ClinCoT: Clinical-Aware Visual Chain-of-Thought for Medical Vision Language Models

Xiwei Liu[1], Yulong Li[1], Xinlin Zhuang[1,2], Xuhui Li[1], Jianxu Chen[1], Haolin Yang[1], Imran Razzak[1], and Yutong Xie[1,†]

[1] Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2] East China Normal University, Shanghai, China
Yutong.Xie@mbzuai.ac.ae [†]Corresponding Author

Abstract. Medical Vision-Language Models have shown promising potential in clinical decision support, yet they remain prone to factual hallucinations due to insufficient grounding in localized pathological evidence. Existing medical alignment methods primarily operate at the response level through preference optimization, improving output correctness but leaving intermediate reasoning weakly connected to visual regions. Although chain-of-thought (CoT) enhances multimodal reasoning, it remains largely text-centric, limiting effective integration of clinical visual cues. To address this gap, we propose ClinCoT, a clinical-aware visual chain-of-thought framework that transforms preference optimization from response-level correction to visual-driven reasoning. We introduce an automatic data generation pipeline that constructs clinically grounded preference pairs through reasoning with hypotheses-driven region proposals. Multiple Med-LLM evaluators rank and assign scores to each response, and these rankings serve as supervision to train the target model. We further introduce a scoring-based margin-aware optimization strategy that incorporates both preference ranking and score difference to refine region-level reasoning trajectories. To maintain alignment as the model's policy evolves during training, we adopt an iterative learning scheme that dynamically regenerates preference data. Extensive experiments on three medical VQA and report generation benchmarks demonstrate that ClinCoT consistently improves factual grounding and achieves superior performance compared with existing preference-based alignment methods. Our code is available in `https://github.com/Xiwei-web/ClinCoT`.

Keywords: Med-VLM · Chain-of-thought · Preference optimization.

1 Introduction

Medical Vision-Language Models (Med-VLMs) have demonstrated promising potential in assisting clinical decision-making, including medical visual question answering (Med-VQA) and radiology report generation [14][17]. However, despite recent advances, Med-VLMs still suffer from a fundamental limitation: insufficient alignment between visual evidence and generated clinical conclusions [33][32]. In particular, models with poor modality alignment may rely

heavily on pretrained language priors while underutilizing localized pathological evidence, leading to hallucinated findings or clinically irrelevant responses [4][11].

To improve factual consistency, recent medical alignment methods adopt preference optimization strategies [1][9]. Clinical-aware preference learning frameworks such as MMedPO [33] construct medically meaningful preference pairs and introduce weighted optimization objectives to emphasize clinically relevant responses. Other alignment approaches refine model outputs through self-rewarding mechanisms [31], self-training with stronger evaluators [24], or modality alignment objectives [32]. While these methods improve response-level factuality and reduce overt hallucination, they largely operate at the output level, treating each response as a monolithic entity. Consequently, they do not explicitly model how localized pathological regions influence intermediate reasoning steps. The lack of region-aware reasoning limits the interpretability and pathological grounding of Med-VLMs, especially in complex diagnostic scenarios.

Beyond preference optimization, Chain-of-thought (CoT) reasoning [26][7][28] has emerged as an effective strategy for improving complex reasoning and stepwise decision-making in both language and vision-language models. In medical contexts, structured diagnostic prompting and reasoning-enhanced training strategies have been shown to improve explanation quality and logical consistency [22][23]. However, most existing CoT approaches remain predominantly text-centric [7]: they guide models to generate sequential reasoning tokens without explicitly restructuring visual attention. This implicitly assumes that the visual encoder uniformly captures all clinically relevant information, which is often unrealistic in medical imaging. Diagnostic reasoning in radiology fundamentally depends on detecting and examining localized abnormalities, such as small nodules, subtle consolidations, or focal fractures [29].

In medical practice, clinicians rarely reason over entire images uniformly. Instead, they form differential diagnostic hypotheses, examine disease-relevant regions, and iteratively refine conclusions based on localized evidence. This naturally raises a fundamental question:

Can preference optimization be extended beyond response-level correction toward hypotheses-driven clinical reasoning?

To address this challenge, we propose ClinCoT, a clinical-aware visual CoT framework that unifies region-level diagnostic hypotheses with margin-aware preference optimization under a coherent reasoning paradigm. Rather than merely improving final outputs, ClinCoT explicitly models how localized pathological evidence shapes intermediate reasoning trajectories, thereby encouraging closer alignment between visual grounding and clinical inference in a process-driven manner. Specifically, we design an automatic two-stage pipeline to construct clinically grounded preference data: 1) Hypotheses-Driven Region Generation: Given a medical image, we generate multiple disease-conditioned regions using a clinical-aware visual tool. Then the target Med-VLM answers the question by jointly processing the original image and each candidate region, forming multiple pathology-aware reasoning chains. 2) Consensus-Weighted Quality Assessment: Multiple medical LLM evaluators score the generated responses. These scores

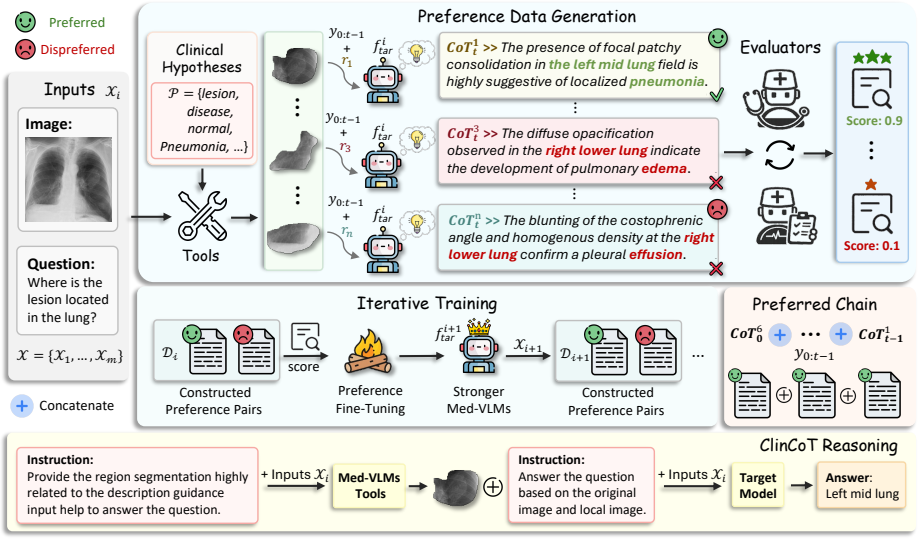

Fig. 1. Overview of the ClinCoT workflow. Given a subset of input pairs $\mathcal{X}_i \subset \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$, ClinCoT first employs a clinical-aware tool with a predefined hypotheses set $\mathcal{P}$ to generate region proposals $\{r_i\}_{i=1}^n$. The target model f_{tar}^i produces region-conditioned reasoning chains $\{y_t^i = CoT_t^i\}_{i=1}^n$ based on the preserved preferred history $y_{0:t-1} = \{CoT_0^6, \ldots, CoT_{t-1}^1\}$, integrating both the original image and the candidate regions. Med-LLM evaluators assign scores to construct preference pairs $\mathcal{D}_i$, distinguishing preferred and dispreferred responses. A consensus-weighted scoring based optimization updates f_{tar}^i to f_{tar}^{i+1} through iterative training. The updated model f_{tar}^{i+1} is then applied to $\mathcal{X}_{i+1}$ to generate new preference pairs $\mathcal{D}_{i+1}$ for the next iteration.

serve as proxies for the clinical validity of the region-conditioned reasoning chains and are adjusted based on evaluator agreement to ensure robust supervision.

Our contributions are three-fold: (1) an automatic clinical hypotheses-driven pipeline for scalable region-level preference data construction; (2) a consensus-weighted scoring based preference optimization with iterative learning for pathology-aware reasoning alignment, enabling finer discrimination of key regions and progressively stabilizing reasoning trajectories; and (3) extensive experiments on multiple medical VQA and report generation benchmarks demonstrating consistent improvements over strong medical baselines.

2 Methodology

Given a target model f_{tar}, a set of evaluator models f_{eval}, and an image-question pair $x = \{x_v, x_q\}$, we illustrate how to construct n preference data points, as shown in Figure 1. Assuming a reasoning chain of T steps, preference data are progressively generated throughout these T stages. At each timestep t, the pipeline consists of three stages: Response Generation, Response Evaluation and

Pair Construction (see Sec. 2.1). Unlike standard Direct Preference Optimization (DPO) [1], we adopt a consensus-weighted scoring based DPO, which not only ranks preference data but also assigns preference scores. This scoring enables more precise optimization based on score differences. During training, the rankings of the preference data act as supervision by minimizing negative log-likelihood loss, while the scores define the decision margin (see Sec. 2.2).

2.1 Preference Data Generation

Response Generation. This stage aims to construct clinically grounded region hypotheses and generate intermediate responses conditioned on localized visual contexts. A 'response' refers to any model output at an intermediate reasoning step, not necessarily the final answer to the question. We denote the response at timestep t as y_t, with the initial input x represented as y_0. Given a medical image x_v and a predefined set of clinical hypotheses $\mathcal{P} = \{p_i\}_{i=1}^n$, we employ a clinical-aware VLM (e.g., MedKLIP [27]) $\mathcal{T}(\cdot)$ to obtain disease-conditioned activation maps, which highlight image regions associated with the clinical concept p_i:

$$h_i = \mathcal{T}(x_v, p_i), \quad i = 1, \ldots, n, \tag{1}$$

where each $h_i \in [0,1]^{H \times W}$ is then used to localize the clinical region via thresholding and connected-component extraction: $r_i = x_v \odot \Phi(h_i)$. Here, $\Phi(\cdot)$ denotes a deterministic region extraction operator with $\odot$ representing element-wise masking. For each region hypothesis r_i, the target Med-VLM f_{tar} generates an intermediate response conditioned on both the global image and the localized region:

$$y_t^i = f_{\mathsf{tar}}(x_v, r_i, x_q \mid y_{0:t-1}), \tag{2}$$

resulting in a set of candidate responses $\{y_t^i\}_{i=1}^n$ that correspond to different region-conditioned visual interpretations of the same image, where $y_{0:t-1}$ denotes the currently preserved reasoning chain from previous timesteps.

Response Evaluation. This stage evaluates the quality of all generated responses by prompting the evaluator model to assign a score between 0 and 1, with higher scores indicating better prediction. The evaluator not only scores each individual response but also counts how this response influences the quality of its next response in the chain. This cumulative evaluation approach helps quantify the impact of each bounded region on the overall reasoning process. At timestep t, the evaluator assigns scores for y_t^i as follows:

$$s_{\mathsf{cur}}^i = f_{\mathsf{eval}}(y_t^i \mid y_{0:t-1}), \quad s_{\mathsf{nxt}}^i = \mathbb{E}_j[f_{\mathsf{eval}}(y_{t+1}^j \mid y_{0:t-1}, y_t^i)], \quad s^i = s_{\mathsf{cur}}^i + \gamma s_{\mathsf{nxt}}^i, \tag{3}$$

where s_{nxt}^i reflects the impact on the next response with expectation $\mathbb{E}[\cdot]$ estimated by randomly sampling j next responses according to current response y_t^i. $\gamma > 0$ is a hyperparameter to combine the current and next response scores, with $\gamma = 0$ at the last step. To mitigate the inherent bias of a single Med-LLM and ensure reliable assessments, we employ a *Consensus-Weighted Scoring* strategy

utilizing two distinct evaluators. We calculate the agreement between the two evaluators to penalize controversial assessments as follows:

$$s^i_{\mathsf{final}} = \left(\frac{s^i_1 + s^i_2}{2}\right) \cdot \exp\left(-|s^i_1 - s^i_2|\right) \tag{4}$$

where s^i_1 and s^i_2 are scores given by two Med-LLM evaluators respectively.
Pair Construction. From these scored reasoning response candidates $\{y^i_t\}^n_{i=1}$ with $\{s^i_{\mathsf{final}}\}^n_{i=1}$, we select k pairs to construct preference comparisons at each timestep t. For each pair, the response with the higher score is treated as the preferred response and concatenated with the historical chain $y_{0:t-1}$ to form the "preferred chain" y^w_t. Similarly, the lower-scoring response is concatenated with the same historical chain to form the "dis-preferred chain" y^l_t. Each pair of chains, together with their corresponding scores (y_w, s_w, y_l, s_l), constitutes one preference training example. The collection of all k such pairs forms the preference dataset $\mathcal{D}_t$ for timestep t. To ensure a stable forward reasoning trajectory, only the highest-scoring response at timestep t is concatenated with the historical chain to form the updated reasoning chain $y_{0:t}$. This updated chain is then used as the input for the next timestep. In other words, although multiple preference pairs are created for learning purposes at each timestep, only the single best-scoring chain is retained to continue the forward reasoning process.

2.2 Preference Fine-tuning

Margin-Aware Optimization. Given an input x, a language model policy π_θ can produce a conditional distribution $\pi_\theta(y \mid x)$ with y as the output text response. Standard DPO [1] reformulates reward learning as policy optimization by reparameterizing the reward under the PPO framework [21]:

$$r(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x). \tag{5}$$

Here π_{ref} denotes the frozen initialization of π_θ controlling the policy regularization, β is a temperature parameter and $Z(x)$ is a y-independent normalization term. Under the Bradley–Terry model [3], DPO minimizes the negative log-likelihood that a preferred response y_w is ranked above the dispreferred one y_l:

$$\begin{aligned} &P(y_w \succ y_l) = \sigma(r(x, y_w) - r(x, y_l))\,, \\ &\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\left[\log P(y_w \succ y_l)\right], \end{aligned} \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid function. In image-level reasoning, different regions contribute unequally to the final decision, motivating finer discrimination between preference pairs. To capture this variability, we introduce a margin term derived from preference scores and formulate a margin-aware objective:

$$\mathcal{L}_{\text{ClinCoT}}(\theta) = -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)} - (g(s_w) - g(s_l))\right)\right]. \tag{7}$$

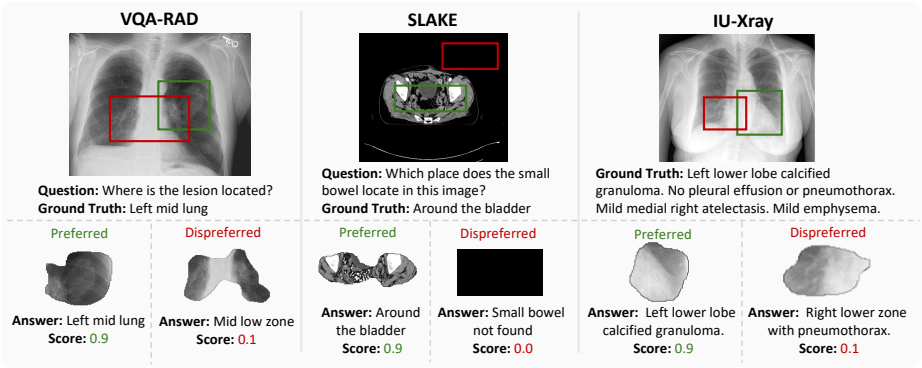

Fig. 2. Examples of generated preference pairs across different datasets. Preferred responses are more consistent with the ground truth and localized visual evidence, while dispreferred responses reflect incorrect localization, missed findings, or hallucinated abnormalities in clinically relevant regions.

Here, the function $g(\cdot)$ is monotonically increasing and maps preference scores into the logit space of the DPO objective, amplifying the key region's influence.

Let $\Delta_r = g(s_w) - g(s_l)$ and define the Gumbel-distributed random variables $R_w \sim \text{Gumbel}\left(r\left(x, y_w\right), 1\right)$ and $R_l \sim \text{Gumbel}\left(r\left(x, y_l\right), 1\right)$. The probability that the winning response is preferred, adjusted by the preference gap, is:

$$\begin{aligned} P\left(R_w - R_l > \Delta_r\right) &= \sigma\left(r\left(x, y_w\right) - r\left(x, y_l\right) - \Delta_r\right) \\ &= \sigma\left(\beta \log \frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\text{ref}}\left(y_w \mid x\right)} - \beta \log \frac{\pi_\theta\left(y_l \mid x\right)}{\pi_{\text{ref}}\left(y_l \mid x\right)} - \Delta_r\right). \end{aligned} \tag{8}$$

This result follows the definition of Gumbel random variables [1] and the Gumbel-max trick [18], with a similar derivation found in ODTO [1]. The Gumbel distribution models the extreme values of a variable, while Δ_r quantifies the degree of preference pairs difference. By maximizing the log-likelihood, we then obtain the adjusted loss function which explicitly optimizes preference learning by distinguishing not only the order but also the magnitude of preference differences.

Iterative Learning. Standard DPO optimizes on a static preference dataset, which may induce distributional mismatch as the model evolves during training. To alleviate this issue, we adopt an iterative preference learning strategy inspired by [30], where `ClinCoT` refines the target model through incrementally updated data generation and optimization. Specifically, the full image-question set $\mathcal{X}$ is partitioned into m disjoint subsets, $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_m\}$, each corresponding to one iteration. Starting from an initial target model f_{tar} and evaluators f_{eval}, iteration i first uses the current model f_{tar}^i to generate preference data $\mathcal{D}_i$ on subset $\mathcal{X}_i$. The model is then updated to f_{tar}^{i+1} by optimizing Eq. (7) on $\mathcal{D}_i$. This procedure repeats for m rounds, yielding the final model f_{tar}^m.

Table 1. Performance comparison on medical VQA and report generation tasks. 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 results and second best results are bold and underlined, respectively.

Models	SLAKE		VQA-RAD		IU-Xray		
	Open	Closed	Open	Closed	BLEU	ROUGE-L	METEOR
LLaVA-Med v1.5	44.26	61.30	29.24	63.97	14.56	10.31	10.95
+ DPO	49.30	62.02	29.76	64.70	16.08	12.95	17.13
+ Self-Rewarding	42.63	61.30	33.29	64.17	14.20	10.38	10.52
+ STLLaVA-Med	48.65	61.75	30.17	64.38	16.11	10.58	10.51
+ POVID	52.43	70.35	31.77	65.07	20.80	24.33	30.05
+ SIMA	51.77	69.10	31.23	64.80	17.11	22.87	29.10
+ FiSAO	52.69	70.46	32.70	64.11	21.06	25.72	30.82
+ MMedPO	**53.99**	73.08	**36.36**	**66.54**	23.49	29.52	34.16
+ **ClinCoT**	53.73±0.3	**73.41**±0.2	33.15±0.2	65.88±0.1	**24.96**±0.2	**29.79**±0.1	**35.44**±0.2
+ SFT	50.45	65.62	31.38	64.26	22.75	28.86	33.66
+ DPO	53.50	69.47	32.88	64.33	23.07	29.97	34.89
+ Self-Rewarding	50.62	65.89	32.69	65.89	22.89	28.97	33.93
+ STLLaVA-Med	52.72	66.69	33.72	64.70	22.79	28.98	34.05
+ POVID	52.18	70.67	32.95	64.97	23.95	29.75	34.63
+ SIMA	51.75	69.28	32.50	64.08	23.90	29.41	34.45
+ FiSAO	52.80	70.82	32.94	65.77	23.57	29.88	35.01
+ MMedPO	55.23	75.24	34.03	67.64	24.00	30.13	35.17
+ **ClinCoT**	**56.13**±0.5	**75.77**±0.2	**34.21**±0.1	**67.89**±0.4	**26.17**±0.2	**31.73**±0.2	**36.59**±0.1

3 Experiments

3.1 Experiment Setup

Evaluation Datasets. To verify the effectiveness of `ClinCoT` in improving factuality, we utilize three medical benchmarks: two medical VQA datasets, i.e., VQA-RAD [13] and SLAKE [16], and one report generation dataset IU-Xray [8].
Implementation Details. We utilize LLaVA-Med-1.5 7B [14] as the target model and apply LoRA fine-tuning [10] for preference optimization with a batch size of 4, a learning rate of 1e-7, $\gamma = 0.3$, $\beta = 0.1$ and train for 3 epochs. Each reasoning chain contains $T = 3$ timesteps, with $k = 2$ per step. Iterative learning of ClinCoT is performed over $m = 4$ rounds. LLaMA3-Med42-7B [5] and BioMistral-7B [12] are used as Med-LLM evaluators. All experiments are implemented using PyTorch 2.1.2 on two NVIDIA RTX A100 GPUs.
Baselines. We compare `ClinCoT` against standard DPO [20] and its variants, including the self-rewarding approach [31] and STLLaVA-Med [24]. The self-rewarding method constructs preference pairs from model-generated responses, while STLLaVA-Med enhances preference selection via GPT-4o for medical alignment. We also evaluate three VLM preference fine-tuning methods originally developed for natural images: POVID [32], SIMA [25], and FiSAO [6], as well as the medical-specific MMedPO [33]. To further evaluate training compatibility,

Table 2. Ablation study on key components of `ClinCoT`. 'w/o ClinCoT' removes intermediate reasoning and directly outputs answers. 'w/ naive DPO' applies standard DPO loss. 'w/o iterative learning' generates all preference pairs in a single pass and trains only once. 'w/o γ model' sets $\gamma = 0$ during response evaluation stage. 'single evaluator' uses a single Med-LLM LLaMA3-Med42-7B as evaluator.

Models	SLAKE		VQA-RAD		IU-Xray		
	Open	Closed	Open	Closed	BLEU	ROUGE-L	METEOR
`ClinCoT` w/o SFT	53.73±0.3	73.41±0.2	33.15±0.2	65.88±0.1	24.96±0.2	29.79±0.1	35.44±0.2
w/o `ClinCoT`	41.09±0.2	56.83±0.1	24.55±0.1	51.38±0.3	17.13±0.2	23.44±0.3	26.85±0.3
w/ naive DPO	50.22±0.1	71.39±0.3	31.06±0.4	62.61±0.1	22.76±0.2	28.51±0.1	33.18±0.4
w/o iterative learning	49.97±0.5	69.83±0.4	30.23±0.4	60.75±0.3	21.84±0.6	26.19±0.3	31.91±0.5
w/o γ	46.53±0.1	65.37±0.2	28.71±0.1	57.69±0.1	20.03±0.3	25.42±0.3	29.89±0.1
single evaluator	52.16±0.7	72.28±0.4	32.34±0.5	64.95±0.2	23.62±0.4	28.81±0.3	34.73±0.2

we assess `ClinCoT` and all baselines under supervised fine-tuning (SFT) settings to examine performance gains across training paradigms.

Evaluation Metrics. For the Med-VQA task, we use accuracy and recall for both closed-ended and open-ended questions. For the report generation task, we use BLEU Score [19], ROUGE-L [15] and METEOR [2] as the metrics.

3.2 Main Results

The overall performance comparisons are reported in Table 1. `ClinCoT` achieves the strongest performance among the compared baselines on report generation, but it underperforms MMedPO on VQA-RAD. This may suggest that response-level clinical weighting remains competitive for short-form VQA answers, where concise linguistic precision dominates. In contrast, `ClinCoT` emphasizes region-conditioned reasoning chains; without prior task adaptation, intermediate reasoning steps may introduce instability. Under SFT-enhanced setting, `ClinCoT` delivers the strongest overall performance. This suggests that SFT may provide a more stable domain-aligned initialization, which facilitates subsequent hypotheses-driven refinement. Once the base model is aligned with domain style, `ClinCoT` refines both answers and reasoning trajectories, yielding more consistent gains. Figure 2 visualizes some preference data from inference process.

3.3 Ablation Study

In Table 2, we present the ablations on key components. **1) Image-level CoT:** Removing CoT causes a significant drop, confirming its necessity. This highlights the potential of our method and suggests future work for improving region accuracy. **(2)Margin-aware DPO :** Results shows consistently degrades performance, indicating that incorporating preference score gaps is crucial for distinguishing subtle differences between reasoning chains. **(3) Iterative learning:** Eliminating iterative updates leads to noticeable declines, suggesting that dynamically generating preference data can help maintain alignment as the model

evolves. **(4) Response evaluation:** Removing the next-step score or using a single evaluator reduces performance, particularly on report generation, showing that stable, consensus-aware scoring improves long-horizon reasoning quality.

4 Conclusion

We propose `ClinCoT` that shifts preference optimization from response-level to hypotheses-driven reasoning. By combining disease-conditioned region proposals, consensus-weighted margin optimization, and iterative learning, ClinCoT aligns intermediate reasoning with localized pathological evidence. Experiments on medical VQA and report generation benchmarks show consistent improvements over strong medical baselines, with further gains under SFT. These results demonstrate that region-level clinical reasoning can be embedded into preference learning to improve factual grounding and reasoning stability in Med-VLMs.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Amini, A., Vieira, T., Cotterell, R.: Direct preference optimization with an offset. arXiv preprint arXiv:2402.10571 (2024)
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
3. 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)
4. Chen, J., Yang, D., Wu, T., Jiang, Y., Hou, X., Li, M., Wang, S., Xiao, D., Li, K., Zhang, L.: Detecting and evaluating medical hallucinations in large vision language models. arXiv preprint arXiv:2406.10185 (2024)
5. Christophe, C., Kanithi, P.K., Raha, T., Khan, S., Pimentel, M.A.: Med42-v2: A suite of clinical llms. arXiv preprint arXiv:2408.06142 (2024)
6. Cui, C., Zhang, A., Zhou, Y., Chen, Z., Deng, G., Yao, H., Chua, T.S.: Fine-grained verifiers: Preference modeling as next-token prediction in vision-language alignment. arXiv preprint arXiv:2410.14148 (2024)
7. Dai, P., Ou, Y., Yang, Y., Jin, Z., Suzuki, K.: Goca: Trustworthy multi-modal rag with explicit thinking distillation for reliable decision-making in med-lvlms. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 251–261. Springer (2025)
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 (2016)
9. 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)

10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
11. Jiang, Y., Chen, J., Yang, D., Li, M., Wang, S., Wu, T., Li, K., Zhang, L.: Medthink: Inducing medical large-scale visual language models to hallucinate less by thinking more. arXiv preprint arXiv:2406.11451 **1**(8) (2024)
12. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: Biomistral: A collection of open-source pretrained large language models for medical domains. arXiv preprint arXiv:2402.10373 (2024)
13. 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)
14. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
15. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
16. 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)
17. Liu, C., Wan, Z., Ouyang, C., Shah, A., Bai, W., Arcucci, R.: Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. arXiv preprint arXiv:2403.06659 (2024)
18. Maddison, C.J., Tarlow, D.: Gumbel machinery (2017), `https://cmaddis.github.io/gumbel-machinery`, available online
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
20. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S.: Direct preference optimization: Your language model is secretly a reward model. NeurIPS (2023)
21. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
22. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)
23. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., et al.: Toward expert-level medical question answering with large language models. Nature medicine **31**(3), 943–950 (2025)
24. Sun, G., Qin, C., Fu, H., Wang, L., Tao, Z.: Stllava-med: Self-training large language and vision assistant for medical. arXiv preprint arXiv:2406.19973 (2024)
25. Wang, X., Chen, J., Wang, Z., Zhou, Y., Zhou, Y., Yao, H., Zhou, T., Goldstein, T., Bhatia, P., Huang, F., et al.: Enhancing visual-language modality alignment in large vision language models via self-improvement. arXiv preprint arXiv:2405.15973 (2024)
26. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)

27. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21372–21383 (2023)
28. Wu, M., Liu, Z., Yan, Y., Li, X., Yu, S., Zeng, Z., Gu, Y., Yu, G.: Rankcot: Refining knowledge for retrieval-augmented generation through ranking chain-of-thoughts. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12857–12874 (2025)
29. 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)
30. Yu, T., Zhang, H., Yao, Y., Dang, Y., Chen, D., Lu, X., Cui, G., He, T., Liu, Z., Chua, T.S., et al.: Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. arXiv preprint arXiv:2405.17220 (2024)
31. Yuan, W., Pang, R.Y., Cho, K., Li, X., Sukhbaatar, S., Xu, J., Weston, J.E.: Self-rewarding language models. In: Forty-first International Conference on Machine Learning (2024)
32. Zhou, Y., Cui, C., Rafailov, R., Finn, C.: Aligning modalities in vision large language models via preference fine-tuning. arXiv preprint arXiv:2402.11411 (2024)
33. Zhu, K., Xia, P., Li, Y., Zhu, H., Wang, S., Yao, H.: Mmedpo: Aligning medical vision-language models with clinical-aware multimodal preference optimization. arXiv preprint arXiv:2412.06141 (2024)