Benchmarking Direct Preference Optimization for Medical Large Vision–Language Models

Dain Kim[1,*] **Jiwoo Lee**[1,*] **Jaehoon Yun**[2,3] **Yong Hoe Koo**[4]
Qingyu Chen[5] **Hyunjae Kim**[5,†] **Jaewoo Kang**[1,2,†]
[1]Korea University [2]AIGEN Sciences [3]Hanyang University College of Medicine
[4]Asan Medical Center, University of Ulsan College of Medicine [5]Yale University
{dain-kim,hijiwoo7,kangj}@korea.ac.kr hyunjae.kim@yale.edu

Abstract

Large Vision-Language Models (LVLMs) hold significant promise for medical applications, yet their deployment is often constrained by insufficient alignment and reliability. While Direct Preference Optimization (DPO) has emerged as a potent framework for refining model responses, its efficacy in high-stakes medical contexts remains underexplored, lacking the rigorous empirical groundwork necessary to guide future methodological advances. To bridge this gap, we present the first comprehensive examination of diverse DPO variants within the medical domain, evaluating nine distinct formulations across two medical LVLMs: LLaVA-Med and HuatuoGPT-Vision. Our results reveal several critical limitations: current DPO approaches often yield inconsistent gains over supervised fine-tuning, with their efficacy varying significantly across different tasks and backbones. Furthermore, they frequently fail to resolve fundamental visual misinterpretation errors. Building on these insights, we present a targeted preference construction strategy as a proof-of-concept that explicitly addresses visual misinterpretation errors frequently observed in existing DPO models. This design yields a 3.6% improvement over the strongest existing DPO baseline on visual question-answering tasks. To support future research, we release our complete framework, including all training data, model checkpoints, and our codebase at https://github.com/dmis-lab/med-vlm-dpo.

1 Introduction

Recent advances in Large Vision-Language Models (LVLMs), which integrate powerful large language models (LLMs) with visual encoders, have greatly improved AI's ability to process and reason over multimodal inputs (Alayrac et al., 2022; Li et al., 2023b; Liu et al., 2023; Zhu et al., 2024; OpenAI, 2023). In the medical domain, these advances have enabled applications such as diagnostic support, clinical question answering, and report generation (Kline et al., 2022; Li et al., 2023a; Chen et al., 2024b; Wu et al., 2025; Xie et al., 2025), but safe deployment remains a critical challenge. For instance, factually incorrect or fabricated outputs, often described as hallucinations, pose particular risks (Maynez et al., 2020; Liu et al., 2024a; Kim et al., 2025). Additionally, errors in interpreting medical images (Jin et al., 2024) may lead to cascading failures in downstream decision-making (Zhang et al., 2024b).

Direct Preference Optimization (DPO) (Rafailov et al., 2023) and its subsequent variants have been explored to improve the reliability of language models (Ethayarajh et al., 2024; Xu et al., 2024). By leveraging preference signals to contrast output pairs, DPO optimizes model parameters to favor safer and more faithful generations. However, while these approaches have been predominantly validated in general-domain language and vision-language tasks (Saeidi et al., 2025; Zhou et al., 2024b; Wang et al., 2024), their performance in high-stakes fields such as medicine remains insufficiently understood. Given the distinct data characteristics and the specific nature of medical errors, general-domain optimizations may not directly translate to reliable clinical performance, necessitating a dedicated validation of preference-based alignment within this specialized context.

In this paper, we present the first comprehensive evaluation of DPO-based alignment for medical LVLMs. We systematically analyze leading multi-modal DPO methods from both general and medical domains, categorizing them into three distinct groups based on their data perturbation strategies: text-only, image-only, and joint text-

[*]These authors contributed equally to this work.
[†]Corresponding authors.

Findings of the Association for Computational Linguistics: EACL 2026, pages 5052–5067
March 24-29, 2026 ©2026 Association for Computational Linguistics

image (Figure 1a). We implement nine DPO formulations atop two representative medical LVLMs: LLaVA-Med (Li et al., 2023a) and HuatuoGPT-Vision (Chen et al., 2024b).

We analyze the models in two stages: a benchmark evaluation and an expert evaluation (Figure 1b). For benchmark evaluation, we first compile five datasets spanning both visual question answering (VQA) and two generation tasks: radiology report generation and image captioning. We evaluate baseline models using accuracy for VQA and employ an LLM-as-a-judge framework (Zheng et al., 2023; Gu et al., 2024) to assess completeness and contradiction in the generation tasks. We observe that all DPO variants consistently improve VQA accuracy. However, similar gains can also be achieved through standard supervised fine-tuning (SFT), making the advantage of DPO less evident in this setting. In the generation tasks, no single method consistently outperforms others across tasks or metrics. For example, a text-only DPO model achieved the highest completeness score on the image captioning dataset with a 3.11% improvement, yet exhibited a 4.81% decrease in report generation performance. Overall, the results suggest that the effects of DPO may not fully align with previous reports of its effectiveness in general-domain (Zhou et al., 2024b; Wang et al., 2024) or early medical-domain studies (Zhu et al., 2025).

To gain deeper insights, we conduct a manual error analysis, aiming to uncover the qualitative limitations underlying our quantitative findings. We observe that a substantial majority of errors originated from the misinterpretation of medical images. Notably, the base LLaVA-Med model exhibits image misunderstandings in 90% of its image captioning outputs (82.5% severe, 7.5% minor) and 97.5% for its report generation outputs (82.5% severe, 15% minor). While a DPO model significantly mitigates the most critical failures—reducing severe interpretation errors from 90% to 50% in image captioning—this improvement is accompanied by a marked increase in minor misinterpretations, which rises from 7.5% to approximately 30%. While encouraging, these results suggest that current DPO formulations merely shift the error profile from severe to minor rather than fully resolving the underlying issues.

To investigate whether these gaps could be narrowed through more targeted alignment, we identify four major categories of visual misinterpretation errors recurring in model outputs. We then tailor the DPO training process by constructing preference data specifically designed to counteract these errors, exploring the feasibility of domain-targeted preference modeling (Figure 1c). While previous experiments showed DPO providing only marginal improvements over SFT in VQA tasks, our specialized DPO approach yields consistent performance gains, outperforming the base LLaVA-Med model by 6.9%, the SFT model by 4.6%, and the best-performing baseline DPO model by 3.6%.

Beyond our initial findings, the observed limitations underscore the need for more rigorous community-wide validation and the development of robust, domain-aware alignment strategies for medical AI. To support these efforts and foster further innovation, we publicly release our entire framework, including the code, curated datasets, trained models, and expert-verified error annotations.[1]

2 Related Work

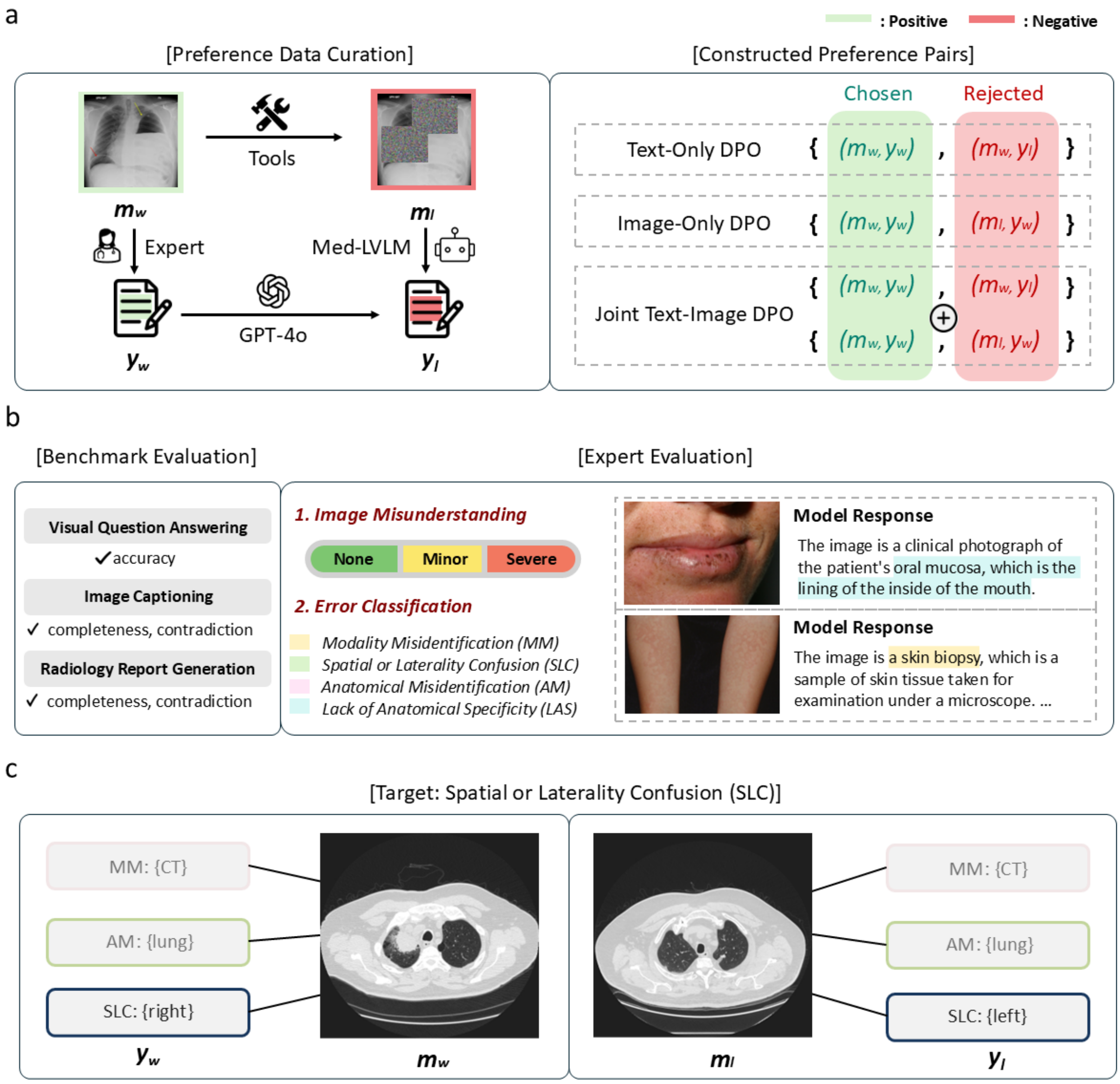
2.1 Large Vision-Language Models in Medicine

Medical LVLMs are adapted from general-purpose models through fine-tuning on biomedical data (Li et al., 2023a; Chen et al., 2024b; Zhang et al., 2024a; Lin et al., 2025; Sellergren et al., 2025), typically using image-caption pairs from PubMed Central for visual alignment and human- or LLM-generated prompts for instruction tuning. Some models further integrate biomedical-specific vision encoders (Lin et al., 2023; Zhang et al., 2025) to enhance domain relevance. While models like LLaVA-Med (Li et al., 2023a) and HuatuoGPT-Vision (Chen et al., 2024b) support broad medical tasks, others are tailored to specific fields such as radiology (Wu et al., 2025; Chen et al., 2024c), surgery (Wang et al., 2025a), pathology (Seyfioglu et al., 2024), and dermatology (Zhou et al., 2024a).

2.2 Direct Preference Optimization for LVLMs

Direct Preference Optimization (DPO) was originally proposed for text-only preference pairs (Rafailov et al., 2023). Subsequent works have extended this framework to multimodal tasks by modifying the definition of the contrastive objective. Specifically, variants differ in the modality being contrasted: some apply DPO to output text only, while others contrast both input and output jointly. HA-DPO (Zhao et al., 2023) and

[1] https://github.com/dmis-lab/med-vlm-dpo


a
: Positive
: Negative
[Preference Data Curation]
Tools
m_w
m_l
Expert
Med-LVLM
GPT-4o
y_w
y_l
[Constructed Preference Pairs]
Chosen
Rejected
Text-Only DPO
{ (m_w, y_w) , (m_w, y_l) }
Image-Only DPO
{ (m_w, y_w) , (m_l, y_w) }
Joint Text-Image DPO
{ (m_w, y_w) , (m_w, y_l) }
⊕
{ (m_w, y_w) , (m_l, y_w) }
b
[Benchmark Evaluation]
Visual Question Answering
✓accuracy
Image Captioning
✓ completeness, contradiction
Radiology Report Generation
✓ completeness, contradiction
[Expert Evaluation]
1. Image Misunderstanding
None
Minor
Severe
2. Error Classification
Modality Misidentification (MM)
Spatial or Laterality Confusion (SLC)
Anatomical Misidentification (AM)
Lack of Anatomical Specificity (LAS)
Model Response
Model Response
c
[Target: Spatial or Laterality Confusion (SLC)]
MM: {CT}
AM: {lung}
SLC: {right}
y_w
m_w
m_l
MM: {CT}
AM: {lung}
SLC: {left}
y_l

The image is a clinical photograph of the patient's oral mucosa, which is the lining of the inside of the mouth.

The image is a skin biopsy, which is a sample of skin tissue taken for examination under a microscope. ...

Figure 1: Overview of the study design. **a.** Evaluated DPO models: Illustration of three DPO configurations (text-only, image-only, and joint text-image) categorized by the modality contrasted during preference learning. **b.** Evaluation framework: A multi-faceted assessment combining automated benchmark evaluation across three core tasks (visual question answering, image captioning, and radiology report generation) and expert qualitative analysis focused on image misunderstanding severity and specific error types (e.g., MM, SLC, AM, LAS). **c.** Targeted preference pair construction: A demonstration of our approach using a SLC example. We construct contrastive pairs by perturbing text keywords and retrieving corresponding "hard-negative" images to improve the model's spatial and anatomical grounding (see Section 5 for details).

HSA-DPO (Xiao et al., 2025) generate rejections by automatically detecting and correcting hallucinated spans, while Silkie (Li et al., 2023c) ranks multiple model outputs to form preference pairs. SIMA (Wang et al., 2025b) further leverages self-feedback, where the model compares and critiques its own outputs, while CLIP-DPO (Ouali et al., 2025) derives preference signals from image-text similarity scores provided by a pretrained CLIP model. In contrast, several methods contrast both outputs and inputs simultaneously. mDPO (Wang et al., 2024) integrates text-based rejections with corrupted inputs, using random cropping as the perturbation. POVID (Zhou et al., 2024b) combines GPT-4-generated hallucinations on the text side with Gaussian-noised images on the visual side. MMedPO (Zhu et al., 2025) likewise merges modalities, treating hallucinated responses as rejections while contrasting original images against ROI-noised counterparts.

Method	Description	Relevant Methods
Text-only Perturbation		
Text-Hallu	y_w corresponds to y. y_l is generated by hallucinating y by GPT-4o.	POVID (Zhou et al., 2024b), MMedPO (Zhu et al., 2025)
Text-Hallu + NLL	Text-Hallu with the addition of NLL loss.	
Text-Noise	y_w corresponds to y. y_l is self-generated from the image m with Gaussian noise.	POVID (Zhou et al., 2024b), STIC (Deng et al., 2024)
Text-Noise + NLL	Text-Noise with the addition of NLL loss.	
IRPO	y_w is a self-generated response closely aligned with y. y_l is a less aligned one.	IRPO (Pang et al., 2024)
Image-only Perturbation		
Image-Noise	m_w corresponds to m. m_l corresponds to m with Gaussian noise.	mDPO (Wang et al., 2024), POVID (Zhou et al., 2024b), MMedPO (Zhu et al., 2025)
Image-ROI	m_w corresponds to m. m_l corresponds to m with Gaussian noise applied to ROI.	
Joint Text-Image Perturbation		
mDPO	y_w and m_w correspond to y and m, respectively. m_l corresponds to m with random cropping applied. y_l is self-generated from m_l.	mDPO (Wang et al., 2024)
MMedPO	y_w corresponds to y, and y_l is generated by GPT-4o. m_w and m_l: Same as Image-ROI.	MMedPO (Zhu et al., 2025)

Table 1: Categorization of DPO methods. y: ground-truth response. m: original image. y_w and y_l: preferred (chosen) and dispreferred (rejected) responses, respectively. m_w and m_l: preferred (chosen) and dispreferred (rejected) images. Relevant methods denote existing frameworks whose core principles were adapted and tuned for medical multimodal alignment. ROI: Regions of Interest.

2.3 Benchmarking Medical LVLMs

A growing body of work has sought to benchmark medical LVLMs across various tasks and dimensions. MultiMedEval (Royer et al., 2024) and Asclepius (Liu et al., 2024b) offer large-scale suites to evaluate accuracy across modalities and specialties, addressing prior issues of fragmented evaluation practices. In parallel, CARES (Xia et al., 2024) introduces a multidimensional framework for trustworthiness, covering trustfulness, fairness, safety, privacy, and robustness. MedHEval (Chang et al., 2025) and Med-HallMark (Chen et al., 2024a) further systematize hallucination evaluation and highlight domain-specific risks like visual misinterpretation and knowledge deficiency. Yet, most benchmarks focus on off-the-shelf models; only MedHEval (Chang et al., 2025) systematically examines inference-time hallucination mitigation. To the best of our knowledge, no prior work has comprehensively evaluated DPO models for medical LVLMs. Our work examines whether and how such preference-based tuning methods affect medical LVLM behavior, through both automatic benchmarks and structured expert assessments.

3 Evaluated DPO Models

Prior work on DPO for LVLMs can be categorized based on the modality being contrasted: text-only, image-only, or joint text-image, as illustrated in Figure 1a. Building on this perspective, we further organize the landscape along two axes: (i) the underlying training objective and (ii) the preference pair curation strategy.

Most of existing approaches were originally developed for the general domain and mainly address issues such as object hallucination (Bai et al., 2024), which are not directly applicable to medical data. To bridge this gap, we adapted these methods to the medical domain while retaining their core principles, resulting in eight domain-specific DPO variants. We also include MMedPO (Zhu et al., 2025), a method developed specifically for medical applications, yielding a total of nine models. Please refer to Table 1 for the full list. In Appendix A, we provide illustrative examples of preference pairs.

3.1 Text-only Perturbation

Let q be a text prompt (i.e., the instruction or query to the model), y_w the preferred (chosen) response, and y_l the dispreferred (rejected) response. The standard DPO objective is:

$$\mathcal{L}_{\text{DPO}} = -\log\sigma\Bigg(\beta\log\frac{\pi_\theta(y_w \mid q)}{\pi_{\text{ref}}(y_w \mid q)} - \beta\log\frac{\pi_\theta(y_l \mid q)}{\pi_{\text{ref}}(y_l \mid q)}\Bigg), \tag{1}$$

where π_θ is the target policy, π_{ref} is a fixed reference model (typically a supervised fine-tuned checkpoint), β is a temperature-like scaling factor, and $\sigma(\cdot)$ denotes the sigmoid function.

In multimodal settings with an input image m and prompt q, this extends to:

$$\mathcal{L}_{\text{DPO}_m} = -\log\sigma\Bigg(\beta\log\frac{\pi_\theta(y_w \mid m, q)}{\pi_{\text{ref}}(y_w \mid m, q)} - \beta\log\frac{\pi_\theta(y_l \mid m, q)}{\pi_{\text{ref}}(y_l \mid m, q)}\Bigg) \tag{2}$$

The text-only models differ in how y_w and y_l are defined. In Text-Hallu, y_w is the ground-truth response y, while y_l is generated by GPT-4o with induced hallucinations. In Text-Noise, $y_w = y$, and y_l is self-generated from a Gaussian-noised image m. In IRPO, y_w and y_l pairs are selected from N=20 self-generated responses (temperature 1.2) ranked by ROUGE-L against the ground truth, with the top-1 and bottom-1 responses chosen. The "+NLL" variants of Text-Hallu and Text-Noise further incorporate a negative log-likelihood term to encourage higher probability for y. The IRPO objective inherently includes an NLL term within its formulation (see Appendix A for details).

3.2 Image-only Perturbation

In this setting, the image m is perturbed, whereas the prompt q and reference response y remain identical across conditions. That is, both (m_w, q) and (m_l, q) are paired with the same y, ensuring that differences arise solely from the image modality. The Conditional Preference Optimization (CoPO) loss (Wang et al., 2024) applies a contrastive objective over image perturbations:

$$\mathcal{L}_{\text{CoPO}} = -\log\sigma\Bigg(\beta\log\frac{\pi_\theta(y_w \mid m_w, q)}{\pi_{\text{ref}}(y_w \mid m_w, q)} - \beta\log\frac{\pi_\theta(y_w \mid m_l, q)}{\pi_{\text{ref}}(y_w \mid m_l, q)}\Bigg). \tag{3}$$

Dataset	**# Ex.**	**# Img.**	**Modality**
Visual Question Answering (VQA)			
VQA-RAD	451	204	Radiology
SLAKE	1,061	96	Radiology
PathVQA	6,719	858	Pathology
Image Captioning			
AMBOSS	164	164	Misc.
Radiology Report Generation			
MIMIC-CXR	1,031	1,031	Chest X-ray

Table 2: Datasets used in the benchmark evaluation. # Ex. and # Img.: the number of examples and images, respectively.

Here, Image-Noise perturbs m with Gaussian noise, while Image-ROI perturbs the ROI (Regions of Interest) extracted using MedCLIP (Wu et al., 2023).

3.3 Joint Text-Image Perturbation

Models in this group combine the objectives defined in Equations 2 and 3. In mDPO, y_w and m_w are the ground-truth response y and original image m, while m_l is a randomly cropped version of m. The rejected response y_l is conditioned on the corrupted image m_l. In MMedPO, y_w and m_w are the ground-truth response y and original image m. The rejected response y_l is generated by GPT-4o. The rejected image m_l are defined as in Image-ROI, with Gaussian noise applied to the ROI. Please refer to Appendix A for the detailed mathematical formulations for mDPO and MMedPO.

4 Experiments

4.1 Tasks and Datasets

The benchmark evaluation includes the following tasks and datasets, along with a description of the metrics used. See Table 2 for a summary.

Visual question answering (VQA) This task is widely used for evaluating medical LVLMs. For datasets, we use VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020), all of which pair radiological/pathological images with clinician-annotated QA pairs. We report accuracy averaged across the three datasets.

Image captioning This task focuses on describing the core visual findings in medical images across various modalities, including radiology, pathology, dermatology, and endoscopy. We utilize image-caption pairs curated by medical experts, sourced from AMBOSS, a medical question bank

	VQA	Image Captioning		Radiology Report Generation	
Model	**Acc (↑)**	**Comp (↑)**	**Cont (↓)**	**Comp (↑)**	**Cont (↓)**
LLaVA-Med					
Base Model	38.7 (-)	9.11 (-)	19.56 (-)	15.83 (-)	11.90 (-)
SFT	41.0 (+2.3)	9.66 (+0.55)	20.43 (+0.87)	10.96 (-4.87)	13.64 (+1.74)
Text-Hallu	41.3 (+2.6)	**12.22 (+3.11)**	13.93 (-5.63)	11.02 (-4.81)	12.39 (+0.49)
+ NLL	**42.0 (+3.3)**	10.67 (+1.56)	19.15 (-0.41)	13.12 (-2.71)	**11.03 (-0.87)**
Text-Noise	39.0 (+0.3)	9.80 (+0.69)	11.63 (-7.93)	17.70 (+1.87)	11.99 (+0.09)
+ NLL	40.6 (+1.9)	9.85 (+0.74)	21.23 (+1.67)	13.13 (-2.70)	13.63 (+1.73)
IRPO	39.0 (+0.3)	9.12 (+0.01)	19.79 (+0.23)	**18.33 (+2.50)**	12.00 (+0.10)
Image-Noise	40.0 (+1.3)	8.22 (-0.89)	**9.35 (-10.21)**	12.24 (-3.59)	12.77 (+0.87)
Image-ROI	41.0 (+2.3)	10.74 (+1.63)	11.38 (-8.18)	11.34 (-4.49)	13.23 (+1.33)
mDPO	41.9 (+3.2)	10.44 (+1.33)	20.58 (+1.02)	12.75 (-3.08)	11.98 (+0.08)
MMedPO	40.1 (+1.4)	11.38 (+2.27)	12.75 (-6.81)	10.82 (-5.01)	12.74 (+0.84)
HuatuoGPT-Vision					
Base Model	49.1 (-)	20.97 (-)	29.58 (-)	21.81 (-)	23.05 (-)
SFT	51.9 (+2.8)	21.03 (+0.06)	27.15 (-2.43)	21.91 (+0.10)	22.76 (-0.29)
Text-Hallu	**53.2 (+4.1)**	21.48 (+0.51)	26.86 (-2.72)	22.34 (+0.53)	22.93 (-0.12)
+ NLL	51.7 (+2.6)	20.39 (-0.58)	28.81 (-0.77)	22.36 (+0.55)	22.60 (-0.45)
Text-Noise	51.1 (+2.0)	20.78 (-0.19)	25.88 (-3.70)	22.50 (+0.69)	25.14 (+2.09)
+ NLL	52.4 (+3.3)	20.89 (-0.08)	25.00 (-4.58)	22.05 (+0.24)	24.15 (+1.10)
IRPO	52.7 (+3.6)	20.97 (+0.00)	26.83 (-2.75)	22.10 (+0.29)	**21.57 (-1.48)**
Image-Noise	49.6 (+0.5)	22.26 (+1.29)	**24.69 (-4.89)**	**22.80 (+0.99)**	22.98 (-0.07)
Image-ROI	52.4 (+3.3)	**23.03 (+2.06)**	28.54 (-1.04)	22.21 (+0.40)	22.81 (-0.24)
mDPO	50.9 (+1.8)	23.02 (+2.05)	25.22 (-4.36)	21.51 (-0.30)	22.42 (-0.63)
MMedPO	51.8 (+2.7)	21.51 (+0.54)	26.90 (-2.68)	22.07 (+0.26)	22.79 (-0.26)

Table 3: Performance comparison of DPO methods and SFT baselines across LLaVA-Med and HuatuoGPT-Vision backbones. Evaluation spans visual question answering (VQA; averaged across the SLAKE, VQA-RAD, and PathVQA datasets), image captioning (the AMBOSS dataset), and radiology report generation (the MIMIC-CXR dataset). We report accuracy (Acc) for VQA, along with completeness (Comp) and contradiction (Cont) for generation tasks (↑: higher is better; ↓: lower is better). Values in parentheses denote the performance delta relative to each SFT base model. For reference, GPT-4.1 achieves a VQA accuracy of 58.1%.

platform designed for licensing/board exam preparation.[2] While the data is licensed, we obtained explicit permission for their use in this study. For metrics, we apply a statement-level, LLM-based evaluation: reference reports are decomposed into atomic clinical statements (Min et al., 2023), and model outputs are classified into entailment, partial entailment, contradiction, or neutral (see Appendix B for details). From this, we compute completeness (proportion of entailed statements) and contradiction scores (proportion of contradicted statements).

Radiology report generation We evaluate models on the MIMIC-CXR dataset (Johnson et al., 2019) using a filtered test set to ensure precise, image-grounded assessment (see Appendix C for details). The same metrics as in image captioning are used—completeness and contradiction.

4.2 Base LVLMs

We use LLaVA-Med v1.5 with a Mistral-7B backbone (Li et al., 2023a) and HuatuoGPT-Vision with a Qwen2-7B backbone (Chen et al., 2024b) as our base models, both of which were pretrained and instruction-tuned on large-scale medical data. We sample 10,000 instructions from each model's training corpus, which serve as a shared foundation for subsequent SFT and DPO preference pair construction. This follows established practice in recent studies, which typically employ between 5,000 and 17,000 preference pairs (Wang et al., 2024, 2025b; Zhou et al., 2024b; Deng et al., 2024).

[2] https://www.amboss.com/us

	Image Misunderstanding (%)			Error-type Distribution (#)			
Model	**None**	**Minor**	**Severe**	**MM**	**SLC**	**AM**	**LAS**
Image Captioning							
Base Model	10.0	7.5	82.5	11	5	6	0
TxtPert-LLM	20.0	30.0	50.0	5	4	11	1
MMedPO	15.0	17.5	67.5	9	4	11	2
SFT	0.0	15.0	85.0	11	2	12	1
Radiology Report Generation							
Base Model	2.5	15.0	82.5	0	3	0	9
TxtPert-LLM	5.0	35.0	57.5	0	1	0	22
MMedPO	2.5	35.0	60.0	0	1	0	14
SFT	5.0	17.5	77.5	0	1	0	24

Table 4: Expert evaluation on the image captioning and report generation tasks with LLaVA-Med. Image misunderstanding is reported as percentages (%), and error-type distribution as counts (out of 40 evaluated cases per dataset). MM: Modality misidentification; SLC: Spatial or laterality confusion; AM: Anatomical misidentification; LAS: Lack of anatomical specificity. Detailed descriptions are provided in the main text.

4.3 Benchmark Results

We conducted bootstrapping with 100 resampling iterations of equal size. Table 3 summarizes model performance across VQA, image captioning, and report generation tasks. Crucially, our results reveal a significant limitation of applying existing DPO strategies to the medical domain. While DPO models improved VQA accuracy (+0.3–3.3% for LLaVA-Med, +0.5–4.1% for HuatuoGPT-Vision), comparable gains were observed with matched SFT baselines. This indicates that the perceived benefits may stem largely from additional training rather than preference optimization itself.

For generation tasks, performance was inconsistent; while some DPO variants improved specific metrics, others showed regression, with deltas often falling within the range of run-to-run variability. Contrary to prior findings reporting uniform improvements (Zhu et al., 2025), our analysis demonstrates that naively transferring DPO methods to complex medical tasks yields unstable and marginal gains.

4.4 Expert Evaluation

Beyond quantitative metrics, we conducted an expert evaluation to investigate the underlying failure modes that the DPO models exhibit in medical tasks. We specifically focused our error analysis on image misunderstanding, as accurate visual recognition serves as the critical foundation for all downstream medical interpretation. Since visual errors often lead to a cascade of incorrect clinical reasoning, two experts categorized these failures by their severity (e.g., none, moderate, or severe) to measure their actual impact. Here, severe errors refer to critical misinterpretations of the image that can significantly compromise downstream reasoning or diagnostic accuracy, whereas minor errors involve subtler inaccuracies that do not substantially alter the clinical meaning. We utilized a curated set of 80 samples (40 from AMBOSS, 40 from MIMIC-CXR) and assessed four models: the base LLaVA-Med, and its SFT baseline, Text-Hallu, and MMedPO. More details are provided in Appendix D.

Table 4 shows that the base LLaVA-Med model exhibited substantial limitations, with 82.5% of image captioning responses and 82.5% of report generation responses containing severe errors, and an additional 7.5% and 15% containing minor errors, respectively. These results indicate that the baseline model lacks sufficient capability for accurate medical image interpretation. Moreover, post-training methods such as SFT and DPO were insufficient in addressing this limitation. In fact, SFT appeared to amplify errors in image captioning, with severe and minor errors increasing to 85% and 15%, respectively. Text-Hallu and MMedPO reduced severe errors in image captioning (to 50% and 67.5%), but simultaneously increased minor errors (from 7.5% to 30% and 17.5%, respectively).

Fine-grained error analysis We further categorized the errors into four major types: (1) Modality misidentification (MM): The model incorrectly identifies the imaging modality, such as mistaking a pathology slide for a clinical photograph. (2) Spatial or laterality confusion (SLC): The model confuses spatial orientation or left/right anatomical

Model	VQA (Pooled)	Subsets			
		MM	SLC	AM	LAS
Base model	38.7	49.8	26.6	40.3	10.0
SFT	41.0	56.2	30.2	43.2	14.2
Text-Hallu	41.3	63.6	30.4	42.2	13.2
+ NLL	42.0	60.1	31.2	43.0	14.9
Text-Noise	39.0	54.2	29.8	41.6	14.8
+ NLL	40.6	56.0	32.3	42.7	13.7
IRPO	39.0	54.1	31.8	40.8	13.9
Image-Noise	40.0	51.8	32.2	43.4	14.9
Image-ROI	41.0	55.5	30.3	42.6	11.7
mDPO	41.9	59.5	32.1	42.7	14.7
MMedPO	40.1	56.7	32.9	41.5	12.3
Ours (DPO)	45.5	**69.4**	35.9	45.5	20.8
Ours (CoPO)	**45.6**	69.0	**36.0**	**45.6**	20.8
Ours (mDPO)	45.4	69.1	35.9	45.5	**20.9**

Table 5: Performance comparison of our DPO models and baseline models on VQA tasks. The left group represents pooled VQA performance (averaged across the full SLAKE, VQA-RAD, and PathVQA datasets). The right group shows accuracy on error-specific subsets, consisting only of VQA items relevant to specific visual recognition errors. MM: Modality misidentification. SLC: Spatial or laterality confusion. AM: Anatomical misidentification LAS: Lack of anatomical specificity.

sides, for instance, describing a left lung lesion as being located in the right lung. (3) Anatomical misidentification (AM): The model misidentifies anatomical structures, such as referring to the lip as intraoral tissue or confusing the arm with the leg. (4) Lack of anatomical specificity (LAS): The model provides overly broad anatomical references relative to the ground truth, such as describing the "right lung" instead of the more precise "right lower lobe." These error types reflect basic visual understanding that should be correctly recognized before any detailed reasoning. However, as shown in Table 4, models frequently made such errors, and post-training methods even amplified specific categories, for example, AM in image captioning and LAS in report generation.

5 Enhanced DPO Training

Manual analysis revealed that the majority of responses contained image misinterpretation errors. Since accurate visual understanding is the foundation for downstream reasoning, such errors often lead to cascading failures in subsequent steps, likely contributing to the overall inconsistency in performance. Unfortunately, existing DPO methods are not well-equipped to address these underlying limitations of the base models. Even MMedPO, which was specifically designed to enhance visual grounding by aligning model attention with clinically critical ROIs in medical images, exhibited significant visual errors in our evaluation. To examine whether this issue is addressable, we explored a straightforward approach by incorporating fine-grained visual error types into the DPO pair construction process. As a proof of concept, we integrated four common categories of model errors (i.e., MM, SLC, AM, and LAS).

5.1 Model

We constructed preference pairs that isolate a single error type while preserving the surrounding clinical context across both text and image modalities. To ensure a fair comparison, we aggregated samples across all error categories to form a single training set of 10k samples, consistent with the size of the dataset used for our baseline models. Using this dataset, we developed and evaluated three distinct DPO configurations to investigate the impact of each modality: text-only (DPO), image-only (CoPO), and joint text-image (mDPO).

Error-type assignment To systematically identify relevant samples, we first defined keyword lists corresponding to each error category (see Appendix E). These lists were then used to map each image to its potential failure modes by identifying relevant clinical terms within the associated instruction and response. Since each sample only pertains to specific error types depending on its content, we tagged each image with only the detected categories. For instance, as illustrated in Figure 1c, an image of a chest CT might be tagged with MM: {CT}, AM: {lung}, and SLC: {right}, while LAS is omitted as no corresponding keywords were matched. These tags serve as ground-truth anchors, allowing us to formulate contrastive pairs by systematically perturbing specific attributes or selectively retrieving images that align with these anchors, all while keeping the rest of the clinical context intact.

Generation of rejected text responses We utilized GPT-4o to generate a corresponding rejected response y_l by perturbing the identified error-type keywords. Following the SLC example in Figure 1c, if the response y describes a finding on the "right," the model was prompted to substitute the target keyword with a plausible but clinically incorrect alternative, resulting in the term "left" in the rejected output y_l. We constrained the generation process to strictly preserve all other clinical details and maintain a consistent length and tone between y_w and y_l.

Retrieval of rejected images The original image, serving as the chosen image m_w, was paired with a hard-negative m_l. We selected m_l by retrieving a sample that differed strictly on the targeted attribute; for instance, as shown in Figure 1c, we selected an image displaying a “left” side pathology instead of “right” while keeping other attributes consistent. This selection process ensured precise supervision by forcing the model to distinguish subtle spatial differences between otherwise nearly identical clinical contexts in images.

5.2 Results

We evaluated our models across three VQA datasets using average accuracy as the primary metric. To conduct a more granular analysis, we applied the same automated classification logic used in our dataset construction to categorize the original questions into the four error types.

As shown in Table 5, the baseline DPO models yielded inconsistent improvements; while they occasionally surpassed SFT in specific categories, they frequently fell short in others. In contrast, our proposed models consistently achieved the highest accuracy across all categories, outperforming both SFT and all other DPO variants. This robust performance demonstrates that targeted preference learning on clinically grounded error types provides more reliable hallucination mitigation than general-purpose DPO strategies.

6 Conclusion

In this work, we examined the efficacy of existing DPO strategies within the medical domain. Our results demonstrated that while DPO provides moderate gains over the base model, these improvements are often indistinguishable from those of matched SFT baselines, suggesting that the benefits may stem primarily from additional supervised training rather than preference optimization itself. Furthermore, we found that DPO’s performance was highly inconsistent across various tasks and datasets, raising concerns regarding its reliability in clinical applications. Most importantly, our expert-driven analysis of visual understanding revealed that DPO-aligned models still exhibit fundamental errors in modality and anatomical identification. These results underscore the limitations of current approaches and call for more advanced, domain-specific methods that prioritize visual grounding. To this end, we provided a proof-of-concept for a targeted preference construction strategy and released our training and evaluation resources to support the development of more robust medical AI.

Limitations

It remains to be verified whether our findings generalize to more recently released models, such as HealthGPT (Lin et al., 2025) or MedGemma (Sellergren et al., 2025). Variations in base architectures and the scale of medical pre-training data across these newer models may influence how they respond to preference optimization. Our evaluation was also constrained to 7B-parameter models, primarily due to computational limitations; however, investigating the scaling effects of preference optimization on larger architectures would be a valuable direction. Furthermore, several specialized models have been developed for radiology report generation (Wu et al., 2025; Chen et al., 2024c; Lee et al., 2024), and evaluating such radiology-specific architectures might have revealed different error patterns. However, as our primary objective was to assess widely adopted models for general medicine, we selected LLaVA-Med and HuatuoGPT-Vision. Also, they represent widely used, open-source benchmarks with transparent training pipelines. This transparency was crucial for ensuring that our comparisons across DPO variants remained controlled and reproducible.

While our analysis primarily focused on visual-level errors, investigating failure modes that occur despite accurate visual recognition remains an intriguing avenue for future research. This includes challenges such as extrinsic hallucinations, overly generic descriptions, logical inconsistency in reasoning, and the degree of alignment with global medical standards. We leave the exploration of these nuanced linguistic and clinical dimensions for future work.

Lastly, our study may not comprehensively cover the full range of real-world clinical scenarios. As such, various types of errors may arise in practical settings that were not captured or analyzed within the scope of this research. Therefore, ongoing efforts toward additional validation are necessary to ensure robustness and reliability in diverse medical contexts.

Acknowledgments

This research was supported by (1) the National Research Foundation of Korea (NRF-2023R1A2C3004176), (2) the Ministry of Health & Welfare, Republic of Korea (HR20C002103), (3) ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201819), (4) the National Research Foundation of Korea(NRF) grant funded by the Korea governmant(MSIT and MOE) (RS-2025-16652968), and (5) the Seoul National University Hospital with support from the Ministry of Science and ICT (RS-2023-00262002)

References

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.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Aofei Chang, Le Huang, Parminder Bhatia, Taha Kass-Hout, Fenglong Ma, and Cao Xiao. 2025. Medheval: Benchmarking hallucinations and mitigation strategies in medical large vision-language models. *arXiv preprint arXiv:2503.02157*.

Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. 2024a. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and 1 others. 2024b. 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.

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

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Y Zou, Kai-Wei Chang, and Wei Wang. 2024. Enhancing large vision language models with self-training on image comprehension. *Advances in Neural Information Processing Systems*, 37:131369–131397.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M Cheung, Robert Chen, Ronald M Summers, Justin F Rousseau, Peiyun Ni, Marc J Landsman, and 1 others. 2024. Hidden flaws behind expert-level accuracy of multimodal gpt-4 vision in medicine. *npj Digital Medicine*, 7(1):190.

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.

Yubin Kim, Hyewon Jeong, Shen Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo R Gameiro, and 1 others. 2025. Medical hallucination in foundation models and their impact on healthcare. *medRxiv*, pages 2025–02.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. 2022. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171.

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.

Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. 2024. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. In *The Twelfth International Conference on Learning Representations*.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023c. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.

Tianwei Lin, Wenqiao Zhang, SIJING LI, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Siliang Tang, and 1 others. 2025. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In *Forty-second International Conference on Machine Learning*.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer.

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.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

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

Jie Liu, Wenxuan Wang, Yihang Su, Jingyuan Huan, Wenting Chen, Yudi Zhang, Cheng-Yi Li, Kao-Jung Chang, Xiaohan Xin, Linlin Shen, and 1 others. 2024b. A spectrum evaluation benchmark for medical multi-modal large language models. *arXiv preprint arXiv:2402.11217*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

OpenAI. 2023. Gpt-4v(ision) system card.

Yassine Ouali, Adrian Bulat, Brais Martinez, and Georgios Tzimiropoulos. 2025. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *Computer Vision – ECCV 2024*, pages 395–413, Cham. Springer Nature Switzerland.

Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37:116617–116637.

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.

Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. 2024. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. In *Medical Imaging with Deep Learning*, pages 1310–1327. PMLR.

Amir Saeidi, Shivanshu Verma, Md Nayem Uddin, and Chitta Baral. 2025. Insights into alignment: Evaluating dpo and its variants across multiple tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 409–421.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. 2024. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088.

Guankun Wang, Long Bai, Wan Jun Nah, Jie Wang, Zhaoxi Zhang, Zhen Chen, Jinlin Wu, Mobarakol Islam, Hongbin Liu, and Hongliang Ren. 2025a. Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

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.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, and 1 others. 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems*, 37:140334–140365.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2025. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25543–25551.

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*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. In *Proceedings of the 41st International Conference on Machine Learning*, pages 55204–55224.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, and 1 others. 2024a. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. 2024b. How language model hallucinations can snowball. In *International Conference on Machine Learning*, pages 59670–59684. PMLR.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2025. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1):AIoa2400640.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, and 1 others. 2024a. Pretrained multimodal large language model enhances dermatological diagnosis using skingpt-4. *Nature Communications*, 15(1):5649.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *12th International Conference on Learning Representations, ICLR 2024*.

Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. 2025. Mmedpo: Aligning medical vision-language models with clinical-aware multimodal preference optimization. In *Forty-second International Conference on Machine Learning*.

A DPO Formulations and Examples

Table A presents the mathematical formulations of IRPO, mDPO, and MMedPO, illustrating how the preference loss is defined for each configuration. Illustrative examples of preference pair curation are presented in Figures A and B.

B Completeness and Contradiction

Each model-generated output is evaluated against a set of atomic statements using GPT-4o based natural language inference (NLI) and classified into one of four classes: entailment, if the model's output supports or conveys the same factual content as the reference; partial entailment, if the output is only partially aligned with the statement, capturing some but not all aspects of the intended meaning; contradiction, if the output directly conflicts with the statement; and neutral, if the response neither confirms nor refutes the information, or fails to address it altogether.

Scores of 1, 0.5, 0, and -1 are assigned to entailment, partial entailment, neutral, and contradiction, respectively. Completeness is the average score across the entailment, partial, and neutral categories, while contradiction is the absolute average of the scores for the contradiction class, both normalized by the total number of reference atomic statements.

C MIMIC-CXR data curation

To enable precise and image-grounded evaluation, we utilized the MIMIC-CXR test set after applying the following filters: (1) only studies with a single frontal chest X-ray image were retained; (2) only the Findings section of each report was used, and reports with extremely short Findings sections were excluded due to insufficient clinical content; and (3) we used GPT-4o to generate modified versions of the reports by removing phrases that required external context—such as prior exams, patient history, institutional conventions, or physician-specific commentary. This selection enables decomposition of reports into atomic, image-verifiable facts for accurate comparison with model outputs.

D Expert Evaluation Details

The models were presented with the following prompt: "Describe the key visual features of the medical image (e.g., shape, size, location, density, contrast). Then, provide the clinical findings." Evaluators measured the accuracy of image understanding by assigning one of three severity levels for image misunderstanding: (1) None: no misinterpretation of critical visual elements, (2) Minor: small inaccuracies that do not substantially affect diagnostic reasoning, and (3) Severe: clear misinterpretation of essential features necessary for accurate clinical inference.

Two annotators with relevant medical backgrounds participated in the expert evaluation. The senior annotator is a licensed physician specializing in Physical Medicine and Rehabilitation, with years of inpatient experience managing complex comorbidities and interpreting diverse clinical data, including imaging. The second annotator is a medical student with prior experience in annotation and manual evaluation across multiple AI projects. Although our benchmarks span multiple medical domains, the evaluation did not require highly specialized expertise from pathologists or radiologists, as the task primarily involved comparing model outputs against available ground truths (e.g., radiology reports for MIMIC-CXR and image captions for AMBOSS).

A calibration session was conducted prior to annotation to align evaluation standards. To quantify annotation consistency, we computed inter-rater reliability (Cohen's κ) over 30 model-generated responses. Overall agreement was 0.9 for MIMIC-CXR and was 0.878 for AMBOSS, indicating strong reliability and consensus.

E Details of Enhanced DPO Experiments

Keyword lists We curated comprehensive keyword lists for each error category to facilitate automated preference pair construction. These keywords serve as the basis for identifying critical clinical entities within the ground-truth instructions and responses. The specific keywords associated with each error type are detailed below, illustrating the scope of our targeted clinical entity extraction.

Modality Misidentification (MM)

Representative keywords:

- CT, computed tomography, MRI, MR, T1, T2, FLAIR, DWI, SWI
- X-ray, radiograph, CXR, ultrasound, US, sonography, echocardiogram, echo
- PET, SPECT, angiography, fluoroscopy, mammography
- fundus, ophthalmoscopy, dermatoscopy, endoscopy,

Method	Objective
IRPO (Pang et al., 2024)	$\mathcal{L}_{\text{IRPO}}(y_w, y_l \mid m, q) = \mathcal{L}_{\text{DPO}}(y_w, y_l \mid m, q) + \alpha \cdot \mathcal{L}_{\text{NLL}}(y_w \mid m, q)$ $= -\log\sigma\Big(\beta \log \frac{\pi_\theta(y_w \mid m,q)}{\pi_{\text{ref}}(y_w \mid m,q)} - \beta \log \frac{\pi_\theta(y_l \mid m,q)}{\pi_{\text{ref}}(y_l \mid m,q)}\Big) - \alpha \cdot \frac{\log \pi_\theta(y_w \mid m,q)}{\lvert y_w \rvert}$
mDPO (Wang et al., 2024)	$\mathcal{L}_{\text{mDPO}} = \mathcal{L}_{\text{DPO}_m} + \mathcal{L}_{\text{CoPO}} + \mathcal{L}_{\text{AncPO}}$ $= -\log\sigma\Big(\beta \log \frac{\pi_\theta(y_w \mid m,q)}{\pi_{\text{ref}}(y_w \mid m,q)} - \beta \log \frac{\pi_\theta(y_l \mid m,q)}{\pi_{\text{ref}}(y_l \mid m,q)}\Big)$ $-\log\sigma\Big(\beta \log \frac{\pi_\theta(y_w \mid m_w,q)}{\pi_{\text{ref}}(y_w \mid m_w,q)} - \beta \log \frac{\pi_\theta(y_w \mid m_l,q)}{\pi_{\text{ref}}(y_w \mid m_l,q)}\Big)$ $-\log\sigma\Big(\beta \log \frac{\pi_\theta(y_w \mid m_w,q)}{\pi_{\text{ref}}(y_w \mid m_w,q)} - \delta\Big)$
MMedPO (Zhu et al., 2025)	$\mathcal{L}_{\text{MMedPO}} = s' \cdot \Big[-\log\sigma\Big(\alpha \log \frac{\pi_\theta(y_w \mid m_w,q)}{\pi_{\text{ref}}(y_w \mid m_w,q)} - \alpha \log \frac{\pi_\theta(y_l \mid m_l,q)}{\pi_{\text{ref}}(y_l \mid m_l,q)}\Big)\Big]$

Table A: Mathematical formulations of IRPO, mDPO, and MMedPO.

colonoscopy, EGD, gastroscopy

- microscopy, H&E, hematoxylin, eosin, electron microscopy, OCT

Anatomical Misidentification (AM)

Representative keywords (examples):

- *Thorax*: lung, lobe, segment, pleura, mediastinum, cardiomediastinum, diaphragm, rib, clavicle
- *Abdomen*: liver, spleen, kidney, adrenal, pancreas, stomach, bowel, colon, rectum
- *Head/Neck*: brain, cerebellum, ventricle, skull, orbit, sinus, maxillary, ethmoid, sphenoid, frontal, nasal, septum, tonsil, pharynx, larynx
- *Extremities/Skin*: arm, leg, hand, foot, femur, humerus, radius, ulna, tibia, fibula, hip, knee, ankle, wrist, skin, dermis, epidermis

Spatial or Laterality Confusion (SLC)

Representative keywords:

- Laterality: left, right, left-sided, right-sided
- Zones: upper, lower, superior, inferior, anterior, posterior, medial, lateral, apical, basal
- Lung subregions: RUL, RML, RLL, LUL, LLL, upper lobe, middle lobe, lower lobe

Lack of Anatomical Specificity (LAS)

Representative keywords:

- Fine-grained: segment numbers (S1, S2, ...), right lower lobe, left upper lobe, quadrant (RUQ, LUQ, RLQ, LLQ), pole
- Broad parents: lung, liver, kidney, sinus, paranasal sinus, brain, hemithorax

VQA Subsets To evaluate the model's robustness against specific types of hallucinations, we constructed specialized evaluation subsets from established VQA benchmarks. By applying the keyword-based classification logic described above, we partitioned original VQA questions into four distinct categories: MM, AM, SLC, and LAS. This fine-grained evaluation framework allows us to analyze whether performance gains are consistent across different clinical dimensions or localized to specific error types. Table B summarizes the statistics of the VQA subsets.

Error Type	SLAKE	VQA-RAD	PathVQA	Total
MM	140	51	366	557
SLC	211	98	213	522
AM	698	267	5268	6233
LAS	83	58	672	813

Table B: Screened question-answer pairs per error type and dataset.

F Hyperparameter Tuning

SFT and DPO rely on different training objectives and, by design, their training data are not identical. Specifically, SFT uses instruction-response pairs, whereas DPO uses preference pairs where the chosen (m_w, y_w) matches the SFT data but requires additional curation of rejected responses. To enable a fair comparison, we therefore conducted independent hyperparameter searches using the SLAKE validation set to identify the best-performing settings for each method (Table C). Based on these analyses, we selected the following settings for our main experiments; for SFT, we used a learning rate 2e-6 with 3 epochs, and for the DPO models, we used a learning rate 1e-7 with 3 epochs.

We further evaluated the effect of varying the

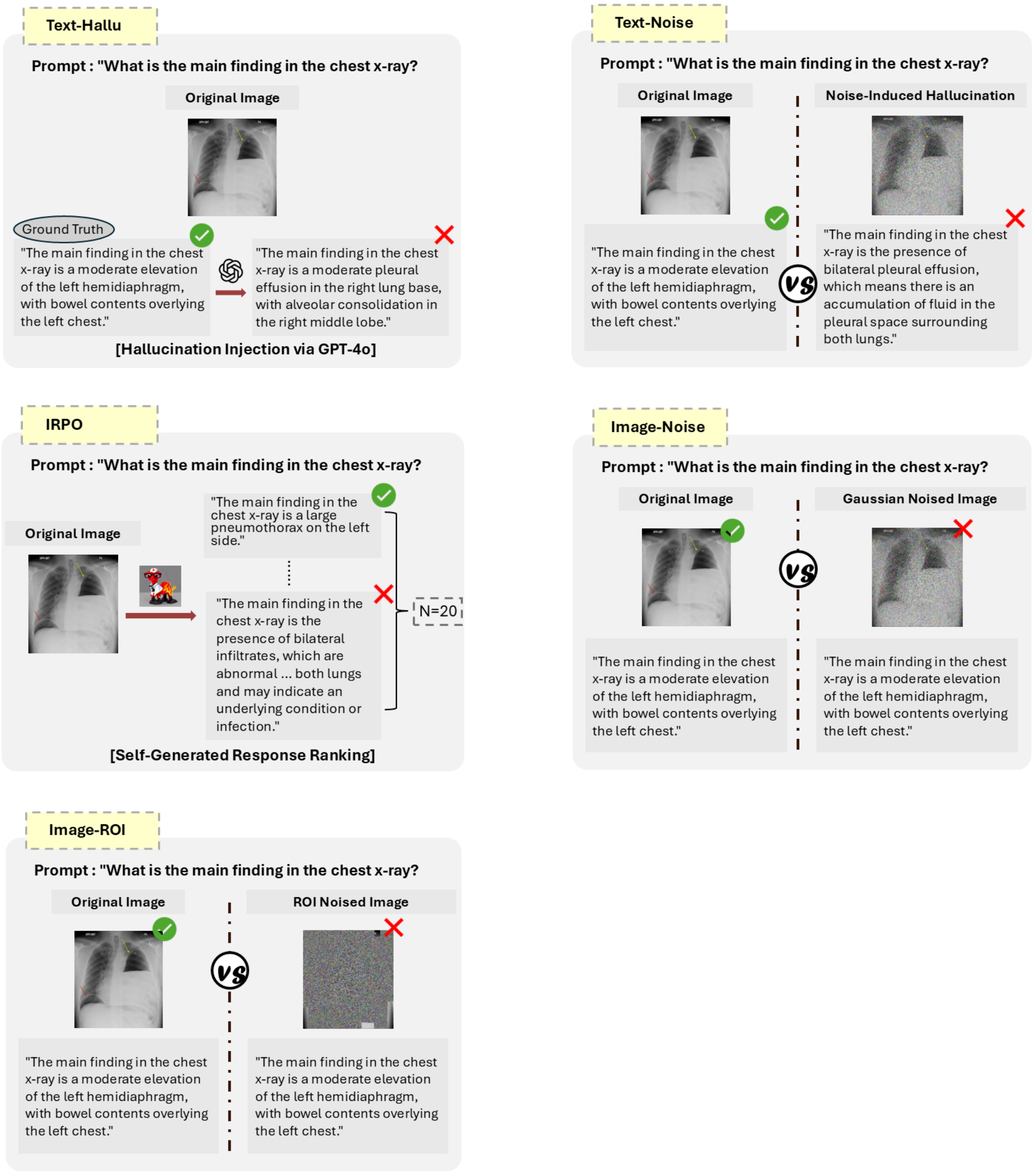
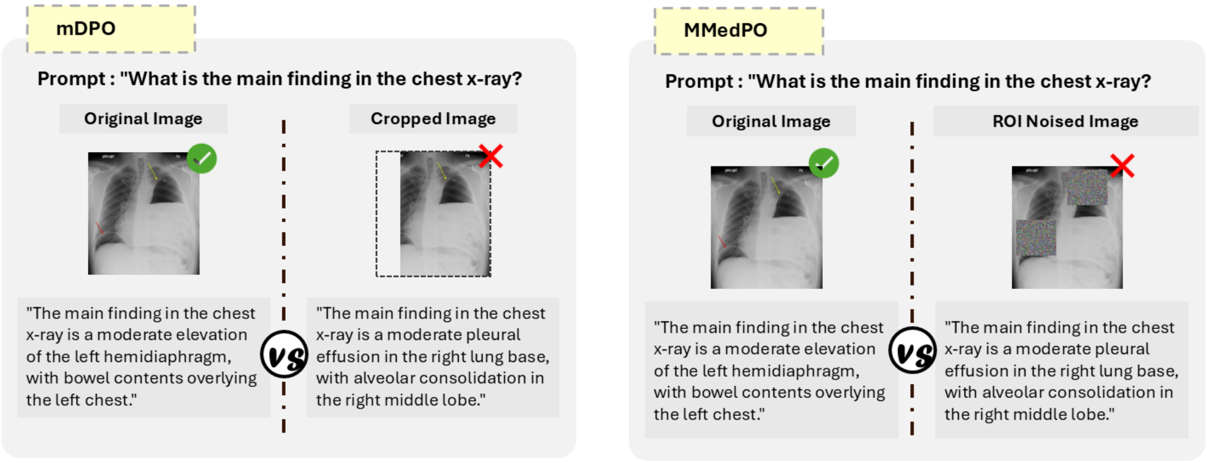

Text-Hallu
Prompt : "What is the main finding in the chest x-ray?
Original Image
Ground Truth
"The main finding in the chest x-ray is a moderate elevation of the left hemidiaphragm, with bowel contents overlying the left chest."
"The main finding in the chest x-ray is a moderate pleural effusion in the right lung base, with alveolar consolidation in the right middle lobe."
[Hallucination Injection via GPT-4o]
Text-Noise
Prompt : "What is the main finding in the chest x-ray?
Original Image
Noise-Induced Hallucination
"The main finding in the chest x-ray is a moderate elevation of the left hemidiaphragm, with bowel contents overlying the left chest."
VS
"The main finding in the chest x-ray is the presence of bilateral pleural effusion, which means there is an accumulation of fluid in the pleural space surrounding both lungs."
IRPO
Prompt : "What is the main finding in the chest x-ray?
Original Image
"The main finding in the chest x-ray is a large pneumothorax on the left side."
"The main finding in the chest x-ray is the presence of bilateral infiltrates, which are abnormal ... both lungs and may indicate an underlying condition or infection."
N=20
[Self-Generated Response Ranking]
Image-Noise
Prompt : "What is the main finding in the chest x-ray?
Original Image
Gaussian Noised Image
VS
"The main finding in the chest x-ray is a moderate elevation of the left hemidiaphragm, with bowel contents overlying the left chest."
"The main finding in the chest x-ray is a moderate elevation of the left hemidiaphragm, with bowel contents overlying the left chest."
Image-ROI
Prompt : "What is the main finding in the chest x-ray?
Original Image
ROI Noised Image
VS
"The main finding in the chest x-ray is a moderate elevation of the left hemidiaphragm, with bowel contents overlying the left chest."
"The main finding in the chest x-ray is a moderate elevation of the left hemidiaphragm, with bowel contents overlying the left chest."

Figure A: Illustrative examples of preference pair curation in text-only and image-only DPO variants.

number of training epochs (Table D), confirming that performance gains are not attributable to additional training alone.


mDPO
Prompt : "What is the main finding in the chest x-ray?
Original Image
Cropped Image
"The main finding in the chest x-ray is a moderate elevation of the left hemidiaphragm, with bowel contents overlying the left chest."
VS
"The main finding in the chest x-ray is a moderate pleural effusion in the right lung base, with alveolar consolidation in the right middle lobe."
MMedPO
Prompt : "What is the main finding in the chest x-ray?
Original Image
ROI Noised Image
"The main finding in the chest x-ray is a moderate elevation of the left hemidiaphragm, with bowel contents overlying the left chest."
VS
"The main finding in the chest x-ray is a moderate pleural effusion in the right lung base, with alveolar consolidation in the right middle lobe."

(a) mDPO

(b) MMedPO

Figure B: Illustrative examples of preference pair curation in joint image-text DPO variants, mDPO and MMedPO.

lr	1 ep	2 ep	3 ep
SFT			
1e-7	0.45	0.45	0.45
5e-7	0.44	0.45	0.45
1e-6	0.42	0.45	0.45
2e-6	0.41	0.45	0.44
2e-5	0.41	0.37	0.37
Text-Hallu + NLL			
2e-8	0.44	0.45	0.45
1e-7	0.46	0.45	0.47
1e-6	0.45	0.47	0.47

Table C: Hyperparameter search results for LLaVA-Med on SLAKE validation set.

Model	1 ep	2 ep	3 ep	4 ep	5 ep
LLaVA-Med					
Base Model	–	–	0.39	–	–
SFT	0.38	0.41	0.41	0.41	0.42
Text-Hallu	0.41	0.42	0.41	0.38	0.35
+ NLL	0.41	0.42	0.42	0.36	0.38
Text-Noise	0.40	0.39	0.39	0.34	0.37
+ NLL	0.41	0.41	0.41	0.36	0.37
IRPO	0.39	0.39	0.39	0.33	0.37
Image-Noise	0.40	0.40	0.40	0.36	0.37
Image-ROI	0.41	0.41	0.41	0.34	0.38
mDPO	0.41	0.42	0.42	0.38	0.39
MMedPO	0.40	0.39	0.40	0.32	0.33
HuatuoGPT-Vision					
Base Model	–	–	0.49	–	–
SFT	0.51	0.50	0.52	0.51	0.52
Text-Hallu	0.52	0.51	0.53	0.51	0.49
+ NLL	0.52	0.51	0.52	0.49	0.50
Text-Noise	0.52	0.50	0.51	0.53	0.51
+ NLL	0.50	0.52	0.52	0.52	0.51
IRPO	0.51	0.50	0.53	0.53	0.51
Image-Noise	0.52	0.51	0.50	0.52	0.52
Image-ROI	0.51	0.51	0.52	0.50	0.50
mDPO	0.52	0.51	0.51	0.54	0.53
MMedPO	0.51	0.53	0.52	0.52	0.53

Table D: Performance comparison across epochs for LLaVA-Med and HuatuoGPT-Vision.