Towards Reliable Medical Large Vision-Language Models via Counterfactual Preference Optimization

Xiaoguang Zhu[1], Naipeng Wang[2], Kartik Patwari[1], Lianlong Sun[3], Chen-Nee Chuah[1], and Chengxin Pang[2⋆]

[1] University of California, Davis, Davis, CA, USA
{xgzhu,kpatwari,chuah}@ucdavis.edu
[2] Shanghai University of Electric Power, Shanghai, China
{naipengwang,chengxinpang}@mail.shiep.edu.cn
[3] University of Rochester, Rochester, NY, USA
lianlongsun@rochester.edu

Abstract. Medical Large Vision-Language Models (Med-LVLMs) have emerged as powerful tools for medical image understanding and automated reporting. However, existing models often suffer from modality bias, where correlations between background context and diagnostic labels cause over-reliance on spurious priors rather than lesion evidence, leading to clinically irrelevant or hallucinated outputs. To address this, we propose a **Co**unterfactual **Med**ical **P**reference **O**ptimization (**CoMedPO**) framework that learns an unbiased policy from a biased reference model. Unlike conventional direct preference optimization (DPO) methods that inherit dataset biases, CoMedPO mitigates spurious correlations while preserving the valuable indirect effect between lesion and background dependencies in medical images. This causal formulation yields a new preference-based loss with theoretical guarantees for bias mitigation and robust policy learning. Our framework is model-agnostic and seamlessly integrates with existing Med-LVLM alignment pipelines. Extensive experiments on Med-VQA and medical report generation demonstrate that CoMedPO consistently improves factual accuracy and clinical relevance, outperforming state-of-the-art DPO variants. These results suggest that CoMedPO provides a principled route toward trustworthy and clinically grounded multimodal medical reasoning. Our code is available at https://github.com/zxgapollo/CoMedPO.

1 Introduction

The emergence of Large Vision Language Models (LVLMs) has advanced multimodal reasoning by enabling unified understanding of visual and textual modalities [20–22, 57]. Trained on large-scale image–text corpora, these models exhibit strong cross-modal generalization and reasoning capabilities across various tasks.

⋆ Corresponding author.

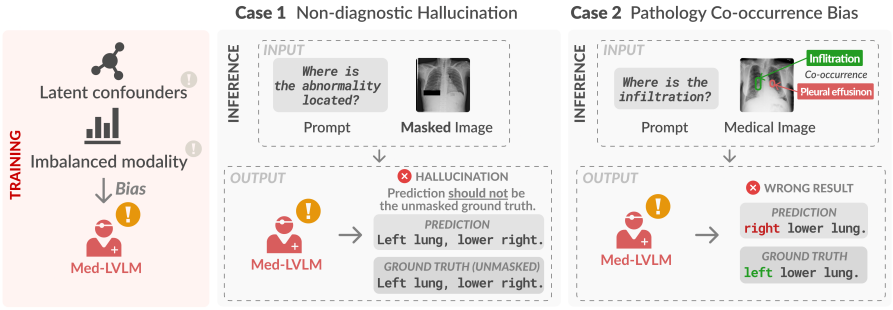

Fig. 1: Effects of spurious correlations on inference. (a) Correct prediction given background image. (b) Co-occurrence of multiple pathologies affects the prediction.

In the medical domain, Medical LVLMs (Med-LVLMs) demonstrate promising capabilities in medical visual question answering [16, 52], radiology report generation [8, 14], and multimodal reasoning [17, 30], offering potential for decision support and automated documentation. Despite these advances, factual reliability remains a critical bottleneck for Med-LVLMs, primarily due to modality misalignment between vision and text [7, 53, 56]. Unlike natural images, medical images often contain highly localized and subtle lesion patterns that are easily obscured by coarse global alignments. As a result, models frequently over-rely on textual priors or contextual shortcuts rather than lesion-based evidence, leading to hallucinations that compromise the factual reliability required for high-stakes clinical decision-making [37].

To address the factuality problem, recent works have adopted Direct Preference Optimization (DPO) to enhance image–text alignment in Med-LVLMs after the supervised fine-tuning (SFT) stage by learning from pairwise clinical preference data [2,36,56]. However, most existing approaches rely on preference-generation pipelines originally designed for general-domain LVLMs [2, 36, 56]. Directly adopting these frameworks for the medical domain neglects the unique constraints of clinical data, where diagnostic precision and lesion localization are paramount. While recent work has begun incorporating clinical relevance by weighting preference samples according to medical significance [58], it heavily relies on statistical associations between textual cues and disease labels, overlooking the spurious causation introduced by biased modality interactions. The absence of an explicit mechanism to disentangle lesion evidence from contextual cues leaves the resulting models vulnerable to shortcut reasoning rather than robust, clinically grounded inference. Consequently, these optimized Med-LVLMs continue to depend on spurious correlations rather than true clinical reasoning, again limiting their factual reliability in high-stakes diagnostic scenarios.

Clinical reliability in Med-LVLMs requires that diagnostic conclusions be supported by lesion evidence, instead of relying on dataset-specific correlations or background cues. However, existing Med-LVLMs often suffer from spurious correlations, where predictions are influenced by latent confounders or imbalanced modality signals. This issue manifests in two primary ways. First, models

develop biased statistical associations with non-diagnostic background regions. As shown in Fig. 1(a), a model may correctly label a disease even when provided only with non-diagnostic background, indicating a dangerous over-reliance on contextual priors instead of lesion evidence. Second, pathology co-occurrence bias leads models to exploit coincidental features where correlated abnormalities lead the model to exploit coincidental features during learning. For instance, as shown in Fig. 1(b), when "pleural effusion" and "infiltration" frequently overlap in training data, the model may incorrectly attribute infiltration features to the presence of an effusion. Such dependencies compromise factual integrity, making the mitigation of spurious modality interactions a prerequisite for trustworthy clinical reasoning.

A principled solution to spurious correlations would be to conduct a randomized controlled trial, where lesion and background factors are independently varied to reveal the true causal effect of lesion evidence on diagnostic outcomes [42]. However, in practice, such balanced sampling is infeasible and requires controlled patient recruitment and extensive expert annotation. Recent work, such as MMedPO [58], attempt to simulate this intervention by introducing local noise into lesion areas. However, this strategy remains sub-optimal as it fails to decouple beneficial background–lesion interactions from harmful ones. In medical imaging, the background context often provides essential structural priors. For example, distinguishing a "pleural effusion" from "infiltration" relies not only on local lesion morphology but also on global spatial constraints like pleural line continuity [50]. Consequently, an effective framework must decouple harmful statistical shortcuts while preserving these valuable structural dependencies.

To overcome the above limitations, we propose **Co**unterfactual **Med**ical **P**reference **O**ptimization framework (**CoMedPO**), which explicitly decouples the harmful direct effect of spurious modality interactions from the valuable indirect effect of lesion and background dependencies. This distinction is critical in the medical domain, where contextual cues often provide essential spatial or anatomical constraints for diagnostic reasoning. Specifically, we formulate a causal graph that connects multimodal inputs to the reward function, enabling theoretical analysis of how biased reference policies influence downstream alignment. Based on this formulation, CoMedPO derives a principled objective that improves robustness even when the SFT-based reference model is biased. Within this causal structure, we explicitly decouple the harmful direct effect, which transmits bias from shortcut correlations, and the valuable indirect effect, which captures the clinically meaningful joint influence of lesion and contextual cues on diagnostic reasoning. During preference construction, we build factual and counterfactual outcome pairs by employing a lesion detection model to identify clinically relevant regions. The factual outcomes preserve the normal lesion–background association, while the counterfactual ones intervene on lesion evidence to isolate indirect effects. In the inference phase, the model is trained on normal preference pairs, ensuring consistency with standard alignment while benefiting from causal debiasing during optimization. CoMedPO can be seamlessly integrated with existing DPO-based alignment methods to further enhance

the performance of Med-LVLMs, serving as a general and model-agnostic optimization paradigm. In summary, our main contributions are threefold:
• We introduce a causal framework for Med-LVLM alignment that exposes how biased reference policies propagate harmful direct effects from spurious causation within modality interactions.
• We propose CoMedPO, a model-agnostic debiasing approach that mitigates spurious causal effects while preserving the valuable indirect effect between lesion and background features. CoMedPO constructs factual–counterfactual outcome pairs based on lesion detection and can be seamlessly integrated with existing DPO methods to enhance robustness.
• Extensive experiments on Med-VQA and medical report generation benchmarks demonstrate that CoMedPO consistently improves factual accuracy and clinical relevance across multiple Med-LVLMs, highlighting its broad applicability and effectiveness in trustworthy multimodal medical reasoning.

2 Related Works

Med-LVLMs Preference Optimization. Preference optimization is widely used to align large vision–language models (LVLMs) with human or model preferences [6, 9, 11, 25, 26, 29, 32, 38, 47], with more details discussed in Appendix S7. Extending such techniques to the medical domain introduces additional challenges, as factual grounding and clinical relevance are essential. Methods such as [3, 10, 41] generate dispreferred responses via GPT-4 or the Med-LVLM for preference optimization. Medical alignment emphasizes clinical grounding and lesion-aware signals. MMedPO [58] proposes clinically aware weighting to preference samples to highlight key medical cues, while MedGR^2 [54] uses generative reward learning for robust medical reasoning. Despite these advances, transferring general-domain pipelines still leaves gaps: models over-rely on textual priors and spurious background associations under imbalanced modality interactions [7, 37]. MMedPO attempts to address these issues by weighting samples based on clinical relevance, but relies on computationally heavy multi-agent scoring and does not preserve useful lesion–background priors. Guided by these observations, our work introduces a causality-aware preference construction to explore lesion evidence while preserving clinically meaningful multimodal cues.
Causal Reference. Causal inference [33] uncovers cause–effect mechanisms through intervention and counterfactual reasoning. Interventions modify distributions for unbiased estimates, while counterfactuals imagine hypothetical outcomes for alternative conditions. Recent learning-based approaches incorporate causal reasoning into deep models to reduce spurious shortcuts and improve generalization across vision tasks [18, 40, 55, 59], including anomaly detection [27], visual recognition [24], recommendation [45, 46], and scene graph generation [23]. In multimodal models, causal modeling is explored to improve robustness and interpretability. Wang *et al.* [43] propose causal reward modeling to prevent reward hacking, but their framework is limited to unimodal text and does not extend to medical multimodal reasoning. Zhou *et al.* [49] integrate structural causal

models into medical VQA to interpret visual–text relations, yet their front-door adjustment requires heavy sampling and fails to preserve lesion–background dependencies, rendering it suboptimal for multimodal alignment. Building on these insights, we go beyond prior counterfactual bias mitigation in discriminative VQA [31] by casting counterfactual reasoning within a reinforcement learning framework for MLLM alignment and establishing a principled connection between causal reward contrasts and preference-based objectives, with theoretical analysis. By constructing factual and counterfactual pairs, our framework suppresses harmful background direct effects while preserving the valuable indirect effect that captures clinically meaningful lesion–background dependencies.

3 Preliminaries

A Med-LVLM combines a visual encoder with an LLM to process multimodal medical inputs. Given a medical image I and a clinical query T, the input is $x = (I, T)$, and the model autoregressively predicts next token probabilities $\pi_\theta(y_i|y_{<i}, x)$, yielding sequence likelihood $\pi_\theta(y|x) = \prod_i \pi_\theta(y_i|y_{<i}, x)$, where y is the generated medical report or answer.

3.1 Preference Optimization

Preference optimization aligns the conditional policy using pairwise preference supervision, x, y_ω, y_l. DPO [36] directly connects policy with the reward difference between preferred and dispreferred responses. To remove the explicit reward model, DPO reformulates this closed-form expression as a supervised learning objective. Given a preference dataset $\mathcal{D} = \{(x^{(i)}, y_w^{(i)}, y_l^{(i)})\}_{i=1}^N$, where y_w and y_l denote the preferred and dispreferred responses for input x, the DPO loss is

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log \sigma\Big(\eta^{-1} \log \tfrac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \eta^{-1} \log \tfrac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Big)\right]. \quad (1)$$

We provide the complete derivation in Appendix S1.1.

3.2 Causal Counterfactual Learning

We adopt the standard structural causal model (SCM) framework [33], where a causal graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consists of chains $A \to M \to O$ and $A \to O$. A is the cause, M the mediator, and O the outcome, the counterfactual outcome is denoted as $O_{a,m} = O(A{=}a, M{=}m)$.

The total causal effect of A is defined as $\text{TE} = O_{a,M_a} - O_{a^*,M_{a^*}}$. TE can be decomposed into the natural direct effect (NDE) and total indirect effect (TIE), where $\text{NDE} = O_{a,M_{a^*}} - O_{a^*,M_{a^*}}$ and $\text{TIE} = O_{a,M_a} - O_{a,M_{a^*}}$. Full derivations and causal semantics are provided in Appendix S1.2.

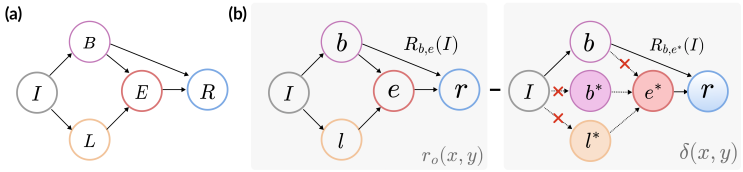

Fig. 2: Comparison of (a) Med-LVLM SCM and (b) counterfactual Med-LVLM SCM.

4 Methodology

4.1 Cause-Effect Look in Preference Optimization

Causal Analysis. As illustrated in Fig. 2, the proposed causal graph for medical multimodal alignment consists of four key components: the medical image input I, its decomposed visual factors including the background B and lesion L, the learned joint features E, and the resulting reward model R. Based on the causal dependencies among these variables, we summarize the major links as follows.
Link $I \rightarrow B \rightarrow R$ captures the harmful shortcut between the original input image and the reward through biased background correlations. Background factors B (e.g., scanner type, imaging device, or patient posture) may be spuriously correlated with certain labels or textual descriptions, leading the model to achieve high reward even when clinically irrelevant regions dominate the decision [51].
Link $B \leftarrow I \rightarrow L$ depicts the generation of both background and lesion representations from the same medical image input. I is processed by the visual encoder to produce background features B and lesion features L together. B corresponds to global background such as lung field structure or imaging device signatures, while L represents suspicious regions detected by attention or segmentation models.
Link $B/L \rightarrow E \rightarrow R$ captures the total causal influence of visual representations on the reward through the embedding E, which aggregates both background and lesion features learned from the visual encoder. In particular, B provides the valuable contextual prior along the good causal link $B \rightarrow E \rightarrow R$, which facilitates reliable recognition when lesion information are subtle or visually confounded by surrounding anatomical structures. For example, in chest radiography, global contextual cues such as lung symmetry, cardiac silhouette, or diaphragm curvature help disambiguate faint opacities and improve diagnostic localization [12]. This clinically grounded pathway allows CoMedPO to preserve joint information while mitigating harmful direct effects from B on R.

Causal Reward. To isolate clinically meaningful signals from biased supervision, we model the causal effect of the background factor B on the reward R through the ensemble mediator E. Specifically, the factual reward generation process under a given input image I can be expressed as:

$$R_{b,e}(I) = R(B = b, E_{b,l} = E(B = b, L = l) \mid I). \tag{2}$$

The factual reward $R_{b,e}(I)$ inherently mixes both the useful lesion-grounded information and the harmful direct effect from B. To quantify these distinct contributions, we first define the Total Effect (TE) as:

$$\mathrm{TE} = R_{b,e}(I) - R_{b^*,e^*}(I), \tag{3}$$

where b^* and e^* correspond to the counterfactual background and its associated ensemble embedding. We next estimate the Natural Direct Effect (NDE) of B for the pure harmful background bias:

$$\mathrm{NDE} = R_{b,e^*}(I) - R_{b^*,e^*}(I). \tag{4}$$

$R_{b,e^*}(I)$ denotes the counterfactual outcome:

$$R_{b,e^*}(I) = R(B = b, E_{b^*,l^*} = E(B = b^*, L = l^*) \mid I). \tag{5}$$

As the indirect effect of embedding E on the link $I \to B/L \to E \to R$ is blocked, the reward model can only be biased relying on the direct effect in the link $I \to B \to R$. To mitigate the background bias in NDE, we compute the Total Indirect Effect (TIE) by subtracting NDE from TE:

$$\mathrm{TIE} = R_{b,e}(I) - R_{b,e^*}(I). \tag{6}$$

Intuitively, TIE represents the reward improvement driven by medically relevant background–lesion interactions. The proof of the identifiability of the causal TIE reward is provided in Appendix S2.

4.2 Theoretical Analysis

To isolate clinically meaningful signals from biased supervision, we model the causal effect of the background factor B on the reward R through the ensemble mediator E. Specifically, the factual reward generation process under a given input image I can be expressed as:

$$R_{b,e}(I) = R(B = b, E_{b,l} = E(B = b, L = l) \mid I). \tag{7}$$

Following Eq. (6), the TIE-form causal reward is expressed as:

$$r_c(x, y) := R_{b,e}(x, y) - R_{b,e^*}(x, y). \tag{8}$$

In standard DPO, the optimization is based on the relative preference difference between a preferred sample y_w and a dispreferred one y_l, where the reward difference $r(x, y_w) - r(x, y_l)$ determines the training direction. However, such rewards may still encode background-related biases that affect both positive and negative samples. To obtain a bias-free causal preference, we substitute r_c into the Bradley–Terry model, the preference probability between (y_w, y_l) becomes

$$P(y_w \succ y_l \mid x) = \sigma(\eta\,[\,r_c(x, y_w) - r_c(x, y_l)\,])\,, \tag{9}$$

where $\sigma(\cdot)$ is the logistic function. The difference $r_c(x, y_w) - r_c(x, y_l)$ serves as a *causal contrast* that reflects the pure clinical contribution through the mediator E. Expanding this difference yields

$$r_c(x, y_w) - r_c(x, y_l) = \underbrace{[R_{b,e}(x, y_w) - R_{b,e}(x, y_l)]}_{\text{factual difference}} - \underbrace{[R_{b,e^*}(x, y_w) - R_{b,e^*}(x, y_l)]}_{\text{counterfactual difference}}. \quad (10)$$

This symmetric structure indicates that both the positive and negative samples are corrected by the same causal mechanism. If the correction were applied only to the preferred response y_w, the residual background effect on y_l would remain, and the pairwise reward difference would still contain the direct bias from B.

In preference optimization, the reference policy π_{ref} is typically obtained from a SFT model that aligns to instruction-following data. However, π_{ref} may carry bias from the underlying training distribution, causing modality shortcuts in medical preference learning. Our goal is to show that the proposed optimization goal can theoretically correct these biases and converge to a causal-optimal policy. Formally, we define the causal-optimal policy and provide a convergence guarantee based on the unbiased TIE reward. Let π_{ref} be an arbitrary reference policy. The observed reward can be decomposed as

$$r_o(x, y) \;=\; r_c(x, y) + \delta(x, y),$$

where r_c is the causal TIE reward and δ denotes the bias reward caused by the direct background effect (Link $B \to R$).

Definition 1 (Causal-optimal policy). *Given a biased reference policy π_{ref} and the unbiased TIE reward $r_c(x, y) = R_{b,e}(I) - R_{b,e^*}(I)$, the* causal-optimal policy *is defined as*

$$\pi_c^\star(y \mid x) \propto \pi_{ref}(y \mid x) \exp\big(\eta\, r_c(x, y)\big).$$

Connection to DPO optimality. DPO admits a closed-form optimum for a given reward $r(x, y)$: $\pi^\star(y \mid x) \propto \pi_{\text{ref}}(y \mid x) \exp\big(\eta\, r(x, y)\big)$. Therefore, substituting the unbiased reward r_c directly yields the *causal-optimal policy*. This shows that our method does not alter the DPO optimization principle, but instead replaces the biased reward with a causally adjusted one.

Theorem 1 (Convergence of causal reward). *Under the assumption that all rewards are bounded by $|r_c|, |\delta| \le R_{\max}$, the bias magnitude satisfies $\|\delta\|_\infty \le \varepsilon$, and the learning rate η is sufficiently small, the policy sequence updated by TIE-DPO satisfies*

$$\mathbb{E}\big[\mathrm{KL}(\pi_{T+1} \| \pi_c^\star)\big] \le \mathrm{KL}(\pi_1 \| \pi_c^\star) - \eta \sum_{t=1}^{T} \mathbb{E}_{x,e}\big[\mathrm{Var}_{\pi_t}(r_c)\big] + O(T\eta^2 R_{\max}^2),$$

which ensures a monotonic improvement toward the causal-optimal policy $\pi_c^\star$.

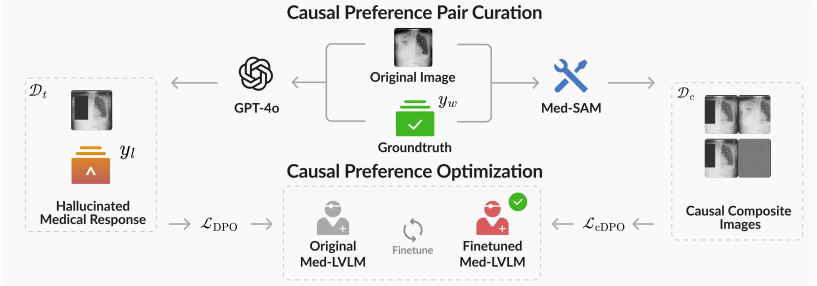

Fig. 3: The overview of CoMedPO that consists of the causal preference pair curation and counterfactual preference optimization.

The bias term δ amplifies spurious correlations between background and reward, causing the exponentiation policy update to deviate from the causal optimum. This does not imply that r_c universally dominates r_o. Rather, under the presence of direct background bias $\delta \neq 0$, optimizing r_c eliminates the exponentiated bias term and therefore produces a policy that is closer to the causal-optimal solution in KL divergence. Full derivations and proofs are provided in Appendix S3.

4.3 Implementation Instantiation

Causal Preference Pair Curation. To realize the causal preference structure defined in Eq.(10), we curate paired data that reflect both factual and counterfactual medical reasoning, where the whole framework is illustrated in Fig. 3. The construction proceeds in two stages.

Generating Hallucinated Medical Responses. Following the causal contrastive formulation, we first obtain dispreference responses y_l by introducing factual inconsistencies into medical reports, following the approach in MMedPO [58]. Specifically, we employ GPT-4o [1] to generate multiple candidate responses for each case sampled from the target Med-LVLMs, and evaluate their factual consistency against the ground-truth response y_w. The candidate exhibiting the highest hallucination score, such as incorrect image interpretation or misleading diagnostic reasoning, is selected as the dispreference response. When no candidate response exhibits sufficient deviation, GPT-4o is prompted to synthesize a dispreference response. The preference pairs constructed are denoted as $\mathcal{D}_t$.

Constructing Causal Composite Images. To explicitly instantiate the causal mechanism illustrated in our causal graph, we construct composite visual inputs that allow controlled intervention on lesion and background area. Specifically, we employ a medical visual localization tool to segment the lesion area from the original image I. Such model-assisted localization provides coarse visual grounding that guides LVLMs toward clinically relevant regions, which has been shown to improve reasoning in medical vision-language models [4, 22, 58]. The lesion mask is then removed and filled with zero-valued pixels to form a background-only image I_b, representing the background variable $B{=}b$ in the structural causal

model. The original image I serves as the factual visual condition associated with the denoted $E{=}e$, while a Gaussian noise-perturbed version of the same image serves as the counterfactual condition $E{=}e^*$.

These components are combined into composite image pairs that preserve the same anatomical context but differ in the lesion evidence, forming the factual and counterfactual visual conditions: $(I_b \oplus I)$ and $(I_b \oplus I_{\text{null}})$, where $\oplus$ denotes image stitching. Each pair is associated with a clinical query prepended by an instruction emphasizing that both images come from the same patient, such as "*Both images belong to the same patient. Please analyze each and then provide a joint clinical conclusion*", where the revised clinical query is denoted as T'. The preference data constructed in this stage is denoted as $\mathcal{D}_c$. By aligning the composite inputs with their preferred and dispreference textual responses, we obtain the overall preference dataset $\mathcal{D}_f = \mathcal{D}_t \cup \mathcal{D}_c = \left\{(x^{(i)}, x^{*(i)}, x_c^{(i)}, x_c^{*(i)}, y_w^{(i)}, y_l^{(i)})\right\}_{i=1}^{N}$, where $x^{(i)} = (I, T)$ and $x^{*(i)} = (I_b, T)$ are the standard preference pairs used in MMedPO, while $x_c^{(i)} = (I_b \oplus I, T')$ and $x_c^{*(i)} = (I_b \oplus I_{\text{null}}, T')$ are the causal pairs. The composite image construction can be interpreted as performing a controlled intervention on lesion evidence while preserving the original background.

Counterfactual Preference Optimization. As the factual and counterfactual policies are latent KL-regularized MaxEnt optima, we substitute the reward-policy duality $R(x, y) = \eta^{-1} \log \frac{\pi(y|x)}{\pi_{\text{ref}}(y|x)} + \eta^{-1} \log Z(x)$ into Eq. (8). The partition functions $Z(\cdot)$ cancel out in the pairwise difference. This derivation relies on the assumption that fitting the same optimized policy π_θ provides a consistent surrogate for the causal reward contrast. Therefore, we approximate the causal reward contrast using the log-likelihood difference of the same policy evaluated under factual and counterfactual inputs. The counterfactual optimization objective is

$$\mathcal{L}_{\text{cDPO}} = -\mathbb{E}_{(x_c, x_c^*, y_w, y_l) \sim \mathcal{D}_f} \left[\log \sigma\big(\eta^{-1} \Lambda_{\text{cDPO}}\big) \right], \tag{11}$$

where

$$\Lambda_{\text{cDPO}} = \left[\log \frac{\pi_\theta(y_w|x_c)}{\pi_{\text{ref}}(y_w|x_c)} - \log \frac{\pi_\theta(y_w|x_c^*)}{\pi_{\text{ref}}(y_w|x_c^*)} \right] - \left[\log \frac{\pi_\theta(y_l|x_c)}{\pi_{\text{ref}}(y_l|x_c)} - \log \frac{\pi_\theta(y_l|x_c^*)}{\pi_{\text{ref}}(y_l|x_c^*)} \right]. \tag{12}$$

Note that the original MMedPO [58] preference optimization goal is kept to ensure Med-LVLM has the ability to make predictions according to the original visual and clinical query input. The optimization objective is

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x, x^*, y_w, y_l) \sim \mathcal{D}_f} \left[\log \sigma\Big(\eta^{-1} \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \eta^{-1} \log \frac{\pi_\theta(y_l|x^*)}{\pi_{\text{ref}}(y_l|x^*)} \Big) \right]. \tag{13}$$

However, we do not employ the clinical relevance score used in MMedPO to reduce the computational workload. Finally, the objective of CoMedPO is the combination of the two preference optimizations:

$$\mathcal{L}_{\text{CoMedPO}} = \mathcal{L}_{\text{DPO}} + \lambda \mathcal{L}_{\text{cDPO}}, \tag{14}$$

where λ is the weight parameter.

Table 1: Performance comparison on medical VQA and report generation tasks covering SLAKE, VQA-RAD, and IU-XRay datasets. For open-ended questions, we report recall (Open); for closed-ended questions, accuracy (Closed). The BLEU score denotes the average of BLEU-1/2/3/4. +SFT indicates that the model is first fine-tuned with SFT before applying the corresponding baselines. The best and second-best results are highlighted in **bold** and underlined, respectively.

Models	SLAKE		VQA-RAD		IU-XRay			MIMIC-CXR		
	Open	Closed	Open	Closed	BLEU	ROUGE-L	METEOR	BLEU	ROUGE-L	METEOR
LLaVA-Med v1.5	47.54	52.99	29.50	64.54	14.56	10.31	10.95	10.25	9.38	7.71
+ STLLaVA-Med	50.48	54.67	35.00	68.53	16.10	10.58	10.51	11.11	9.29	7.72
+ DPO	49.84	53.59	32.00	66.53	16.08	12.95	17.13	11.19	9.45	7.80
+ mDPO	51.20	60.29	35.00	68.13	18.43	22.38	26.52	11.55	9.62	7.90
+ SPPO	52.62	67.58	38.50	68.93	21.66	27.30	32.10	12.32	10.19	8.83
+ MMedPO	51.67	58.25	36.50	67.33	23.49	29.52	34.16	**12.85**	11.13	10.03
+ **CoMedPO w/ mDPO**	51.67	65.07	38.50	68.53	22.67	30.65	33.24	12.31	10.58	9.88
+ **CoMedPO w/ SPPO**	**53.73**	69.02	**42.50**	69.72	21.46	27.17	32.23	12.54	10.56	9.78
+ **CoMedPO**	53.11	**71.17**	40.00	**71.31**	**23.78**	**32.68**	**35.26**	12.65	**11.61**	**10.37**
+ SFT	52.62	58.01	31.50	64.14	22.75	28.86	33.66	12.39	10.21	8.75
+ STLLaVA-Med	54.53	61.72	33.50	70.12	22.97	28.98	34.05	12.21	10.12	8.98
+ DPO	53.49	64.23	33.00	64.94	23.82	29.97	34.89	12.37	10.38	9.10
+ mDPO	53.10	63.88	34.50	67.33	23.57	28.35	33.13	11.77	9.45	8.91
+ SPPO	58.57	66.99	41.00	73.31	23.78	30.05	34.78	12.87	12.32	9.45
+ MMedPO	56.60	67.82	37.00	67.73	24.00	30.13	35.17	13.28	13.22	10.20
+ **CoMedPO w/ mDPO**	57.56	69.02	40.00	71.31	25.28	33.69	37.42	12.79	11.35	10.04
+ **CoMedPO w/ SPPO**	61.66	70.93	**44.00**	74.10	**27.33**	**34.38**	38.11	12.84	12.60	11.63
+ **CoMedPO**	**62.17**	**74.40**	42.50	**74.90**	26.23	34.13	**38.25**	**14.91**	**14.58**	**12.22**

5 Experiments

Datasets & Metrics. We evaluate on VQA-RAD [15], SLAKE [19], IU-Ray [8], and MIMIC-CXR [14]. We report Accuracy/Recall for VQA, and BLEU/ROUGE-L/METEOR for reports. Clinical correctness is assessed via Lingshu [48]. Details are in Appendix S4.

Implementation. We use LLaVA-Med-1.5-mistral-7B [17] with Med-SAM [28] for localization. Optimization employs LoRA with rank 128 and scaling 256. Dropout is set to 0.05 and no LoRA bias is used. Training runs for three epochs with a per-device batch size of two and gradient accumulation of one. We set the learning rate to 1e-6 with cosine scheduling and a warmup ratio of 0.03. Weight decay is zero. We save at the end of each epoch, keeping a single retained checkpoint. λ is set to be 1.0 according to ablation study. All experiments are conducted on 8 NVIDIA A100 GPUs. All methods use the standard single-image and prompt at inference.

Baselines. Comparisons include DPO [36], STLLaVA-Med [41], MMedPO [58], mDPO [44], and SPPO [47]. Descriptions are in Appendix S5.

5.1 Main Results

We note that the SLAKE and VQA-RAD test sets used in MMedPO differ from the official splits. For fair comparison, we therefore re-evaluate LLaVA-Med [17], STLLaVA-Med [41], DPO [36], and MMedPO [58] using the Lingshu [48] toolkit, while official results are retained for IU-XRay and MIMIC-CXR. CoMedPO is

introduced as a causal extension of the unweighted MMedPO and is further applied to mDPO [44] and SPPO [47] to assess the generality of the proposed counterfactual preference optimization framework.

Comparison with Baseline Methods. As reported in Table 1, CoMedPO achieves consistent improvements over all baseline preference optimization methods across both medical VQA and report-generation tasks. When compared with MMedPO [58], CoMedPO exhibits clear gains on multiple benchmarks. For example, on SLAKE, CoMedPO improves open-ended recall from 51.67% to 53.11% and closed-ended accuracy from 58.25% to 71.17%. On IU-XRay, CoMedPO also yields consistent gains, improving ROUGE-L from 29.52% to 32.68%. These representative improvements indicate that CoMedPO achieves better alignment of medical reasoning across diverse datasets.

Comparison with Baseline Methods Enhanced by SFT. As shown in Table 1, CoMedPO continues to outperform all SFT-enhanced baselines, with more improvements than in the non-SFT setting. Relative to MMedPO [58] after supervised fine-tuning, CoMedPO achieves larger gains across both VQA and report-generation benchmarks. For example, on SLAKE-Closed, performance increases from 67.82% to 74.40%, and on VQA-RAD closed-ended accuracy from 67.73% to 74.90%. On IU-XRay, CoMedPO further improves ROUGE-L from 30.13% to 34.13%. These results suggest that SFT provides a stronger reference policy that amplifies the benefits of counterfactual preference optimization. By decoupling background-related bias from the preference signal, CoMedPO is able to extract more reliable supervision from the improved SFT initialization.

Applicability to Different Preference Optimization Algorithms. Beyond enhancing MMedPO, the proposed causal adjustment is also applied to mDPO and SPPO to assess its generality. In both cases, the causal variants exhibit consistent improvements over their baselines. For mDPO [44], CoMedPO increases SLAKE closed-ended accuracy from 60.29% to 65.07% and improves IU-XRay ROUGE-L from 22.38% to 30.65%. For SPPO, CoMedPO further raises VQA-RAD open-ended recall from 38.50% to 42.50% and achieves stronger report-generation results, using the second-round SPPO outputs for evaluation. These observations indicate that the proposed causal correction provides a transferable enhancement that strengthens preference optimization algorithms.

5.2 Empirical Causal Validation

We empirically validate the causal objective of CoMedPO through controlled interventions and distribution-shift evaluations. Additional analyses and extended results are reported in Appendix S6.2.

Background Shortcut Sensitivity. To directly probe whether models rely on spurious background cues, we evaluate performance under background-only inputs. As shown in Part I of Table 2, MMedPO^s achieves substantially higher accuracy than CoMedPO^s when only background information is provided (e.g., VQA-RAD Closed: 65.34% vs. 53.78%). The degraded performance of CoMedPO under background-only inputs demonstrates that the learned policy no longer

Table 2: Causal validation and ablation study on VQA-RAD and IU-XRay. All models are trained with LLaVA-Med v1.5 as the backbone. s denotes the model enhanced with SFT, t denotes the model under testing and $*$ denotes the model trained on the other same task dataset. All models use Med-SAM segmentation unless specified.

Methods	VQA-RAD		IU-XRay		
	Open	Closed	BLEU	ROUGE-L	METEOR
DPO^s	33.00	64.94	23.83	29.97	34.89
MMedPO^s	37.00	67.73	24.00	30.13	35.17
CoMedPO^s	42.50	74.90	26.23	34.13	38.25
I. Background Shortcut Sensitivity (Lower is better)					
DPO^s_t w/ Background Only	34.50	63.74	21.56	27.22	32.39
MMedPO^s_t w/ Background Only	26.50	65.34	15.77	10.36	10.05
CoMedPO^s_t w/ Background Only	22.00	53.78	13.29	9.45	10.58
II. Cross-Dataset Generalization					
DPO$^s_*$	32.50 (-0.50)	60.96 (-3.98)	19.82 (-4.01)	23.97 (-6.00)	26.89 (-8.00)
MMedPO$^s_*$	33.50 (-3.50)	62.15 (-5.58)	21.17 (-2.83)	27.26 (-2.87)	31.84 (-3.33)
CoMedPO$^s_*$	40.50 (-2.00)	72.51 (-2.39)	25.32 (-0.91)	32.43 (-1.70)	36.55 (-1.70)
III. Loss Function Variants					
CoMedPO^s w/o $\mathcal{L}_{\text{DPO}}$	29.50	57.77	14.11	10.65	10.37
CoMedPO^s w/ Causal-weighted $\mathcal{L}_{\text{DPO}}$	37.50	70.52	24.25	28.40	34.64
CoMedPO^s w/o Counterfactual Query	42.00	72.91	26.56	34.02	36.68
IV. Inference Input Consistency					
CoMedPO^s_t w/ Composite Input Only	37.50	69.32	22.35	28.69	32.57
CoMedPO^s_t w/ Composite + Matched Prompt	43.43	73.00	25.67	33.62	35.05
CoMedPO^s_t w/ Random Composite	37.00	67.33	21.24	27.67	31.82

relies on spurious background correlations. This behavior is consistent with our causal objective, which explicitly removes the direct effect of B on the reward.

Cross-Dataset Generalization. We evaluate robustness under distribution shift by training on SLAKE and testing on VQA-RAD for VQA, and training on MIMIC-CXR and testing on IU-XRay for report generation. As shown in Part II of Table 2, CoMedPO exhibits consistently smaller performance drops than DPO and MMedPO across all metrics. This indicates that CoMedPO reduces dataset-specific shortcut reliance and learns more transferable representations, leading to stronger robustness under cross-dataset shifts.

5.3 Ablation Study

In Table 2, we select SFT-enhanced MMedPO as the baseline to conduct ablation studies on VQA-RAD and IU-XRay datasets. Additional details regarding multi-round training and compatibility analysis are provided in Appendix S6.3.

Loss Function Variants. As reported in Part III of Table 2, we evaluate several loss function variants to understand their individual contributions. Removing the standard DPO term $\mathcal{L}_{\text{DPO}}$ leads to a clear performance drop compared to the standard model, where VQA-RAD open-ended recall decreases significantly from 42.50% to 29.50%, indicating that learning from factual preference pairs remains essential. Applying causal weights to the original DPO loss, the details of which are provided in Appendix S6.3, improves over the SFT baseline but remains weaker than the full CoMedPO model. Furthermore, instead of intro-

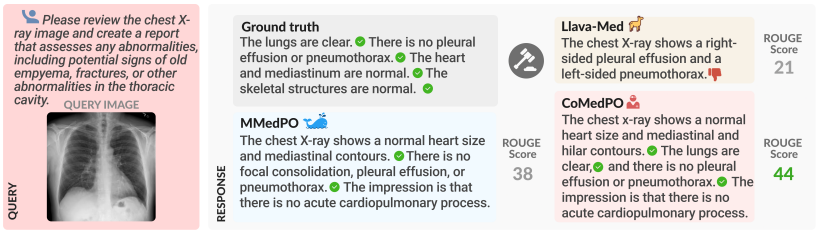

Fig. 4: CoMedPO generates clinically consistent and comprehensive radiology reports compared with baseline models.

ducing a counterfactual clinical query, we use the original standard prompt T for the composite preference pairs. This variant yields comparable performance, indicating that our framework is relatively stable with respect to prompt design. **Inference Input Consistency.** In Part IV of Table 2, we further evaluate the effect of different input configurations to verify our causal mechanism. We specifically test using the composite image as the sole visual input, which pairs a standard single-image prompt with the composite image. While this variant performs better than the background-only inputs evaluated in Part IV, yielding 37.50% versus 22.00% on VQA-RAD Open-ended recall, its overall performance remains lower than the full CoMedPO model. This is likely because the presence of explicit background channels alongside a prompt-visual mismatch still introduces confounding information. To address this, we introduce a matched prompt specifically designed to align with the composite image. When combining the composite input with this matched prompt, performance rebounds significantly, reaching 43.43% on VQA-RAD Open-ended recall and 33.62% on IU-XRay ROUGE-L, which even surpasses the standard model on the open-ended task. These observations are consistent with our causal graph formulation, where the background can act as the biased term, confirming that maintaining input consistency is crucial for effective causal reasoning. Moreover, We added a Random Composite control, where the causally constructed I_b is replaced by a random patch while keeping the same composite format, prompt, and training recipe. This variant is slightly better than DPO on VQA-RAD but remains clearly worse than CoMedPO, especially on IU-XRay. This suggests that random composition may introduce some augmentation benefit, but it breaks the intended causal formulation and can undesirably reinforce background information. Therefore, CoMedPO's gain is not merely due to augmentation or visual diversity, but comes from the causally grounded construction.

5.4 Qualitative Analysis

To qualitatively assess the clinical reliability of CoMedPO, we examine representative examples comparing its outputs with those produced by LLaVA-Med, and MMedPO. More cases are in Appendix S6.4 and S6.5. As illustrated in Fig. 4, LLaVA-Med produces an erroneous description, incorrectly suggesting bilateral pleural effusions in the report generation task. MMedPO improves factual accuracy but still omits clinically relevant details and exhibits partially incomplete

reasoning. In contrast, CoMedPO provides a more comprehensive and clinically aligned report. It correctly identifies normal cardiac size, mediastinal contours, and clear lung fields, fully matching the ground-truth interpretation and achieving the highest ROUGE score.

6 Conclusion

This work presents CoMedPO, a counterfactual preference optimization framework that reduces background-induced shortcuts in medical large vision language model. Using counterfactual reasoning and isolating background-related evidence, CoMedPO produces a more reliable preference reward and yields consistent improvements in both medical VQA and report-generation tasks. Theoretical analysis further shows that CoMedPO converges closer to the causal-optimal policy even when the reference model is biased, reinforcing its validity as a principled approach for medical alignment. Our findings also indicate that accurate lesion localization contributes meaningfully to downstream clinical reasoning. A natural direction for future research is to integrate lesion grounding more tightly with medical preference learning, for example through joint optimization strategies or cooperative game design that strengthen the alignment between visual evidence and clinical decision-making.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Hatfield-Dodds, Z., et al.: Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 (2022)
3. Banerjee, O., Zhou, H.Y., Wu, K., Adithan, S., Kwak, S., Rajpurkar, P.: Direct preference optimization for suppressing hallucinated prior exams in radiology report generation. In: Machine Learning for Healthcare Conference. PMLR (2024)
4. Bannur, S., Bouzid, K., Castro, D.C., Schwaighofer, A., Thieme, A., Bond-Taylor, S., Ilse, M., Pérez-García, F., Salvatelli, V., Sharma, H., et al.: Maira-2: Grounded radiology report generation. arXiv preprint arXiv:2406.04449 (2024)
5. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika **39**(3/4), 324–345 (1952)
6. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems (2017)
7. Cui, C., Zhou, Y., Yang, X., Wu, S.F., Zhang, L., Zou, J., Yao, H.: Holistic analysis of hallucination in gpt-4v(ision): Bias and interference challenges. arXiv preprint arXiv:2311.03287 (2023)
8. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2015)

9. Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., Kiela, D.: Kto: Model alignment as prospect theoretic optimization. arXiv preprint arXiv:2402.01306 (2024)
10. Hein, D., Chen, Z., Ostmeier, S., Xu, J., Varma, M., Reis, E.P., Michalson, A.E., Bluethgen, C., Shin, H.J., Langlotz, C., et al.: Preference fine-tuning for factuality in chest x-ray interpretation models without human feedback (2024)
11. Hong, J., Lee, N., Thorne, J.: Reference-free monolithic preference optimization with odds ratio. arXiv preprint arXiv:2403.07691 **6** (2024)
12. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
13. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., Langlotz, C.P., Rajpurkar, P.: Radgraph: Extracting clinical entities and relations from radiology reports (2021)
14. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
15. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
16. Lau, J.J., Gayen, S., Abacha, A.B., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5**, 180251 (2018)
17. Li, C., Wong, C.Y., Zhang, S., Usuyama, N., Poon, H., Gao, J.: Llava-med: Training a large language and vision assistant for biomedicine in one day. In: NeurIPS Datasets and Benchmarks Track (2023)
18. Li, S., Xue, F., Liu, K., Guo, D., Hong, R.: Multimodal graph causal embedding for multimedia-based recommendation. IEEE Transactions on Knowledge and Data Engineering **36**(12), 8842–8858 (2024)
19. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 1650–1654. IEEE (2021)
20. Liu, H., Li, C., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (2024)
21. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
22. Liu, J., Gong, L., Guo, J., Wu, J., Sun, L., Bi, Y., Patwari, K., Chen, B., Zhang, L., Zhou, W., et al.: Multimodal large language models in medicine and nursing: A survey. Authorea Preprints
23. Liu, L., Sun, S., Zhi, S., Shi, F., Liu, Z., Heikkilä, J., Liu, Y.: A causal adjustment module for debiasing scene graph generation. IEEE Trans. Pattern Anal. Mach. Intell. (2025)
24. Liu, R., Liu, H., Li, G., Hou, H., Yu, T., Yang, T.: Contextual debiasing for visual recognition with causal mechanisms. In: CVPR. pp. 12755–12765 (2022)
25. Liu, S., Fang, W., Hu, Z., Zhang, J., Zhou, Y., Zhang, K., Tu, R., Lin, T.E., Huang, F., Song, M., et al.: A survey of direct preference optimization. arXiv preprint arXiv:2503.11701 (2025)

26. Liu, T., Zhao, Y., Joshi, R., Khalman, M., Saleh, M., Liu, P.J., Liu, J.: Statistical rejection sampling improves preference optimization. In: The Twelfth International Conference on Learning Representations (2024)
27. Liu, Y., Wang, H., Wang, Z., Zhu, X., Liu, J., Sun, P., Tang, R., Du, J., Leung, V.C., Song, L.: Crcl: Causal representation consistency learning for anomaly detection in surveillance videos. IEEE Trans. Image Process. (2025)
28. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
29. Meng, Y., Xia, M., Chen, D.: Simpo: Simple preference optimization with a reference-free reward. Advances in Neural Information Processing Systems **37**, 124198–124235 (2024)
30. Moor, M., Huang, Q., Wu, S.H., Dalmia, M., Leskovec, J., Rajpurkar, P.: Medflamingo: A multimodal medical few-shot learner. In: Machine Learning for Health (ML4H). pp. 353–367 (2023)
31. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12700–12710 (2021)
32. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in neural information processing systems **35**, 27730–27744 (2022)
33. Pearl, J.: Causality. Cambridge university press (2009)
34. Pearl, J.: Interpretation and identification of causal mediation. Psychological methods **19**(4), 459 (2014)
35. Pearl, J., Glymour, M., Jewell, N.P.: Causal inference in statistics: A primer. John Wiley & Sons (2016)
36. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
37. Royer, C., Menze, B., Sekuboyina, A.: Multimedeval: A benchmark and toolkit for evaluating medical vision language models. arXiv preprint arXiv:2402.09262 (2024)
38. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. In: arXiv preprint arXiv:1707.06347 (2017)
39. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
40. Shao, F., Luo, Y., Gao, F., Yang, Y., Xiao, J.: Knowledge-guided causal intervention for weakly-supervised object localization. IEEE Transactions on Knowledge and Data Engineering **36**(11), 6477–6489 (2024)
41. Sun, G., Qin, C., Fu, H., Wang, L., Tao, Z.: Stllava-med: Self-training large language and vision assistant for medical question-answering. arXiv preprint arXiv:2406.19973 (2024)
42. Thirunavukarasu, A.J., Ting, D.S.J., Elangovan, K., Gutierrez, L., F., T.T., Ting, D.S.W.: Large language models in medicine. Nature Medicine **29**(8), 1930–1940 (2023)
43. Wang, C., Zhao, Z., Jiang, Y., Chen, Z., Zhu, C., Chen, Y., Liu, J., Zhang, L., Fan, X., Ma, H., et al.: Beyond reward hacking: Causal rewards for large language model alignment. arXiv preprint arXiv:2501.09620 (2025)

44. Wang, F., Zhou, W., Huang, J.Y., Xu, N., Zhang, S., Poon, H., Chen, M.: mdpo: Conditional preference optimization for multimodal large language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 8078–8088 (2024)
45. Wang, X., Li, Q., Yu, D., Cui, P., Wang, Z., Xu, G.: Causal disentanglement for semantic-aware intent learning in recommendation. IEEE Transactions on Knowledge and Data Engineering **35**(10), 9836–9849 (2022)
46. Wei, Y., Wang, X., Nie, L., Li, S., Wang, D., Chua, T.S.: Causal inference for knowledge graph based recommendation. IEEE Transactions on Knowledge and Data Engineering **35**(11), 11153–11164 (2022)
47. Wu, Y., Sun, Z., Yuan, H., Ji, K., Yang, Y., Gu, Q.: Self-play preference optimization for language model alignment. In: The Thirteenth International Conference on Learning Representations (2025)
48. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
49. Xu, Z., Li, Q., Nie, W., Wang, W., Liu, A.: Structure causal models and llms integration in medical visual question answering. IEEE Transactions on Medical Imaging (2025)
50. Yamada, A., Taiji, R., Nishimoto, Y., Itoh, T., Marugami, A., Yamauchi, S., Minamiguchi, K., Yanagawa, M., Tomiyama, N., Tanaka, T.: Pictorial review of pleural disease: Multimodality imaging and differential diagnosis. Radiographics **44**(4), e230079 (2024)
51. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS medicine **15**(11), e1002683 (2018)
52. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
53. Zheng, X., Liao, C., Fu, Y., Lei, K., Lyu, Y., Jiang, L., Ren, B., Chen, J., Wang, J., Li, C., et al.: Mllms are deeply affected by modality bias. arXiv preprint arXiv:2505.18657 (2025)
54. Zhi, H., Chen, P., Yao, H.: Medgr2: Generative reward learning for robust medical reasoning. arXiv preprint arXiv:2501.05173 (2025)
55. Zhou, K., Jiang, M., Gabrys, B., Xu, Y.: Learning causal representations based on a gae embedded autoencoder. IEEE Transactions on Knowledge and Data Engineering (2025)
56. Zhou, Y., Cui, C., Rafailov, R., Finn, C., Yao, H.: Aligning modalities in vision large language models via preference fine-tuning. In: ICLR 2024 Workshop on Reliable and Responsible Foundation Models (2024)
57. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations (2024)
58. Zhu, K., Xia, P., Li, Y., Zhu, H., Wang, S., Yao, H.: Mmedpo: Aligning medical vision language models with clinical-aware multimodal preference optimization. In: International Conference on Machine Learning (ICML) (2025)
59. Zhu, X., Sun, L., Liu, Y., Jiang, P., Srivatsa, U., Chiamvimonvat, N., Filkov, V.: Causal debiasing medical multimodal representation learning with missing modalities. arXiv preprint arXiv:2509.05615 (2025)