Hallucination Mitigating for Medical Report Generation

Anonymous ACL submission

Abstract

In the realm of medical report generation (MRG), the integration of natural language processing has emerged as a vital tool to alleviate the workload of radiologists. Despite the impressive capabilities demonstrated by large vision language models (LVLMs) in understanding natural language, their susceptibility to generating plausible yet inaccurate claims, known as "hallucinations", raises concerns—especially in the nuanced and critical field of medical. In this work, we introduce a framework, **K**nowledge-**E**nhanced with Fine-Grained **R**einforced Rewards **M**edical Report Generation (KERM), to tackle the issue. Our approach refines the input to the LVLM by first utilizing MedCLIP for knowledge retrieval, incorporating relevant lesion fact sentences from a curated knowledge corpus. We then introduce a novel purification module to ensure the retrieved knowledge is contextually relevant to the patient's clinical context. Subsequently, we employ fine-grained rewards to guide these models in generating highly supportive and clinically relevant descriptions, ensuring the alignment of model's outputs with desired behaviors. Experimental results on IU-Xray and MIMIC-CXR datasets validate the effectiveness of our approach in mitigating hallucinations and enhancing report quality.

1 Introduction

Generating radiology reports from medical images represents a critical endeavor within the realm of medical imaging. The task of manually composing such reports by radiologists is not only time-consuming and labor-intensive but also demands a high level of expertise. Consequently, there is a burgeoning interest in methods for automatically generate medical reports for an X-ray, promising solutions that can alleviate these challenges and enhance the overall efficiency of the diagnostic process (Chen et al., 2020; Li et al., 2023b; Yang et al., 2021).

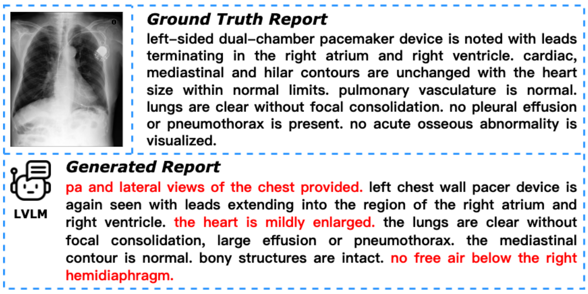


Figure 1: An example of the report generated by the LVLM, where the terms marked in red are hallucinations.

The recent advancements in large language models (LLMs) (Touvron et al., 2023; Ouyang et al., 2022) have inspired the development of large vision-language models (LVLMs) (Dai et al., 2023; Li et al., 2022), which aim to pair these powerful LLMs with image information, building a bridge between the visual and the textual, thus enabling robust comprehension and reasoning across modalities. However, when applying LVLMs to medical report generation, we encountered several challenges, particularly the phenomenon of "hallucinations", where the model generates false yet seemingly plausible information. For instance, as illustrated in Figure 1, the ground truth report describes a patient "with a dual-chamber pacemaker", and the report generated by the LVLM incorrectly suggests "mild enlargement of the heart" as well as some extraneous terms, which are not present in the ground truth. Such hallucinations can lead to misdiagnosis and inappropriate treatment plans, with potentially severe consequences for patient care. Prior methods for mitigating LVLMs' hallucinations have focused on refining the training data and adjusting the model architecture (Liu et al., 2023a; Lee et al., 2023). However, these approaches have not fully addressed the issue, primarily because they neglect the scarcity of high-quality annotations in medical training datasets. The specificity and precision required for medical reports are difficult to achieve without expert knowledge, which can result in model generating incorrect information. This issue stems from the insufficient guidance provided by a lack of accurate and detailed annotations. Moreover, the long-tail problem is prevalent in medical datasets, with common conditions being overrepresented and rare ones underrepresented. This imbalance may cause the model's outputs to deviate from the expected medical findings.

To address these challenges, we propose a new framework, called **K**nowledge-**E**nhanced with Fine-Grained **R**einforced Rewards **M**edical Report Generation (KERM). It efficiently and substantially enhances the visual grounding of LVLMs beyond pretrained baselines such as LLaVA (Liu et al., 2023b), while simultaneously preserving their capability to generate accurate and detailed descriptions. Given a pretrained LVLM (e.g., LLaVA), firstly, we conduct a knowledge corpus, including medical literature and clinical guidelines selected from public datasets such as MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019), and enhance the model's input by retrieving external knowledge sources through MedCLIP (Wang et al., 2022c) and introduces a purification module to refine the relevance of retrieved knowledge to the patient's specific clinical context. We provide the necessary external knowledge to ground the LVLM's understanding, thereby improving the accuracy and relevance of the generated reports. Secondly, we employ fine-grained reward modeling by conducting a dual-level assessment to align the model's output with desired behaviors and mitigate the occurrence of hallucinations. At the disease label level, we evaluate the model's output against known medical labels, ensuring that the diagnoses mentioned are consistent with the image content. At the sentence description level, we utilize GPT-3.5 to scrutinize the coherence and plausibility of the generated sentences, penalizing deviations from the expected medical findings, even if they are not outright incorrect. This encourages the model to generate reports that are not only factually accurate but also aligned with the typical patterns observed in medical practice. Experimental results on a public dataset, MIMIC-CXR (Johnson et al., 2019), confirm the validity and effectiveness of our proposed approach.

Overall, the main contributions of this work are:

- We introduce a knowledge-enhanced approach, which integrates a curated knowledge corpus sourced from public datasets. It can fortifies the LVLM's input with external knowledge, ensuring that the generated medical reports are grounded in accurate and relevant medical information, thereby enhancing the model's ability to produce reliable and detailed descriptions.

- We develop fine-grained reinforced reward modeling that penalizes hallucinatory content from the perspectives of disease-level and sentence-level respectively, promoting outputs that closely align with medical norms and mitigating the occurrence of hallucinations.

- We conduct comprehensive experiments to demonstrate the effectiveness of our proposed method, which outperforms existing methods on both Natural Language Generation and clinical efficacy metrics.

2 Related Work

2.1 Medical Report Generation

The domain of Medical Report Generation (MRG) in medical artificial intelligence (AI) has surged recently. Early research (Allaouzi et al., 2018) drew inspiration from image captioning models, using deep Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) in an encoder-decoder format (Vinyals et al., 2014).Several studies introduced auxiliary classification tasks to predict medical abnormalities (Shin et al., 2016; Wang et al., 2018) , enhancing structured guidance for report generation. The attention mechanism improved the integration of visual and linguistic modalities in MRG systems (Jing et al., 2017; Chen et al., 2020).

To bridge visual observations and medical domain knowledge, numerous visionand- language pre-training methods have been devised to incorporate domain-specific knowledge (Li et al., 2020, 2023b).Generative language modeling evolved from RNNs to transformer architectures, including Large Language Models (LLMs) like LLaMA (Touvron et al., 2023), improving clinical accuracy. Some studies used reinforcement learning (RL) to optimize clinical relevance (Liu et al., 2019; Miura et al., 2020). However, reliance on models like CheXbert or RadGraph for clinical entity extraction complicates optimization.

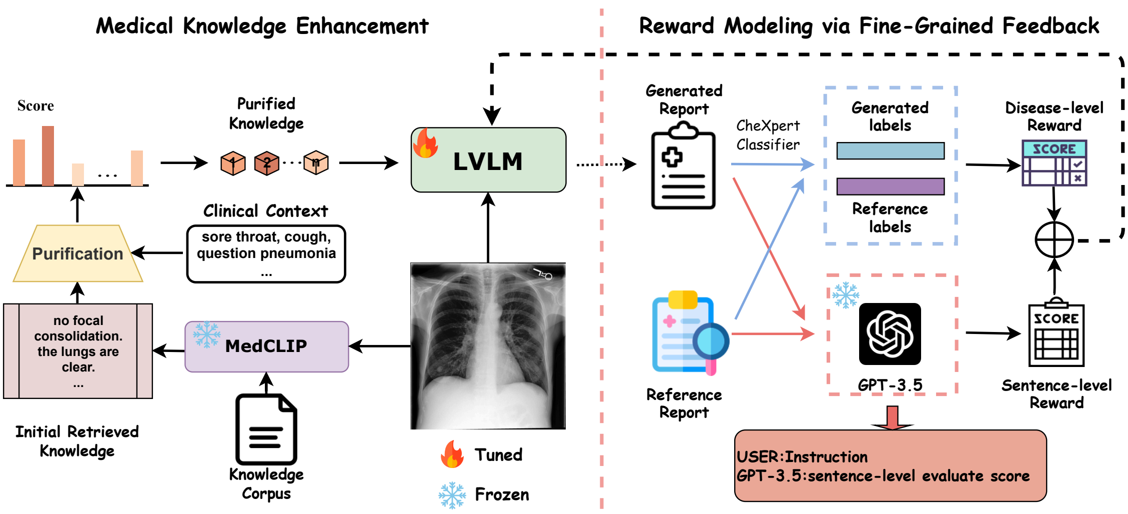


Figure 2: Overview of KERM. We first retrieve the knowledge from our constructed Knowledge Corpus to enhance the image representation as additional input. During the training period, we employ CheXpert to obtain disease labels, applying penalties to hallucinatory content at both the disease and sentence levels. This reward is then feedback to the LVLM, thereby guiding the model's performance.

2.2 Large Vision-Language Models

In recent years, the integration of large language models (LLMs) into multimodal domains has garnered considerable attention (Ouyang et al., 2022; Touvron et al., 2023). This surge has led to the development of large vision-language models (LVLMs) powered by LLMs (Ye et al., 2023; Dai et al., 2023; Li et al., 2022), enabling comprehension of multimodal inputs and performance of diverse tasks under instructions.

LVLMs typically follow a paradigm where a multimodal alignment module comprehends inputs, followed by a LLM generating responses. For instance, mPLUG-Owl (Ye et al., 2023) pre-trains the encoder and alignment module and finetunes LLaMa (Touvron et al., 2023) using low-rank adaption. Conversely, LLaVA (Liu et al., 2023b) pretrains only the alignment network and finetunes it alongside Vicuna (Peng et al., 2023) based on constructed instructions. MiniGPT-4 (Zhu et al., 2023) focuses on finetuning the cross-modal alignment network while freezing other modules.

Recent advancements also include the development of multimodal biomedical chatbots and generalist models. ELIXR, based on the BLIP-2 framework (Li et al., 2023a), trains for contrastive and generative tasks on X-ray image-report pairs, although its evaluation remains private due to the proprietary PaLM-2 model. In contrast, Med-PaLM (Tu et al., 2023) proposes a private, PaLM-based generalist model demonstrating impressive performance across various medical tasks and image types, including VQA, image classification, and report generation. However, neither prioritizes the generation and comprehension of X-ray reports, and they appear to lack clinical accuracy, leading to hallucinations, when evaluated for medical image interpretation.

3 Method

In this section, we will introduce the detailed implementations of our proposed **K**nowledge-**E**nhanced with Fine-Grained **R**einforced Rewards **M**edical Report Generation (KERM). We first introduce the overview of our model, then present the proposed modules, Medical Knowledge Enhancement(MKE) and Reward Modeling via Fine-Grained Feedback(RM), respectively.

3.1 Overview

The overall architecture of our framework is illustrated in Figure 2. It's based on a LVLM, composed of a Medical Knowledge Enhancement branch and a Reward Modeling via Fine-Grained Feedback branch. Given an input medical image I, the system processes it through a visual encoder to obtain image features F_I. These features, along with the retrieved knowledge, are then input into the LVLM to generate a descriptive medical report $R = \{y_1, y_2, \dots, y_n\}$, where y_i is a token and n is the length of the report. We formulate our approach as:

$$K_{retrieved} = \mathrm{MKE}(I, C), \tag{1}$$
$$R = \mathrm{LVLM}((F_I, K_{retrieved})). \tag{2}$$

where $\mathrm{MKE}(\cdot)$ represents the Medical Knowledge Enhancement branch. $K_{retrieved}$ stands for the knowledge retrieved by MedCLIP that is most relevant to the image, with C representing the Knowledge Corpus. The final report R is obtained by decoding the internal states of the LVLM, which are influenced by both the image features and the external knowledge.

Given the ground truth report $R^* = \{y_1^*, y_2^*, \ldots, y_n^*\}$, we can train the model by minimizing a combined loss function that includes cross-entropy loss for language generation and a reinforcement loss guided by the fine-grained rewards:

$$\mathcal{L}_{RL} = \mathrm{RM}(R, R^*) \tag{3}$$
$$\mathcal{L}_{\mathrm{CE}}(\theta) = -\sum_{i=1}^{n} \log p_\theta(y_i = y_i^* | y_{1:i-1}^*, I) \tag{4}$$
$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{RL} \tag{5}$$

where $\mathrm{RM}(\cdot)$ denotes the Reward Modeling via Fine-Grained Feedback branch, and $\mathcal{L}_{RL}$ is the reinforcement loss based on the rewards which we will explain in Section 3.3.3.

3.2 Medical Knowledge Enhancement

To generate accurate radiology reports from medical images, understanding the medical context and relationships depicted in the images is crucial. This requires not only visual recognition but also the ability to interpret the significance of visual features in relation to medical knowledge. Inspired by (Li et al., 2023c) , we first construct a medical knowledge corpus and then utilize a pretrained multimodal model MedCLIP (Wang et al., 2022c) to retrieve relevant facts for each image view, and then apply a purification module to refine the relevance of retrieved knowledge to the patient's specific clinical context. At each step t, the input image with its retrieved knowledge are fed into the LVLM to ground the model's understanding so as to guide better report generation.

3.2.1 Knowledge Corpus Construction

The knowledge base serves as a repository of medical facts that describe the visual content of medical images. To compile a comprehensive and diverse set of medical descriptions, we parse region descriptions from the medical imaging datasets MIMIC-CXR and CheXpert, focusing on their training sets. After removing duplicates, we construct a knowledge corpus consisting of 100k facts expressed in medical language descriptions, which serve as a Knowledge Corpus for our proposed KERM framework.

3.2.2 Knowledge Retrieval

Our objective is to associate each medical image with relevant facts that enhance the model's understanding of the visual content. We employ a pretrained model MedCLIP, which includes an image encoder and a text encoder that map images and text into a shared embedding space. The text encoder is used to encode all facts in the knowledge corpus as search keys, while the image encoder processes the related images as queries. We then identify the facts with the highest cosine similarity scores to the image queries. For each image, we retain the top-10 facts with the highest scores as the initial retrieval knowledge.

3.2.3 Purification Module

Given the high stakes in medical report generation, it is imperative that the knowledge items selected are not only accurate but also highly pertinent to the patient's clinical narrative, including indications and medical history. Therefore, we propose a purification module in our to distill the most contextually relevant knowledge from the initial top-k retrieval result, ensuring that the retrieved facts are optimally aligned with the patient's specific clinical context. Specially, we construct a context embedding E_C that encapsulates the clinical needs and historical features of the patient derived from their *indications* and *clinical history*. Let $K = \{k_1, k_2, \ldots, k_t\}$ represent the initial top-k retrieved facts, each fact k_i is encoded into an embedding E_{k_i} to facilitate the calculation of its similarity to the context vector. Then we computes the cosine similarity between these vectors to quantify the relevance score s_i for each fact, leveraging this score to re-rank the items and prioritize those most contextually aligned with the patient's clinical narrative. The top-5 items, deemed most relevant based on these scores, are selected to form the purified knowledge set K', informing the report generation process.

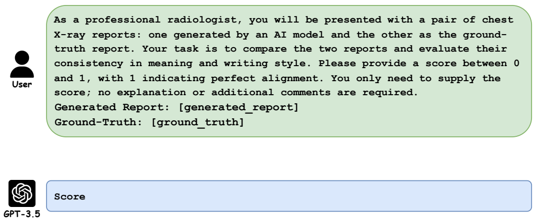


Figure 3: The prompt for generating sentence-level score that scored by GPT-3.5.

3.3 Reward Modeling via Fine-Grained Feedback

In our approach to enhancing the accuracy and coherence of medical report generation, we have developed a novel reinforcement learning strategy that incorporates dual-level reward modeling. This strategy is meticulously designed to mitigate of hallucinations by providing granular feedback at both the disease label and sentence description levels.

3.3.1 Disease-level Reward

We employ the CheXPert (Irvin et al., 2019) labeling tool to label generated reports and the reference reports in 14 different medical terminologies. We calculate the F1 score as the disease-level reward score $\mathbf{R}_{dis}$ for each label to assess the alignment between the model's output and the actual medical findings. The F1 score is a robust measure that balances the trade-off between precision and recall, ensuring that the model's predictions are not only correct but also comprehensive. TP (true positives), FP (false positives), and FN (false negatives) are used to calculate this score, representing correct diagnoses, incorrect diagnoses, and missed diagnoses, respectively.

3.3.2 Sentence-level Reward

At the sentence level, we leverage the advanced language understanding capabilities of GPT-3.5 to assess the coherence and plausibility of the generated sentences. We provide GPT-3.5 with sentence pairs, where one is from the generated report and the other from the reference report, along with detailed evaluation instruction as shown in Figure 3. GPT-3.5 scores the similarity between these pairs ranging from 0 to 1 , with a score closer to 1 indicating a higher degree of coherence and plausibility. This score, $\mathbf{R}_{sen}$, serves as the sentence-level reward.

3.3.3 Reinforcement Algorithm Loss

Since the decoded text cannot provide gradient information for model training, we harness the Reinforce Algorithm (Sutton et al., 1999) to design a loss function aimed at achieving these goals. At each training step, we sample text sequences from the probability distribution $\mathbf{p}$, which is derived from the softmax function applied to the LVLM's logits. The cumulative reward for each sequence is a weighted blend of $\mathbf{R}_{dis}$ and $\mathbf{R}_{sen}$, with a hyperparameter α adjusting the emphasis between disease label and sentence description assessments.The loss function of reinforcement algorithm, which incorporates these reward scores, denoted as $\mathcal{L}_{RL}$:

$$R_t = (1-\alpha)\, R_{dis,t} + \alpha R_{sen,t} \tag{6}$$

$$\mathcal{L}_{RL} = \textstyle\sum_{t=1}^{T} p \cdot R_t \cdot log\,(a_t \mid s_t) \tag{7}$$

where $\mathbf{T}$ represents the length of the generated text, $\mathbf{a_t}$ is the token sampled at step t, $\mathbf{s_t}$ is the corresponding state, α represents hyperparameter, and $\mathbf{R_t}$ represents the reward obtained for the current text.

4 Experiment

4.1 Dataset

We evaluate our proposed KERM on two widely-used radiology reporting benchmark, IU-Xray (Demner-Fushman et al., 2015) and MIMIC-CXR (Johnson et al., 2019), to verify the model's effectiveness. To ensure a fair comparison, we adopt the settings in (Chen et al., 2020) for report preprocessing.

IU-Xray is a publicly available radiological dataset collected by Indiana University, with 7,470 frontal and lateral-view chest X-ray images and 3,955 reports. The reports include *impression*, *findings*, *comparison*, and *indication* sections. Following (Li et al., 2018), we excluded images without reports and there are 5,910 images and 2,955 reports left for this study. Following (Chen et al., 2020), we split the data into training/validation/test set by 7:1:2 of the dataset, and took the *impression* and the *findings* sections as the target captions to be generated.

MIMIC-CXR is the largest radiology image dataset so far, sourcing from the Beth Israel Deaconess Medical Center between 2011-2016. We followed (Liu et al., 2021) to adopt an alpha version

Dataset	Model	NLG Metrics						CE Metrics		
		BL-1	BL-2	BL-3	BL-4	MTR	RG-L	P	R	F1
IU-Xray	HRGR	0.438	0.298	0.208	0.151	-	0.322	-	-	-
	CoAtt	0.455	0.288	0.205	0.154	-	0.369	-	-	-
	PKERRG	0.450	0.301	0.213	0.158	-	0.384	-	-	-
	CMAS-RL	0.464	0.301	0.210	0.154	-	0.362	-	-	-
	R2Gen	0.470	0.304	0.219	0.165	0.187	0.371	-	-	-
	CMN	0.475	0.309	0.222	0.170	0.191	0.375	-	-	-
	PPKED	0.483	0.315	0.224	0.168	0.190	0.376	-	-	-
	Multicriteria	0.496	0.319	0.241	0.175	-	0.377	-	-	-
	KM	0.496	0.327	0.238	0.178	-	0.381	-	-	-
	KERM	**0.511**	**0.333**	**0.249**	**0.182**	**0.197**	**0.388**	-	-	-
MIMIC-CXR	CCR	0.313	0.206	0.146	0.103	-	**0.306**	-	-	-
	Multicriteria	0.351	0.223	0.157	**0.118**	-	0.287	-	-	-
	R2Gen	0.353	0.218	0.145	0.103	0.142	0.277	0.333	0.273	0.276
	CMN	0.353	0.218	0.148	0.106	0.142	0.278	0.334	0.275	0.278
	PPKED	0.360	0.224	0.149	0.106	0.149	0.284	-	-	-
	KM	0.363	0.228	0.156	0.115	-	0.284	**0.458**	0.348	0.371
	KERM	**0.378**	**0.235**	**0.157**	0.109	**0.152**	0.283	0.394	**0.436**	**0.415**

Table 1: Comparisons of our model with previous studies on the IU X-Ray and MIMIC-CXR test set with respect to natural language generation (NLG) and clinical efficacy (CE) metrics. BL-n denotes BLEU score using up to n-grams; MTR and RG-L denote METEOR and ROUGE-L, respectively. P, R and F1 represent precision, recall and F1-score, respectively. KERM is our proposed model. Best results are in bold.

of 473, 057 Chest X-ray images and 206, 563 reports from 63, 478 patients. Each study comprises multiple sections, including *comparison*, *clinical history*, *indication*, *reasons for examination*, *impressions*, and *findings*. We adopted the official split of training/validation/test set, and took the *findings* section as the target captions to be generated.

4.2 Baselines and Evaluation Metrics

Baselines we compare our KERM with a wide range of existing state-of-the-art MRG systems on the benchmark, including R2Gen (Chen et al., 2020), HRGR (Li et al., 2018), CoAtt (Jing et al., 2017), PKERRG (Wang et al., 2022a), CMAS-RL (Jing et al., 2019), CMN (Chen et al., 2022), CCR (Liu et al., 2019), PPKED (Liu et al., 2021), KM (Yang et al., 2021) and Multicriteria (Wang et al., 2022b) . Since we follow the same settings, we directly cite the results from original papers.

Evaluation Metrics We utilize automatic Natural Language Generation (NLG) evaluation metrics such as CIDEr (Vedantam et al., 2014), ROUGE-L (Lin, 2004), and BLEU (Papineni et al., 2002), which quantify the correlation between two text sequences statistically. However, these metrics, which are limited to n-grams of up to 4, may not fully capture the nuances of disease states due to the prevalence of negations in medical language, where negation cues and disease terms can be spatially distant within a sentence. To address this, we incorporate medical abnormality detection as an additional metric. Specifically, we assess the generated reports against the ground truth by comparing the CheXpert (Irvin et al., 2019) labeled annotations for certain categories within the 14 diseases. For this comparison, we calculate the F1-Score, precision, and recall for all models, ensuring a comprehensive evaluation of their performance.

4.3 Implementation Details

In our experiments, we adopt the pretrained Med-CLIP(Wang et al., 2022c) to retrieve facts for each image. And we employ the LVLM, LLaVA-1.5-7b (Liu et al., 2023b) as the backbone, and then we employ LoRA-tuning (Hu et al., 2021) and deepspeed zero stage 3 to conduct minimal training on the model for 1 epoch. The learning rate is set as 2e-4 and the optimizer is AdamW (Loshchilov and Hutter, 2017) with a weight decay of 0.02. During the training phase, we initiate a warm-up ratio of 0.03, after which we apply the cosine schedule to decay the learning rate. We set α to 0.4, based on a hyperparameter search (see Supplemental Material). All of the experiments are conducted on 8

NVIDIA GeForce RTX3090 GPUs.

4.4 Results and Discussion

4.4.1 Main Results

Table 1 presents the comparison results across both Natural Language Generation (NLG) and clinical efficacy (CE) metrics on both MIMIC-CXR and IU X-Ray. On IU X-Ray, our method significantly outperforms methods in previous studies in all NLG metrics. Specifically, KERM achieves BL-4 score of 0.182, MTR score of 0.197, and RG-L score of 0.388. This demonstrates that our model excels not only in generating accurate words and phrases but also in constructing coherent long sentences and maintaining logical flow between sentences. On MIMIC-CXR, it is observed that our method surpasses existing methods in most NLG metrics and achieves comparable performance to the state-of-the-art in BL-4 and MTR. This indicates a robust capability in capturing the nuances of medical language and adhering to clinical standards. The RG-L metric may not be optimal because the order of lesions or sentences in the reports generated by our model does not strictly align with the ground-truth order. In the three CE metrics, our method significantly outperforms previous methods, which indicates that our model predicts much fewer false positive and false negative diseases, respectively. Although our method has a lower precision compared to the KM method, it exceeds KM in the more comprehensive F1-score metric. The significant improvements in CE metrics are a direct result of our approach, which enriches the model's understanding by retrieving factual knowledge from a comprehensive corpus. This is complemented by a fine-grained reward model that penalizes inaccuracies and deviations, ensuring the generation of contextually appropriate and clinically sound reports.

Settings	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
Base	0.445	0.295	0.210	0.162	0.320	0.372
w/MKE	0.475	0.308	0.222	0.170	0.330	0.385
w/RM	0.455	0.302	0.217	0.165	0.325	0.380
KERM	**0.511**	**0.333**	**0.249**	**0.182**	**0.197**	**0.388**

Table 2: The comparison of natural language generation (NLG) metrics on IU X-Ray dataset. "$w/(\cdot)$" means the application of the module.

4.4.2 Ablation study

In this section, we conduct ablation studies on IU-Xray and MIMIC-CXR datasets to investigate the contribution of each component in our proposed KERM. Table 3 presents the quantitative analysis of KERM on MIMIC-CXR across both NLG and CE metrics. And cmeasuring descriptive accuracy is reported in Table 2. Our base model is LLaVA-1.5-7b.

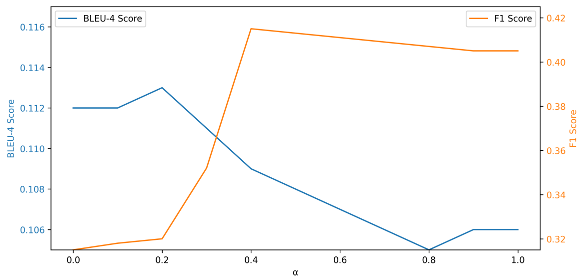


Figure 4: Analysis of the hyperparameter α with respect to F1 and BLEU-4 on MIMIC-CXR dataset.

Effect of The Components and Submodules It can be observed that adding MKE(Medical Knowledge Enhancement) and RM(Reward Modeling via Fine-Grained Feedback) on both the MIMIC-CXR and IU X-Ray datasets individually, in comparison to the baseline model, leads to significant improvements on all metrics. This observation indicates the effectiveness of both modules. MKE exhibits greater enhancement compared to RM. This might stem from the fact that the knowledge, obtained through retrieval, are more closely related to the current image. These knowledge contain additional detailed information, such as position and existence. Incorporating fine-grained rewards shows substantial growth, with the introduction of reward scores effectively mitigating the issue of hallucinations. This encourages the model to focus on avoiding inaccuracies and deviations.

Furthermore, comparing (c) and (d) in Table 3, it is observed that R_{dis} brings more improvement than R_{sen} on the NLG metrics, while the opposite is true on the CE metrics. We speculate the reason is that disease-level reward can more effectively improve the model to identify the existence of diseases and sentence-level reward promotes outputs that closely align with medical norms. Ultimately, the integration of such three improvements yields the best overall performance.

Ultimately, the integration of MKE and RM, as seen in the KERM model, yields the best overall performance on both datasets. This synergistic effect results in highly accurate and clinically relevant medical reports, reflecting the model's enhanced diagnostic capabilities and the reliability of

Settings	MKE	R_{dis}	R_{sen}	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	P	R	F1
Base	✗	✗	✗	0.337	0.203	0.132	0.098	0.131	0.273	0.296	0.163	0.153
(a)	✓	✗	✗	0.361	0.222	0.149	0.103	0.142	0.278	0.332	0.264	0.297
(b)	✗	✓	✓	0.352	0.216	0.144	0.101	0.135	0.275	0.322	0.253	0.282
(c)	✓	✗	✓	0.370	0.231	0.154	**0.112**	0.145	**0.285**	0.359	0.280	0.315
(d)	✓	✓	✗	0.368	0.223	0.145	0.106	0.141	0.279	0.363	0.282	0.317
KERM	✓	✓	✓	**0.378**	**0.235**	**0.157**	0.109	**0.152**	0.283	**0.394**	**0.436**	**0.415**

Table 3: Quantitative analysis of proposed method on MIMIC-CXR dataset. MKE, R_{dis} and R_{sen} represent Medical Knowledge Enhancement, disease-level and sentence-level feedback, respectively.

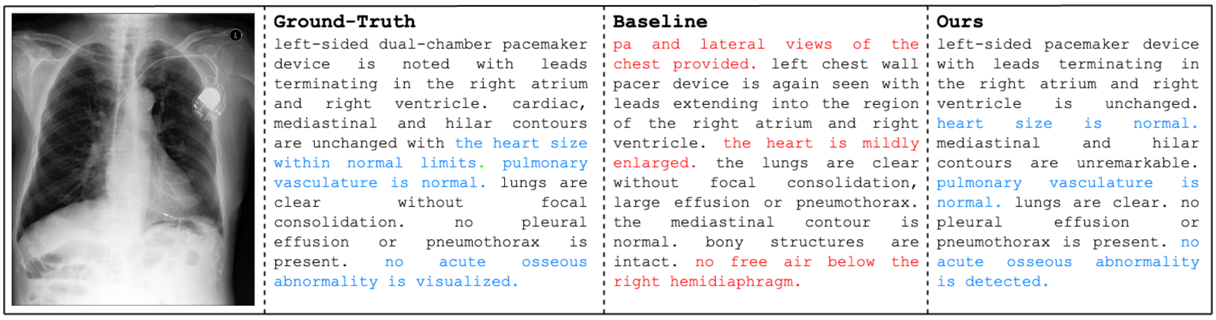


Figure 5: Illustrations of reports from ground truth, ours and Base. For better visualization, different colors highlight different medical terms. The terms marked in red are hallucinations, the terms marked in blue means descriptions included in Ground-Truth but not mentioned in the base model.

its generated radiology reports.

Hyperparameter Analysis We also conduct an ablation study on the hyperparameter α to investigate at which value can better enhance the model's performence of generating accurate and consistent report on MIMIC-CXR dataset. As is shown in Figure 4, α is analyzed with values ranging from 0 to 1 in terms of F1 and BLEU-4 scores. Overall, the performance remains stable across a wide range of α, as the fluctuations of F1 and BLEU-4 are within 10% and 1.2%, respectively. $\alpha = 0.4$ performs better in F1 and BLEU-4 scores, which is the value we used in the experiments.

4.4.3 Case Study

To further investigate the effectiveness of our method, we provide a qualitative comparison to the base model (LVLM) in Figure 5, where different colors on the texts indicate different medical terms(more cases can be seen in Appendix A.1). It is observed that our model generates descriptions that closely align with the ground-truth report in terms of content flow. Furthermore, as shown in Figure 5, we have found that KERM covers almost all of the necessary medical terms and abnormalities in the ground-truth reports, this comprehensive coverage is a significant improvement over the base model, which often misses crucial medical details. The performance of KERM proves that the reports generated from our model are comprehensive and accurate compared to the base model, effectively alleviating hallucinations.

5 Conclusions and Future Work

In this paper, we introduce KERM, a new framework designed to enhance the accuracy and reliability of radiology report generation from medical images. KERM addresses the critical challenge of hallucinations in the LVLM by retrieving fact knowledge from a comprehensive corpus and introducing a purification module to ensure contextual relevance, which enriches the model's understanding. This approach is complemented by fine-grained reward modeling, which penalizes both disease-level inaccuracies and sentence-level deviations from the expected medical findings. Our method's effectiveness is validated through extensive experiments, showcasing its potential to significantly improve the diagnostic process. In the future, we plan to develop more comprehensive evaluation metrics to better assess hallucinations in medical reports.

6 Limitations

While our KERM framework has demonstrated significant improvements in the accuracy and reliability of medical report generation, there are several limitations that warrant discussion. Firstly, the performance of KERM is inherently dependent on the quality and comprehensiveness of the knowledge corpus used for knowledge retrieval. Should the corpus lack certain medical facts or contain outdated information, it could potentially lead to omissions or inaccuracies in the generated reports.

Secondly, the Purification module, although designed to enhance the contextual relevance of the retrieved knowledge, may not always perfectly align with the specific nuances of each patient's clinical narrative. This could be due to the complexity of medical cases and the variability in how clinical history is documented.

Additionally, our framework's reliance on fine-grained rewards for guiding the generation process assumes that the reward model accurately reflects all aspects of clinical relevance and accuracy. However, the model's ability to capture the full spectrum of medical knowledge and the subtleties of medical language is subject to the training data and the design of the reward system.

Moreover, while our experiments on IU-Xray and MIMIC-CXR datasets have shown promising results, the external validity of our approach may be limited. The generalizability of KERM to other datasets or different medical domains requires further investigation, as the model's performance could vary with changes in data distribution or clinical presentation.

Lastly, the computational expense associated with training and deploying large vision language models like those used in KERM cannot be overlooked. The resource-intensive nature of our approach may pose challenges for implementation in settings with limited computational resources.

In future work, we aim to address these limitations by expanding the knowledge corpus, refining the Purification module, enhancing the reward modeling, and conducting additional experiments across diverse datasets to ensure broader applicability and robustness of our framework.

7 Ethics Considerations

The development and application of our KERM framework are grounded in a commitment to ethical standards, particularly concerning the handling of sensitive medical data. Our work strictly adheres to the deidentification protocols and usage policies associated with the IU X-Xray and MIMIC-CXR dataset, ensuring that all patient information remains confidential and is used solely for research purposes.

A critical aspect of our ethical considerations involves the responsible use of large language models (LLMs), such as the gpt-3.5-turbo model deployed on the Azure OpenAI platform. We acknowledge the financial implications of utilizing cloud-based services, recognizing that the cost per thousand tokens can create barriers to access and scalability, potentially limiting the equitable use of advanced AI in medical applications.

Moreover, we are vigilant about the risks associated with LLMs, including the potential for "hallucinations"— the generation of false or misleading information. In the context of medical report generation, where accuracy is paramount, we have implemented strategies to minimize these risks. Our approach prompts the LLM to rephrase existing medical content into coherent and stylistically consistent prose, rather than creating new medical content. This method is designed to leverage the strengths of LLMs in language generation while reducing the likelihood of introducing inaccuracies.

In conclusion, our ethical considerations are integral to the design and implementation of the KERM framework. We remain dedicated to the responsible use of AI in medicine, prioritizing accuracy, patient confidentiality, and the avoidance of misinformation in medical report generation.

References

Imane Allaouzi, M. Ben Ahmed, B. Benamrou, and M. Ouardouz. 2018. Automatic caption generation for medical images. In *Proceedings of the 3rd International Conference on Smart City Applications*, SCA '18, New York, NY, USA. Association for Computing Machinery.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2022. Cross-modal memory networks for radiology report generation. *ArXiv*, abs/2204.13258.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. *ArXiv*, abs/2010.16056.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza M. Rodriguez, Sameer Kiran 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 : JAMIA*, 23 2:304–10.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Jeremy A. Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David Andrew Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, C. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and A. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*.

Baoyu Jing, Zeya Wang, and Eric P. Xing. 2019. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In *Annual Meeting of the Association for Computational Linguistics*.

Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2017. On the automatic generation of medical imaging reports. In *Annual Meeting of the Association for Computational Linguistics*.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. *ArXiv*, abs/2311.07362.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023b. Dynamic graph enhanced contrastive learning for chest x-ray report generation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3334–3343.

Mingjie Li, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. 2020. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, 26:253 – 270.

Xiangyang Li, Zihan Wang, Jiahao Yang, Yaowei Wang, and Shuqiang Jiang. 2023c. Kerm: Knowledge enhanced reasoning for vision-and-language navigation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2583–2592.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. *ArXiv*, abs/1805.08298.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13748–13757.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew B. A. McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. *ArXiv*, abs/1904.02633.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *ArXiv*, abs/2304.08485.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Yasuhide Miura, Yuhao Zhang, C. Langlotz, and Dan Jurafsky. 2020. Improving factual completeness and consistency of image-to-text radiology report generation. *ArXiv*, abs/2010.10042.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *ArXiv*, abs/2304.03277.

Hoo-Chang Shin, Kirk Roberts, Le Lu, Dina Demner-Fushman, Jianhua Yao, and Ronald M. Summers. 2016. Learning to read chest x-rays: Recurrent neural cascade model for automated image annotation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2497–2506.

Richard S. Sutton, David A. McAllester, Satinder Singh, and Y. Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Neural Information Processing Systems*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David J. Fleet, P. A. Mansfield, Sushant Prakash, Renee C Wong, Sunny Virmani, Christopher Semturs, Seyedeh Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Joëlle K. Barral, Dale R. Webster, Greg S Corrado, Yossi Matias, K. Singhal, Peter R. Florence, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards generalist biomedical ai. *ArXiv*, abs/2307.14334.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and D. Erhan. 2014. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Song Wang, Liyan Tang, Mingquan Lin, George L. Shih, Ying Ding, and Yifan Peng. 2022a. Prior knowledge enhances radiology report generation. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2022:486–495.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9049–9058.

Zhanyu Wang, Hongwei Han, Lei Wang, Xiu Li, and Luping Zhou. 2022b. Automated radiographic report generation purely on transformer: A multicriteria supervised approach. *IEEE Transactions on Medical Imaging*, 41:2803–2813.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022c. Medclip: Contrastive learning from unpaired medical images and text. In *Conference on Empirical Methods in Natural Language Processing*.

Shuxin Yang, Xian Wu, Shen Ge, S.kevin Zhou, and Li Xiao. 2021. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *ArXiv*, abs/2304.10592.

A Appendix

A.1 More Cases.

More cases can seen in Figure 6.

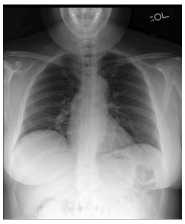
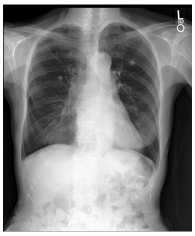
Input Images	Ground-Truth	Base	Ours
	the lungs are clear. the cardiomediastinal silhouette is within normal limits. no acute osseous abnormalities. no free intraperitoneal air identified.	the lung volumes are low. the heart is normal in size. the cardiomediastinal silhouette is normal. imaged osseous structures are intact.	the lungs are clear without focal consolidation, effusion, or edema. the cardiomediastinal silhouette is within normal limits. no acute osseous abnormalities. no free air below the right hemidiaphragm is seen.
	cardiac silhouette size is normal. the aortic knob is calcified. the mediastinal and hilar contours are within normal limits. pulmonary vasculature is not engorged. lungs are clear without focal consolidation. there is continued blunting the right costophrenic angle, likely pleural thickening, unchanged. no pleural effusion or pneumothorax is present. anterior wedge compression deformity of l<num> vertebral body is unchanged.	the lungs are clear without focal consolidation. no pleural effusion or pneumothorax is seen. the cardiac and mediastinal silhouettes are unremarkable.	heart size is normal. the mediastinal and hilar contours are normal. the pulmonary vasculature is normal. lungs are clear. no pleural effusion or pneumothorax is seen. there are no acute osseous abnormalities.
	semi upright ap and lateral views of the chest provided. a right ij access dialysis catheter is seen with its tip extending to the low svc. the heart is mildly enlarged. patient's leftward rotation limits evaluation. there is no focal consolidation, effusion or pneumothorax. vascular calcification is noted along the descending thoracic aorta. bony structures are intact appear	the heart is mildly enlarged. the mediastinal and hilar contours appear unchanged. there is a persistent moderate-sized pleural effusion on the right, which is probably similar in size, allowing for differences in technique. a small pleural effusion is suspected on the left. there is no pneumothorax. the lungs appear clear.	ap upright and lateral views of the chest provided. a right ij access dialysis catheter is again seen with its tip in the region of the cavoatrial junction. the heart is mildly enlarged. the lungs are clear without focal consolidation, large effusion or pneumothorax. mediastinal contour is stable. bony structures are intact.

Figure 6: Qualitative examples of ground truth, ours and Base. Blue font indicates consistent content with the ground-truth while red font indicates hallucinations.