# ANALYZING AND IMPROVING FINE-GRAINED PREFERENCE OPTIMIZATION IN MEDICAL LVLMS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Preference optimization is increasingly used to post-train medical large vision-language models (LVLMs), yet it operates at a much coarser granularity than the one that defines clinical correctness. Whether one response is clinically better than another usually comes down to a few decisive phrases, such as an anatomical laterality or a lesion attribute, and to whether each is supported by the image region the question concerns. Direct Preference Optimization (DPO) and its variants, by contrast, reduce the comparison to a single response-level scalar, an objective structurally unable to represent which tokens carry clinical meaning. We show that this mismatch is costly: of the objectives we compare, response-level DPO places the least reward on the phrases that decide clinical correctness, and substituting supervised references for preferred responses opens a stylistic gap that the model exploits as a reward-hacking shortcut, raising preference accuracy but not clinical accuracy. To capture the full feedback in a clinical comparison, we propose Fine-grained Regularized Medical Preference Optimization (FIRE-MPO). Preference pairs are built by minimally editing the model's own generations, so that preferred and rejected responses differ only on clinically decisive spans, and each is paired with a lesion-corrupted image withholding the supporting visual evidence. Because span-localized rewards are sparse, we stabilize optimization with a bidirectional token-wise KL regularizer. Across medical visual question answering and report generation, FIRE-MPO outperforms DPO, RRPO, and competing fine-grained objectives on two popular LVLMs, one medical and one general-purpose, while strengthening visual grounding and placing more reward on the medical phrases than the alternatives.

## 1 INTRODUCTION

In medicine, the clinical correctness of a response can depend on only a few critical details, e.g., an anatomical laterality (Sangwaiya et al., 2009; Luetmer et al., 2013), a lesion attribute (Hong et al., 2005), or a measurement (MacMahon et al., 2017; Eisenhauer et al., 2009), each of which must be supported by the relevant region of the image (Bruno et al., 2015; Waite et al., 2017). Preference optimization provides a way to align models with such judgments and has become an increasingly common approach to post-training (Rafailov et al., 2023; Ethayarajh et al., 2023; Azar et al., 2024). Direct Preference Optimization (DPO), in particular, learns directly from preferred and rejected responses without requiring an explicit reward model (Hein et al., 2025b; Zhu et al., 2024; Savage et al., 2025; Sun et al., 2024). However, clinical judgments are often much more specific than the response-level preference that DPO learns from.

Consider two responses for a given chest X-ray that differ only in whether an opacity is localized to the *left* or *right* lower lobe. The clinically relevant supervision is not merely that one response is preferred, but that *left* is the decisive phrase and is grounded in a specific region of the image. DPO and its variants compress the preference between two responses into a single response-level scalar. In doing so, they discard much of the information carried by a clinical comparison: not only which response is superior, but which phrases make it so and which regions of the image provide the supporting evidence.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This limitation has two measurable consequences, which we establish empirically in Section 2 before introducing our method. The first is *reward misallocation*. Because a response-level margin is spread across every token of the response, nothing in the optimization determines whether the learned reward settles on the clinical content at all. Indeed, when we measure the token-level implicit reward on radiology reports, DPO places little of it on the medical phrases that decide clinical correctness. The second is *stylistic reward hacking*. Because expert preference annotation is expensive, medical DPO commonly uses the supervised reference as the preferred response (Puli et al., 2024; Zhu et al., 2024). References and model samples, however, differ systematically in length and structure, and a response-level objective cannot separate this stylistic gap from the clinical correction it is meant to reward. The model exploits the gap it can see, so preference accuracy rises while clinical accuracy stalls. Neither failure is resolved by collecting more preference data or scaling the backbone: both stem from the mismatch between the granularity of the supervision and that of clinical correctness, which the remainder of this paper addresses.

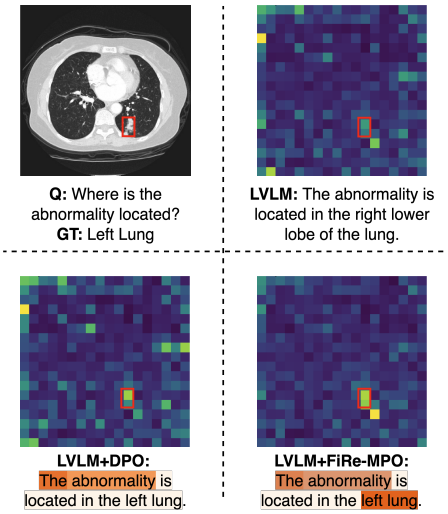



Figure 1: Reward assignment and visual attention in DPO vs. FIRE-MPO. DPO inherits the diffused attention of the base model and fails to prioritize the clinically decisive "left lung" tokens (darker indicates higher implicit reward).

FIRE-MPO addresses this mismatch by matching the granularity of the supervision to that of clinical correctness, in both the text and the image, drawing on two lines of work that have so far been developed separately. On the textual side, we follow fine-grained preference objectives from the general domain, RRPO (Sarkar & Etemad, 2025) and Mask-DPO (Gu et al., 2025), which restrict the reward to the tokens that differ between the two responses. To make the differing tokens coincide with the clinically decisive ones, we construct preference pairs by minimally editing the model's own generations: only the clinically erroneous spans are replaced, so the preferred and rejected responses share the model's phrasing and differ solely in clinical content. Applying the reward to these spans removes the stylistic gap that off-policy pairs expose and concentrates the learning signal on the decisive phrases. On the visual side, we follow image-conditioned preference objectives, mDPO (Wang et al., 2024) and its medical counterpart MMedPO (Zhu et al., 2024), which contrast the original image with a corrupted one. We present each pair with a lesion-corrupted image from which the evidence supporting the edited spans has been removed, and reverse the preference, so the model is penalized for committing to the correct answer once the evidence for it is gone. This discourages correct guesses made for the wrong reasons. Because a reward carried by a handful of tokens is far sparser than a response-level margin, we anchor the policy with a bidirectional token-wise KL regularizer, which allows new clinical concepts to be learned without collapsing the model's generative diversity. We evaluate FIRE-MPO on two tasks: medical visual question answering and chest X-ray report generation.

Our contributions are threefold:

- **A diagnosis of the granularity mismatch.** We argue that response-level preference optimization does not reflect medical correctness, and show that DPO places little reward on medical phrases and, with supervised references as preferred responses, learns to reward style instead.
- **A preference objective matched to clinical granularity.** Inspired by fine-grained textual preference methods, we propose FIRE-MPO, which rewards only the clinically decisive spans, ties them to their visual evidence through lesion-corrupted images, and stabilizes training with a bidirectional token-wise KL regularizer.
- **Consistent empirical gains.** On HuatuoGPT-Vision-7B and Qwen3-VL-4B-Instruct, evaluated on medical VQA (SLAKE, VQA-RAD) and chest X-ray report generation (IU-Xray), FIRE-MPO attains the best average score against five preference-optimization baselines, improving on DPO by 5.3 and 4.7 points, respectively, while placing more reward on medical phrases and keeping attention on the queried image region.

# 2 THE GRANULARITY MISMATCH IN MEDICAL PREFERENCE OPTIMIZATION

In this section, we empirically examine the granularity mismatch. After introducing the objectives underlying our analysis (Section 2.1), we investigate two consequences of this mismatch: first, the implicit reward induced by a response-level objective may not concentrate on clinicaly relevant content (Section 2.2); second, standard medical preference pair construction may reward stylistic differences rather than clinical correctness (Section 2.3).

## 2.1 PRELIMINARIES

**DPO.** Given a vision-language pair $\{v, q\}$, DPO (Rafailov et al., 2023) seeks to align the policy model $\pi_\theta$ by increasing the likelihood of a chosen response $y^+$ relative to a rejected one $y^-$. This alignment is achieved by maximizing the reward margin between $\pi_\theta$ and a fixed reference model $\pi_{\text{ref}}$ via the following objective:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}\left[\log \sigma\left(r_\theta(v, q, y^+) - r_\theta(v, q, y^-)\right)\right], \tag{1}$$

where the implicit reward is formulated as $r_\theta(v, q, y) = \beta \log \frac{\pi_\theta(y|v,q)}{\pi_{\text{ref}}(y|v,q)}$. Here, $\beta$ serves as a hyperparameter that controls the strength of the constraint toward the reference distribution.

However, this formulation relies on a *coarse-grained* reward mechanism. Since $r_\theta(v, q, y)$ is calculated at the response level, the optimization process penalizes or rewards all tokens in $y$ uniformly. This approach is often suboptimal for fine-grained alignment, as the semantic divergence between $y^+$ and $y^-$ may be concentrated within only a few critical tokens. Furthermore, the reliance on a response-level reward means that for extended outputs, the gradients for $\mathcal{L}_{\text{DPO}}$ can become disproportionately large. Such volatility often leads $\pi_\theta$ to diverge, resulting in a loss of capabilities and performance degradation (Xu et al., 2024; Yan et al., 2024; Cho et al., 2025).

Several recent methods localize reward signals to the token or span level (Sarkar & Etemad, 2025; Gu et al., 2025; Zeng et al., 2024; Zhu et al., 2025); we discuss these approaches, and position FIRE-MPO relative to the most closely related method, RRPO (Sarkar & Etemad, 2025), in Section 5.

## 2.2 WHERE THE IMPLICIT REWARD IS SPENT

A response-level objective assigns a single scalar margin to the entire response without specifying whether the preference for the chosen response over the rejected one is driven by clinically meaningful content, such as *left lower lobe*, or generic phrasing, such as *the image shows*. Since the DPO margin depends only on the *sum* of per-token log-ratios, we ask which tokens produce it. Log-ratios of different tokens within a response are not directly comparable (Rafailov et al., 2024), but the log-ratio of the *same* token under the *same* prefix is comparable across models that share $\pi_{\text{ref}}$. We therefore let three aligned models, DPO, MMedPO (Zhu et al., 2024), which weights each pair by clinical significance but still assigns one scalar per response, and FIRE-MPO (Section 3), score the same fixed text: the ground-truth IU-Xray reports (Demner-Fushman et al., 2016) and a medically incorrect counterpart of each. For each report, we record which model assigns the most reward, $\log(\pi_\theta/\pi_{\text{ref}})$, to its medical tokens, using non-medical tokens as a control (full protocol in Appendix E).

Table 1: **Token-level reward allocation across preference objectives.** Entries report the percentage of IU-Xray reports on which each method assigns the highest reward $\log(\pi_\theta/\pi_{\text{ref}})$ to medical and non-medical tokens.

| Method | Correct $y^+$ | | Incorrect $y^-$ | |
|---|---|---|---|---|
| | Med. ↑ | Non-med. ↓ | Med. ↑ | Non-med. ↓ |
| DPO | 14.0 | 36.2 | 16.5 | 34.8 |
| MMedPO | 27.0 | 33.4 | 25.1 | 33.6 |
| FIRE-MPO | **59.0** | 30.4 | **58.4** | 31.6 |

As shown in Table 1, both response-level objectives place little of their reward on clinical content. On correct reports, DPO and MMedPO assign the highest reward to medical tokens in only $14\%$ and $27\%$ of cases, while DPO is the most frequent winner on non-medical tokens. On incorrect responses, they likewise rarely penalize the erroneous medical tokens most. By rewarding only the clinically decisive spans and anchoring the remaining tokens to $\pi_{\text{ref}}$ with its KL regularizer, FIRE-MPO (Section 3) reverses this pattern: it wins on medical tokens in $59.0\%$ of correct and $58.4\%$ of incorrect reports while winning least often on non-medical tokens, so its advantage is specific to clinical content rather than a uniformly larger departure from $\pi_{\text{ref}}$.

### 2.3 Reward Hacking in Medical Preference Learning

In medical DPO, expert preference annotation is often expensive, so a common alternative is to use the fixed SFT reference response as the preferred answer (Puli et al., 2024; Zhu et al., 2024). This construction introduces an off-policy distribution shift due to the distinct style of Ground Turth (GT) medical responses, which are often shorter and more structured, whereas model responses are typically longer and more explanatory. This can confound clinical correctness with response style, a failure mode we refer to as stylistic reward hacking.

This issue is particularly problematic in clinical alignment because the decisive difference between two responses is often localized to a small span, such as an anatomical side, lesion attribute, or abnormality label. If the preferred and rejected responses also differ in length, syntax, or format, the optimization signal no longer isolates the clinically meaningful error.

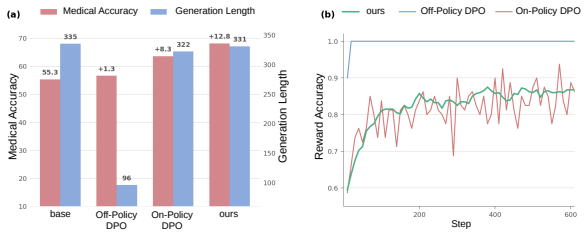



Figure 2: Ground-truth references as the preferred response reward style, not correctness. (a) Off-policy DPO copies the reference length without gaining medical accuracy. (b) Reward accuracy rises regardless: the signature of stylistic reward hacking.

Figure 2 illustrates this effect by comparing model outputs and reward accuracy under `off-policy` and `on-policy` preference construction. Under off-policy DPO, high reward accuracy together with low medical accuracy indicates this reward hacking: the model becomes better at satisfying the preference objective without necessarily correcting the clinically relevant visual finding. In contrast, on-policy preference construction keeps the response style closer to the model's own outputs.

## 3 Method

FiRe-MPO is based on a single principle, where the supervision provided to an objective should match the granularity of the judgment it is intended to reproduce.Section 2 established the structure of this judgment, a verdict determined by a small number of clinically decisive spans and supported by localized regions of the image, as well as how a response-level objective captures this structure. We next formalize both components and introduce an objective designed to incorporate them. Preference pairs are built to differ only in clinically decisive spans, and the ranking loss rewards only those spans. Each pair is accompanied by a lesion-corrupted counterpart in which the visual evidence supporting those spans is removed, and the ranking loss is inverted accordingly. Because localizing the reward in this manner makes the training signal sparse, the token-wise regularizer introduced below is essential for effective optimization. Our objective is most closely related to RRPO (Sarkar & Etemad, 2025); Section 5 positions our approach relative to RRPO and details the key differences.

**Ranking Loss.** Given the subtle visual nuances in the medical domain, our objective is to prevent the model from hallucinating correct-sounding responses when essential visual evidence is absent or corrupted. We achieve this by structuring the ranking loss around a dual-constraint. First, we define preference pairs $(q, v, y^+) \succ (q, v, y^-)$, which align $\pi_\theta$ toward medically accurate answers when supported by valid visual evidence $v$. Second, we introduce pairs $(q, v', y^-) \succ (q, v', y^+)$, which compel the model to refrain from providing a definitive response when the underlying visual evidence is corrupted ($v'$). This ensures that the model prioritizes

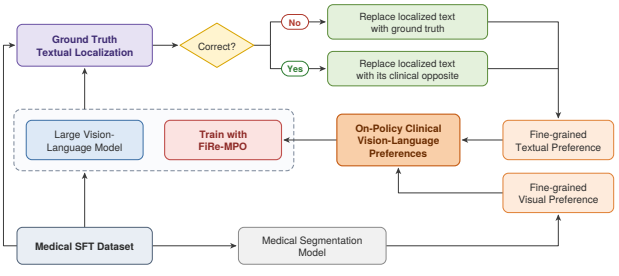



Figure 3: An overview of the fine-grained on-policy visual-textual preference creation process

visual evidence over linguistic priors, effectively penalizing correct guesses made for the wrong reasons.

Figure 3 illustrates the overall flow of our data and training pipeline. First, we obtain a response from $\pi_\theta$ for a given $\{v, q\}$. If the output contains clinically designated errors, we substitute only the erroneous medical statements with ground-truth information to form $y^+$, while the original output serves as $y^-$. Conversely, if the initial response is correct, we use them as $y^+$ and generate $y^-$ by altering the response with plausible medical counterfactuals. In both scenarios, the model's original linguistic style is preserved to allow on-policy alignment. We construct the first set of pairs as $(q, v, y^+)$ and $(q, v, y^-)$. Next, we construct the second set of pairs as $(q, v', y^-)$ and $(q, v', y^+)$, where we obtain $v'$ by applying targeted Gaussian noise to specific lesions or anatomical landmarks on $v$ relevant to the query. Note that in both cases, we provide a fine-grained reward only on the clinically divergent tokens between $y^+$ and $y^-$ by utilizing the ranking loss defined as:

$$\begin{aligned}\mathcal{L}^{(\text{Rank})}_{\text{FiRe-MPO}}(\pi_\theta; \pi_{\text{ref}}) = -\,\mathbb{E}\Bigg[\log\sigma\left(\sum_i r_\theta(v, q, y_i^+) - \sum_i r_\theta(v, q, y_i^-)\right) \\ + \gamma \log\sigma\left(\sum_i r_\theta(v', q, y_i^-) - \sum_i r_\theta(v', q, y_i^+)\right)\Bigg],\end{aligned} \tag{2}$$

**KL Regularizer.** Restricting the reward to the spans on which $y^+$ and $y^-$ differ is what makes the signal clinically precise, but it also makes it sparse: a margin previously carried by an entire response is now carried by a handful of tokens. The regularizer is therefore not an auxiliary refinement but the condition under which fine-grained supervision is usable at all. Removing it costs 22.28 points of Green Score on IU-Xray (Table 3), a collapse an order of magnitude larger than the effect of any other design choice we ablate.

The direction of the anchor decides which failure the model is exposed to. The forward term $\mathbb{D}_{\text{KL}}(\pi_{\text{ref}} \parallel \pi_\theta)$, as used in RRPO, reduces at the token level to $\arg\max_\theta \mathbb{E}_{y\sim\pi_{\text{ref}}}[\log \pi_\theta(y \mid x)]$, so it requires $\pi_\theta$ to place mass wherever $\pi_{\text{ref}}$ does. This preserves the linguistic variety and general clinical knowledge of the base model, but it is mean-seeking, and it penalizes the policy for moving mass onto the new concepts that domain adaptation requires, low-precision modes included. The reverse term $\mathbb{D}_{\text{KL}}(\pi_\theta \parallel \pi_{\text{ref}})$ decomposes as $\arg\max_\theta \mathbb{E}_{y\sim\pi_\theta}[\log \pi_{\text{ref}}(y \mid x)] + \mathcal{H}(\pi_\theta(y \mid x))$ and acts as an on-policy regularizer, penalizing the policy for straying into regions the reference finds unlikely. This tail suppression is what prevents a medical LVLM from emitting spuriously attractive but clinically incorrect tokens, but on its own its mode-seeking behaviour collapses generation into repetitive responses. The two failure modes are opposite in character, and in a clinical setting neither is acceptable on its own.

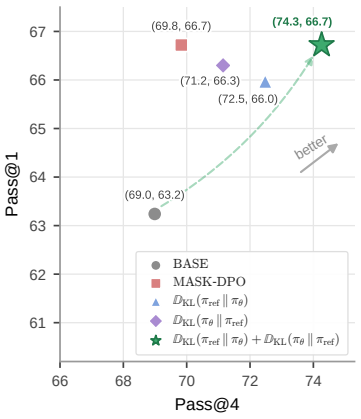



Figure 4: Pass@1 vs. Pass@4 under different regularizers. Bidirectional KL attains the highest clinical precision while retaining distributional richness.

Figure 4 shows the trade-off directly: neither force in isolation is sufficient for high-stakes alignment, since one preserves coverage at the cost of precision and the other buys precision by discarding the interpretive variety that complex medical reasoning needs. We therefore replace the unidirectional constraint with a bidirectional, token-wise regularizer that applies both constraints at the token level:

$$\begin{aligned}\mathcal{L}^{(\text{KL})}_{\text{FiRe-MPO}}(\pi_\theta; \pi_{\text{ref}}) = -\,\mathbb{E}\Bigg[\sum_t \Big(\lambda \mathbb{D}_{\text{KL}}\big(\pi_{\text{ref}}(y_t|v, q, y_{<t}) \parallel \pi_\theta(y_t|v, q, y_{<t})\big) \\ + (1-\lambda)\mathbb{D}_{\text{KL}}\big(\pi_\theta(y_t|v, q, y_{<t}) \parallel \pi_{\text{ref}}(y_t|v, q, y_{<t})\big)\Big)\Bigg],\end{aligned} \tag{3}$$

where $\lambda$ is a hyperparameter that dictates the relative influence of each constraint.

Table 2: Comparison of preference-optimization losses on medical VQA and report-generation benchmarks. SLAKE and VQA-RAD report Closed and Open accuracy; IU-Xray reports the Green Score, Precision, and Recall. The Average column aggregates across all reported metrics. Higher is better; best and second best per backbone in **bold** and underline, respectively. Text* denotes fine-grained text. All methods are reproduced by us under a single unified experimental protocol.

| Models | Text* Pref. | Img. Pref. | SLAKE Closed | SLAKE Open | VQA-RAD Closed | VQA-RAD Open | IU-Xray Green Score | IU-Xray Prec. | IU-Xray Rec. | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| HuatuoGPT-Vision-7B | - | - | 70.33 | 47.49 | 68.58 | 40.53 | 71.55 | 76.44 | 58.47 | 61.21 |
| + DPO | × | × | 79.40 | 55.38 | 71.65 | 39.47 | 71.66 | 74.07 | 59.32 | 64.46 |
| + mDPO | × | ✓ | 73.59 | 56.10 | 72.80 | 40.58 | 73.17 | 67.80 | 62.03 | 64.82 |
| + MMedPO | × | ✓ | 79.40 | 59.11 | 71.50 | 41.20 | 74.50 | 68.00 | **68.50** | 66.40 |
| + RRPO | ✓ | × | 79.12 | 59.11 | 71.26 | 43.68 | 78.84 | 77.12 | 64.24 | 68.15 |
| + MASK-DPO | ✓ | × | **81.32** | 57.10 | **73.18** | 43.68 | 58.42 | 58.98 | 55.59 | 61.52 |
| **+ FIRE-MPO (Ours)** | ✓ | ✓ | **81.32** ↑ 10.9 | **61.26** ↑ 13.7 | 71.26 ↑ 2.68 | **45.26** ↑ 4.73 | **80.70** ↑ 9.15 | **82.54** ↑ 6.10 | 63.56 ↑ 5.09 | **69.71** ↑ 8.50 |
| Qwen3-VL-4B-Instruct | - | - | 76.64 | 56.24 | 64.75 | 40.00 | 72.18 | 85.08 | 71.18 | 63.24 |
| + DPO | × | × | **82.41** | 50.78 | 67.43 | 38.94 | 71.11 | 82.71 | 71.18 | 62.73 |
| + mDPO | × | ✓ | 77.19 | 56.24 | 68.58 | 40.00 | 71.66 | 82.71 | 71.18 | 63.87 |
| + MMedPO | × | ✓ | 78.00 | 57.50 | 69.20 | 40.60 | 73.00 | 74.50 | 72.37 | 64.90 |
| + RRPO | ✓ | × | 77.19 | 60.11 | 67.43 | **41.57** | 73.27 | 85.08 | 72.37 | 65.26 |
| + MASK-DPO | ✓ | × | 80.49 | 59.54 | 69.34 | 39.94 | 75.11 | 79.83 | **76.27** | 66.12 |
| **+ FIRE-MPO (Ours)** | ✓ | ✓ | 81.32 ↑ 4.68 | **60.25** ↑ 4.01 | **71.26** ↑ 6.51 | 41.05 ↑ 1.05 | **76.22** ↑ 4.04 | **89.53** ↑ 4.45 | 73.89 ↑ 2.71 | **67.41** ↑ 4.17 |

**Final Loss.** The complete FIRE-MPO objective is thus formulated by unifying the ranking loss with the KL regularizer:

$$\mathcal{L}_{\text{FiRe-MPO}}(\pi_\theta; \pi_{\text{ref}}) = \mathcal{L}^{(\text{Rank})}_{\text{FiRe-MPO}}(\pi_\theta; \pi_{\text{ref}}) + \alpha \cdot \mathcal{L}^{(\text{KL})}_{\text{FiRe-MPO}}(\pi_\theta; \pi_{\text{ref}}), \tag{4}$$

where $\alpha$ is a loss coefficient to control the influence of fine-grained reward and model divergence.

# 4 EXPERIMENTS

In this section, we present our training data construction process, experimental setup and results.

## 4.1 SETUP

**Base Models.** We evaluate FIRE-MPO on two open-source backbones spanning distinct points in the model design space: `HuatuoGPT-Vision-7B` (Chen et al., 2024), a medical LVLM pre-trained on large-scale medical image-text pairs, and `Qwen3-VL-4B-Instruct` (Bai et al., 2025), a general-purpose model with a high-resolution visual adapter. We compare against response-level DPO, mDPO (Wang et al., 2024), which adds a visual preference component, MMedPO (Zhu et al., 2024), and RRPO (Sarkar & Etemad, 2025) and MASK-DPO (Gu et al., 2025), which localize the textual signal, with RRPO also regularizing the policy toward the reference model.

**Training Datasets.** We use three benchmarks spanning distinct clinical modalities and tasks: VQA-RAD (Lau et al., 2018), radiology VQA over head, chest, and abdomen images; SLAKE (Liu et al., 2021), of which we use the English portion and the visual-textual pairs only, without the semantic labels or knowledge graph; and IU-Xray (Demner-Fushman et al., 2016) for chest X-ray report generation. Dataset sizes and splits are provided in Appendix C.

**Implementation Details.** We fine-tune using LoRA (Hu et al., 2022), applied to the LLM component only, while freezing the vision encoder and cross-modal projectors to preserve pre-trained perceptual features. Preference pairs are built using `GPT-4o-mini` for VQA extraction and `Gemini 3 Flash Preview` for report refinement, with constraints ensuring that only clinical accuracy changes while the target response style remains unchanged. Corrupted images are produced by localizing the queried concept with MedSAM3-v1 (Liu et al., 2025a) and adding targeted noise to the localized and segmented region. The default hyperparameters are $\lambda = 0.5$, $\alpha = 0.01$, $\gamma = 0.1$; more details on the corrupted image generation pipeline and optimization setting is provided in Appendix C.

**Evaluation Benchmarks and Metrics.** For the Med-VQA and report generation tasks, all evaluations are conducted on the standard in-distribution test splits of their respective datasets. For Med-

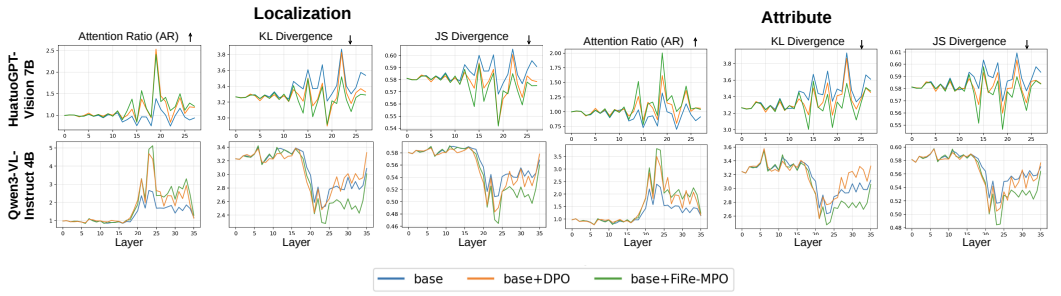



Figure 5: FIRE-MPO attends to the image region a clinical claim depends on. Each panel tracks one grounding metric layer by layer, measured on localization and attribute tokens. Higher Attention Ratio and lower KL/JS divergence both mean attention stays on the queried region; the gap widens in the deeper layers, where the prediction is formed.

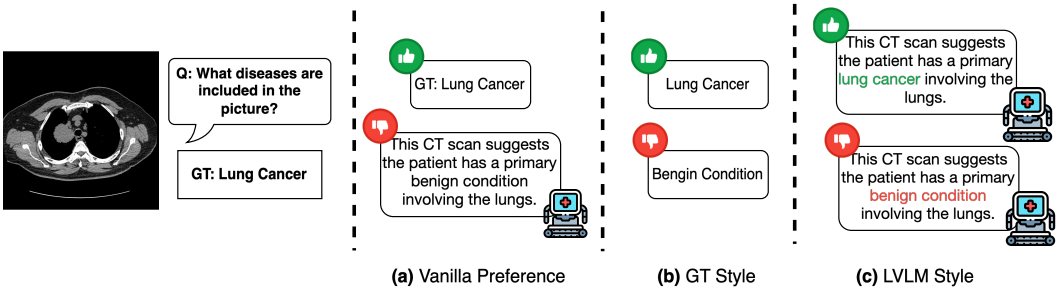



Figure 6: Preference-pair construction in the textual-style ablation. (a) Vanilla: the SFT ground truth is preferred over the model's own output, so the pair also differs in style. (b) GT style: both responses follow the short reference format. (c) LVLM style (ours): only the clinically incorrect span of the model's response is edited, so the pair differs in clinical content alone. (b) and (c) correspond to the first and third rows of Table 6; (a) is compared against both in Appendix F.

VQA, we report accuracy for both Closed (binary/categorical) and Open (free-form) questions on VQA-RAD and SLAKE. Given the question, its ground truth, and corresponding model response, `GPT-4o-mini` is used as the judge. For medical semantic consistency on IU-Xray, we utilize the GREEN Score (Ostmeier et al., 2024), a specialized metric that prioritizes medical correctness over surface-level n-gram overlap, and report the accuracy, precision, and recall of the generated reports against the ground-truth test set. To assess visual grounding, we evaluate on VGMED (Liu et al., 2026), a grounding-centric extension of SLAKE dataset designed specifically to measure model sensitivity to subtle yet diagnostically decisive pathological features. Following the benchmark protocol, we measure the Attention Ratio (AR) (Khayatkhoei et al., 2025), KL Divergence, and Jensen-Shannon (JS) divergence (Liu et al., 2026) for critical clinical tokens (Localization and Attribute).

### 4.2 MAIN RESULTS

**Quantitative Performance.** Table 2 reports the results for all preference-optimization strategies across the medical VQA and report-generation benchmarks. FIRE-MPO attains the best overall average on both backbones, and the gains are concentrated where clinical correctness is hardest to fake: the open-ended and report-generation splits improve the most, while the closed-ended splits improve the least. The same performance ordering holds across both the medical-specialized `HuatuoGPT-Vision-7B` and the general-purpose `Qwen3-VL-4B-Instruct`, suggesting that the gains are not specific to a particular backbone or its prior medical knowledge. On IU-Xray, FIRE-MPO achieves the highest GREEN Score, the metric most closely aligned with clinical correctness, on both backbones. Qualitative examples are given in Section H in the Appendix.

Table 3: **Ablation on the KL regularization.** Removing either KL reduces performance, with the largest degradation observed when forward KL is removed. The default bidirectional regularizer achieves the best overall performance.

| Loss Term | Slake | VQA-RAD | IU-XRAY | Average |
|---|---|---|---|---|
| $\mathcal{L}^{(\text{Rank})}$ | 65.40 | **60.75** | 58.42 | 61.52 |
| + $\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ (Reverse KL) | 66.35 | 59.65 | 80.98 | 68.99 |
| + $\mathbb{D}_{\text{KL}}(\pi_{\text{ref}} \| \pi_\theta)$ (Forward KL) | 65.97 | 59.65 | 78.84 | 68.15 |
| + $\mathbb{D}_{\text{KL}}(\pi_{\text{ref}} \| \pi_\theta) + \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ | **68.14** | 60.31 | **80.70** | **69.72** |

Table 4: **Exploring of image corruption strategies.** Corrupting only the queried lesion beats every global perturbation, so the visual signal must be as localized as the textual one.

| Corruption Type | Slake | VQA-RAD | IU-XRAY | Average |
|---|---|---|---|---|
| Cropped | 56.13 | 58.76 | 70.13 | 61.67 |
| Image Level Noise | 66.35 | 59.65 | 79.10 | 68.37 |
| Full Black | 66.72 | 55.28 | 60.22 | 60.74 |
| Lesion-Based* | **68.14** | **60.31** | **80.70** | **69.72** |

Table 5: **Ablation on the visual loss.** $v'$ denotes a corrupted image. Contrasting against corrupted images hurts performance, while penalizing correct responses given corrupted images yields the best average.

| Loss Term | Slake | VQA-RAD | IU-XRAY | Average |
|---|---|---|---|---|
| $(v, q, y^+) \succ (v, q, y^-)$ | 67.48 | 58.76 | **81.79** | 69.34 |
| + $(v, q, y^+) \succ (v', q, y^+)$ | 63.05 | 58.09 | 77.73 | 66.29 |
| + $(v, q, y^+) \succ (v', q, y^-)$ | 65.60 | 59.64 | 79.48 | 68.24 |
| + $(v', q, y^-) \succ (v', q, y^+)$ | **68.14** | **60.31** | 80.70 | **69.72** |

Table 6: **Exploring preference data construction strategy under different preference loss.** Every construction (rows) is trained with every loss (columns) under identical settings.

| | Slake | | |
|---|---|---|---|
| **Preference Data** | **DPO** | **MMedPO** | **FIRE-MPO** |
| GT style pairs | 62.0 | 63.0 | 63.8 |
| MMedPO style pairs | 63.4 | 65.8 | 66.5 |
| FIRE-MPO style pairs | **64.3** | **66.3** | **68.1** |

**Visual Grounding.** In Figure 5, we analyze grounding layer by layer for localization and attribute tokens. The unaligned backbones and response-level DPO show similar patterns: attention to the queried region decreases in the deeper layers, where the final prediction is formed. In contrast, FIRE-MPO better preserves attention to the relevant region across layers, with lower KL and JS divergence on both backbones. We also observe a similarity between DPO and the unaligned baseline, suggesting that because a response-level objective does not explicitly supervise the correspondence between clinically relevant tokens and their supporting image regions optimizing it alone doesn't improve visual grounding.

### 4.3 Ablation Study

We ablate the three core design choices of FIRE-MPO on `HuatuoGPT-Vision-7B`: the bidirectional token-wise KL regularizer (Table 3), the lesion-targeted image corruption (Table 4), and the formulation of the visual preference pair (Table 5). We then decouple the contributions of the preference-data and training objective by evaluating every preference-data construction strategy with every preference loss (Table 6).

**Bidirectional KL Regularization.** Table 3 indicates the contribution of the KL regularization terms. Removing either direction reduces overall performance. Forward KL has the largest contribution, particularly on IU-Xray (long-form report generation), where removing it reduces the score from $80.70$ to $58.42$. Reverse KL has a smaller but consistent effect across the three benchmarks. Removing the KL regularizer entirely also reduces performance, with the largest drop again observed on IU-Xray.

**Visual-contrastive Preference Pairs.** Table 4 shows that lesion-based corruption consistently outperforms global perturbations, particularly cropping and full-image masking: removing the evidence for a specific finding links that evidence to its claim, whereas global corruption discards visual information indiscriminately. Both alternative visual preference terms in Table 5 also underperform ours, so contrasting a clean image with its corrupted counterpart is not sufficient; the pair must also capture how removing the evidence changes the correctness of the clinical answer.

**Decoupling the Preference Framework from the Loss.** Table 6 disentangles the contribution of the preference data construction from that of the preference optimization objective. We train each of the three preference-pair constructions (rows: GT-style, MMedPO-style, and FIRE-MPO-style; examples in Figure 6) with each of the three preference losses (columns: DPO, MMedPO, and FIRE-MPO), while keeping all other settings fixed. This allows us to vary the data construction while holding the loss fixed (within each column), and vary the loss while holding the preference pairs fixed (within each row). As shown in Table 6, FIRE-MPO-style pairs achieve the best performance under all three losses, indicating that the proposed pair construction provides a stronger preference signal independent of the optimization objective. Similarly, for every pair construction, FIRE-MPO

achieves the highest performance among the three losses. The best result (68.1) is obtained when the proposed pair construction and loss are used together, showing that the two components provide complementary improvements. Appendix F provides the per-split results.

**Reward Localization on Clinical Tokens.** We next examine where each preference objective places its reward using the diagnostic introduced in Section 2.2. As shown in Table 1, FIRE-MPO assigns the highest reward to the medical tokens of correct reports in $59\%$ of cases, compared with $27\%$ for MMedPO and $14\%$ for DPO, while winning only $30.4\%$ of reports on non-medical tokens. Its reward therefore concentrates on clinical content rather than moving uniformly away from the reference model.

## 5 Related Work

**General Domain Alignments.** Several methods address the sparsity of a response-level margin. The closest to ours is RRPO (Sarkar & Etemad, 2025), which rewards only the sub-sequences that differ between the preferred and rejected responses and adds a forward token-wise KL for stability. FIRE-MPO departs from it in three respects. First, RRPO uses dataset references as preferred responses and template-matches the rejected ones to them, whereas both of our responses derive from minimal span edits of the model's own generations. Second, RRPO perturbs the image at random only to elicit rejected responses, whereas we corrupt the lesion the question concerns and reverse the preference on the corrupted image. Third, RRPO's regularizer is forward-only, whereas ours is bidirectional (Section 3). Token-level variants of DPO (Zeng et al., 2024; Zhu et al., 2025) likewise redistribute the reward across tokens, while Mask-DPO (Gu et al., 2025) and ASPO (Wang et al., 2025) operate at sentence level to mask noise and weight segments adaptively. OPA-DPO (Yang et al., 2025) instead attacks the distribution shift, sampling preferred responses from the model itself. All of these localize the textual signal alone, leaving the dependence of a clinical token on a specific image region unexpressed.

**Multimodal Direct Preference Learning.** CHiP (Fu et al., 2025) optimizes textual and visual representations jointly through a cross-modal hierarchical objective, and SymPO (Shukla et al., 2025) uses symmetric image-text pairs to push the model off its linguistic priors. Both rely on coarse image-level perturbations, which cannot represent the minute lesions and focal anomalies that decide a diagnosis.

**Preference Learning in Clinical Settings.** CheXalign (Hein et al., 2025a) scores radiology reports against references automatically, Rad-DPO (Puli et al., 2024) targets VQA hallucination through data construction, MMedPO (Zhu et al., 2024) weights each pair by clinical significance with one scalar per response, and R-DPO (Liu et al., 2025b) refines factuality recursively. Each adapts the data or the weighting while retaining the response-level DPO schema for reward assignment, so none can express which spans decide correctness or which visual evidence licenses them. That is the gap FIRE-MPO addresses.

## 6 Conclusion

We identified a mismatch between the granularity at which clinical correctness is defined and the granularity a response-level preference objective is able to express, and we showed that the mismatch is measurable rather than rhetorical: the implicit reward learned by a response-level objective need not settle on clinical content, and substituting supervised references for preferred responses turns a stylistic gap into a reward-hacking shortcut. FIRE-MPO closes that gap by matching the supervision to the judgement, with preference pairs that differ only on clinically decisive spans, lesion-corrupted counterparts that withhold the evidence those spans depend on, and a bidirectional token-wise regularizer that keeps so sparse a signal stable. Across VQA-RAD, SLAKE, and IU-Xray, FIRE-MPO outperforms state-of-the-art alignment methods, concentrates its reward on the tokens a clinician would check, and does so without the stylistic imitation that off-policy construction invites. The broader point is that the limiting factor in clinical alignment may be less the quality of the preference data than the resolution at which an objective is able to read it.

## ETHICS STATEMENT

The broader impact of this work lies in its potential to enhance the reliability and factual consistency of AI-driven medical diagnostics. Our framework advances the development of more trustworthy medical AI tools with direct implications for diagnostic accuracy and medical image faithfulness. Improved alignment between model outputs and clinically decisive visual evidence could reduce hallucinated findings in radiology reports and medical VQA, supporting clinicians in time-sensitive diagnostic workflows. However, ethical considerations remain paramount. The aligned model can still suffer from critical errors as of the base model. The responsible use requires safeguards against over-reliance on AI-generated medical advice in high-stakes decision-making contexts. Future societal benefits may include reduced diagnostic errors, improved radiological workflow efficiency, and broader access to AI-assisted clinical interpretation in resource-limited settings, but these gains must be pursued alongside rigorous prospective validation and ongoing ethical oversight to ensure that advances in fine-grained medical alignment genuinely serve the best interests of patients and healthcare providers.

## REPRODUCIBILITY STATEMENT

We have taken several steps to ensure that the results reported in this paper can be reproduced. The complete FIRE-MPO objective is specified in Equations 2–4, and a self-contained reference implementation of the loss is provided as pseudocode in Appendix A. All base models, training datasets, and evaluation benchmarks are publicly available and are described in Section 4.1, together with the licenses of every asset we use (Appendix C). The full set of training hyperparameters, the LoRA configuration, and the hardware used for training and inference are reported in Appendix C. The preference-data construction pipeline is described in Section 3, and the exact prompt templates used for VQA answer localization, report refinement, and fine-grained difference masking are reproduced verbatim in Appendix G.

## AI USE STATEMENT

Generative AI tools are a component of the method itself rather than only an authoring aid, and we disclose both uses here. As described in Section 4.1, we use `GPT-4o-mini` to localize and extract answers from open-ended VQA generations and to mask fine-grained differences between responses, `Gemini 3 Flash Preview` to refine radiology reports during preference-pair construction, and `GPT-4o-mini` again as the judge for Med-VQA accuracy at evaluation time; the corresponding prompts are given in Appendix G. Separately, we used a large language model for minor writing assistance (grammar, phrasing, and LaTeX formatting) on text that was drafted by the authors. Generative AI was not used to formulate research ideas or design the methodology; its role was limited to code implementation and minor refactoring. All AI-assisted output, including the constructed preference data and every claim in this paper, was reviewed by the authors, and we take full responsibility for the final content of this work.

## REFERENCES

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Michael A. Bruno, Eric A. Walker, and Hani H. Abujudeh. Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. *RadioGraphics*, 35(6):1668–1676, 2015. doi: 10.1148/rg.2015150023.

J. Harry Caufield, Yichao Zhou, Yunsheng Bai, David A. Liem, Anders O. Garlid, Karol E. Watson, Alex A. T. Bui, Wei Wang, and Peipei Ping. A comprehensive typing system for information extraction from clinical narratives. *medRxiv*, 2019. doi: 10.1101/19009118.

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

Jay Hyeon Cho, JunHyeok Oh, Myunsoo Kim, and Byung-Jun Lee. Rethinking dpo: The role of rejected responses in preference misalignment. *arXiv preprint arXiv:2506.12725*, 2025.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Elizabeth A. Eisenhauer, Patrick Therasse, Jan Bogaerts, Lawrence H. Schwartz, Daniel Sargent, Robert Ford, Janet Dancey, Susan Arbuck, Steve Gwyther, Margaret Mooney, Larry Rubinstein, Lalitha Shankar, Lori Dodd, Richard Kaplan, Denis Lacombe, and Jaap Verweij. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*, 45(2):228–247, 2009. doi: 10.1016/j.ejca.2008.10.026.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization, 2024. *URL https://arxiv. org/abs/2402.01306*, 14, 2023.

J. Fu et al. CHiP: Cross-modal hierarchical DPO. *arXiv preprint arXiv:2501.16629*, 2025.

Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. Mask-dpo: Generalizable fine-grained factuality alignment of llms. *arXiv preprint arXiv:2503.02846*, 2025.

D. Hein et al. CheXalign: Preference fine-tuning in chest X-ray interpretation models without human feedback. In *Proceedings of the Association for Computational Linguistics (ACL)*, 2025a.

Dennis Hein, Zhihong Chen, Sophie Ostmeier, Justin Xu, Maya Varma, Eduardo Pontes Reis, Arne Edward Michalson, Christian Bluethgen, Hyun Joo Shin, Curtis Langlotz, and Akshay S Chaudhari. CheXalign: Preference fine-tuning in chest X-ray interpretation models without human feedback. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27679–27702, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1342. URL https://aclanthology.org/2025.acl-long.1342/.

Angela S. Hong, Eric L. Rosen, Mary S. Soo, and Jay A. Baker. BI-RADS for sonography: Positive and negative predictive values of sonographic features. *American Journal of Roentgenology*, 184 (4):1260–1265, 2005. doi: 10.2214/ajr.184.4.01841260.

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

Mahyar Khayatkhoei, Prateek Chhikara, Filip Ilievski, et al. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 2018.

Anglin Liu, Rundong Xue, Xu R Cao, Yifan Shen, Yi Lu, Xiang Li, Qianqian Chen, and Jintai Chen. Medsam3: Delving into segment anything with medical concepts. *arXiv preprint arXiv:2511.19046*, 2025a.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 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.

Y. Liu et al. R-DPO: Recursive direct preference optimization for medical vision-language models. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2025b.

Y. Liu et al. Analyzing visual grounding in medical LVLMs. *arXiv preprint arXiv:2603.14323*, 2026.

Marianne T. Luetmer, Christopher H. Hunt, Robert J. McDonald, Brian J. Bartholmai, and David F. Kallmes. Laterality errors in radiology reports generated with and without voice recognition software: Frequency and clinical significance. *Journal of the American College of Radiology*, 10 (7):538–543, 2013. doi: 10.1016/j.jacr.2013.02.017.

Heber MacMahon, David P. Naidich, Jin Mo Goo, Kyung Soo Lee, Ann N. C. Leung, John R. Mayo, Atul C. Mehta, Yoshiharu Ohno, Charles A. Powell, Mathias Prokop, Geoffrey D. Rubin, Cornelia M. Schaefer-Prokop, William D. Travis, Paul E. Van Schil, and Alexander A. Bankier. Guidelines for management of incidental pulmonary nodules detected on CT images: From the Fleischner Society 2017. *Radiology*, 284(1):228–243, 2017. doi: 10.1148/radiol.2017161659.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 374–390, 2024.

A. Puli et al. Rad-DPO: Reducing hallucinations in medical visual question answering. *arXiv preprint arXiv:2406.06496*, 2024.

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

Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. From $r$ to $q^*$: Your language model is secretly a q-function. In *First Conference on Language Modeling*, 2024.

Minal J. Sangwaiya, Sanjay Saini, Michael A. Blake, Keith J. Dreyer, and Mannudeep K. Kalra. Errare humanum est: Frequency of laterality errors in radiology reports. *American Journal of Roentgenology*, 192(5):W239–W244, 2009. doi: 10.2214/AJR.08.1778.

Pritam Sarkar and Ali Etemad. Self-alignment of large video language models with refined regularized preference optimization. *arXiv preprint arXiv:2504.12083*, 2025.

Thomas Savage, Stephen P Ma, Abdessalem Boukil, Ekanath Rangan, Vishwesh Patel, Ivan Lopez, and Jonathan Chen. Fine-tuning methods for large language models in clinical medicine by supervised fine-tuning and direct preference optimization: Comparative evaluation. *Journal of Medical Internet Research*, 27:e76048, 2025.

S. Shukla et al. SymPO: Symmetric preference optimization for vision-language models. *arXiv preprint arXiv:2506.11712*, 2025.

Guohao Sun, Can Qin, Huazhu Fu, Linwei Wang, and Zhiqiang Tao. Self-training large language and vision assistant for medical question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 20052–20060, 2024.

Stephen Waite, Jinel Scott, Brian Gale, Travis Fuchs, Srinivas Kolla, and Deborah Reede. Interpretive error in radiology. *American Journal of Roentgenology*, 208(4):739–749, 2017. doi: 10.2214/AJR.16.16963.

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

H. Wang et al. ASPO: Adaptive sentence-level preference optimization for multimodal large language models. In *Findings of the Association for Computational Linguistics (ACL)*, 2025.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024.

Yuzi Yan, Yibo Miao, Jialian Li, Yipin Zhang, Jian Xie, Zhijie Deng, and Dong Yan. 3d-properties: Identifying challenges in dpo and charting a path forward. *arXiv preprint arXiv:2406.07327*, 2024.

X. Yang et al. OPA-DPO: On-policy alignment for preference optimization. *arXiv preprint arXiv:2501.09695*, 2025.

Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024.

K. Zhu et al. MMedPO: Aligning medical vision-language models with clinical-aware multimodal preference optimization. *arXiv preprint arXiv:2412.06141*, 2024.

Mingkang Zhu, Xi Chen, Zhongdao Wang, Bei Yu, Hengshuang Zhao, and Jiaya Jia. Tgdpo: Harnessing token-level reward guidance for enhancing direct preference optimization. *arXiv preprint arXiv:2506.14574*, 2025.

## A FiRe-MPO Pseudocode

```
def FiRe-MPO(
    self,
    # phrase-level logprobs
    chosen_phrase_logps, rejected_phrase_logps,
    ref_chosen_phrase_logps, ref_rejected_phrase_logps,

    # v visual pair at phrase level: (m_l, y_l) vs (m_l, y_w)
    v_chosen_phrase_logps, v_rejected_phrase_logps,
    ref_v_chosen_phrase_logps, ref_v_rejected_phrase_logps,

    # token logits for phrase-level KL
    chosen_logits, ref_chosen_logits, phrase_mask, ref_phrase_mask,

    has_rejected_image,
    beta, alpha, gamma, lambda,  # hyperparams
):

    # -------- 1) Base phrase preference --------
    # margins computed from phrase-only logprobs
    base_margin = beta * (
        (chosen_phrase_logps - rejected_phrase_logps)
        - (ref_chosen_phrase_logps - ref_rejected_phrase_logps)
    )
    base_loss = -F.logsigmoid(base_margin)  # [B]

    # -------- 2) Visual preference (v only, phrase-level) --------
    v_margin = beta * (
        (v_chosen_phrase_logps - v_rejected_phrase_logps)
        - (ref_v_chosen_phrase_logps - ref_v_rejected_phrase_logps)
    )
    v_loss = -F.logsigmoid(v_margin)      # [B]

    loss = base_loss + gamma * v_loss

    # -------- 3) token-wise KL --------
    logp_pi  = F.log_softmax(chosen_logits, dim=-1)       # [B,S,V]
    logp_ref = F.log_softmax(ref_chosen_logits, dim=-1)   # [B,S,V]

    # forward KL(ref || pi), then keep phrase positions only
    tkl_tok = (logp_ref.exp() * (logp_ref - logp_pi)).sum(dim=-1) # [B,S]
    tkl = (tkl_tok * phrase_mask.float()).sum(dim=-1)             # [B]

    # reverse KL(pi || ref), also phrase-only
    rtkl_tok = (logp_pi.exp() * (logp_pi - logp_ref)).sum(dim=-1) # [B,S]
    rtkl = (rtkl_tok * ref_phrase_mask.float()).sum(dim=-1)       # [B]

    mixed_tkl = lambda * tkl + (1.0 - lambda) * rtkl
    loss = loss + alpha * mixed_tkl

    return loss.mean()
```

## B Limitations

While FiRe-MPO demonstrates consistent gains across medical VQA and report generation benchmarks, several limitations remain. First, the on-policy preference construction pipeline depends on external models, GPT-4o-mini for VQA extraction and MedSAM3-v1 for lesion segmentation, introducing potential error propagation when the task is more niche. Second, the lesion-targeted corruption strategy requires reliable segmentation masks, which may be unavailable or imprecise for rare pathologies or underrepresented imaging modalities. Third, our evaluation is confined to radiology, and it remains an open question whether the framework generalizes to other clinical

Table 7: Details of training hyperparameters.

| | **HuatuoGPT-Vision-7B** | **Qwen3-VL-4B** |
|---|---|---|
| Is Medically SFT | ✓ | × |
| Trainable module | LoRA in LLM and everything else is kept frozen | |
| LoRA setup | rank=128, alpha=256, dropout=0.05 | |
| Learning rate | 1e-6 | 1e-6 |
| Learning rate scheduler | Cosine | Cosine |
| Optimizer | AdamW | AdamW |
| Weight decay | 0.0 | 0.0 |
| Warmup ratio | 0.0 | 0.0 |
| Epoch | 1 | 1 |
| Batch size per GPU | 4 | 4 |
| Batch size (total) | 8 | 8 |
| $\beta$ (loss coefficient) | 0.1 | 0.1 |
| $\alpha$ (loss coefficient) | 0.01 | 0.01 |
| $\gamma$ (loss coefficient) | 0.1 | 0.1 |
| $\lambda$ (loss coefficient) | 0.5 | 0.5 |
| Memory optimization | unsloth | unsloth |

imaging domains such as pathology, dermatology, or ophthalmology, which exhibit substantially different visual grounding characteristics.

# C Implementation Details.

Training was performed on an H100 80GB GPU, while inference utilized two L40s.

## C.1 Lesion-Targeted Image Corruption

The corrupted image $v'$ used in the second term of Eq. 2 is built in three steps, so that it differs from $v$ only in the evidence the question depends on.

**Segmentation prompt.** MedSAM3 is a text-prompted segmentation model, so each question is first converted into a short description of the region it concerns. `GPT-4o-mini` receives the question and its ground-truth answer and returns a noun phrase naming that region, e.g., "brain infarction" or "chest tube", using the template in Figure 10. Some questions concern the image as a whole rather than a specific region, e.g., the imaging modality or the patient's sex; for these holistic questions the model returns `none`.

**Localization.** The phrase is passed to MedSAM3-v1 (Liu et al., 2025a) at an input resolution of $1008$, with a detection threshold of $0.5$ and an NMS IoU threshold of $0.5$. All predicted instance masks are merged into a single region.

**Corruption.** Pixel-wise Gaussian noise with $\sigma = 45$ (on the 0–255 intensity scale) is added inside the region, and the result is clipped to the valid range; pixels outside the region are left unchanged. For holistic questions, and for the cases in which MedSAM3 detects no region, the relevant evidence is not confined to one area, so weaker noise with $\sigma = 12$ is added to the whole image instead. The noise is drawn with a seed derived from the sample identifier, so each $v'$ is deterministic. Corrupted images are generated once per question and shared by both backbones.

**Licenses of existing assets used.**

- HuatuoGPT-Vision-7B (Apache License 2.0): `https://huggingface.co/FreedomIntelligence/HuatuoGPT-Vision-7B`
- Qwen3-VL-4B-Instruct (Apache License 2.0): `https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct`
- d4data/biomedical-ner-all (Apache License 2.0), used to tag clinical tokens for the reward-localization analysis in Table 1: `https://huggingface.co/d4data/biomedical-ner-all`

## D EXAMPLES OF CORRUPTED IMAGES

Figure 7 shows training samples from VQA-RAD and SLAKE next to the corrupted images $v'$ produced by the procedure in Appendix C.1, together with the question, the MedSAM3 segmentation prompt, and the preference pair. For the first four questions, the prompt names a localized finding, and the noise covers only the region MedSAM3 segments for it: the ring-enhancing lesion, the intraventricular mass, the hepatic lesion, and the infarcted areas. The rest of the image is left unchanged, so $v'$ withholds exactly the evidence that separates $y^+$ from $y^-$. The last question, about the imaging modality, is holistic; its prompt is `none`, and weaker noise is added to the whole image instead. On $v$ the loss prefers $y^+$, and on $v'$ the preference is reversed (Eq. 2).

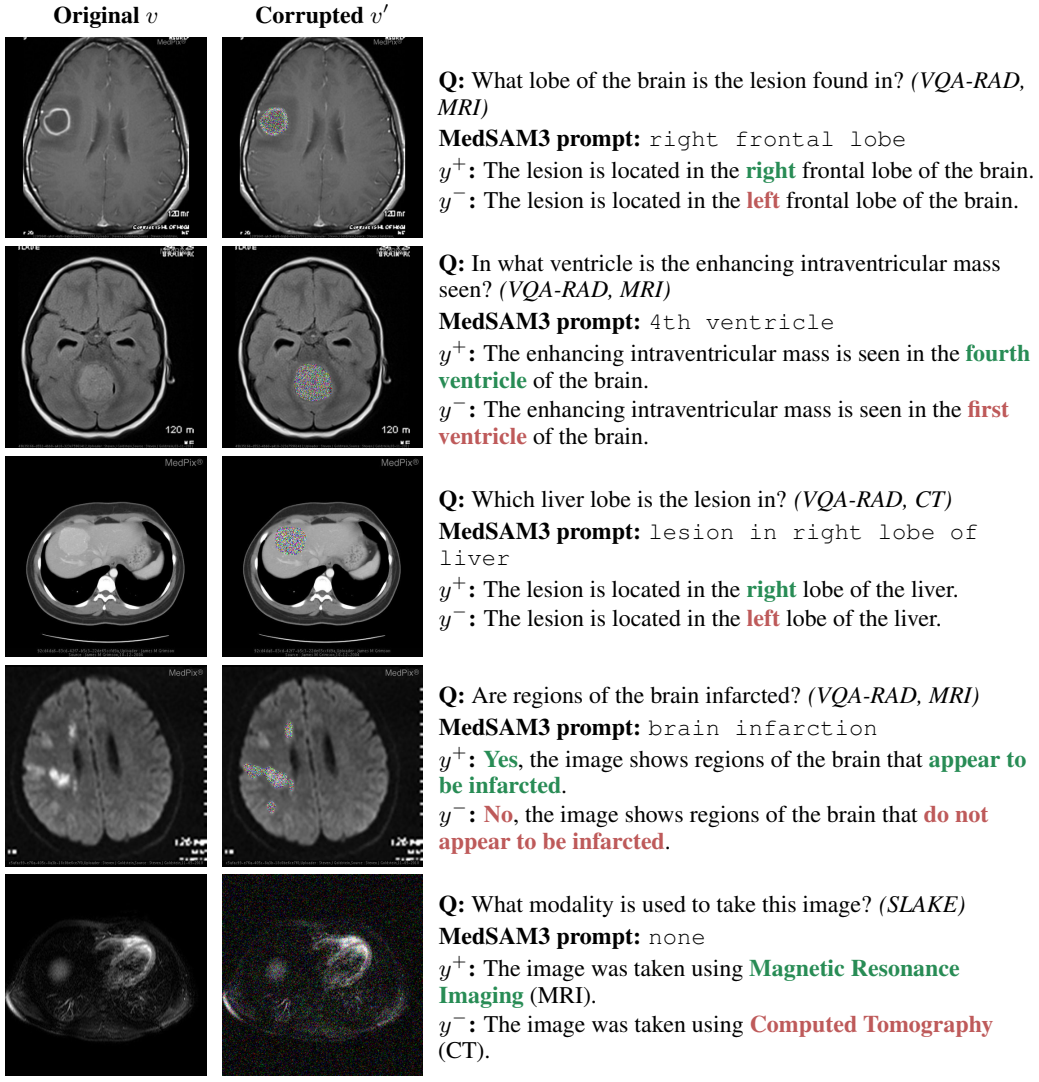



Figure 7: Original and corrupted images for five training questions. The decisive spans of the preferred response $y^+$ and the rejected response $y^-$ are shown in green and red. For the four localized questions, Gaussian noise ($\sigma = 45$) is added inside the region MedSAM3 segments for the prompt; for the holistic modality question, weaker noise ($\sigma = 12$) is added to the whole image.

## E REWARD LOCALIZATION ANALYSIS

This section details the analysis summarized in Table 1 (Section 2.2).

Table 8: Comparison of preference-pair construction strategies on medical VQA benchmarks. Cl, Op, and Ov denote closed-ended, open-ended, and overall performance, respectively.

| Models | SLAKE | | | VQA-RAD | | |
|---|---|---|---|---|---|---|
| | Cl | Op | Ov | Cl | Op | Ov |
| Style Agnostic | 73.63 | 47.77 | 56.64 | 70.88 | 38.94 | 57.42 |
| GT Style | **80.76** | 52.36 | 62.11 | **73.56** | **44.21** | **61.20** |
| LVLM Style | 79.40 | **55.38** | **63.62** | 71.65 | 39.47 | 58.09 |

**Which comparisons are valid.** The DPO loss (Eq. 1) depends on a response only through the sum of its per-token log-ratios, $\sum_t \log \frac{\pi_\theta(y_t|v,q,y_{<t})}{\pi_{\text{ref}}(y_t|v,q,y_{<t})}$, so a positive margin does not reveal which tokens produce it. Individual terms cannot be compared across positions: in the token-level view of DPO (Rafailov et al., 2024), each term equals a per-token reward plus a difference of value functions,

$$\beta \log \frac{\pi_\theta(y_t \mid s_t)}{\pi_{\text{ref}}(y_t \mid s_t)} = r(s_t, y_t) + V(s_{t+1}) - V(s_t), \qquad s_t = (v, q, y_{<t}), \tag{5}$$

and the value terms change with the position in the sequence. Across models, however, the same token $y_t$ under the same prefix $s_t$ is directly comparable: both conditionals are normalized over the vocabulary at that step, and all aligned models share the same $\pi_{\text{ref}}$. We therefore never compare tokens with one another, only models with one another on a fixed token. We report $\log(\pi_\theta/\pi_{\text{ref}})$ without $\beta$, so the comparison does not depend on each method's choice of $\beta$.

**Protocol.**

- **Text.** Every model scores the same fixed text by teacher forcing: each ground-truth IU-Xray report ($y^+$) and a medically incorrect counterpart ($y^-$) taken from the rejected responses of our preference pairs (Section 3). No model scores its own generations and no external judge is involved.
- **Medical tokens.** Each report is tagged with `d4data/biomedical-ner-all`, a DistilBERT model fine-tuned for biomedical named-entity recognition on MACCROBAT (Caufield et al., 2019). Tokens inside a clinical entity, e.g., findings, anatomical locations, and measurements, are labelled medical; all remaining tokens are non-medical.
- **Score.** For each report and each model, we sum the log-ratios over the medical tokens and, separately, over the non-medical tokens. Since all models are scored on the same token set, summing and averaging select the same winner.

**Metrics.** For each token set, the win rate of a method is the percentage of reports on which it is the winner among the three methods. On $y^+$ the winner is the model that assigns the highest reward; on $y^-$, whose medical tokens are erroneous, it is the model that assigns the lowest. With three methods, chance is $33.3\%$ and each column of Table 1 sums to $100\%$. A method that concentrates its reward on clinical content should win above chance on medical tokens but not on non-medical ones; a method that simply drifts further from $\pi_{\text{ref}}$ would win on both.

## F Textual Preference-Pair Style Ablation

Table 6 decouples the preference-data framework from the preference loss using a single aggregate score per cell. Here we report the finer-grained comparison of the textual pair-construction strategies alone, with closed-ended and open-ended performance separated, on both medical VQA benchmarks. All rows use the same loss and differ only in how the preferred and rejected responses are written; the three settings are illustrated in Figure 6.

The style-agnostic setting, which does not explicitly control for stylistic differences between preferred and rejected responses, performs worst on both datasets, at $56.64\%$ overall on SLAKE and $57.42\%$ on VQA-RAD. Both GT Style and LVLM Style improve substantially over it, indicating that preserving a coherent response structure during preference construction is beneficial regardless of which style is chosen. Between the two, LVLM Style is stronger on SLAKE, driven by a large gain on the open-ended split ($55.38\%$ vs. $52.36\%$), where responses are long enough for stylistic shortcuts to matter; GT Style retains an advantage on VQA-RAD, whose answers are short enough that the stylistic gap between reference and on-policy responses is small.

```
{Stage 1: Prompt used in VQA answer localization.}
You are evaluating a model's answer to a medical question.


Question: {question}
Correct answer (ground truth): {answer}
Model's full answer: {output}

Perform these four steps and respond with a single JSON object only
(no other text):

1. **answer_sentence**: The minimal phrase or clause from the
model's answer that states the final answer, using the model's
exact wording. Preserve the model's language so the final answer
is visible in context. Do not include extra clauses that follow.
2. **final_answer**: Locate and extract the final answer to the
question from the model's answer. Use as few words as possible (the
first occurrence of the conclusive answer).
3. **is_correct**: true if the final answer is correct given the
ground truth, false otherwise.
4. **opposite_answer**: Modify the answer sentence as little as
possible to make a medically plausible alternative for the question.

Respond only with valid JSON in this exact shape:
{"answer_sentence": "...", "final_answer": "...", "is_correct":
true or false, "opposite_answer": "..."}


{Stage 2: Prompt used in VQA answer difference localization.}
You are given two sentences that are the same except for some phrases.
Your task is to identify the phrases in each sentence that differs
from the other.

Sentence A: {sentence_a}
Sentence B: {sentence_b}

Wrap only the differing phrases in `<mask` `</mask` tags. The
differing phrases should be as short as possible.
```

Figure 8: Prompt template used for VQA task answer localization and Ground Truth incorporation.

## G PROMPT TEMPLATES

The instructions used for processing the open-ended generated responses are detailed below. We utilize a multi-stage approach to first evaluate correctness (Stage 1) and then identify and incorporate fine-grained clinical differences (Stage 2). Figure 10 gives the prompt used to obtain the segmentation target for image corruption (Appendix C.1).

```
% {Stage 1:  Prompt used in Report generation modification}

**Task:** Rewrite **Report A** to make a medically difficult
preference pair.

**Instruction:**
* **Change Logic**:  Use the facts in **Report B** to replace a few
medical words in Report A to create a fine-grained preferred pair.
* **all_correct**:  If all the fact in Report B are already present
in Report A, change a few words yourself to make a rejected pair.

**Constraints:**
* **Structural Locking:** The format of Report A should be preserved.
* **Token Parsimony:** At most **two or three words per sentence**
can be changed.
* **Clinical Significance:** The rejected response should be
clinically different with Report A. Changing synonyms (e.g.  "clear"
vs "normal", "healthy" vs "unremarkable") is not accepted.
* **Output Format:** Return a JSON object containing keys
`all_correct` and `rewritten_report_a`.

**Input Data:**
* **Report A:** {report_a}
* **Report B:** {report_b}

% {Stage 2:  Prompt used in Report Generation answer difference
localization}


CHEST_MASK_PROMPT = """Compare two reports and put the differing
phrases inside `<mask` `</mask` tags.

Output:
Return **only** a valid JSON object with two keys:  `report_a_masked`
and `report_b_masked`.  Each key should be a string with the
differing phrases wrapped in `<mask` `</mask` tags.

Report A:
{report_a}

Report B:
{rewritten_report_a}
"""
```

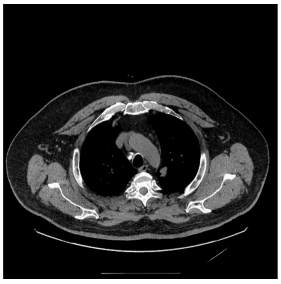
Figure 9: Prompt template used for Report Generation task answer localization and Ground Truth incorporation.

```
{Prompt used to generate the MedSAM3 segmentation prompt.}
Your task is to generate the prompt to a medical segmentation model.
The model expects short text prompts to ground on the image.

Question: {question}
Answer: {answer}

Considering the question and the answer, respond with exactly one
line: only the text that should be passed to the segmentation model
as its prompt.
If there is no reasonable target to segment for this Q&A (e.g. the
answer is only a modality like "MRI" or "CT" or it depends on the
whole image rather than a specific part), respond with exactly: none

Examples:
Question: What is the modality of the image? Answer: MRI ->
none
Question: Does the picture contain liver? Answer: Yes ->
"liver"
Question: Where is/are the abnormality located? Answer: Left
lung -> "left lung"
Question: are regions of the brain infarcted? Answer: yes ->
"brain infarction"
Question: is there a pneumothorax? Answer: no -> "lungs"
Question: is there a chest tube placed? Answer: yes -> "chest
tube"
Question: what is the gender of this patient? Answer: man ->
none
Question: in what vascular territory is the lesion located?
Answer: right mca -> "lesion in right mca"
...
```

Figure 10: Prompt template used to generate the text prompt for MedSAM3 segmentation. SLAKE and VQA-RAD use the same instruction with dataset-specific in-context examples (12 and 18, respectively); a representative subset is shown.

## H MORE CASES

Figures 11–17 compare the base model, DPO, and FIRE-MPO on the same image and question. ✓ and ✗ mark agreement with the ground-truth answer.



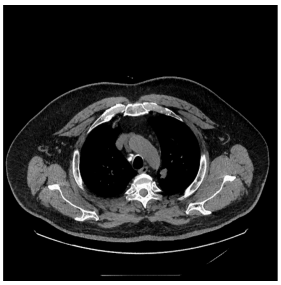
**Question:** What is the main organ in this image?
**Ground truth:** Lung

**HuatuoGPT-Vision-7B** ✓
The main organ in the image is the lungs, which occupy most of the space within the thoracic cavity.

**+ DPO** ✓
Lung

**+ FIRE-MPO** ✓
The main organ in the image is the lungs, which occupy most of the space within the thoracic cavity.

Figure 11: HuatuoGPT-Vision-7B, chest CT. All three models answer correctly; DPO collapses to a one-word answer, while FIRE-MPO keeps the base model's full explanation.



**Question:** What is the main organ in this image?
**Ground truth:** Lung

**Qwen3-VL-4B-Instruct** ✗
The main organ in the image is the **heart**.
In this axial CT scan of the thorax, the heart is centrally located within the mediastinum, visible as a dense, slightly oval structure with a central cavity (the cardiac silhouette). It is surrounded by the lungs (which appear dark/black due to air content) and is positioned anterior to the vertebral column. The heart is the primary organ in this cross-section, and its position and density are consistent with normal anatomical presentation.
Other visible structures include:
- Lungs (air-filled, dark areas)
- Spine (vertebral body at the center-bottom)
- Rib cage (white bony structures)
- Major vessels (e.g., aorta, pulmonary arteries) — though less prominent than the heart

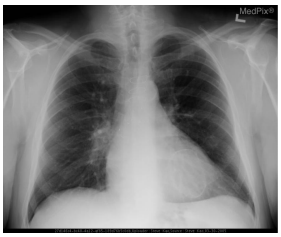
Thus, the heart is the main organ depicted.

**+ DPO** ✓
The main organ in the image is the Lung.

**+ FIRE-MPO** ✓
The main organs visible in the image are:
- **Lungs** (the large dark areas on the left and right sides of the image)
- **Thoracic spine** (the bony structure in the center-bottom, showing the vertebral body)
- **Rib cage** (the bony structures curving around the top and sides)
- **Heart** (visible as a small structure in the center, between the lungs)
- **Esophagus** (a small tubular structure, visible as a dark line just to the left of the trachea)

The **lungs** are the largest and most prominent organs.

Figure 12: Qwen3-VL-4B-Instruct, same image and question as Figure 11. The base model names the heart; both DPO and FIRE-MPO correct this, and FIRE-MPO retains the structured, descriptive answer style of the base model.



**Question:** Is there lung consolidation?
**Ground truth:** No

**HuatuoGPT-Vision-7B** ✗
Yes, there is evidence of bilateral lung consolidation visible on this chest X-ray. The consolidation appears as areas of increased opacity within both lung fields, indicating an abnormal accumulation of fluid or inflammatory material within the lung parenchyma. This finding is consistent with various pulmonary pathologies, such as pneumonia, atelectasis, or other conditions that can lead to alveolar filling and impaired gas exchange. Further clinical correlation and additional imaging studies may be necessary to determine the underlying cause of the observed lung consolidation.

**+ DPO** ✓
No

**+ FIRE-MPO** ✓
No, there is no evidence of consolidation in the lungs on this chest X-ray. The lung fields appear clear and well-aerated without any areas of increased opacity that would suggest consolidation.

Figure 13: HuatuoGPT-Vision-7B, chest X-ray. The base model hallucinates bilateral consolidation; FIRE-MPO gives the correct answer and grounds it in the appearance of the lung fields.



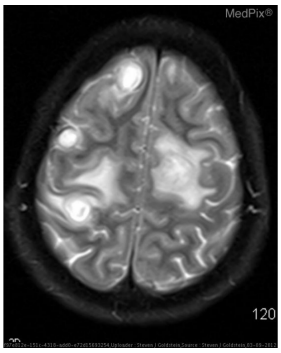
**Question:** Are the costophrenic angles preserved?
**Ground truth:** Yes

**HuatuoGPT-Vision-7B** ✓
Yes, the costophrenic angles are preserved in this chest X-ray image. The costophrenic angles are the areas where the diaphragm meets the ribs and lungs, and their preservation indicates that there is no pleural effusion or other abnormal fluid accumulation in the pleural spaces.

**+ DPO** ✗
No

**+ FIRE-MPO** ✓
Yes, the costophrenic angles appear to be preserved in this chest X-ray image. The costophrenic angles are the areas where the diaphragm meets the ribs and lungs, and their preservation is an important finding that can help rule out conditions like pleural effusions or other abnormalities in these regions.

Figure 14: HuatuoGPT-Vision-7B, chest X-ray. DPO flips a correct base answer to an incorrect one; FIRE-MPO preserves it.



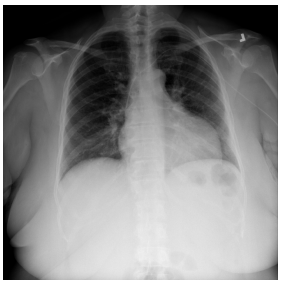

**Question:** Are the lesions ring-enhancing?
**Ground truth:** Yes

**HuatuoGPT-Vision-7B** ✓
Yes, the lesions in the image show ring enhancement, which is indicative of an active inflammatory or infectious process.

**+ DPO** ✗
No

**+ FIRE-MPO** ✓
Yes, the lesions in the image appear to be ring-enhancing.

Figure 15: HuatuoGPT-Vision-7B, brain MRI. As in Figure 14, DPO overturns a correct base answer, while FIRE-MPO keeps it.



**Question:** Which organs appear in pairs?
**Ground truth:** Lungs

**HuatuoGPT-Vision-7B** ✗
In the chest X-ray image, the lungs and kidneys appear in pairs.

**+ DPO** ✓
Lungs

**+ FIRE-MPO** ✓
In the chest X-ray image, the lungs appear in pairs.

Figure 16: HuatuoGPT-Vision-7B, chest X-ray. FIRE-MPO removes only the clinically incorrect span ("and kidneys") from the base response and leaves the rest unchanged.



**Question:** Where is the liver in this image?
**Ground truth:** Left

**HuatuoGPT-Vision-7B** ✗
In the provided CT scan image, the liver can be identified on the right side of the abdomen, just below the diaphragm.

**+ DPO** ✓
Left

**+ FIRE-MPO** ✓
In the CT scan image, the liver is located on the left side of the image.

Figure 17: HuatuoGPT-Vision-7B, abdominal CT. The base model answers in patient-relative terms ("right side of the abdomen"), whereas the question and ground truth refer to the image; FIRE-MPO answers in image coordinates.