HSCR: Hierarchical Self-Contrastive Rewarding for Aligning Medical Vision Language Models

Songtao Jiang[1], **Yan Zhang**[2], **Yeying Jin**[3], **Zhihang Tang**[1]
Yangyang Wu[1], **Yang Feng**[4], **Jian Wu**[1,5], **Zuozhu Liu**[1,5†]

[1]Zhejiang University, [2]Byte Dance, [3]National University of Singapore
[4]Angelalign Inc., China [5]Zhejiang Key Laboratory of Medical Imaging Artificial Intelligence
Correspondence[†]: zuozhuliu@intl.zju.edu.cn

Abstract

Medical Vision-Language Models (Med-VLMs) have achieved success across various tasks, yet most existing methods overlook the modality misalignment issue that can lead to untrustworthy responses in clinical settings. In this paper, we propose Hierarchical Self-Contrastive Rewarding (HSCR), a novel approach that addresses two critical challenges in Med-VLM alignment: 1) Cost-effective generation of high-quality preference data; 2) Capturing nuanced and context-aware preferences for improved alignment. HSCR first leverages the inherent capability of Med-VLMs to generate dispreferred responses with higher sampling probability. By analyzing output logit shifts after visual token dropout, we identify modality-coupled tokens that induce misalignment and derive an implicit alignment reward function. This function guides token replacement with hallucinated ones during decoding, producing high-quality dispreferred data. Furthermore, HSCR introduces a multi-level preference optimization strategy, which extends beyond traditional adjacent-level optimization by incorporating nuanced implicit preferences, leveraging relative quality in dispreferred data to capture subtle alignment cues for more precise and context-aware optimization. Extensive experiments across multiple medical tasks, including Med-VQA, medical image captioning and instruction following, demonstrate that HSCR not only enhances zero-shot performance but also significantly improves modality alignment and trustworthiness with just 2,000 training entries. Code is released on https://github.com/jiangsongtao/HSCR.

1 Introduction

Medical Vision-Language Models (Med-VLMs) have shown strong performance in tasks like medical visual question answering (Med-VQA) (Moor et al., 2023; Saab et al., 2024; Gai et al., 2025; Li et al., 2024a; Jiang et al., 2025; Xu et al., 2025) by integrating pre-trained vision encoders into large language models (LLMs), allowing access to visual information. However, limited quality and quantity of paired multimodal medical training data often lead to modality misalignment (Jiang et al., 2024c; Krieger, 1992). As a result, medical VLMs may hallucinate image contents, favoring text-based preferences over actual visual content. This misalignment undermines trustworthiness, posing challenges for reliable applications of these models in high-stakes medical scenarios (Liu et al., 2023a).

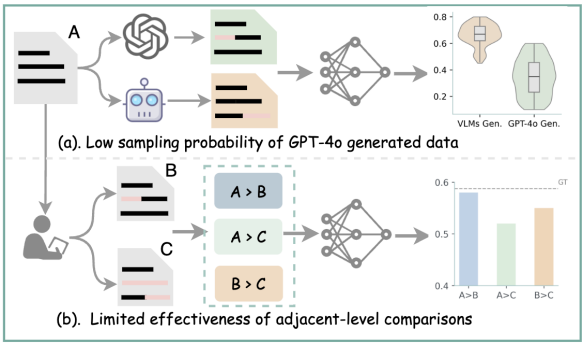

Figure 1: Two key challenges in preference optimization for Med-VLMs. See Appendix A.1 for details.

Recent work in multimodal learning has explored preference optimization methods to improve modality alignment, such as Reinforcement Learning with Human Feedback (Sun et al., 2023; Song et al., 2024) and Direct Preference Optimization (Rafailov et al., 2024; Jiang et al., 2024a; Zhou et al., 2024a; Zhang et al., 2024). While these approaches show promise in general domains, their application to Med-VLMs faces two critical challenges (Figure 1): (1) limited sampling probabilities for preferred/dispreferred responses during optimization, and (2) reduced effectiveness of adjacent-level comparisons in weakly trained Med-VLMs. These issues stem from significant gaps in data quality, scale, and distribution between general VLMs and Med-VLMs, compounded by reliance

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13853–13868
July 27 - August 1, 2025

on manual annotations or synthetic preference data from larger VLMs like GPT-4o (Hurst et al., 2024).

For the first challenge, human-annotated or GPT-4o-generated preference data exhibit limited sampling probabilities during Med-VLM optimization. This inefficiency arises from a misalignment between the decoding behavior of Med-VLMs and the external preference data distribution (e.g., GPT-4o), primarily due to divergent training data sources and objectives between GPT-4o and Med-VLMs. Consequently, preferred responses are rarely sampled during optimization, resulting in weak reward signals and suboptimal alignment performance (Zhou et al., 2024b; Azar et al., 2024). For the second challenge, adjacent-level comparisons between correct and incorrect responses often exhibit pronounced disparities, causing Med-VLMs to easily saturate their ability to distinguish preferred and dispreferred outputs. This phenomenon limits their capacity to learn nuanced preferences, as large preference gaps also obscure clear optimization directions during training (Zhou et al., 2024c). The issue is exacerbated in medical domains, where subtle distinctions between plausible responses require finer-grained learning.

To address these challenges, we propose Hierarchical Self-Contrastive Rewarding (HSCR), a novel preference optimization method for Med-VLM alignment. HSCR encompasses three steps: token-level self-contrastive rewarding for data generation, similarity-aware preference re-ranking, and multi-level preference optimization. First, to generate preference data with enhanced sampling probabilities, we leverage the inherent capabilities of Med-VLMs to produce misaligned responses, eliminating the need for external resources or annotations. These misaligned responses naturally exhibit higher sampling probabilities, making them effective examples of dispreferred data. In particular, inspired by the masking strategies in Masked Autoencoders (He et al., 2022) and Vision Transformers (Dosovitskiy, 2020),we introduce visual token dropout to expose inherent misalignment in the model's generation behavior. By analyzing the resulting differences in logits, we can identify strongly modality-coupled tokens that are prone to inducing hallucinations due to significant shifts in their logits. These logit differences are formed as an implicit reward function to replace sensitive tokens with hallucinated ones. Afterwards, to ensure that response rankings accurately reflect semantic differences, we compute the semantic similarity between dispreferred and preferred responses, re-ranking them accordingly. This process yields rank-based preference lists.

Finally, unlike existing preference optimization techniques that focus solely on categorizing correct responses as preferred and hallucinated responses as dispreferred (i.e., explicit preference), our study reveals that varying levels of incorrectness in dispreferred responses can provide richer and more nuanced preference signals. Specifically, we designed an implicit optimization objective which encourages the model to discern differences in the degree of incorrectness among dispreferred responses. This multi-level joint preference optimization approach not only captures broad, high-level alignment signals through explicit preference but also learns subtle and intricate preferences through implicit preference. HSCR achieves performance improvements across a wide range of Med-VQA, captioning, and instruction-following tasks. Notably, it achieves state-of-the-art (SOTA) zero-shot performance on Rad-VQA (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020). Code and datasets will be released.

2 Preliminaries

Direct Preference Optimization (DPO) streamlines alignment by directly optimizing the policy π_η using preference data $\mathcal{P}$, eliminating the need for an explicit reward model (Rafailov et al., 2024). DPO links the reward function $g(x, y)$ to the policy:

$$g(x,y) = \gamma \log \frac{\pi_\eta(y|x)}{\pi_{\text{base}}(y|x)} + \gamma \log Z(x), \quad (1)$$

where $Z(x)$ is the partition function. The optimization objective is:

$$\mathcal{J}_{\text{DPO}}(\pi_\eta; \pi_{\text{init}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{P}} \left[\log \sigma \left(\gamma \log \frac{\pi_\eta(y_w|x)}{\pi_{\text{init}}(y_w|x)} - \gamma \log \frac{\pi_\eta(y_l|x)}{\pi_{\text{init}}(y_l|x)} \right) \right], \quad (2)$$

where y_w and y_l denote preferred and less preferred outputs, respectively. This approach enhances efficiency and precision in alignment (Dong et al., 2024; Pal et al., 2024). Details are in Appendix A.2.

3 Hierarchical Self-Contrastive Rewarding (HSCR)

In this section, we delve into the HSCR framework (see Fig. 2). First, we construct preference datasets

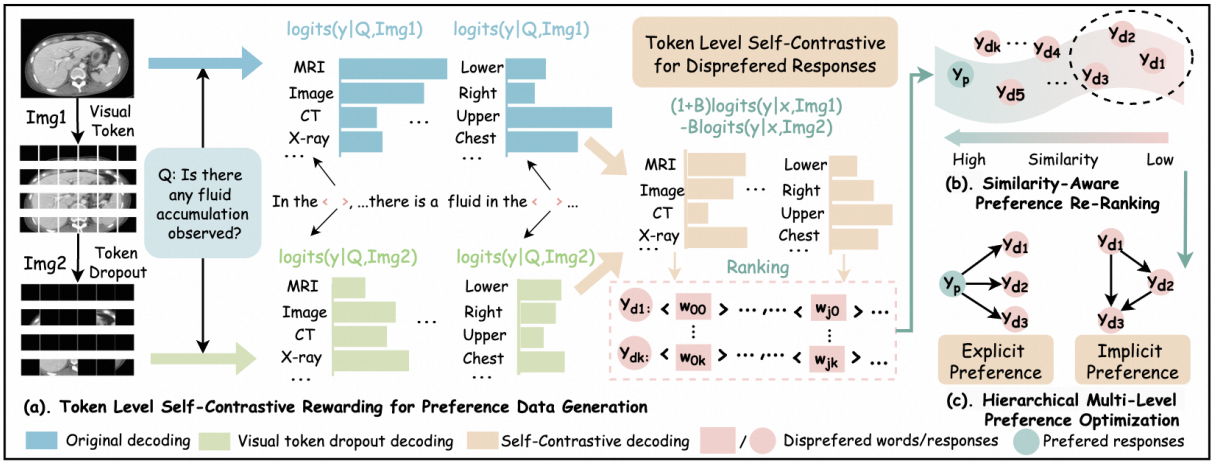

Figure 2: Overview of the HSCR pipeline, comprising token-level self-contrastive rewarding for data generation (Methods 3.1), similarity-aware preference re-ranking for quality control (Methods 3.2), and hierarchical multi-level preference optimization (Methods 3.3).

without external tools through self-contrastive rewarding. Next, we perform semantic similarity-based checking and re-ranking of the preference data. Finally, we optimize a supervised fine-tuned medical VLM, modelled as a policy π_θ parameterized by θ, using multi-level preference optimization to achieve enhanced alignment.

3.1 Token-Level Self-Contrastive Rewarding for Data Generation

For data generation, we treat the ground truth as the preferred response y_w and aim to construct dispreferred responses y_l that reflect the inherent misalignment or unreliable behavior of VLMs. To achieve this, we adopt a two-step approach. First, to expose potential misalignment, we employ a visual token dropout strategy (see Fig. 2 a) to disrupt the image modality. Specifically, given the original visual token i, we apply a 70% dropout rate to obtain i'. Combined with the given textual query x, we compute the token logits output by the VLM as $\text{logit}_\theta(y \mid i, x)$ and $\text{logit}_\theta(y \mid i', x)$, respectively. Second, to identify tokens prone to causing misalignment, we locate the top-n tokens with the largest logit differences between the two distributions. These tokens, which exhibit strong modality coupling, are often error-prone:

$$P_{\text{diff}} = \text{Softmax}\left[(1+\beta)\cdot \text{logit}_\theta(y \mid i, x) - \beta \cdot \text{logit}_\theta(y \mid i', x)\right], \tag{3}$$

where β controls the contrast strength between distributions, with higher values enhancing the distinction between the two. After identifying these sensitive tokens, we generate token-level dispreferred responses by replacing them with incorrect tokens through contrastive decoding. Specifically, for each sensitive token, we decode based on the logit differences from P_{diff} in ascending order, substituting them with tokens that exhibit lower logit differences. These substituted tokens, which are weakly correlated with the actual visual information, often correspond to hallucinated outputs of the model (Leng et al., 2024). By replacing all sensitive tokens, we generate a set of dispreferred responses with varying degrees of incorrectness:

$$\{y_{l1}, y_{l2}, \ldots, y_{lk}\} \sim \left(\prod_{t=1}^{T} P_{\text{diff}}(y_t \mid y_{<t}, i, i', x)\right), \tag{4}$$

where y_t is the token at position t, $y_{<t}$ denotes preceding tokens, and T is the sequence length (Hyperparameters are in Section 4.1).

3.2 Similarity-Aware Preference Re-Ranking

After applying the approach in Section 3.1, we obtain a set of candidate dispreferred responses. However, these responses may exhibit varying degrees of semantic similarity to the preferred response. For instance, some dispreferred responses might be partially correct or contextually relevant but less precise, while others could be entirely unrelated or misleading. To ensure accurate preference ordering, we introduce a similarity-aware re-ranking module to refine the responses (see Fig. 2 b). Specifically, we compute the semantic similarity (Corley and Mihalcea, 2005) between each dispreferred response y_{lk} and the preferred response y_w, denoted as $\text{sim}(y_{lk}, y_w)$. The responses are then re-ranked in descending order of similarity. From the re-

ranked list, we select j responses such that their similarity differences with respect to y_w are at least 0.1 for optimization. By explicitly integrating semantic similarity into the ranking process, this approach enables a more nuanced distinction between responses of varying quality. Consequently, the final dispreferred responses $\{y_{l1}, y_{l2}, \ldots, y_{lj}\}$ will be used for subsequent preference optimization.

3.3 Hierarchical Multi-Level Preference Optimization

To comprehensively optimize both explicit and implicit preferences in training data, we propose Multi-Level Preference Optimization (MLPO), a novel framework that enhances the model's ability to distinguish not only between preferred and dispreferred responses but also among dispreferred responses at varying levels. Unlike traditional binary preference optimization methods (Yu et al., 2024a), which rely on coarse-grained comparisons, MLPO enables fine-grained preference learning through two key components: *Explicit Preference Learning* and *Implicit Preference Learning*, each performing multi-level optimization (see Fig. 2 c).

Explicit Preference Learning. Explicit preferences focus on the distinction between the preferred response y_w and each dispreferred response y_{lj}. To quantify this, we compute the loss iteratively for y_w and all k dispreferred outputs:

$$L_{\text{E}} = -\sum_{j=1}^{k} \mathbb{E}_{(x,y_w,y_{lj})\sim D}\left[\log\sigma\left(\gamma\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{sft}}(y_w|x)} - \gamma\log\frac{\pi_\theta(y_{lj}|x)}{\pi_{\text{sft}}(y_{lj}|x)}\right)\right]. \tag{5}$$

Here, L_{E} encourages the model to assign higher probabilities to the preferred response y_w compared to each dispreferred response y_{lj}. The term γ acts as a temperature parameter, controlling the strength of the preference signal. By iterating over all the dispreferred responses, the model learns to explicitly prioritize y_w over suboptimal alternatives.

Implicit Preference Learning. Implicit preferences focus on the relative quality among dispreferred responses (e.g., high-ranked vs. low-ranked). To capture these nuanced differences, we compute the loss for all dispreferred pairs (y_{lj}, y_{lm}):

$$L_{\text{I}} = -\sum_{j=1}^{k}\sum_{m=j+1}^{k} \mathbb{E}_{(x,y_{lj},y_{lm})\sim D}\left[\log\sigma\left(\gamma\log\frac{\pi_\theta(y_{lj}|x)}{\pi_{\text{sft}}(y_{lj}|x)} - \gamma\log\frac{\pi_\theta(y_{lm}|x)}{\pi_{\text{sft}}(y_{lm}|x)}\right)\right]. \tag{6}$$

Here, L_{I} encourages the model to distinguish between dispreferred responses based on their relative quality. By comparing all possible pairs (y_{lj}, y_{lm}), the model learns to implicitly rank dispreferred responses, ensuring that higher-quality but still suboptimal responses are prioritized over lower-quality ones. The total loss is defined as the sum of the explicit and implicit losses:

$$L_{\text{HSCR}} = L_{\text{E}} + L_{\text{I}}. \tag{7}$$

This loss function enables the model to simultaneously capture both explicit and implicit preferences. By integrating these two levels of preference learning, our HSCR effectively captures both high-level distinctions and fine-grained nuances.

4 Experiments

4.1 Experiements Setup

Training Settings. Following previous research (Li et al., 2024a), we adopt CLIP-ViT-L/14@336px (Radford et al., 2021) as the visual encoder to extract visual features from medical images, Mistral-7B (Jiang et al., 2023) as the text encoder, and a two-layer MLP with a GeLU (Hendrycks and Gimpel, 2016) activation function as the projector to align the text and visual encoders.

Details of Hyperparameters We sample 2,000 entries from the dataset used in instruction-tuning stage to construct preference datasets as described in Methods 3.1. For this process, we set the parameters as follows: $j = 3$, $\beta = 0.9$, $n = 10$, and $\gamma = 0.1$. Multi-level preference optimization is performed using the constructed implicit and explicit preference datasets, with LoRA (Hu et al., 2021) applied at a rank of 16. Training is conducted for 2 epochs with a learning rate of 5e-7 and without weight decay. The training hyperparameters are in Table 7.

Evaluation Datasets. We evaluate HSCR on all benchmarks from LLaVA-Med (Li et al., 2024a), including Rad-VQA (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020), covering both open-ended and closed-ended

Method	RAD-VQA		SLAKE		PathVQA	
	Open	Closed	Open	Closed	Open	Closed
Representative & SoTA methods reported in the literature (Non-VLMs Based Methods)						
VL Encoder–Decoder (Bazi et al., 2023)	-	82.47	-	-	-	85.61
Q2ATransformer (Liu et al., 2023b)	-	81.20	-	-	54.85	88.85
Prefix T. Medical LM (van Sonsbeek et al., 2023)	-	-	-	82.01	-	87.00
PubMedCLIP (Eslami et al., 2023)	-	80.00	-	82.50	-	-
BiomedCLIP (Zhang et al., 2023b)	-	79.80	-	89.70	-	-
M2I2 (Li et al., 2022)	-	83.50	-	91.10	-	88.00
BiomedGPT-S (Zhang et al., 2023a)	13.40	57.80	66.50	73.30	10.70	84.20
BiomedGPT-M (Zhang et al., 2023a)	53.60	65.07	78.30	86.80	12.5	85.70
CLIP-ViT w/ GPT2-XL (Radford et al., 2021)	-	-	84.30	82.10	40.0	87.00
Zero-shot results						
GPT-4o (Hurst et al., 2024)	51.6	63.97	59.06	71.63	24.14	75.97
LLaVA1.5 (Liu et al., 2024a)	23.63	50.74	35.23	52.16	11.85	52.76
Med-Flamingo (Moor et al., 2023)	10.32	52.21	8.46	37.02	1.23	45.59
PMC-VQA (Zhang et al., 2023c)	6.26	41.54	7.29	33.89	1.02	40.10
SQ-LLaVA (Sun et al., 2025)	23.91	52.57	40.04	57.45	12.24	53.79
ST-LLaVA (Sun et al., 2024)	33.81	59.16	40.13	55.53	10.38	52.05
VCD (Leng et al., 2024)	30.54	55.88	42.92	56.93	9.13	58.16
LiPO (Liu et al., 2024c)	31.85	57.37	43.18	58.13	9.37	60.17
LLaVA-Med1.5 (Li et al., 2024a)	32.31	56.62	42.45	56.49	10.01	59.75
HSCR (Ours)	**35.92(+3.61)**	**60.13(+3.51)**	**45.32(+2.87)**	**63.46(+6.97)**	**12.36(+2.35)**	**64.17(+4.42)**

Table 1: Performance on Med-VQA tasks. **Bold** denotes the best performance,underlined denotes the second-best.

	Question Types		Domains					Overall
	Conversation	Description	CXR	MRI	Histology	Gross	CT	
LLaVA	39.4	26.2	41.6	33.4	38.4	32.9	33.4	36.1
LLaVA-Med								
Pretrain	22.6	25.2	25.8	19.0	24.8	24.7	22.2	23.3
SFT (10K)	42.4	32.5	46.1	36.7	43.5	34.7	37.5	39.9
SFT (60K)	53.7	36.9	57.3	39.8	49.8	47.4	52.4	49.4
SFT (60K-IM)	55.1	36.4	56.2	40.4	52.7	51.8	50.1	50.2
LLaVA-Med1.5								
SFT (60K-IM)	58.6	42.5	59.6	46.5	58.8	52.8	53.8	54.4
HSCR (2K)	**59.4 (+0.8)**	**52.9 (+10.4)**	**62.0 (+2.4)**	**47.9 (+1.4)**	**65.1 (+6.3)**	**53.5 (+0.7)**	**59.5 (+5.7)**	**57.7 (+3.3)**

Table 2: Performance on captioning and instruction-following tasks. IM denotes visual instruction-following data enhanced with figure references from PubMed Central articles (Pringle and Wyatt, 2006).

settings (details in Appendix A.3). For captioning and instruction-following tasks, we use LLaVA-Med's evaluation datasets, which include unseen image-caption pairs from PMC-15M (Zhang et al., 2023b), generating conversation and detailed description questions across five medical modalities and multi-turn dialogues. This ensures a thorough evaluation of medical VLMs across diverse domains and tasks.

Evaluation Metrics. For Med-VQA tasks, we follow prior work (Sun et al., 2024; Zhang et al., 2023c), using accuracy for closed-set (yes/no) questions and recall for open-set (free-form) questions. For captioning and instruction-following tasks, we employ GPT-4 (OpenAI, 2023) as an automated evaluator. GPT-4 generates a reference answer based on the image and caption, against which the candidate model's response is scored. Each response receives an overall score on a 10-point scale, normalized for comparability, along with a detailed explanation justifying the rating.

Baselines. We select a variety of strong baselines. For general-purpose VLMs, we include models such as LLaVA1.5 (Liu et al., 2024a) and SQ-LLaVA (Sun et al., 2025), as they represent top-performing VLMs trained on general-domain instruction-following datasets without incorporating medical data. For medical-specific VLMs, we select leading models, including LLaVA-Med1.5 (Li et al., 2024a), Med-Flamingo (Moor et al., 2023), and PMC-VQA (Zhang et al., 2023c), which are specifically trained on medical instruction-following datasets and demonstrate superior performance on medical tasks. Additionally, for methods aimed at enhancing the alignment of medical VLMs, we include the latest approaches such as ST-LLaVA (Sun et al., 2024), VCD (Leng et al., 2024) and LiPO (Liu et al., 2024c).

4.2 Main Results

Performance on Med-VQA Tasks. As shown in Table 1, our method achieves SOTA performance in the zero-shot setting among open-source models. Specifically, in the closed setting, our method performs nearly on par with GPT-4 on RAD-VQA and achieves a notable 7% performance improve-

Method	RAD-VQA		SLAKE		PathVQA	
	Open	Closed	Open	Closed	Open	Closed
LLaVA-Med1.5 (Li et al., 2024a)	32.31	56.62	42.45	56.49	10.01	59.75
GPT-4o Generated (Hurst et al., 2024)	33.14 (+0.83)	57.20 (+0.58)	41.83 (-0.62)	57.96 (+1.47)	9.55 (-0.46)	60.35 (+0.60)
HSCR (Ours)	**35.92 (+3.61)**	**60.13 (+3.51)**	**45.32 (+2.87)**	**63.46 (+6.97)**	**12.36 (+2.35)**	**64.17 (+4.42)**

Table 3: Ablation of preference data construction.

Explicit Pref.	Implicit Pref.	RAD-VQA		SLAKE		PathVQA	
		Open	Closed	Open	Closed	Open	Closed
×	×	32.31	56.62	42.45	56.49	10.01	59.75
✓	×	33.69 (+1.38)	58.31 (+1.69)	43.14 (+0.69)	57.78 (+1.29)	10.46 (+0.45)	60.05 (+0.30)
×	✓	34.13 (+1.82)	57.65 (+1.03)	43.95 (+1.50)	60.32 (+3.83)	11.24 (+1.23)	62.12 (+2.37)
✓	✓	**35.92 (+3.61)**	**60.13 (+3.51)**	**45.32 (+2.87)**	**63.46 (+6.97)**	**12.36 (+2.35)**	**64.17 (+4.42)**

Table 4: Ablation study on the impact of explicit and implicit preferences.

Mask Strategy	RAD-VQA		SLAKE		PathVQA	
	Open	Closed	Open	Closed	Open	Closed
Baseline	32.31	56.62	42.45	56.49	10.01	59.75
Pixel-Level Mask	33.15 (+0.84)	57.17 (+0.55)	42.65 (+0.20)	57.49 (+1.00)	10.45 (+0.44)	60.79 (+1.04)
Patch-Level Mask	34.03 (+1.72)	57.91 (+1.29)	43.13 (+0.68)	58.44 (+1.95)	11.01 (+1.00)	61.83 (+2.08)
Latent Space Mask	33.45 (+1.14)	58.79 (+2.17)	43.56 (+1.11)	60.32 (+3.83)	11.23 (+1.22)	62.77 (+3.02)
Visual Token Dropout	**35.92 (+3.61)**	**60.13 (+3.51)**	**45.32 (+2.87)**	**63.46 (+6.97)**	**12.36 (+2.35)**	**64.17 (+4.42)**

Table 5: Performance comparison of different masking strategies.

ment over LLaVA-Med1.5 on the SLAKE dataset. More importantly, for challenging open-set questions, our approach consistently demonstrates performance gains with only 2,000 training entries (See Appendix A.5 for more discussion).

Performance on Captioning and Instruction-Following Tasks. As demonstrated in Table 2, our method achieves promising improvements in both captioning and instruction-following capabilities for medical VLMs. It enhances conversational and descriptive performance across diverse medical modalities, consistently surpassing LLaVA-Med1.5. Notably, with only 2,000 training samples, our approach delivers more substantial performance gains compared to scaling supervised fine-tuning (SFT) data from 10k to 50k entries, achieving a 10.4% improvement in captioning accuracy versus a 4.4% gain from data expansion. Furthermore, our method demonstrates consistent performance improvements across all five medical modalities, highlighting its robustness and adaptability. This not only elevates overall performance but also enhances the trustworthiness and reliability of the model in real-world medical applications.

5 Ablation and Analysis

Ablation on Preference Dataset Construction. We evaluate our token-level preference datasets against GPT-4o-generated preferences (Hurst et al., 2024) in a binary setting (Table 3). While GPT-4o preferences yield modest gains in closed-ended tasks (+1.47% on SLAKE, +0.58% on RAD-VQA), they degrade open-ended performance (-0.62% on SLAKE, -0.46% on PathVQA). This highlights the limitations of external preference reward signals from stronger models in guiding medical VLMs, particularly for complex open-ended queries. The key issue stems from the mismatch between external preferences and inherent VLM misalignment biases, leading to suboptimal optimization.

In contrast, HSCR constructs preference datasets by exposing and leveraging misalignment responses inherent to medical VLMs. By precisely correcting these intrinsic misalignment behaviors, HSCR significantly enhances both the performance and trustworthiness of medical VLMs.

Ablation on Implicit and Explicit Preferences. As shown in Table 4, we analyze the effects of explicit and implicit preferences on medical VLMs. Implicit preferences outperform explicit ones by leveraging fine-grained reward gradients to provide nuanced, context-aware guidance, achieving sig-

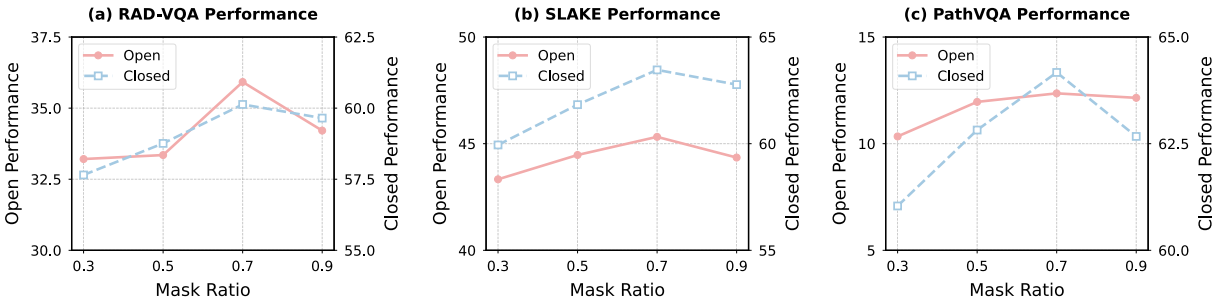

Figure 3: Performance comparison with different mask ratios. Details in Appendix Table 6.

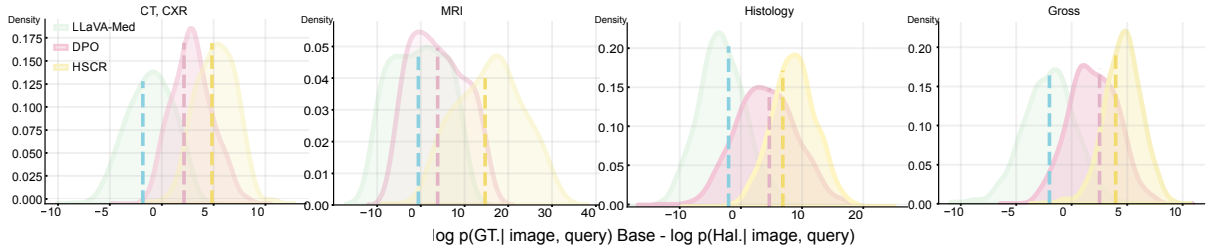

Figure 4: The x-axis represents $\log p(\text{Ground Truth Response}) - \log p(\text{Hallucinatory})$ across five different medical modalities. The dashed line in the figure indicates the median value. A larger value signifies a stronger ability of the VLMs to distinguish hallucinatory responses.

nificant gains: 1.82% on RAD-VQA, 3.83% on SLAKE, and 2.37% on PathVQA. In contrast, explicit preferences, while beneficial, are more effective for high-level, coarse-grained optimization, resulting in comparatively modest improvements. The combination of both approaches yields optimal results, with improvements of up to 3.61% on RAD-VQA, 6.97% on SLAKE, and 4.42% on PathVQA. This demonstrates their complementary roles: *explicit preferences offer broad, high-level direction, while implicit preferences capture subtle, task-specific signals*, enhancing the robustness and trustworthiness of medical VLMs.

Ablation on Mask Ratio. We conducted an ablation study to investigate the impact of different mask ratios on model performance in Figure 3. Our findings indicate that the proposed method exhibits robustness to variations in the mask ratio, consistently outperforming the baseline across all tested ratios. Notably, the performance gains are more substantial for mask ratios exceeding 0.5. This phenomenon can be attributed to the necessity of a sufficiently high mask ratio to effectively disrupt visual information, thereby eliciting meaningful misalignment responses. Based on our empirical results, we adopted a mask ratio of 0.7 for our experiments, as it consistently yields optimal performance across a wide range of scenarios.

Effect of Mask Strategies. In this ablation study, we fixed the mask ratio at the optimal value of 70% and systematically evaluated various levels of mask application (see Appendix A.4 for detailed implementation). As shown in Table 5, While all four disruption methods demonstrate performance improvements over the baseline, visual token dropout emerges as the most effective strategy. We observe a positive correlation between the proximity of the disruption operation to the LLM backbone's input and the overall model performance. This trend can be explained by the fact that visual token dropout directly eliminates portions of visual information before it reaches the LLM, thereby more effectively triggering the inherent misalignment mechanisms of the underlying language model.

How does HSCR improve modality alignment? As illustrated in Figure 5, HSCR significantly enhances modality alignment by generating more accurate and detailed responses that are strongly relevant to the image content. In the example from RAD-VQA, HSCR correctly identifies the brain MRI image as normal, providing a precise and trustworthy explanation, while LLaVA-Med1.5 incorrectly identifies an abnormality. The attention map (Vaswani, 2017) of vision tokens, highlighted by the black box, demonstrates that HSCR places greater focus on the image modality, ensuring that the generated responses are well-grounded in visual evidence. This improved attention to image details

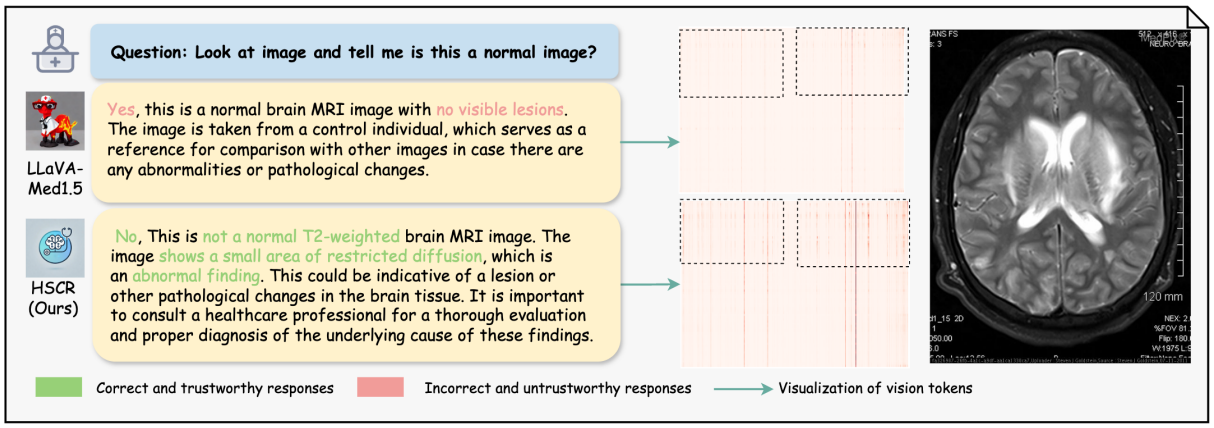

Figure 5: Zero-shot comparison of LLaVA-Med1.5 (Li et al., 2024a) and HSCR. HSCR generates more accurate, detailed, and image-relevant responses, with attention maps (highlighted in black) showing stronger focus on the image modality, achieving superior alignment and trustworthiness. Additional cases are in Appendix 8.

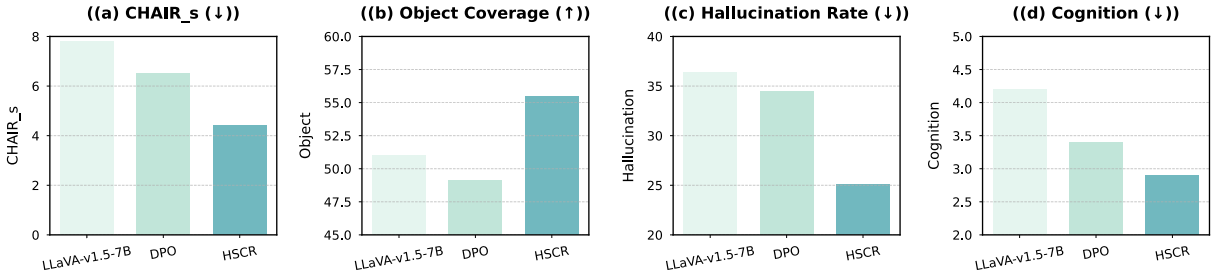

Figure 6: Performance on general multimodal tasks. Details in Appendix Table 8.

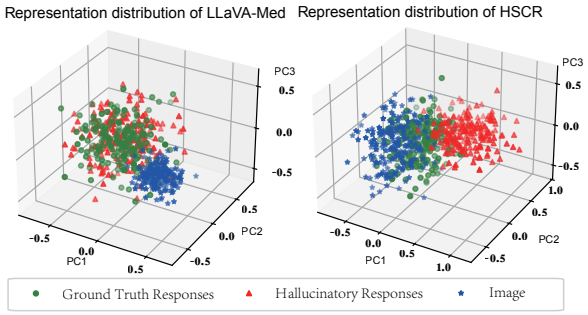

Figure 7: Representation distribution comparison.

allows HSCR to achieve superior modality alignment, resulting in more reliable and high-quality answers. By effectively bridging the gap between visual and textual modalities, HSCR not only enhances the accuracy but also boosts the overall trustworthiness of VLMs in medical applications.

HSCR Enhances Representation Learning in Medical VLMs. Figure 7 illustrates the comparative analysis of sample embeddings in Rad-VQA, showcasing correct responses, hallucinated responses, and corresponding image embeddings. The visualization reveals two critical insights: First, the baseline LLaVA-Med model demonstrates significant limitations in distinguishing between correct and hallucinated responses, accompanied by a pronounced modality gap between text and image representations in the latent space. This is evidenced by the substantial distance between embeddings of different modalities. Second, after incorporating HSCR, we observe a notable improvement in the alignment between correct response embeddings and image embeddings, with these representations converging more closely in the latent space. This enhanced clustering indicates improved modality alignment and demonstrates the model's increased capability to discriminate between factual and hallucinated responses. These findings collectively suggest that HSCR effectively augments the representation learning capacity of medical VLMs, leading to strengthened cross-modal alignment and enhanced model trustworthiness.

Generalizability of HSCR to General Multimodal Tasks To assess the generalizability of our approach beyond medical domains, we integrated HSCR with the general-purpose VLM LLaVA-v1.5 (Liu et al., 2024b) and conducted preference optimization. The model was subsequently evaluated on the widely-used AMBER

benchmark (Wang et al., 2023) for general multimodal tasks (implementation details provided in Appendix A.6). As illustrated in Figure 6, HSCR demonstrates significant effectiveness in enhancing model trustworthiness and modality alignment. Notably, our method outperforms the current mainstream approach DPO, indicating its superior capability in mitigating hallucination issues. These results substantiate that HSCR is not only effective for medical VLMs but also demonstrates robust performance in general-domain tasks.

6 Related Work

Preference Optimization in VLMs. Vision Language Models (VLMs) have achieved significant success in a wide range of tasks (Jiang et al., 2024b; Lin et al., 2024; Li et al., 2024b). Preference optimization has emerged as a critical technique for enhancing the trustworthiness of VLMs. Existing approaches can be broadly categorized into two paradigms: manually constructed preference pairs and automatically generated ones using MLLMs. Methods like RLHF-V (Yu et al., 2024a; Sun et al., 2023) utilize human feedback to refine hallucinated captions, transforming rejected responses into preferred ones. Meanwhile, POVID (Zhou et al., 2024a; Pi et al., 2024) introduces diffusion noise to images, enabling MLLMs to autonomously generate hallucinated content as rejected responses, eliminating the need for human intervention. AMP (Zhang et al., 2024) enhances stability by producing multiple candidate responses, and RLAIF-V (Yu et al., 2024b) improves DPO by aggregating high-quality responses from multiple MLLMs. However, these methods are resource-intensive and have not been validated in resource-constrained clinical settings.

In the medical field, recent concurrent works have also begun exploring preference optimization for alignment. For instance, ST-LLaVA (Sun et al., 2024) employs a self-training (Rosenberg et al., 2005) paradigm to generate preference data and uses GPT-4o (Hurst et al., 2024) to score and create binary preference datasets. However, this approach incurs additional costs and relies on external tools like GPT-4o. Similarly, MMedPO (Zhu et al., 2024b) leverages multi-agent systems to construct preference data, but this requires significant computational resources and memory overhead. In clinical medical settings, there is a pressing need for a cost-efficient method to implement preference optimization, enhancing both the trustworthiness and modality alignment of medical VLMs.

7 Conclusion

This paper proposes Hierarchical Self-Contrastive Rewarding (HSCR) to address modality misalignment in medical VLMs. By leveraging visual token dropout, alignment rewards, and dual preference optimization, HSCR improves modality alignment, trustworthiness, and zero-shot performance with minimal training data. Our work paves the way for developing trustworthy medical AI systems.

8 Limitations

While our work demonstrates promising results, there are some limitations that warrant further exploration. First, the limited availability of high-quality and diverse medical data continues to constrain the development of medical Vision-Language Models (VLMs). This scarcity impacts the generalizability and robustness of the models, particularly in addressing rare or complex medical scenarios. Second, although our evaluation covers extensive experimental benchmarks, it is primarily conducted in controlled research settings. This leaves room for further validation through integration with clinical workflows and real-world trials to better assess the practicality, reliability, and safety of the proposed methods in healthcare applications.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 12326612, 62476241), the Natural Science Foundation of Zhejiang Province, China (Grant No. LZ23F020008), and the Zhejiang University-Angelalign Inc. R&D Center for Intelligent Healthcare.

References

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. 2024. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

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

Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. Vision–language model for visual question answering in medical imagery. *Bioengineering*.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Courtney D Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment*, pages 13–18.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. 2024. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*.

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

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163.

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

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

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*.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Songtao Jiang, Yuan Wang, Sibo Song, Yan Zhang, Zijie Meng, Bohan Lei, Jian Wu, Jimeng Sun, and Zuozhu Liu. 2025. Omniv-med: Scaling medical vision-language model for universal visual understanding. *arXiv preprint arXiv:2504.14692*.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Yeying Jin, and Zuozhu Liu. 2024a. Modality-fair preference optimization for trustworthy mllm alignment. *arXiv preprint arXiv:2410.15334*.

Songtao Jiang, Yan Zhang, Chenyi Zhou, Yeying Jin, Yang Feng, Jian Wu, and Zuozhu Liu. 2024b. Joint visual and text prompting for improved object-centric perception with multimodal large language models. *arXiv preprint arXiv:2404.04514*.

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. 2024c. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3843–3860.

W Bradley Knox and Peter Stone. 2011. Augmenting reinforcement learning with human feedback. In *ICML 2011 Workshop on New Developments in Imitation Learning (July 2011)*, volume 855.

Nancy Krieger. 1992. Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. *American journal of public health*, 82(5):703–710.

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.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. 2022. Self-supervised vision-language pretraining for medical visual question answering. *arXiv preprint arXiv:2211.13594*.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

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.

Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, et al. 2023a. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, et al. 2024c. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*.

Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. 2023b. Q2atransformer: Improving medical vqa via an answer querying decoder. *arXiv preprint arXiv:2304.01611*.

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

OpenAI. 2023. GPT-4 technical report. https://arxiv.org/abs/2303.08774. *Preprint*, arXiv:2303.08774.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*.

Baolin Peng, Linfeng Song, Ye Tian, Lifeng Jin, Haitao Mi, and Dong Yu. 2023. Stabilizing rlhf through advantage model and selective rehearsal. *arXiv preprint arXiv:2309.10202*.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. 2024. Strengthening multimodal large language model with bootstrapped preference optimization. *arXiv preprint arXiv:2403.08730*.

Roger Kevin Pringle and Lawrence H Wyatt. 2006. The appropriate use of radiography in clinical practice: a report of two cases of biomechanical versus malignant spine pain. *Chiropractic & Osteopathy*, 14(1):1–8.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998.

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

Guohao Sun, Can Qin, Jiamian Wang, Zeyuan Chen, Ran Xu, and Zhiqiang Tao. 2025. Sq-llava: Self-questioning for large vision-language assistant. In *European Conference on Computer Vision*, pages 156–172. Springer.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Tim Van Erven and Peter Harremos. 2014. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820.

Tom van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. 2023. Open-ended medical visual question answering through prefix tuning of language models. *arXiv preprint arXiv:2303.05977*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Qinwei Xu, Xingkun Xu, Chenyi Zhou, Zuozhu Liu, Feiyue Huang, Shaoxin Li, Lifeng Zhu, Zhian Bai, Yuchen Xu, and Weiguo Hu. 2025. Towards normalized clinical information extraction in chinese radiology report with large language models. *Expert Systems with Applications*, 271:126585.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Kai Zhang, Jun Yu, Zhiling Yan, Yixin Liu, Eashan Adhikarla, Sunyang Fu, Xun Chen, Chen Chen, Yuyin Zhou, Xiang Li, et al. 2023a. Biomedgpt: A unified and generalist biomedical generative pre-trained transformer for vision, language, and multimodal tasks. *arXiv preprint arXiv:2305.17100*.

Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Yifan Sun, Haocheng Feng, et al. 2024. Automated multi-level preference for mllms. *arXiv preprint arXiv:2405.11165*.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. 2023b. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023c. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

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

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024b. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*.

Zhanhui Zhou, Jie Liu, Jing Shao, Xiangyu Yue, Chao Yang, Wanli Ouyang, and Yu Qiao. 2024c. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10586–10613.

Banghua Zhu, Michael I Jordan, and Jiantao Jiao. 2024a. Iterative data smoothing: Mitigating reward overfitting and overoptimization in rlhf. *arXiv preprint arXiv:2401.16335*.

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

A Appendix

A.1 Details of Figure 1

In Figure 1(a), we construct dispreferred responses by applying GPT-4o (Hurst et al., 2024) and LLaVA-Med1.5 (Li et al., 2024a) separately to paired images and captions from the PMC-15M dataset (Zhang et al., 2023b). Preference optimization is then conducted following prior works (Sun et al., 2023; Zhou et al., 2024a; Jiang et al., 2024a). The bar chart on the right illustrates the sampling probability distribution of preference datasets constructed using these two methods. It is evident that preference pairs generated by GPT-4o exhibit significantly lower sampling probabilities compared to those generated by medical VLMs, with most probabilities falling below 0.5. In contrast, the majority of VLM-generated data have sampling probabilities exceeding 0.6. This discrepancy suggests that external preferences derived from more powerful models are inconsistent with the inherent preferences embedded in medical VLMs' own generations, making them suboptimal for preference optimization datasets.

In Figure 1(b), we manually modify the original image-caption pairs from the PMC-15M dataset to introduce varying levels of quality, thereby creating multi-level dispreferred responses. These dispreferred responses, alongside the original correct captions, are used to construct preference pairs for preference optimization. When evaluated on the PathVQA (He et al., 2020) benchmark, the results are unsatisfactory, with no significant correlation observed between model performance and the quality levels of dispreferred responses. This indicates that the model fails to effectively learn distinctions between high-quality preference datasets.

A.2 Details of Preliminaries

Reinforcement Learning from Human Feedback (RLHF) Reinforcement Learning from Human Feedback (RLHF) is a widely used framework for aligning models with human preferences. It involves training a reward model f_ξ on pairwise preference data (Bai et al., 2022; Knox and Stone, 2011; Christiano et al., 2017). The reward model is optimized using a cross-entropy loss function:

$$\mathcal{J}_{\text{Reward}} = -\log\left(\sigma\left(f_\xi(x, y_w) - f_\xi(x, y_l)\right)\right), \tag{8}$$

where $\sigma(\cdot)$ is the logistic sigmoid function, and $f_\xi(x, y_w)$ and $f_\xi(x, y_l)$ represent the rewards assigned to the preferred and less preferred outputs, respectively. Once the reward model is trained, the policy π_η is optimized to maximize the expected reward while maintaining proximity to a reference policy π_{base}. This is achieved through the following objective:

$$\max_{\pi_\eta} \mathbb{E}_{x\sim\mathcal{P}, y\sim\pi_\eta(y|x)}\left[f_\xi(x, y) - \gamma D_{\text{KL}}(\pi_\eta(y|x) \parallel \pi_{\text{base}}(y|x))\right], \tag{9}$$

where γ controls the trade-off between reward maximization and regularization, and D_{KL} is the Kullback-Leibler divergence (Van Erven and Harremos, 2014), which ensures that policy remains stable and does not deviate excessively from the reference policy (Peng et al., 2023; Zhu et al., 2024a).

Direct Preference Optimization (DPO) Direct Preference Optimization (DPO) offers a streamlined alternative to RLHF by directly optimizing the policy π_η using preference data $\mathcal{P}$, thereby bypassing the need for an explicit reward model (Rafailov et al., 2024). DPO establishes a direct relationship between the reward function $g(x, y)$ and the policy π_η, expressed as:

$$g(x, y) = \gamma \log \frac{\pi_\eta(y|x)}{\pi_{\text{base}}(y|x)} + \gamma \log Z(x), \tag{10}$$

where $Z(x)$ is the partition function that ensures normalization. By substituting this relationship into the reward model loss, DPO formulates the following optimization objective:

$$\mathcal{J}_{\text{DPO}}(\pi_\eta; \pi_{\text{init}}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{P}}\left[\log\sigma\left(\gamma \log \frac{\pi_\eta(y_w|x)}{\pi_{\text{init}}(y_w|x)} - \gamma \log \frac{\pi_\eta(y_l|x)}{\pi_{\text{init}}(y_l|x)}\right)\right], \tag{11}$$

where y_w and y_l denote the preferred and less preferred outputs, respectively. This approach eliminates the complexity of training a separate reward model, enabling more efficient and precise alignment with human preferences (Dong et al., 2024; Pal et al., 2024).

In the context of Vision Language Models (VLMs), DPO typically involves combining the image m and textual query t into a unified input x (Jiang et al., 2024a; Yu et al., 2024b; Sun et al., 2023; Zhu et al., 2024b). The preferred responses y_w are selected for their accuracy and relevance, while the less preferred responses y_l often contain errors or irrelevant information.

A.3 Details of Evaluation Datasets

VQA-RAD (Lau et al., 2018) is a medical visual question answering dataset consisting of 3,515 question-answer pairs. The questions span 11 distinct categories and include both closed-ended (e.g., yes/no) and open-ended (e.g., descriptive) types, providing a comprehensive evaluation of model capabilities in medical contexts.

SLAKE (Liu et al., 2021) is a multimodal dataset designed for medical VQA, featuring over 7,000 question-answer pairs. It includes detailed annotations such as semantic segmentation masks and object detection bounding boxes, enabling fine-grained visual understanding. The dataset covers diverse anatomical regions, including the brain, neck, chest, abdomen, and pelvic cavity, and supports both English and Chinese. For consistency, our experiments focus exclusively on the English subset.

PathVQA (He et al., 2020) is a pathology-focused dataset containing 4,998 images paired with 32,799 question-answer pairs. The questions address various attributes, including spatial location, morphological features (e.g., shape, color), and pathological characteristics. These questions are categorized into open-ended and closed-ended types, offering a robust benchmark for evaluating model performance in pathology-related reasoning tasks.

A.4 Comparison with LiPO

While both our method and LiPO (Liu et al., 2024c) leverage list preference datasets for optimization, they differ significantly in scope, data generation, and training requirements. LiPO primarily focuses on RLHF for text-only models and tasks, whereas our approach, HSCR, is specifically designed for Vision-Language Models (VLMs) in medical multimodal tasks, aiming to enhance trustworthiness and modality alignment. In terms of preference data generation, LiPO employs prompt-based methods

Mask Ratio	RAD-VQA		SLAKE		PathVQA	
	Open	Closed	Open	Closed	Open	Closed
Baseline	32.31	56.62	42.45	56.49	10.01	59.75
0.3	33.21 (+0.90)	57.65 (+1.03)	43.33 (+0.88)	59.94 (+3.45)	10.34 (+0.33)	61.04 (+1.29)
0.5	33.35 (+1.04)	58.76 (+2.14)	44.47 (+2.02)	61.83 (+5.34)	11.96 (+1.95)	62.82 (+3.07)
0.9	34.21 (+1.90)	59.65 (+3.03)	44.35 (+1.90)	62.77 (+6.28)	12.15 (+2.14)	62.67 (+2.92)
0.7	35.92 (+3.61)	60.13 (+3.51)	45.32 (+2.87)	63.46 (+6.97)	12.36 (+2.35)	64.17 (+4.42)

Table 6: Performance comparison with different mask ratios.

Config	HSCR Training Config
Deepspeed	Zero2
Image encoder	CLIP-ViT-L/14@336px
Feature select layer	-2
Image projector	2 Linear layers with GeLU
Epoch	2
Learning rate	5e-7
Learning rate schedule	Cosine
Weight decay	0.0
Text max length	4096
Batch size per GPU	2
GPU	8 × 3090-24G
Precision	Bf16

Table 7: Our experimental hyperparameters.

to create sentence-level dispreferred responses that differ substantially from the preferred ones, enabling easier distinction early in optimization. In contrast, HSCR generates token-level preference data through visual token dropout and contrastive decoding, producing dispreferred responses that remain largely similar to the preferred ones while introducing critical divergences that reflect misalignment, thereby enabling finer-grained optimization. Regarding training requirements, LiPO relies on a reward model (T5-XXL, 11B parameters) trained with human-annotated datasets for ranking preference lists, which introduces significant computational overhead and depends heavily on large-scale human annotations—a resource particularly scarce in the medical domain. In contrast, HSCR eliminates the need for external MLLMs or human annotations by leveraging the model's internal behavior to generate and rank preference data, achieving superior performance with minimal cost and requiring only 2,000 training samples.

A.5 Details of Mask Strategies

Pixel-level Masking: In this approach, we directly crop 70% of the content from the original image, leaving only 30% of the pixels as input to the CLIP visual encoder. This technique is commonly employed in many previous VLMs.

Patch-level Masking: Given that the patch size of the CLIP visual encoder in LLaVA-Med is 14, we divide the original image into patches of size 14 and randomly discard 70% of them. As a result, only the remaining 30% of patches are fed into the Vision Transformer (ViT).

Latent-space Masking: In this strategy, we modify the attention mask within the CLIP visual encoder of LLaVA-Med by randomly setting 70% of the mask values to -inf, simulating a latent-space level mask.

Visual Token Dropout: This technique disrupts the latent space after the projector in the VLM. Specifically, the image is first encoded by the ViT, then transformed into visual tokens by an MLP projector, and finally concatenated with text tokens before being input into the LLM backbone. We drop 70% of the visual tokens to induce misalignment responses in the VLMs as they are restricted in their access to visual tokens.

Method	$CHAIR_s \downarrow$	Object Coverage ↑	Hallucination Rate ↓	Cognition ↓
LLaVA-v1.5-7B	7.8	51.0	36.4	4.2
LLaVA-v1.5-7B + DPO	6.5	49.1	34.5	3.4
LLaVA-v1.5-7B + HSCR	**4.4**	**55.5**	**25.1**	**2.9**

Table 8: Performance comparison on general multimodal benchmark.

A.6 Implementation Details of Figure 7

We evaluated LLaVA-v1.5-7B on the widely adopted AMBER benchmark (Wang et al., 2023), maintaining consistent experimental parameters with our main experiments. The preference datasets were constructed using HSCR based on the RLHF-V dataset (Yu et al., 2024a). The AMBER benchmark comprehensively assesses VLM trustworthiness through four key metrics:

- **CHAIR**: Quantifies object hallucination frequency in generated captions (lower values indicate better performance).
- **Object Coverage**: Measures the proportion of image objects accurately described in captions (higher values indicate better performance).
- **Hallucination Rate**: Evaluates the frequency of hallucinated objects in generated descriptions (lower values indicate better performance).
- **Cognition**: Measures the degree to which hallucinations in VLMs align with human cognitive patterns. This metric evaluates whether the model's hallucination tendencies resemble those observed in human cognition processes (lower values indicate better performance).

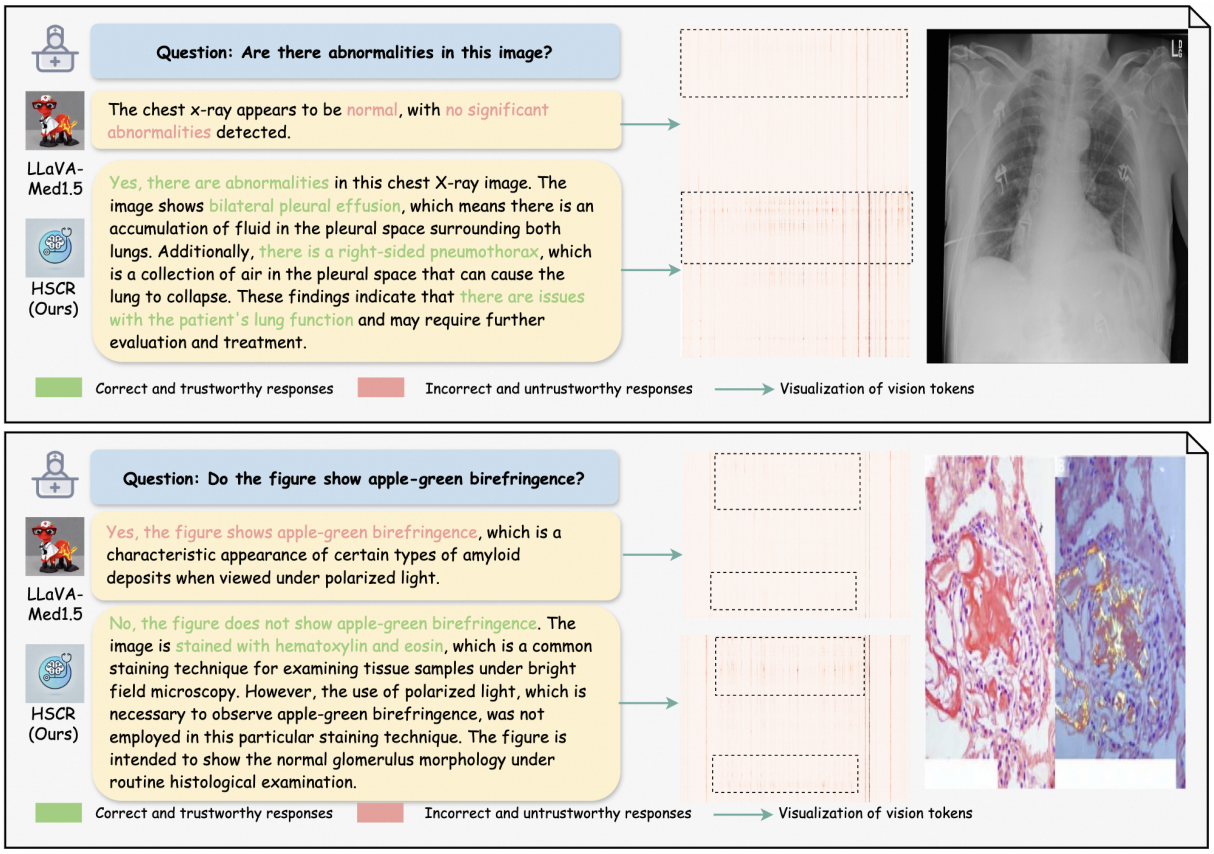

Figure 8: More cases on zero-shot comparison of LLaVA-Med1.5 (Li et al., 2024a) and HSCR.