilon, 1+\varepsilon \Big) A_i \Big) \\ - \beta\, D_{\text{KL}}\big(\pi_\theta \,\big\|\, \pi_{\text{ref}}\big) \Bigg], \end{aligned} \quad (11)$$

where ε is the clipping hyperparameter and β controls the KL divergence penalty against a reference policy π_{ref}. Our implementation and hyperparameter settings follow the original work (Shao et al. 2024).

Experiment

In our experiments, we aim to answer three core questions:

Q1 Does our proposed model outperform contemporary baselines, including MRG-specific methods and MLLMs-based MRG methods? ***A1: Main Results.***

Q2 Does our hierarchical reward design, emphasizing multiple linguistic granularities, enhance the generation of coherent, consistent, and semantically aligned medical reports to the ground truth? ***A2: Ablation Study.***

Q3 Does our Human-inspired Dynamic Reward Adjustment strategy effectively align the generation process of HiMed-RL with the diagnostic workflow of human radiologists, from basic facts interpretation to comprehensive report synthesis? ***A3: Discussion and Case Study.***

Experiment Settings

Baselines and Datasets. We benchmarked HiMed-RL against four categories of baselines: (1) general multimodal models such as Qwen2.5-VL (Bai et al. 2025) and InternVL3 (Zhu et al. 2025); (2) medically fine-tuned models including LLaVA-Med (Li et al. 2023) and MedGemma (Sellergren et al. 2025); (3) reinforcement learning-based methods like QoQMed-VL (Dai et al. 2025b) and MedVLM-R1 (Pan et al. 2025); (4) MRG-specific methods like R2Gen (Chen et al. 2020), RGRG (Tanida et al. 2023). We used MIMIC-CXR (Johnson et al. 2019), Chexpert (Irvin et al. 2019), and IU-Xray (Demner-Fushman et al. 2015) for in-domain evaluation, and Padchest-GR (de Castro et al. 2025) for out-of-domain generalization test (see ***Appendix A.1*** for details).

Evaluation Metrics. To ensure a thorough and multi-faceted evaluation, we adopt a comprehensive suite of metrics organized across three distinct levels of granularity for a complementary and detailed assessment of report quality. Specifically, at the token-level, we employ BLEU-1 through BLEU-4 to measure n-gram precision and fluency. For concept-level, METEOR can consider precise word matching, including synonyms and rewriting, and provide a more detailed report quality assessment (Banerjee and Lavie 2005). At the semantic level, clinical accuracy was assessed using the RaTEScore metric, which is specifically designed for the evaluation of medical reports. This metric prioritizes the accuracy of medical entities, such as anatomical details and diagnostic findings (Zhao et al. 2024).

Implementation Details. We implement our HiMed-3B model and its training pipeline using the `verl` framework. The backbone is initialized from Qwen2.5-VL-Instruct-3B, and then undergoes supervised fine-tuning on the target medical datasets as a cold-start phase. For the reinforcement learning stage, we adopt GRPO algorithm. The learning rate is set to 1×10^{-6}, with a mini-batch size of 32. At each GRPO step, we sample a group of $G = 16$ candidate reports for each set of medical images. The PPO clipping threshold is set to $\varepsilon = 0.2$, and the KL-divergence penalty weight is set to $\beta = 0.1$. For the semantic-level reward, we employ `qwen3-30b-a3b` as the LLM verifier. All experiments are conducted on a cluster of eight NVIDIA A100 GPUs. Hyperparameters settings are given in ***Appendix A.8***.

Components			Performance Metrics		
$w/\mathbb{R}_{\text{semantic}}$	$w/\mathbb{R}_{\text{concept}}$	$w/\mathbb{R}_{\text{token}}$	MIMIC-CXR	CheXpert	IU-Xray
			0.050	0.054	0.073
		✓	0.192	0.129	0.263
	✓		0.191	0.117	0.289
✓			0.101	0.125	0.288
	✓	✓	0.258	0.138	0.301
✓	✓	✓	**0.271**	**0.157**	**0.319**

Table 1: Ablation study of different components performance. We use Rouge-L to evaluate each datasets.

Configuration	ROUGE-L	METEOR	RATE
No Adjustment (Fixed)			
$\alpha_1 = 0.5, \alpha_2 = 0.5$	0.264	0.211	0.525
Linear Decay Adjustment			
$T = 5k, \alpha_{\text{min}} = 0.1$	0.266	0.214	0.536
$T = 20k, \alpha_{\text{min}} = 0.1$	0.269	0.218	0.539
$T = 10k, \alpha_{\text{min}} = 0.0$	0.268	0.216	0.541
$T = 10k, \alpha_{\text{min}} = 0.3$	**0.272**	0.219	0.535
$T = 10k, \alpha_{\textbf{min}} = 0.1$	0.271	**0.220**	**0.544**

Table 2: Impact of Human-inspired Dynamic Reward Adjustment hyperparameters on the MIMIC-CXR. Our final configuration is highlighted.

Main Results

We compare our proposed method with SOTA approaches, which are categorized into general-purpose, medical-domain, RL-based models, and MRG-specific methods. As shown in ***Table 3***, our proposed HiMed-RL achieves the best performance across the majority of metrics. Notably, our model achieves the highest scores on the semantic-level RATE, indicating that it not only generates linguistically fluent and contextually relevant reports but, more importantly, accurately preserves the essential medical findings and impressions. Furthermore, our HiMed-RL model secures this SOTA performance with a relatively compact size of 3B parameters, surpassing various models with significantly larger parameter counts and thus highlighting the efficiency of our training strategy. Moreover, we compare our method with MRG-specific baselines. (Details analysis and results in ***Appendix A.6***) The results confirm their tendency to overfit on token-level details, which leads to a significant degradation in overall semantic quality.

Ablation Study

Ablation studies are illustrated in ***Figure 4***, we first evaluated the performance of the different training stages. We

Dataset	Method	Token-level				Concept-level				Semantic-level
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	METEOR	RATE
MIMIC-CXR	Lingshu-7B*	0.202	0.100	0.048	0.026	0.303	0.098	0.289	0.183	0.532
	MedGemma-27B*	0.228	0.113	0.052	0.019	0.247	0.061	0.233	0.215	0.520
	HuatuoGPT-V-7B*	0.226	0.100	0.035	0.010	0.230	0.045	0.214	0.196	0.489
	HuatuoGPT-V-34B*	0.225	0.096	0.035	0.010	0.231	0.045	0.216	0.199	0.490
	Qwen2.5-VL-7B‡	0.214	0.094	0.034	0.008	0.242	0.050	0.226	0.190	0.472
	BiMediX2-8B*	0.159	0.050	0.005	0.001	0.186	0.025	0.173	0.118	0.446
	LLaVA-Med-7B*	0.042	0.012	-	-	0.150	0.023	0.138	0.066	0.425
	HealthGPT-14B*	0.205	0.084	0.018	0.002	0.205	0.039	0.193	0.172	0.473
	InternVL3-8B‡	0.189	0.085	0.025	0.003	0.234	0.052	0.222	0.185	0.493
	InternVL3-14B‡	0.223	0.101	0.043	0.010	0.226	0.048	0.210	0.219	0.487
	QoQMed-VL-7B§	0.122	0.056	0.022	0.006	0.185	0.037	0.174	0.188	0.495
	MedVLM-R1-2B§	0.184	0.067	0.013	0.002	0.192	0.027	0.178	0.145	0.417
	HiMed-3B§	0.292	0.156	0.088	0.052	0.289	0.079	0.271	0.220	0.544
CheXpert	Lingshu-7B*	0.179	0.071	0.027	0.011	0.194	0.033	0.181	0.163	0.439
	MedGemma-27B*	0.117	0.047	0.018	0.006	0.165	0.033	0.153	0.183	0.471
	HuatuoGPT-V-7B*	0.130	0.047	0.016	0.003	0.157	0.026	0.145	0.180	0.435
	HuatuoGPT-V-34B*	0.093	0.031	0.010	0.002	0.151	0.024	0.141	0.151	0.431
	Qwen2.5-VL-7B‡	0.046	0.019	0.007	0.002	0.103	0.017	0.097	0.135	0.427
	BiMediX2-8B*	0.094	0.022	0.002	-	0.099	0.011	0.092	0.120	0.353
	LLaVA-Med-7B*	0.133	0.040	0.003	-	0.136	0.019	0.122	0.138	0.404
	HealthGPT-14B*	0.119	0.041	0.012	0.003	0.158	0.026	0.146	0.165	0.452
	InternVL3-8B‡	0.074	0.029	0.008	0.002	0.116	0.021	0.111	0.175	0.441
	InternVL3-14B‡	0.080	0.028	0.008	0.002	0.120	0.019	0.114	0.173	0.432
	QoQMed-VL-7B§	0.033	0.015	0.005	0.001	0.090	0.014	0.086	0.115	0.454
	MedVLM-R1-2B§	0.039	0.012	0.002	0.001	0.116	0.014	0.110	0.095	0.399
	HiMed-3B§	0.182	0.077	0.032	0.017	0.163	0.048	0.157	0.219	0.487
IU-Xray	Lingshu-7B*	0.362	0.203	0.130	0.085	0.386	0.138	0.359	0.312	0.590
	MedGemma-27B*	0.268	0.134	0.074	0.038	0.300	0.081	0.282	0.282	0.609
	HuatuoGPT-V-7B*	0.134	0.061	0.025	0.006	0.215	0.040	0.207	0.266	0.544
	HuatuoGPT-V-34B*	0.131	0.063	0.031	0.014	0.232	0.051	0.221	0.266	0.582
	Qwen2.5-VL-7B‡	0.057	0.027	0.011	0.004	0.158	0.031	0.154	0.181	0.569
	BiMediX2-8B*	0.094	0.018	0.002	0.001	0.100	0.009	0.096	0.134	0.420
	LLaVA-Med-7B*	0.116	0.036	0.003	-	0.140	0.018	0.131	0.142	0.432
	HealthGPT-14B*	0.126	0.053	0.022	0.010	0.207	0.040	0.193	0.227	0.541
	InternVL3-8B‡	0.082	0.037	0.018	0.007	0.149	0.034	0.142	0.233	0.561
	InternVL3-14B‡	0.100	0.042	0.017	0.004	0.166	0.032	0.158	0.242	0.556
	QoQMed-VL-7B§	0.034	0.018	0.009	0.003	0.137	0.029	0.132	0.133	0.572
	MedVLM-R1-2B§	0.184	0.067	0.013	0.001	0.153	0.021	0.149	0.145	0.531
	HiMed-3B§	0.407	0.259	0.171	0.115	0.320	0.089	0.319	0.411	0.611

Table 3: The performance of different models on the MIMIC-CXR, CheXpert and IU-Xray datasets. Model types are separated into: ‡General Model, *Medical Model, §RL-Based Model. Higher scores are better for all metrics.

observed significant performance gains after both the initial cold-start SFT and the subsequent HiMed-RL fine-tuning. Notably, performance is enhanced during the RL phase, which demonstrates the powerful advantage of our training strategy in enhancing the quality of reports. Next, our ablation study on the reward components in ***Table 1*** reveals two key findings. First, each reward level individually provides an effective optimization signal, targeting distinct granularities: fluency (token), factual consistency (concept), and diagnostic logic (semantic). Second, combining these components demonstrates a strong synergistic effect, yielding the best performance across three datasets.

Discussion

Human-inspired Dynamic Reward Adjustment Analysis. We analyze the impact of this module's hyperparameters on HiMed-3B's performance, as shown in ***Table 2***. The fixed-weight configuration yields ROUGE-L of 0.264, METEOR of 0.211, and RATE of 0.525. In contrast, the optimal configuration ($T = 10k, \alpha_{\text{min}} = 0.1$) achieves ROUGE-L of 0.271, METEOR of 0.220, and RATE of 0.544, with improvements of 2.7%, 4.3%, and 3.6%, respectively. These findings validate that a balanced transition from factual accuracy to semantic coherence, as facilitated by $T = 10k$ and $\alpha_{\text{min}} = 0.1$, best mimics clinical reasoning, enhancing report quality and mitigating hallucinations.

Method	BLEU-1	METEOR	RATE
Qwen2.5VL-7B	0.024	0.067	0.348
InternVL3-8B	0.010	0.004	0.027
Lingshu	0.074	0.149	0.331
HiMed-3B (Ours)	**0.121**$_{+63.5\%}$	**0.178**$_{+19.5\%}$	**0.371**$_{+12.1\%}$

Table 4: Out-of-distribution performance on the Pad-Chest dataset. We compare HiMed-3B, against several strong baselines. Evaluation is conducted using: BLEU-1 ($\mathbb{R}_{\text{token}}$), METEOR ($\mathbb{R}_{\text{concept}}$), and RATE ($\mathbb{R}_{\text{semantic}}$).

Generalization Study. The generalization test is a core experiment in deep learning. For the bilingual Padchest-GR dataset, we specifically utilized the English subset to ensure linguistic consistency with our training data. As shown

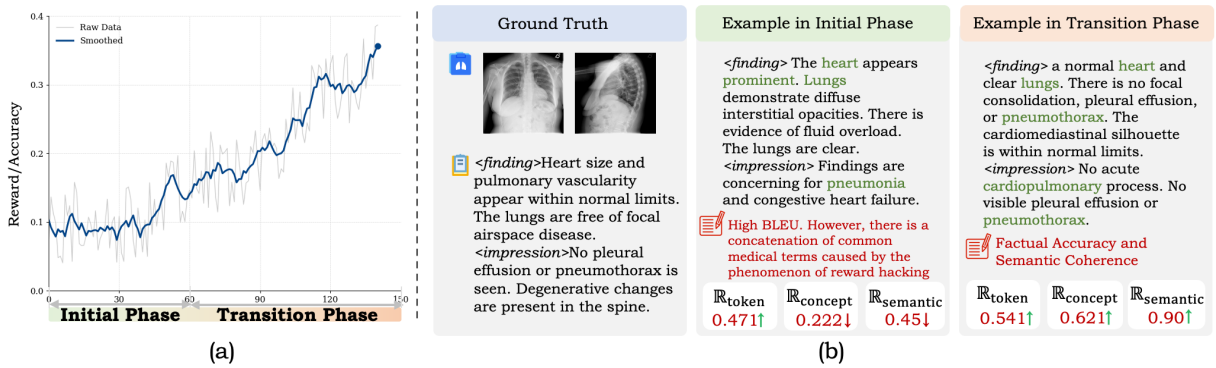

Figure 3: (a) Accuracy reward training curve. (b) Comparison of generated reports between the initial and transition phases.

in ***Table 4***, HiMed-RL achieves significant improvements against three strong baselines across all metrics, affirming its impressive generalization capabilities on unseen data. Results indicate that our strategy enables MLLMs to master intrinsic MRG patterns rather than rote memorization, ensuring practical value for real-world clinical applications.

LLM Verifier Analysis. The effects of different LLM verifiers are presented in ***Table 5***. First, compared to the baseline without a verifier, introducing an LLM to provide semantic-level reward signals significantly enhances the quality of the final generated reports. Second, across all three datasets, Qwen3-a3b (30B) comprehensively outperforms the smaller models. This indicates that a more capable and larger-scale LLM, when acting as a "semantic referee," provides more precise and effective reward signals, thereby improving report generation performance.

Case Study

Figure 3 (b) presents a case comparison that highlights the efficacy of our proposed adjustment strategy. In the initial training phase, the generated report achieves a high token-level reward ($\mathbb{R}_{\text{token}}$), yet it contains severe contradictions and factual errors. For instance, it simultaneously claims the presence of "diffuse interstitial opacities" and that the "lungs are clear." Furthermore, the impression of "pneumonia and congestive heart failure" is entirely inconsistent with the ground truth. We attribute this phenomenon to reward hacking, wherein the model naively stacks professional medical terminology to maximize superficial rewards, while disregarding semantic coherence and factual alignment.

LLM verifier	Params	MIMIC-CXR	CheXpert	IU-Xray
w/o $\mathbb{R}_{semantic}$	–	0.258	0.138	0.301
Llama3.2	8B	0.267	0.134	0.311
Qwen2.5-I	7B	0.266	0.135	0.313
InternVL2	7B	0.265	0.133	0.310
Qwen3-a3b	30B	**0.271**	**0.157**	**0.319**

Table 5: Analysis on the LLM verifier for the semantic-level reward ($\mathbb{R}_{\text{semantic}}$). The best judge verifier is highlighted. We use Rouge-L to compare the performance.

Turning to the transition phase, after adjusting the weights of the semantic-level rewards, the generated report becomes coherent and highly consistent with the ground truth, achieving high scores across all three reward levels. This comparison reveals the essence of our hierarchical reward learning paradigm: by optimizing these reward signals, we guide the model beyond superficial mimicry towards a genuine understanding of the medical concepts within the image, thereby producing logically sound and factually accurate reports.

The reward curve in ***Figure 3 (a)*** corroborates this qualitative analysis. The model's accuracy reward consistently improves throughout training. After a period of slow and unstable improvement in the initial phase, the performance accelerates significantly during the transition phase.

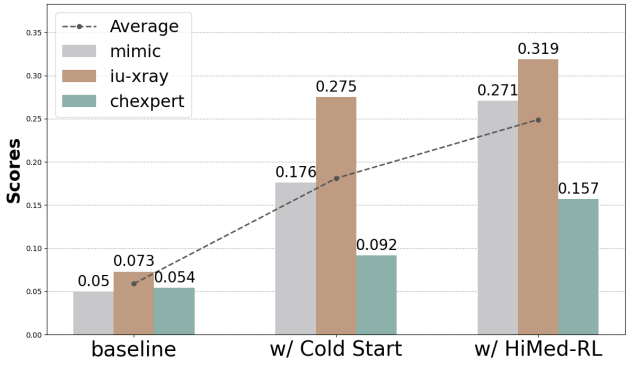

Figure 4: Ablation study results showcasing performance improvements in HiMed-3B across initial cold start and RL fine-tuning stages.

Conclusion

We proposed HiMed-RL, a hierarchical reward learning framework, which leverages reward signals from multiple linguistic granularities to guide the generation of high-quality reports, and its effectiveness is demonstrated through comprehensive experiments. The insight of our work is that by coordinating complementary reward signals with a dynamic learning strategy, we can impel an MLLMs to transition from superficial mimicry to genuine understanding.

Acknowledgments

This work was supported by the National Key R&D Program of China (Grant No. 2024YFC3308304), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant No. 2025C01128), the National Natural Science Foundation of China (Grant No. 62476241), and the Natural Science Foundation of Zhejiang Province, China (Grant No. LZ23F020008) and the ZJU-Angelalign R&D Center for Intelligence Healthcare.

References

Aggarwal, P.; and Welleck, S. 2025. L1: Controlling How Long A Reasoning Model Thinks With Reinforcement Learning. arXiv:2503.04697.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Banerjee, S.; and Lavie, A. 2005. 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*, 65–72.

Bu, S.; Song, Y.; Li, T.; and Dai, Z. 2024. Dynamic knowledge prompt for chest x-ray report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5425–5436.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Cohen, J. F.; Korevaar, D. A.; Altman, D. G.; Bruns, D. E.; Gatsonis, C. A.; Hooft, L.; Irwig, L.; Levine, D.; Reitsma, J. B.; De Vet, H. C.; et al. 2016. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ open*, 6(11): e012799.

Dai, W.; Chen, P.; Ekbote, C.; and Liang, P. P. 2025a. QoQ-Med: Building Multimodal Clinical Foundation Models with Domain-Aware GRPO Training. arXiv:2506.00711.

Dai, W.; Chen, P.; Ekbote, C.; and Liang, P. P. 2025b. QoQ-Med: Building Multimodal Clinical Foundation Models with Domain-Aware GRPO Training. *arXiv preprint arXiv:2506.00711*.

de Castro, D. C.; Bustos, A.; Bannur, S.; Hyland, S. L.; Bouzid, K.; Wetscherek, M. T.; Sánchez-Valverde, M. D.; Jaques-Pérez, L.; Pérez-Rodríguez, L.; Takeda, K.; et al. 2025. Padchest-gr: A bilingual chest X-ray dataset for grounded radiology report generation. *NEJM AI*, 2(7): AIdbp2401120.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S.; Thoma, G. R.; and McDonald, C. J. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310.

Fan, Z.; Liang, C.; Wu, C.; Zhang, Y.; Wang, Y.; and Xie, W. 2025. ChestX-Reasoner: Advancing Radiology Foundation Models with Reasoning through Step-by-Step Verification. arXiv:2504.20930.

Guo, L.; Tahir, A. M.; Zhang, D.; Wang, Z. J.; Ward, R. K.; et al. 2024. Automatic medical report generation: Methods and applications. *APSIPA Transactions on Signal and Information Processing*, 13(1).

Hou, W.; Cheng, Y.; Xu, K.; Hu, Y.; Li, W.; and Liu, J. 2024. ICON: Improving Inter-Report Consistency in Radiology Report Generation via Lesion-aware Mixup Augmentation. *arXiv preprint arXiv:2402.12844*.

Huang, S.; Luo, H.; Jing, H.; Zhang, Q.; Chang, L.; Feng, Y.; Lin, X.; Qin, C.; Chen, H.; Jia, S.; et al. 2025. NEED: Cross-Subject and Cross-Task Generalization for Video and Image Reconstruction from EEG Signals. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Huang, Z.; Zhang, X.; and Zhang, S. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19809–19818.

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

Jiang, S.; Wang, Y.; Song, S.; Hu, T.; Zhou, C.; Pu, B.; Zhang, Y.; Yang, Z.; Feng, Y.; Zhou, J. T.; et al. 2025. Hulu-Med: A Transparent Generalist Model towards Holistic Medical Vision-Language Understanding. *arXiv preprint arXiv:2510.08668*.

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

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317.

Kyung, S.; Seo, J.; Lim, H.; Kim, D.; Park, H.; Sung, J.; Kim, J.; Jo, W.; Nam, Y.; and Kim, N. 2025. MedRegion-CT: Region-Focused Multimodal LLM for Comprehensive 3D CT Report Generation. *arXiv preprint arXiv:2506.23102*.

Langlotz, C. P. 2006. RadLex: a new method for indexing online educational materials.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Li, W.; Han, G.; Wu, Y.; Huang, I.-C.; and Huang, X. 2025. Joint Imbalance Adaptation for Radiology Report Generation. *Journal of Healthcare Informatics Research*, 1–23.

Li, Y.; Wang, Z.; Liu, Y.; Wang, L.; Liu, L.; and Zhou, L. 2024. KARGEN: Knowledge-Enhanced Automated Radiology report generation using large language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 382–392. Springer.

Liu, J.; Wang, Y.; Du, J.; Zhou, J. T.; and Liu, Z. 2024. Medcot: Medical chain of thought via hierarchical expert. *arXiv preprint arXiv:2412.13736*.

Ma, C.; Ji, Y.; Ye, J.; Zhang, L.; Chen, Y.; Li, T.; Li, M.; He, J.; and Shan, H. 2025. Towards interpretable counterfactual generation via multimodal autoregression. *arXiv preprint arXiv:2503.23149*.

Pan, J.; Liu, C.; Wu, J.; Liu, F.; Zhu, J.; Li, H. B.; Chen, C.; Ouyang, C.; and Rueckert, D. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Sellergren, A.; Kazemzadeh, S.; Jaroensri, T.; Kiraly, A.; Traverse, M.; Kohlberger, T.; Xu, S.; Jamil, F.; Hughes, C.; Lau, C.; et al. 2025. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and Explainable Region-guided Radiology Report Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7433–7442. IEEE.

Wang, Y.; Liu, J.; Gao, S.; Feng, B.; Tang, Z.; Gai, X.; Wu, J.; and Liu, Z. 2025. V2T-CoT: From Vision to Text Chain-of-Thought for Medical Reasoning and Diagnosis. *arXiv preprint arXiv:2506.19610*.

Wu, J.; Wu, X.; Zheng, Y.; and Yang, J. 2024. Medkp: Medical dialogue with knowledge enhancement and clinical pathway encoding. *arXiv preprint arXiv:2403.06611*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; and Bengio, Y. 2016. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. arXiv:1502.03044.

Xu, Z.; Xu, W.; Wang, R.; Chen, J.; Qi, C.; and Lukasiewicz, T. 2023. Hybrid reinforced medical report generation with m-linear attention and repetition penalty. *IEEE Transactions on Neural Networks and Learning Systems*.

Zeng, Y.; Liu, G.; Ma, W.; Yang, N.; Zhang, H.; and Wang, J. 2024. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*.

Zhang, X.; Shi, Y.; Ji, J.; Zheng, C.; and Qu, L. 2025a. MEP-Net: Medical Entity-Balanced Prompting Network for Brain CT Report Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 25940–25948.

Zhang, X.; Wang, Y.; Feng, Z.; Chen, R.; Zhou, Z.; Zhang, Y.; Xu, H.; Wu, J.; and Liu, Z. 2025b. Med-U1: Incentivizing Unified Medical Reasoning in LLMs via Large-scale Reinforcement Learning. *arXiv preprint arXiv:2506.12307*.

Zhang, Y.; Yuan, K.; Lu, H.; Yue, Y.; Chen, J.; and Wu, K. 2025c. MedTVT-R1: A Multimodal LLM Empowering Medical Reasoning and Diagnosis. arXiv:2506.18512.

Zhao, W.; Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. RaTEScore: A Metric for Radiology Report Generation. arXiv:2406.16845.

Zheng, C.; Ji, J.; Shi, Y.; Zhang, X.; and Qu, L. 2024. See detail say clear: Towards brain CT report generation via pathological clue-driven representation learning. *arXiv preprint arXiv:2409.19676*.

Zhou, Y.; Huang, L.; Zhou, T.; Fu, H.; and Shao, L. 2021. Visual-textual attentive semantic consistency for medical report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3985–3994.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Zou, J.; Li, Q.; Lian, C.; Liu, L.; Yan, X.; Wang, S.; and Qin, J. 2025. CorBenchX: Large-Scale Chest X-Ray Error Dataset and Vision-Language Model Benchmark for Report Error Correction. arXiv:2505.12057.