Beyond Surface Features: Advancing Medical Vision-Language Alignment via Dynamic Evidence-Guided Preference Optimization

Zixuan Huang[2,3]*, Zhihong Zhu[4]*, Xiaolong Liu[5], Yanchao Hao[5], Manman Zhang[5], Zheng Wei[5], Bowen Xing[6], Xian Wu[5], Ye Li[2,3], Fen Miao[7]†, Yefeng Zheng[1]†

[1]Medical Artificial Intelligence Lab, Westlake University [2]University of Chinese Academy of Sciences
[3]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
[4]Peking University [5]Tencent [6]University of Science and Technology Beijing
[7]Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China
zhengyefeng@westlake.edu.cn fenmiao@uestc.edu.cn

Abstract

Medical large Vision-Language Models (Med-LVLMs) have shown strong potential in multimodal clinical applications such as medical visual question answering and report generation. However, Med-LVLMs remain challenged by hallucinations caused by modality misalignment, where models prioritize textual knowledge over visual evidence and generate outputs that conflict with medical images. To mitigate this issue, recent studies have explored preference optimization to improve image–text alignment, achieving promising results. Despite these advances, existing preference-based methods still face two limitations in medical settings: (1) overfitting to superficial cues, and (2) pseudo convergence of the preference signal. In this paper, we propose Dynamic Evidence-Guided Preference Optimization (DEPO), a new framework that enables evidence-aware and adaptive preference learning for Med-LVLMs. DEPO introduces Multi-Modal Evidence Perturbation (MEP) to suppress non-causal textual and visual shortcuts, and Dispreferred Evidence Resampling (DER) to continuously update dispreferred responses as hallucination patterns evolve. Experiments on multiple medical VQA and report generation benchmarks demonstrate consistent improvements over existing methods, with strong robustness across datasets and architectures.

1 Introduction

Building upon the success of Large Language Models (LLMs) (Vaswani et al., 2017; Devlin et al., 2019; Brown et al., 2020; Achiam et al., 2023; Guo et al., 2025; Qiu et al., 2025), Large Vision-Language Models (LVLMs) extend language understanding to the visual modality (Radford et al., 2021; Dosovitskiy, 2020; Alayrac et al., 2022; Liu et al., 2023; Zhang et al., 2024a) and have demonstrated promising potential in medical tasks such as report generation (Zhou et al., 2022; Jin et al., 2024; Messina et al., 2024) and visual question answering (VQA) (Shaaban et al., 2025; Yan et al., 2025; Hu et al., 2024). However, Med-LVLMs remain challenged by hallucinations, where models tend to generate descriptions that appear superficially fluent but are factually incorrect or ungrounded in the visual imagery (Li et al., 2023; Tian et al., 2023). In such high-stakes medical environments, these errors severely compromise the reliability and widespread adoption of LVLM-powered systems.

To tackle this issue, recent studies have explored preference optimization for improving alignment between medical images and textual outputs (Yuan et al., 2024; Zhu et al., 2024; Sun et al., 2024). Self-Rewarding (Yuan et al., 2024) iteratively constructs preference pairs, while STLLaVA-Med (Sun et al., 2024) further employs GPT-based filtering to refine preference selection for Med-LVLM fine-tuning. MMedPO (Zhu et al., 2024) explicitly incorporates clinical relevance into preference construction and applies Direct Preference Optimization (DPO) (Rafailov et al., 2023), achieving promising results.

Despite these advances, existing preference-based methods still face two challenges, as illustrated in Figure 1: ***(1) Overfitting to superficial cues***. Because preference optimization relies on static preference data, models can be driven to attend to shallow, non-causal patterns. Prior studies have shown that preference-based training may assign high preference to tokens that lack causal relevance (Zhang et al., 2025; Liu et al., 2024; Huang et al., 2024; Li et al., 2025), such as templated report wording or habitual phrasing. Consequently, the model tends to respond based on linguistic statistics rather than clinically grounded visual evidence. ***(2) Pseudo convergence of the reward signal***. More critically, static preference datasets cannot adapt to the evolving distribution of halluci-

*These authors contributed equally to this work. Z. Huang contributed to this work as a visiting student at Westlake University.
†Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26125–26137
July 2-7, 2026

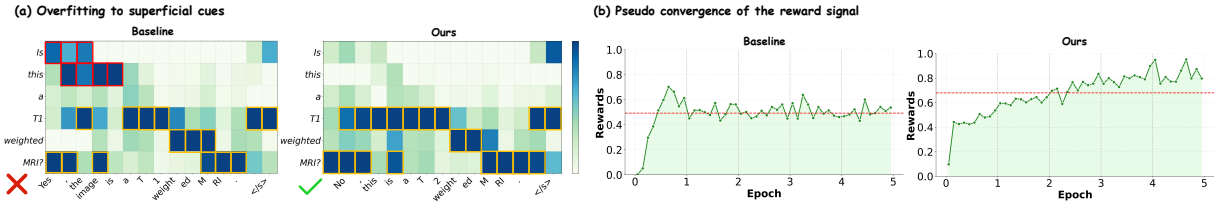

Figure 1: Motivation illustration. (a) An example of ***overfitting to superficial textual cues*** in preference-based optimization, where the baseline (Li et al., 2023) assigns high preference to spurious linguistic patterns instead of clinically relevant visual evidence, leading to incorrect predictions. (b) ***Pseudo convergence of the reward signal*** under static preference supervision (Rafailov et al., 2023), where the reward appears to stabilize early despite the persistence of hallucinations, indicating a lack of sustained and informative training signals.

nations during training. Although the model may reduce the likelihood of previously dispreferred responses, it can still generate new hallucinations that remain unpenalized. As a result, the reward signal appears to converge, while fixed dispreference feedback fails to provide sustained and informative gradients, preventing the model from learning core visual clinical cues.

In this paper, we introduce a new Dynamic Evidence-guided Preference Optimization (DEPO) framework for Med-LVLMs. ***To address the first challenge***, we introduce Multi-Modal Evidence Perturbation (MEP) that jointly regularizes textual and visual evidence. On the textual side, we apply sentence-level rephrasing and adaptive token-level perturbations to weaken reliance on templated phrasing and non-causal lexical patterns. On the visual side, we adopt curriculum-based visual perturbations that progressively mask background shortcuts and distort diagnostic regions, forcing the model to ground predictions in robust medical image evidence. ***To address the second challenge***, we propose Dispreferred Evidence Resampling (DER), which updates dispreference signals after each training epoch to track newly emerging hallucination patterns. Instead of optimizing against static dispreferred responses, DER continuously replaces outdated dispreferred answers with harder hallucinated samples, ensuring that preference signals remain informative throughout training.

Our contributions can be summarized as follows: (1) We propose Dynamic Evidence-guided Preference Optimization (DEPO), a new multi-modal preference optimization framework for Med-LVLMs. (2) We propose two corresponding mechanisms in DEPO: Multi-Modal Evidence Perturbation (MEP) to address overfitting to superficial cues, and Dispreferred Evidence Resampling (DER) to address pseudo convergence of the reward signal. (3) Experiments including quantitative and qualitative evaluations as well as cross-dataset and cross-architecture studies, demonstrate the effectiveness and generalizability of the proposed framework.

2 Methodology

In this section, we present Dynamic Evidence-guided Preference Optimization (DEPO), a new multi-modal preference optimization framework designed to mitigate hallucinations in Med-LVLMs. DEPO integrates multi-modal evidence perturbation and dispreferred evidence resampling within a unified DPO formulation, as shown in Figure 2.

2.1 Preliminaries

In Med-LVLMs, aligning generated clinical text with visual evidence is critical to mitigate hallucinations (Zhu et al., 2024). Preference optimization offers an alignment framework by learning from pairwise response supervision (Bai et al., 2022).

Formally, given a medical image x_i and a textual query x_t, a Med-LVLM parameterized by θ defines a conditional distribution $\pi_\theta(y \mid x)$ over textual responses y, where $x = (x_i, x_t)$. Preference supervision is constructed as pairs (y_w, y_l) for the same input, yielding a dataset $\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^{N}$. Direct Preference Optimization (DPO) (Rafailov et al., 2023) aligns the model by encouraging higher likelihood for preferred responses relative to a fixed reference policy π_{ref} obtained via supervised fine-tuning, which can be defined as follows:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x_t,x_i,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\alpha \log \frac{\pi_\theta(y_w|x_t,x_i)}{\pi_{\text{ref}}(y_w|x_t,x_i)} - \alpha \log \frac{\pi_\theta(y_l|x_t,x_i)}{\pi_{\text{ref}}(y_l|x_t,x_i)}\right)\right], \tag{1}$$

where α is a temperature parameter that controls the strength of the preference signal.

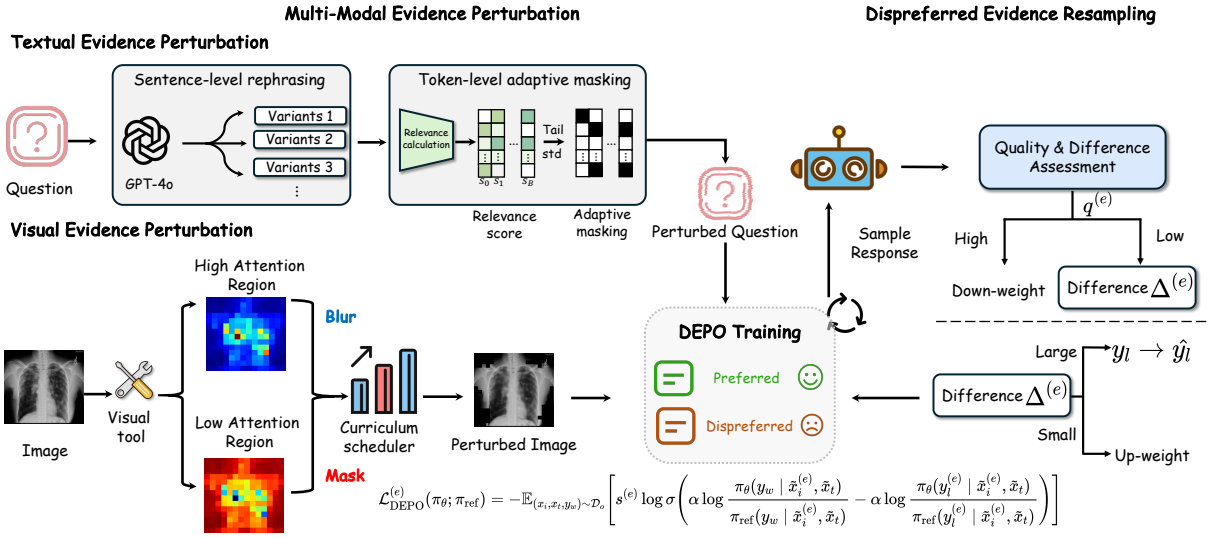

Figure 2: Overall framework of DEPO. DEPO comprises (1) Multi-Modal Evidence Perturbation (MEP) and (2) Dispreferred Evidence Resampling (DER), which jointly strengthen medical evidence grounding.

2.2 Multi-Modal Evidence Perturbation

To address overfitting to superficial cues induced by static preference data, we introduce Multi-Modal Evidence Perturbation (MEP). MEP perturbs both the textual query and the medical image during preference optimization, thereby preventing the model from relying on fixed surface patterns or non-causal correlations and encouraging preference learning to be grounded in clinically meaningful visual evidence. Concretely, during training we replace (x_t, x_i) in Eq. (1) with perturbed inputs $(\tilde{x}_t, \tilde{x}_i)$. In the following, we describe the construction of textual and visual perturbations.

Textual Evidence Perturbation. In preference optimization, the textual query paired with each medical image is typically fixed across training iterations. When the model repeatedly observes the same wording, it can latch onto superficial lexical patterns and reduce its reliance on visual evidence, which aggravates modality misalignment and leads to fluent yet visually ungrounded outputs (Zhang et al., 2025). To break this shortcut, MEP performs a sentence-level semantic perturbation by sampling alternative formulations of the same clinical intent. Specifically, we construct a semantic-preserving rephrasing set:

$$\mathcal{Q}(x_t) = \mathrm{LLM}(x_t; K_q) = \{\bar{x}_t^{(1)}, \ldots, \bar{x}_t^{(K_q)}\}, \tag{2}$$

and randomly sample $\bar{x}_t \sim \mathcal{Q}(x_t)$ as the training query, so that the preference signal is observed under diverse yet equivalent expressions.

Sentence variation alone is insufficient because weakly grounded tokens can still act as spurious cues. Let $\bar{x}_t = \{q_j\}_{j=1}^{L}$ be the query, $F_v(x_i) = \{v_r\}_{r=1}^{M}$ denote visual token features and $F_t(q_j) = t_j$ denote the feature of token q_j. We assign each token q_j an evidence-grounding score g_j that measures how strongly it is supported by the image:

$$g_j = \max_{r \in \{1,\ldots,M\}} \mathrm{Cosine}(v_r, t_j). \tag{3}$$

A larger g_j indicates that token q_j is better supported by some image evidence. We perturb the N_t least grounded tokens by selecting

$$\mathcal{I}_t = \underset{\mathcal{I} \subseteq \{1,\ldots,L\}, |\mathcal{I}| = N_t}{\arg\min} \sum_{j \in \mathcal{I}} g_j. \tag{4}$$

For $j \in \mathcal{I}_t$, we replace q_j with a mask token and obtain the perturbed query $\tilde{x}_t$.

To accommodate varying grounding strength across queries, we adapt the perturbation budget using the uncertainty of low-evidence tokens. Mathematically, let $g_{(1)} \le \cdots \le g_{(L)}$ be the sorted scores and consider the lowest-K tail. The tail dispersion Δg and the budget N_t can be calculated as follows:

$$\Delta g = \mathrm{Std}\left(\{g_{(1)}, \ldots, g_{(K)}\}\right),$$
$$N_t = \mathrm{clip}\left(\left\lfloor \omega \cdot \frac{1}{\Delta g + \epsilon} \right\rfloor + 1,\ N_{\min},\ N_{\max}\right). \tag{5}$$

where $\mathrm{clip}(\cdot; N_{\min}, N_{\max})$ bounds the masking budget to a reasonable range for stability, and ϵ is a small constant to avoid numerical issues.

In this way, perturbations are strengthened under uncertain grounding, which encourages preference learning to rely on visual evidence.

Visual Evidence Perturbation. To suppress the shortcuts caused by medical images, MEP perturbs the image according to a query-conditioned evidence map that highlights regions relevant to the clinical question. We obtain a relevance map $\mathbf{A} \in [0,1]^{H\times W}$ from a frozen medical localizer (e.g., MedKLIP (Wu et al., 2023)) as follows:

$$\mathbf{A} = \mathrm{MedKLIP}(x_i, \bar{x}_t). \tag{6}$$

Larger A_p indicates higher relevance; using $\mathbf{A}$, we split the image into high/low-evidence regions at epoch e using thresholds $\tau_{\text{high}}^{(e)}$ and $\tau_{\text{low}}^{(e)}$ as:

$$\mathcal{R}_{\text{high}}^{(e)} = \{p \mid A_p \geq \tau_{\text{high}}^{(e)}\}, \quad \mathcal{R}_{\text{low}}^{(e)} = \{p \mid A_p \leq \tau_{\text{low}}^{(e)}\}. \tag{7}$$

To prevent the model from adapting to a fixed perturbation pattern, we progressively increase the perturbation difficulty over training by linearly scheduling the thresholds:

$$\begin{aligned} \tau_{\text{high}}^{(e)} &= \tau_{\text{high}}^{(0)} + \frac{e}{E}\big(\tau_{\text{high}}^{(E)} - \tau_{\text{high}}^{(0)}\big), \\ \tau_{\text{low}}^{(e)} &= \tau_{\text{low}}^{(0)} + \frac{e}{E}\big(\tau_{\text{low}}^{(E)} - \tau_{\text{low}}^{(0)}\big). \end{aligned} \tag{8}$$

The perturbed image $\tilde{x}_i^{(e)}$ is then defined as:

$$\tilde{x}_i^{(e)}(p) = \begin{cases} \mathrm{Blur}(x_i)(p), & p \in \mathcal{R}_{\text{high}}^{(e)}, \\ \mathrm{Mask}(x_i)(p), & p \in \mathcal{R}_{\text{low}}^{(e)}, \\ x_i(p), & \text{otherwise.} \end{cases} \tag{9}$$

2.3 Dispreferred Evidence Resampling

To maintain informative dispreference feedback throughout training, we further propose Dispreferred Evidence Resampling (DER). DER dynamically refreshes dispreferred answers and adjusts sample weights as hallucination patterns evolve (Li et al., 2025).

Concretely, for each training example, DER maintains a dispreferred answer $y_l^{(e)}$ and a sample weight $s^{(e)}$ at epoch e. After each epoch, DER draws a new candidate dispreferred answer $\hat{y}_l^{(e)}$ from the current policy under the perturbed input $(\tilde{x}_t, \tilde{x}_i^{(e)})$. Its evidence-grounded quality is evaluated by a frozen verifier $\mathcal{V}(\tilde{x}_i^{(e)}, \tilde{x}_t, y) \in [0,1]$, which can be formulated as:

$$q^{(e)} = \mathcal{V}\Big(\tilde{x}_i^{(e)}, \tilde{x}_t, \hat{y}_l^{(e)}\Big). \tag{10}$$

A high $q^{(e)}$ indicates that the current pair is already well separated under the verifier, so continuing to emphasize this example may lead to overfitting; accordingly, DER down-weights it. Conversely, when $q^{(e)}$ is low, $\hat{y}_l^{(e)}$ is likely a hallucinated response and can serve as a hard negative.

To distinguish whether $\hat{y}_l^{(e)}$ corresponds to a genuinely new hallucination mode or a previously observed one, DER measures its deviation from the previous dispreferred answer:

$$\Delta^{(e)} = \Delta\Big(y_l^{(e)}, \hat{y}_l^{(e)}\Big), \tag{11}$$

where $\Delta(\cdot,\cdot)$ can be instantiated as edit distance or a sentence-level semantic distance. DER then updates $(y_l^{(e)}, s^{(e)})$ as:

$$(y_l^{(e+1)}, s^{(e+1)}) = \begin{cases} (y_l^{(e)},\ s^{(e)}\downarrow), & q^{(e)} \geq \tau_q, \\ (\hat{y}_l^{(e)},\ s^{(e)}), & q^{(e)} < \tau_q \ \wedge\ \Delta^{(e)} > \tau_\Delta, \\ (y_l^{(e)},\ s^{(e)}\uparrow), & q^{(e)} < \tau_q \ \wedge\ \Delta^{(e)} \leq \tau_\Delta, \end{cases} \tag{12}$$

where τ_q controls the verifier strictness and τ_Δ controls the novelty threshold.

2.4 Training Objective

Following the standard DPO formulation, DEPO integrates MEP and DER within a unified objective. Specifically, MEP perturbs the input pair (x_t, x_i) into $(\tilde{x}_t, \tilde{x}_i^{(e)})$, while DER dynamically updates the dispreferred answer $y_l^{(e)}$ and assigns a sample weight $s^{(e)}$ at epoch e. The final objective is:

$$\mathcal{L}_{\text{DEPO}}^{(e)}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x_i,x_t,y_w)\sim\mathcal{D}_o}\Bigg[s^{(e)} \log \sigma\Bigg(\alpha \log \frac{\pi_\theta(y_w \mid \tilde{x}_i^{(e)}, \tilde{x}_t)}{\pi_{\text{ref}}(y_w \mid \tilde{x}_i^{(e)}, \tilde{x}_t)} - \alpha \log \frac{\pi_\theta(y_l^{(e)} \mid \tilde{x}_i^{(e)}, \tilde{x}_t)}{\pi_{\text{ref}}(y_l^{(e)} \mid \tilde{x}_i^{(e)}, \tilde{x}_t)}\Bigg)\Bigg]. \tag{13}$$

where $\mathcal{D}_o$ denotes the offline preference dataset. During training, DEPO maintains an epoch-dependent dispreferred response $y_l^{(e)}$ and weight $s^{(e)}$ for each input pair.

3 Experiment

3.1 Experimental Settings

Datasets. Building on previous works (Lin et al., 2025; Chang et al., 2025), we evaluated our method using two open medical VQA datasets and two report generation datasets. (1) VQA-RAD (Lau et al., 2018) is a radiology VQA dataset featuring diverse, clinician-generated questions across 11 classes. (2) SLAKE (Liu et al., 2021) is a bilingual radiology dataset with questions spanning various

Methods	Medical VQA				Report Generation					
	SLAKE		VQA-RAD		MIMIC-CXR			IU-Xray		
	Open	Closed	Open	Closed	BLEU	ROUGE-L	METEOR	BLEU	ROUGE-L	METEOR
LLaVA-Med v1.5	44.26	61.30	29.24	63.97	10.25	9.38	7.71	14.56	10.31	10.95
+Self-Rewarding	42.63	61.30	33.29	64.17	10.78	9.27	7.73	14.20	10.38	10.52
+DPO	49.30	62.02	29.76	64.70	11.19	9.45	7.80	16.08	12.95	17.13
+POVID	52.43	70.35	31.77	65.07	11.21	9.66	7.84	20.80	24.33	30.05
+SIMA	51.77	69.10	31.23	64.80	11.16	9.58	7.49	17.11	22.87	29.10
+FiSAO	52.69	70.46	32.70	64.11	11.32	9.68	7.62	21.06	25.72	30.82
+STLLaVA-Med	48.65	61.75	30.17	64.38	11.11	9.29	7.72	16.11	10.58	10.51
+MMedPO	53.99	73.08	36.36	66.54	12.85	11.13	10.03	23.49	29.52	34.16
+DEPO (Ours)	**66.68**†	**78.12**†	**38.21**†	**69.12**†	**13.05**†	**15.94**†	**14.49**†	**24.05**†	**31.25**†	**35.49**†
+SFT	50.45	65.62	31.38	64.26	12.39	10.21	8.75	22.75	28.86	33.66
+Self-Rewarding	50.62	65.89	32.69	65.89	12.15	10.05	8.77	22.89	28.97	33.93
+DPO	53.50	69.47	32.88	64.33	12.37	10.38	9.10	23.07	29.97	34.89
+POVID	52.18	70.67	32.95	64.97	11.85	10.45	9.05	23.95	29.75	34.63
+SIMA	51.75	69.28	32.50	64.08	12.44	10.25	9.02	23.90	29.41	34.45
+FiSAO	52.80	70.82	32.94	65.77	12.97	10.69	9.39	23.57	29.88	35.01
+STLLaVA-Med	52.72	66.69	33.72	64.70	12.21	10.12	8.98	22.79	28.98	34.05
+MMedPO	55.23	75.24	34.03	67.64	13.28	13.22	10.20	24.00	30.13	35.17
+DEPO (Ours)	**68.24**†	**80.53**†	**39.20**†	**70.96**†	**13.33**	**16.10**†	**14.42**†	**24.15**†	**31.33**†	**35.59**†

Table 1: Result comparison on medical VQA and report generation tasks. Following Li et al. (2023), we report accuracy for closed and Recall for open questions. "+SFT" indicates that the model is first fine-tuned via supervised fine-tuning (SFT) before applying the corresponding baseline. The best results are highlighted in **bold**. "†" denotes the improvements over the best baseline are statistically significant with $p < 0.01$ under t-test.

human body parts. For a fair comparison, we only evaluated on its English subset. (3) MIMIC-CXR (Johnson et al., 2019) is a large public dataset of chest radiographs in DICOM format with free-text radiology reports. (4) IU-XRay (Demner-Fushman et al., 2015) is a dataset that contains chest X-ray images with diagnostic reports. The dataset preprocessing, preference construction, and sample weights initialization follow Zhu et al. (2024).

Evaluation Metrics. For VQA, we evaluate performance using accuracy and recall for closed-ended and open-ended questions. For report generation, we adopt BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), following Zhu et al. (2024).

Baselines. Following previous works, we compare our method with DPO (Rafailov et al., 2023) and representative variants. Self-Rewarding (Yuan et al., 2024) constructs preference pairs by itself, whereas STLLaVA-Med (Sun et al., 2024) leverages GPT to further filter and refine preference selection for Med-LVLM fine-tuning. MMedPO (Zhu et al., 2024) considers the clinical relevance of preference samples to enhance base model. In addition, we include three preference-based VLM fine-tuning approaches developed for natural images, namely POVID (Zhou et al., 2024), FiSAO (Cui et al., 2024), and SIMA (Wang et al., 2025b).

Implementation Details. We adopt LLaVA-Med v1.5 7B (Li et al., 2023) as our base model. Preference optimization is performed via LoRA (Hu et al., 2022) with batch size 4, learning rate 5×10^{-7}, for 5 epochs. To curate preference pairs, GPT-4o is used to generate rephrasing. We also utilize attention maps of MedKLIP (Wu et al., 2023) as prior knowledge to guide textual and visual tokens perturbation. All results are obtained by averaging the scores over five runs with different random seeds.

3.2 Main Results

As shown in Table 1, we compare the proposed DEPO with the original LLaVA-Med v1.5 and a suite of preference-optimization baselines on both medical VQA and report generation. To provide a complete picture, we report results under two training regimes: (i) applying each method *directly* on LLaVA-Med v1.5, and (ii) applying each method on an SFT-initialized LLaVA-Med v1.5.

For Medical VQA, DEPO achieves the best performance in both regimes. Without SFT, DEPO surpasses the strongest baseline (+MMedPO) by an average of 14.30% on open questions (Recall; aver-

Methods	Medical VQA				Report Generation					
	SLAKE		VQA-RAD		MIMIC-CXR			IU-Xray		
	Open	Closed	Open	Closed	BLEU	ROUGE-L	METEOR	BLEU	ROUGE-L	METEOR
w/o TEP	68.01 (↓0.23)	78.37 (↓2.16)	36.00 (↓3.20)	69.12 (↓1.84)	12.67 (↓0.66)	14.36 (↓1.74)	13.00 (↓1.42)	23.29 (↓0.86)	29.31 (↓2.02)	35.01 (↓0.58)
w/o Sentence-level Semantic Perturbation	68.05 (↓0.19)	78.35 (↓2.18)	38.21 (↓0.99)	69.53 (↓1.43)	13.31 (↓0.02)	14.66 (↓1.44)	13.06 (↓1.36)	23.48 (↓0.67)	29.09 (↓2.24)	35.26 (↓0.33)
w/o Adaptive Token-level Perturbation	67.96 (↓0.28)	79.23 (↓1.30)	37.63 (↓1.57)	69.50 (↓1.46)	12.92 (↓0.41)	15.04 (↓1.06)	13.08 (↓1.34)	23.48 (↓0.67)	29.61 (↓1.72)	35.02 (↓0.57)
w/o VEP	66.60 (↓1.64)	79.57 (↓0.96)	38.41 (↓0.79)	68.38 (↓2.58)	13.24 (↓0.09)	12.27 (↓3.83)	11.91 (↓2.51)	23.63 (↓0.52)	30.33 (↓1.00)	35.25 (↓0.34)
w/o DER	66.56 (↓1.68)	78.61 (↓1.92)	38.16 (↓1.04)	66.54 (↓4.42)	12.40 (↓0.93)	15.08 (↓1.02)	13.97 (↓0.45)	23.35 (↓0.80)	29.63 (↓1.70)	34.89 (↓0.70)
Remove All	53.50 (↓14.74)	69.47 (↓11.06)	32.88 (↓6.32)	64.33 (↓6.63)	12.37 (↓0.96)	10.38 (↓5.72)	9.10 (↓5.32)	23.07 (↓1.08)	29.97 (↓1.36)	34.89 (↓0.70)
Full model	**68.24**	**80.53**	**39.20**	**70.96**	**13.33**	**16.10**	**14.42**	**24.15**	**31.33**	**35.59**

Table 2: Ablation study under the SFT setting. ↓ are reported as Full − Ablation, and 'w/o' is short for 'without'.

aged over SLAKE and VQA-RAD) and 5.39% on closed questions (Accuracy). With SFT initialization, DEPO further improves, outperforming the best +SFT baseline by 19.37% (open) and 5.97% (closed). Notably, the larger gains on open questions suggest that DEPO more effectively mitigates shortcut learning via evidence-guided perturbations and dynamic resampling. ***For report generation***, DEPO also yields consistent improvements over the strongest baseline. In the direct setting, it achieves average gains of 1.97% in BLEU, 24.54% in ROUGE-L, and 24.18% in METEOR across MIMIC-CXR and IU-Xray; with SFT initialization, it still improves upon the best +SFT baseline by 0.50% (BLEU), 12.88% (ROUGE-L), and 21.28% (METEOR) on average. These gains indicate that our preference optimization better aligns generation with clinically grounded descriptions.

3.3 Quantitative Analysis

3.3.1 Ablation Study

We perform ablation studies of the key components of the proposed DEPO in Table 2. From the results, we can observe that: ***(1) Removing Textual Evidence Perturbation (TEP)*** yields consistent degradation on both VQA and report generation. The largest drops occur on VQA-RAD Open and MIMIC-CXR report generation, suggesting that the model becomes more sensitive to query phrasing and relies more heavily on linguistic shortcuts rather than stable visual evidence. ***(2) Ablating Visual Evidence Perturbation (VEP)*** also causes a clear performance decline, most notably on MIMIC-CXR ROUGE-L and METEOR. VQA accuracy similarly drops (e.g., SLAKE Open 68.24→66.60). This indicates increased susceptibility to spurious visual cues and weaker visual grounding, which is particularly detrimental for long-form report generation. ***(3) Disabling Dispreferred Evidence Resampling (DER)*** results in consistent performance drops across all datasets and tasks. This supports our motivation that dynamically refreshing dispreferred answers and reweighting samples is crucial for maintaining informative preference signals as the policy evolves.

Complementary Roles of Sentence- and Token-level Augmentation. To further examine how TEP reduces textual shortcuts, we ablate its two components in Table 2. The results indicate that the two perturbation levels play complementary roles. Sentence-level semantic perturbation helps reduce overfitting to fixed question templates and recurring phrasing by introducing semantically equivalent clinical queries with varied wording, while adaptive token-level perturbation weakens lexical shortcuts by masking tokens with limited visual support. Empirically, token-level perturbation has a larger impact on open-ended VQA, whereas sentence-level perturbation yields comparable or stronger gains on several closed-ended and report-generation metrics. These results suggest that the two mechanisms address textual bias at different levels and together improve evidence-grounded reasoning.

Hyper-parameter Analysis. We further conduct a sensitivity analysis of DER threshold τ_q in both tasks. From Figure 3, we observe a consistent 'inverted-U' trend: when τ_q is too small, a large fraction of samples are judged as unreliable, leading to under-training and degraded performance; when τ_q is too large, many ambiguous samples are retained with insufficient suppression, weakening DER's fil-

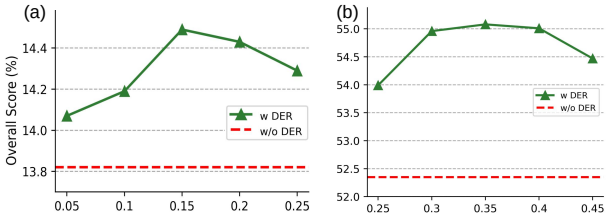

Figure 3: Effect of DER threshold τ_q on VQA (VQA-RAD) (a) and report generation (MIMIC-CXR) (b).

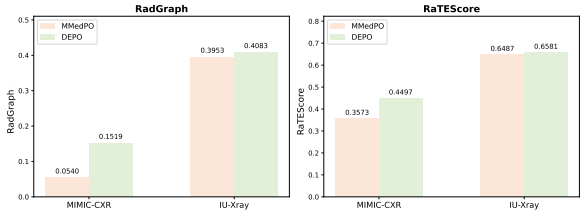

Figure 4: Comparison of RadGraph and RaTEScore Results on MIMIC-CXR and IU-Xray.

tering effect. In the two boundary cases, the sample weights tend to shift almost globally rather than being adaptively redistributed, which makes DER's expected update mechanism nearly ineffective; consequently, the performance becomes closer to the setting w/o DER. In between, a moderate τ_q yields the best overall results by balancing noise suppression and effective data utilization.

3.3.2 Clinical Hallucination Evaluation

To further evaluate hallucinations in report generation, we additionally assess DEPO using two clinically oriented metrics, RadGraph and RaTEScore. Unlike general text-overlap metrics, these metrics are more sensitive to whether a generated report remains clinically faithful and factually grounded.

DEPO shows a consistent advantage on both datasets, with particularly large gains on MIMIC-CXR. Compared with MMedPO, DEPO improves RadGraph and RaTEScore by 181.30% and 25.86%, respectively, on MIMIC-CXR, and also achieves consistent improvements of 3.29% and 1.45% on IU-Xray. These results indicate that the improvement of DEPO is not limited to surface-level text quality, but extends to the generation of more clinically faithful content. Overall, DEPO more effectively suppresses hallucinations and generates reports that are more factually consistent with medical evidence.

3.3.3 Compatibility Analysis

A natural question is whether the proposed method is compatible with different medical tasks under a unified LVLM backbone. To this end, we conduct compatibility experiments using LLaVA-Med++ (Xie et al., 2025), which is pretrained on MedTrinity-25M as a shared backbone for medical multimodal understanding. From the results, we observe consistent performance improvements across the two tasks after applying our method. This indicates that the proposed approach does not rely on task-specific architectural modifications, but can be seamlessly integrated into a single medical LVLM to support heterogeneous clinical reasoning and generation objectives. Overall, these results demonstrate the compatibility of our method within other unified Medical LVLM frameworks.

Method	SLAKE	VQA-RAD	IU-Xray	MIMIC-CXR
w/o DEPO	49.87	54.38	24.88	8.08
w DEPO	**52.78**	**58.88**	**28.82**	**12.91**

Table 3: Compatibility analysis of DEPO on LLaVA-Med++.

3.4 Qualitative Analysis and Case Study

Qualitative Assessment of Radiographic Interpretations. Figure 5 provides a qualitative comparison of model responses for chest X-ray interpretation under the same input query. The ground-truth report focuses on clinical findings, including clear lung fields, normal cardiac silhouette, and the absence of pleural effusion or mediastinal abnormalities. As shown in the figure, baseline models such as LLaVA-Med and DPO primarily generate generic descriptions that emphasize the imaging modality or broadly describe anatomical structures, without meaningful findings. In contrast, DEPO produces a clinically precise and concise description that accurately captures key radiographic features consistent with the ground truth, including normal heart size, clear lungs, and the absence of pleural effusion or pneumothorax. This comparison demonstrates that DEPO is more effective at extracting and articulating clinically meaningful information from medical images.

Qualitative Assessment of Medical VQA. Figure 1(a) provides a qualitative comparison of model responses for Medical VQA under the same input query. For LLaVA-Med-1.5, the generation exhibits an *attention shift* toward superficial tokens (e.g., function words and early query tokens such as *Is* and *this*), rather than medicine-critical semantics (e.g., *MRI*). As a result, the model fails to establish

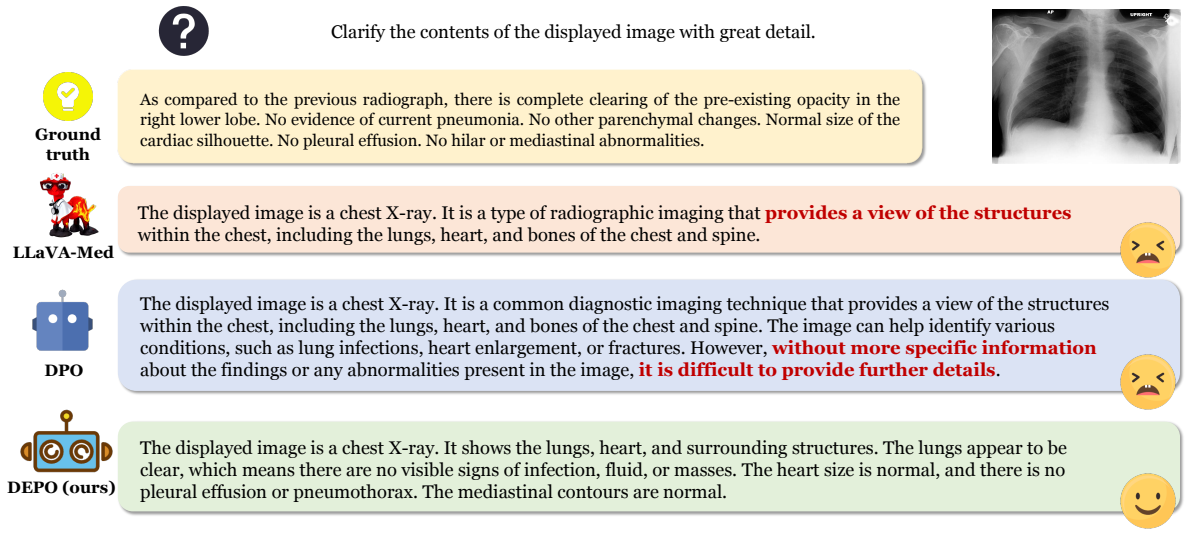

Figure 5: Qualitative comparison. DEPO generates clinically precise and detailed descriptions consistent with ground truth, whereas baseline models produce vague or generic responses with limited diagnostic value.

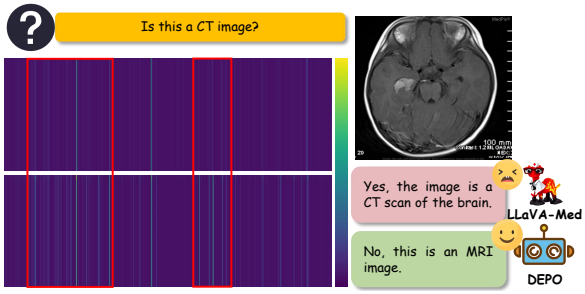

Figure 6: Visualization of image-token attention maps.

stable associations between key question tokens and discriminative visual tokens, and the decision is dominated by weak textual priors, producing an incorrect "Yes" answer. In contrast, DEPO maintains stronger alignment on modality-relevant tokens and concentrates attention on a compact set of informative image tokens, which supports evidence-based reasoning and yields the correct response.

Improvement of Visual Understanding. Figure 6 visualizes the image-token attention maps. Compared with LLaVA-Med, whose attention over image tokens is relatively sparse and scattered, DEPO yields a more concentrated and structured attention distribution. This suggests that DEPO strengthens visual grounding by increasing the contribution of discriminative image tokens to the generated answer. As a result, the model is more likely to base its prediction on modality-specific imaging characteristics rather than defaulting to textual priors, improving medical visual understanding.

4 Related Work

Factuality Issues in Med-LVLMs. Medical LVLMs integrate visual encoders with large language models (LLMs) to support tasks such as medical VQA and report generation (Zhu et al., 2025b). Early studies focused on cross-modal fusion and alignment architectures (Cong et al., 2022; Nguyen et al., 2021). With the emergence of general-purpose LVLMs, medical variants such as LLaVA-Med (Li et al., 2023) and HuatuoGPT-Vision (Chen et al., 2024b) further improve performance through domain-specific pretraining and instruction tuning (Zhu et al., 2025a). Recent work explores retrieval augmentation, agent-based reasoning, multi-step inference, and knowledge graphs to enhance factual accuracy (Xiao et al., 2025; Gai et al., 2025; Wang et al., 2025a). Despite these advances, most methods still rely on static supervision signals, which can promote shortcut learning from superficial textual patterns or spurious visual cues, resulting in fluent but ungrounded hallucinations.

In this work, we introduce Multi-Modal Evidence Perturbation, an evidence-driven augmentation strategy that shifts the model from superficial patterns toward clinically grounded semantics. MEP applies sentence-level rephrasing and adaptive token-level perturbations to regularize textual evidence, and randomly masks low-attention regions while progressively blurring high-attention diagnostic regions to regularize visual evidence.

Preference Optimization in Med-LVLMs. Preference optimization has emerged as an effective

paradigm for mitigating hallucinations by aligning model outputs with relative preferences rather than absolute supervision. Existing approaches to preference fine-tuning generally fall into two categories: methods that rely on human feedback (Bai et al., 2022; Rafailov et al., 2023), and those that leverage feedback generated by large models or the target model itself (Lee et al., 2023; Zhou et al., 2025). More recently, preference optimization has been adapted to medical vision-language models (Sun et al., 2024; Banerjee et al., 2024) by constructing preferred and dispreferred responses using GPT-based annotators or Med-LVLMs themselves.

Despite their effectiveness, most existing methods rely on static preference pairs and fixed sample importance, limiting their ability to adapt to evolving error patterns and suppress newly emerging hallucinations. Although MMedPO introduces clinically informed weighting, it remains static during training and fails to capture dynamic model failures. To address this limitation, we propose Dispreferred Evidence Resampling, which dynamically updates dispreferred responses and their weights to track evolving failure patterns.

5 Conclusion

In this paper, we presented DEPO, a dynamic evidence-guided preference optimization framework for Med-LVLMs. DEPO addresses two key issues, namely overfitting to superficial cues and pseudo convergence of the reward signal. It incorporates Multi-Modal Evidence Perturbation to regularize textual and visual evidence, and Dispreferred Evidence Resampling to dynamically update dispreferred responses and sample weights based on evolving model behavior. Experimental results demonstrate consistent improvements on medical VQA and report generation benchmarks, with enhanced robustness and generalization across datasets and model architectures.

Limitation

Across two tasks, we observe that the model achieves substantial improvements on majority outcomes in the datasets, while the performance gains on minority disease categories remain limited. We attribute this behavior to two main factors. First, existing medical datasets exhibit class imbalance, with a large proportion of samples corresponding to common conclusions. Second, in order to maintain overall performance stability, the model tends to adopt conservative diagnostic behaviors for other classes, limiting improvements on minority classes. In future work, we plan to explicitly address long-tail bias and conservative generation to improve the recognition and expression of all classes.

Acknowledgments

This work was supported by the Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (2024R01007), the "Pioneer" and "Leading Goose" Research and Development Program of Zhejiang (2025C02077), and the Basic Research Project of Shenzhen under Grant JCYJ20220818101216034.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Oishi Banerjee, Hong-Yu Zhou, Kay Wu, Subathra Adithan, Stephen Kwak, and Pranav Rajpurkar. 2024. Direct preference optimization for suppressing hallucinated prior exams in radiology report generation. In *Machine Learning for Healthcare Conference*. PMLR.

Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72.

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

Aofei Chang, Le Huang, Alex James Boyd, Parminder Bhatia, Taha Kass-Hout, Cao Xiao, and Fenglong

Ma. 2025. Focus on what matters: Enhancing medical vision-language models with automatic attention alignment tuning. *arXiv preprint arXiv:2505.18503*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and 1 others. 2024a. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, and 1 others. 2024b. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Fuze Cong, Shibiao Xu, Li Guo, and Yinbing Tian. 2022. Caption-aware medical vqa via semantic focusing and progressive cross-modality comprehension. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3569–3577.

Chenhang Cui, An Zhang, Yiyang Zhou, Zhaorun Chen, Gelei Deng, Huaxiu Yao, and Tat-Seng Chua. 2024. Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment. In *The Thirteenth International Conference on Learning Representations*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. 2025. Medthink: A rationale-guided framework for explaining medical visual question answering. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7438–7450.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

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

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimed-vqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

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

Alistair EW 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(1):317.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

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

Shiyu Li, Yang Tang, Ruijie Liu, Shi-Zhe Chen, and Xi Chen. 2025. Conan-embedding-v2: Training an llm from scratch for text embeddings. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15011–15027.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, and 1 others. 2025. Healthgpt: A medical large vision-language model

for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Pablo Messina, René Vidal, Denis Parra, Álvaro Soto, and Vladimir Araujo. 2024. Extracting and encoding: Leveraging large language models and medical knowledge to enhance radiological text representation. *arXiv preprint arXiv:2407.01948*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pages 353–367. PMLR.

Hoang TN Nguyen, Dong Nie, Taivanbat Badamdorj, Yujie Liu, Yingying Zhu, Jason Truong, and Li Cheng. 2021. Automated generation of accurate\& fluent medical x-ray reports. *arXiv preprint arXiv:2108.12126*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, and 1 others. 2025. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. *arXiv preprint arXiv:2505.06708*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Mai A Shaaban, Tausifa Jan Saleem, Vijay Ram Kumar Papineni, and Mohammad Yaqub. 2025. Motor: Multimodal optimal transport via grounded retrieval in medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 459–469. Springer.

Guohao Sun, Can Qin, Huazhu Fu, Linwei Wang, and Zhiqiang Tao. 2024. Stllava-med: Self-training large language and vision assistant for medical question-answering. *arXiv preprint arXiv:2406.19973*.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, and 1 others. 2023. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Futian Wang, Yuhan Qiao, Xiao Wang, Fuling Wang, Yuxiang Zhang, and Dengdi Sun. 2025a. R2genkg: Hierarchical multi-modal knowledge graph for llm-based radiology report generation. *arXiv preprint arXiv:2508.03426*.

Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Taha Kass-Hout, and 1 others. 2025b. Enhancing visual-language modality alignment in large vision language models via self-improvement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 268–282.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21372–21383.

Ziyan Xiao, Ruiyang Zhang, Yushi Feng, Lingting Zhu, Liang Peng, and Lequan Yu. 2025. A dynamic agent framework for large language model reasoning for medical and visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1143–1152.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and 1 others. 2025. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. In *The Thirteenth International Conference on Learning Representations*.

Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. 2025. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19188–19205.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *Forty-first International Conference on Machine Learning*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644.

Kejia Zhang, Keda Tao, Zhiming Luo, Chang Liu, Jiasheng Tang, and Huan Wang. 2025. Tars: Minmax token-adaptive preference strategy for hallucination reduction in mllms. *arXiv e-prints*, pages arXiv–2507.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024b. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277.

Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. 2022. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

Yiyang Zhou, Zhaoyang Wang, Tianle Wang, Shangyu Xing, Peng Xia, Bo Li, Kaiyuan Zheng, Zijian Zhang, Zhaorun Chen, Wenhao Zheng, Xuchao Zhang, Chetan Bansal, Weitong Zhang, Ying Wei, Mohit Bansal, and Huaxiu Yao. 2025. Anyprefer: An agentic framework for preference data synthesis. In *The Thirteenth International Conference on Learning Representations*.

Kangyu Zhu, Ziyuan Qin, Huahui Yi, Zekun Jiang, Qicheng Lao, Shaoting Zhang, and Kang Li. 2025a. Guiding medical vision-language models with diverse visual prompts: Framework design and comprehensive exploration of prompt variations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11726–11739.

Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. 2024. Mmedpo: Aligning medical vision-language models with clinical-aware multimodal preference optimization. *arXiv preprint arXiv:2412.06141*.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, QingqingLong QingqingLong, Yefeng Zheng, and Xian Wu. 2025b. Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769.

A Additional Cross-Architecture Generalization Results

To further address the concern regarding limited empirical support for broad transferability, we extend the compatibility analysis to more architectures beyond the LLaVA-Med family. In addition to LLaVA-Med++, we evaluate DEPO on HuatuoGPT-V, Med-Flamingo, RadFM, and MedVInT across two datasets, VQA-RAD and MIMIC-CXR. The corresponding backbone architectures are summarized in Table 4, and the quantitative results are reported in Table 5 and Table 6. As shown in these results, DEPO consistently improves performance across all evaluated architectures, which provides additional evidence for the generality of the proposed framework.

Model	Architecture
LLaVA-Med++	LLaVA-Med
HuatuoGPT-V(Chen et al., 2024a)	LLaVA-v1.5-LLaMA3-8B
Med-Flamingo(Moor et al., 2023)	OpenFlamingo-9B
RadFM(Wu et al., 2025)	LLaMA-13B
MedVInT(Zhang et al., 2024b)	PMC-LLaMA and PMC-CLIP

Table 4: Model architectures used in the additional cross-architecture generalization experiments.

Model	w/o DEPO	w/ DEPO	Δ
LLaVA-Med v1.5	54.38	58.88	+4.50
HuatuoGPT-V(Chen et al., 2024a)	52.76	56.62	+3.86
Med-Flamingo(Moor et al., 2023)	46.10	49.37	+3.27
RadFM(Wu et al., 2025)	44.82	48.03	+3.21
MedVInT(Zhang et al., 2024b)	47.27	50.85	+3.58

Table 5: Cross-architecture generalization results on VQA-RAD.

Model	w/o DEPO	w/ DEPO	Δ
LLaVA-Med++	8.08	12.91	+4.83
HuatuoGPT-V(Chen et al., 2024a)	8.01	13.24	+5.23
Med-Flamingo(Moor et al., 2023)	7.81	12.53	+4.72
RadFM(Wu et al., 2025)	7.52	11.95	+4.43
MedVInT(Zhang et al., 2024b)	7.73	11.83	+4.10

Table 6: Cross-architecture generalization results on MIMIC-CXR.

Hyperparameter	SLAKE	VQA-RAD	MIMIC-CXR	IU-Xray
Base Model	LLaVA-Med-7B	LLaVA-Med-7B	LLaVA-Med-7B	LLaVA-Med-7B
Visual Encoder	CLIP-ViT-L/14	CLIP-ViT-L/14	CLIP-ViT-L/14	CLIP-ViT-L/14
Max Token Length	64	64	1024	1024
Optimizer	AdamW	AdamW	AdamW	AdamW
LoRA Rank	128	128	128	128
LoRA Alpha	256	256	256	256
Multimodal Projector Learning Rate	2e-5	2e-5	2e-5	2e-5
Learning Rate	5e-7	5e-7	5e-7	5e-7
LR Scheduler	Cosine Decay	Cosine Decay	Cosine Decay	Cosine Decay
Warmup Ratio	0.03	0.03	0.03	0.03
Batch Size	4	4	4	4
Epochs	5	5	5	5
Weight Decay	0.05	0.05	0.05	0.05
Verify rule	Recall	Recall	BLEU	BLEU
Update range	Only open	Only open	All	All
α	0.1	0.1	0.1	0.1
τ_q	0.2	0.2	0.15	0.35
τ_Δ	0.8	0.8	0.8	0.8
Dynamic Masking Budgets	1-3	1-3	1-3	1-3
DER Weighting (0.1 - 0.3)	± 0.2	± 0.2	± 0.2	± 0.2
ω	0.1	0.1	0.1	0.1
Precision	bf16	bf16	bf16	bf16
Hardware	4×4090 (24G)	4×4090 (24G)	4×4090 (24G)	4×4090 (24G)

Table 7: Detailed hyperparameter settings for different datasets.

▶ Rewriting prompt

You are a careful medical editor. Your job is to rewrite a given medical question without changing its truth conditions or correct answer. Please rewrite the following medical question in 10 different ways, but ensure that the rewritten questions share exactly the same answer as the original question. Do not change the answer type (e.g., Yes/No, numeric, categorical) or the scope of the question. Only rephrase the wording while keeping the underlying meaning and answer consistent. Output 10 rewritten versions.
Original question: {original_question}
Format: 10 different rewritten versions, each on a separate line, numbered 1-10

B Additional Implementation Details

To improve reproducibility, we provide the detailed hyperparameter settings used in Table 7.

C Rewriting Prompt

For completeness, we provide the prompt used to generate semantically equivalent rewrites of each medical question: