MANY VIEWS, FEW INSIGHTS: RETHINKING THE ALIGNMENT OF VISION LANGUAGE MODEL FOR CLINICAL MULTI-IMAGES ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical diagnosis routinely requires synthesizing evidence across multiple images of the same patient, yet current vision-language models (VLMs) fall far short, largely because they fail to integrate evidence across images. The standard alignment recipe to align VLM on clinical multi-images is supervised fine-tuning (SFT) optionally followed by reinforcement learning (RL). This baseline recipe carries several under-studied free choices, including whether to add an RL stage, the SFT epoch count, and the LoRA rank. We systematically compare these choices, evaluating each configuration on in-domain and out-of-domain clinical benchmarks, and distill several **insightful findings** showing that existing recipes leave substantial performance on the table for assessing multiple images. We trace this gap to a specific training-time failure: SFT-only models show little hallucination on in-domain benchmarks but hallucinate substantially more on out-of-domain benchmarks, pointing to non-visual shortcuts that break under distribution shift. We call this supervision-modality mismatch. Building on these findings, we propose **MULTI-MULTI** for clinical **multi**-modal and **multi**-image analysis, which decomposes multi-image reasoning into an explicit, per-image-grounded structure and introduces Image-Swap Supervision (ISS) which pairs a case's images with a mismatched clinical history and mixed into fine-tuning samples so each reasoning step is recoverable only from the images. Despite this lightweight recipe, the resulting fine-tuned 8B open-source model reaches in-domain accuracy comparable to zero-shot open models two orders of magnitude larger, and closes part of the gap to leading proprietary models.

1 INTRODUCTION

Clinical diagnosis routinely requires synthesizing evidence across multiple images of the same patient: radiologists compare a current chest X-ray against a prior study, read paired frontal and lateral views together, or integrate findings across the standard views of a mammography exam (Geras et al., 2017; Rubin et al., 2018). Therefore, any vision-language model (VLM) intended for clinical use should be able to reason jointly across such image sets, not just interpret each image in isolation. Yet current VLMs are far from this bar. On MedThinkVQA, a benchmark of clinical cases averaging 6.6 images each, even the strongest frontier models reach only 54.9–57.2% accuracy, and the majority of the errors trace to failures in reading and integrating evidence across images rather than in medical reasoning itself (Yao et al., 2026). This mirrors a broader pattern in general-domain VLMs, where benchmarks such as MuirBench and MMIU similarly show sharp degradation once a task requires comparing or combining information across several images rather than processing one (Wang et al., 2025; Meng et al., 2025). In other words, today's VLMs may see many views but extract few insights.

The standard recipe for adapting a general-domain VLM to a specialized domain is post-training *alignment*: supervised fine-tuning (SFT) on domain-specific instruction data, often followed by reinforcement learning (RL) that optimizes the model against a reward signal rather than only imitating labeled examples (Ouyang et al., 2022). Recent reasoning-focused work replaces learned reward models with verifiable, rule-based rewards optimized via Group Relative Policy Optimization (GRPO) (Shao et al., 2024; DeepSeek-AI, 2025). This SFT-then-RL recipe has already been

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.

applied to single-image medical visual reasoning by Med-R1 and MedVLM-R1 (Lai et al., 2025; Pan et al., 2025). This recipe carries several free choices whose effects are rarely studied together including whether an RL stage is added at all, how many SFT epochs precede it, and what LoRA rank is allocated to the update (Hu et al., 2022). Critically, none of this prior alignment work addresses the multi-image case, each treats a clinical case as a single image. Whether SFT-only or SFT+RL alignment works better, and what number of epochs or LoRA rank is optimal, etc. has not been tested when a case includes multiple images.

In this work, we systematically revisit this question for the multi-image clinical setting. We compare SFT-only alignment against SFT followed by GRPO-based RL, and, within each, vary the number of SFT epochs and the LoRA rank used for fine-tuning, holding the base VLM, data, and multi-image input format fixed across conditions. We evaluate every configuration under a consistent protocol on both in-domain and out-of-domain clinical benchmarks. For in-domain evaluation we use MedThinkVQA (Yao et al., 2026), whose cases average 6.6 images each and directly test cross-view diagnostic reasoning; for out-of-domain generalization we use MedXpertQA (Zuo et al., 2025), an expert-level, multimodal medical reasoning benchmark spanning 17 specialties that was constructed independently of our training data and setup. This combination lets us separate genuine gains in multi-image alignment from benchmark-specific overfitting. Across this comparison, we distill **six findings** showing that existing alignment recipes leave substantial, previously unexplained performance on the table, which motivates a closer look at *why* they fail before proposing a fix.

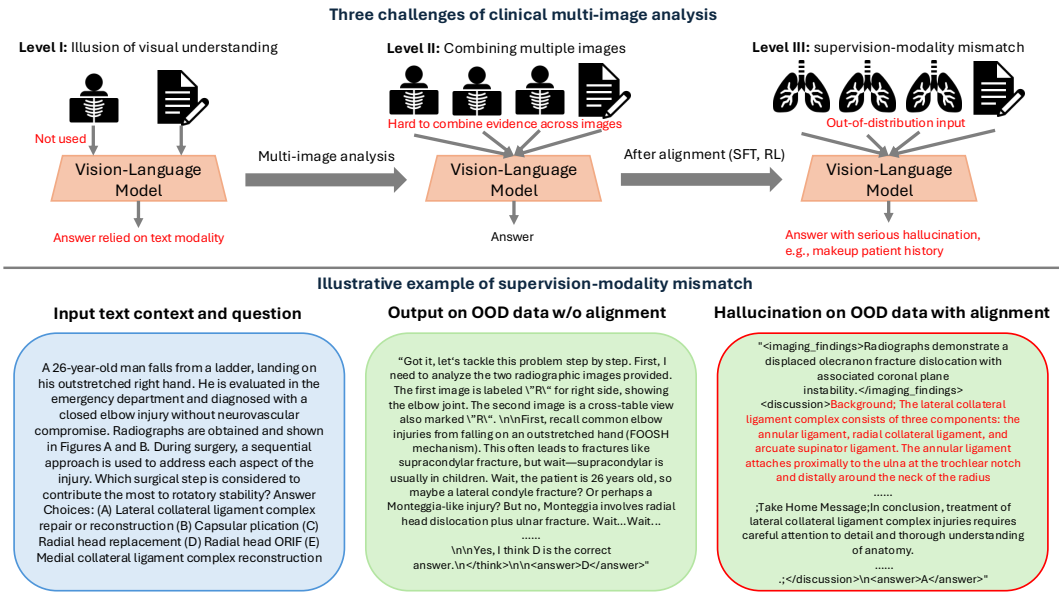


Figure 1: Three challenges of clinical multi-image analysis (*top*), illustrated by a concrete supervision-modality mismatch (*bottom*). *Top:* Level I, the model ignores the image and answers from text alone; Level II, combining evidence across multiple images is hard; Level III, after alignment (SFT, RL) the model responds to out-of-distribution input with confident hallucination rather than uncertainty. *Bottom:* on the same out-of-distribution case, the unaligned model's raw output reasons through the given images toward an answer, while the aligned model instead returns a well-formatted report that fabricates imaging findings never given in the input (highlighted).

As shown in Figure 1, recent work has established two distinct failure modes in how VLMs handle visual evidence in medical reasoning: models generate confident, fluent diagnostic claims even when images are entirely absent Asadi et al. (2026), and models fail to reliably ground, align, and compose evidence across multiple views even when the relevant images are present (Yao et al., 2026). We identify a third, more elusive training-time failure: standard caption-supervised SFT can teach a model to under-rely on genuinely present images without the images ever being absent. We see this directly in our own results: the SFT-only models perform well with little to no hallucination in-domain (MedThinkVQA), yet hallucinate substantially more out-of-domain (MedXpertQA). If the model were simply unable to use images, hallucination should be similarly frequent in-domain.

Instead, the gap suggests the model succeeded in-domain by leaning on non-visual regularities of the training distribution, such as case phrasing and label priors, that exist in domain. However, this breaks down into fluent but ungrounded reasoning once those regularities no longer transfer in OOD. We call this *supervision-modality mismatch*: the model learns to produce reasoning that is correctly formatted around the visual evidence but not actually derived from it.

To address this, we propose **MULTI-MULTI**, a two component method for clinical **multi**-modal and **multi**-image analysis. First, we decompose multi-image medical reasoning into an explicit, per-image-grounded structure (per-image findings, cross-image composition, and answer justification) that makes visual reliance inspectable at each step rather than just the final answer. Second, we show this structure alone does not guarantee genuinely image-derived content, and introduce Image-Swap Supervision (ISS): a counterfactual data-construction procedure that pairs a case's images with a mismatched clinical history from a different sample, guiding each reasoning step to be recoverable only from the images. Together, these two components turn the decomposed structure from a descriptive scaffold into one that is both verifiable and enforceable. This is achieved through SFT-stage data construction alone, without reinforcement learning or preference-based post-training. Empirically, the resulting 8B open-source model approaches the performance of 397B-scale models and leading closed-source models.

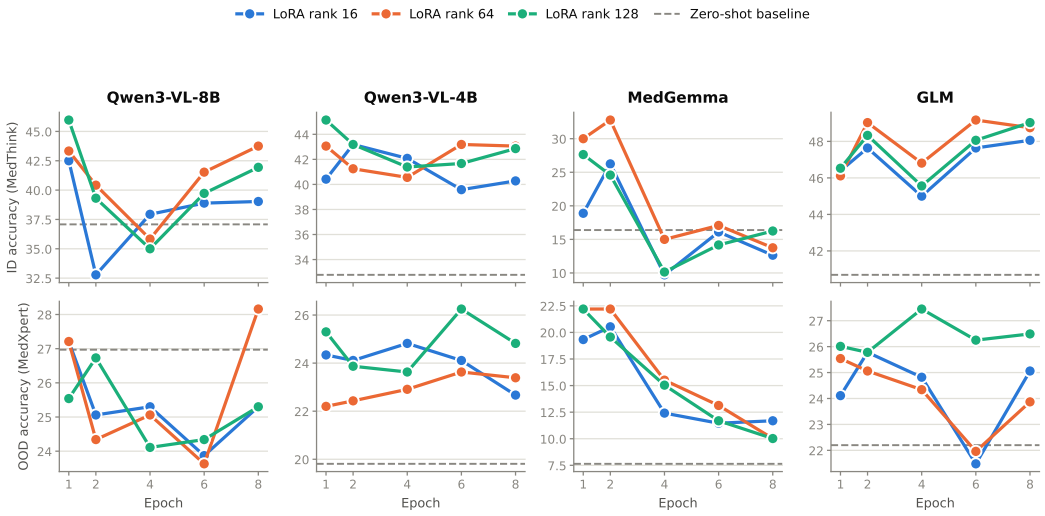


Figure 2: ID (MedThinkVQA, top row) and OOD (MedXpertQA, bottom row) accuracy across training epoch, for LoRA ranks 16/64/128 (color) and four backbones (columns).

2 VLM Alignment on Clinical Multi-images

2.1 Experimental Setup

Data. We source alignment supervision from the training split of MedThinkVQA (Yao et al., 2026), which contains 7347 training samples. Each training case pairs its set of images with a human-authored findings section (per-image observations) and discussion section (the synthesis and final diagnostic reasoning); we use this findings-plus-discussion trace as the target output for supervised fine-tuning, so every SFT example carries a full clinical rationale grounded in that case's own images rather than only a terminal answer label. For in-domain (ID) evaluation we use the MedThinkVQA test split, comprising 720 samples. For out-of-domain (OOD) evaluation we use MedXpertQA (Zuo et al., 2025), filtered to its multi-image examples, yielding 419 samples averaging 3.03 images per case (Table 3); note that this figure describes the filtered multi-image subset we evaluate on, not MedXpertQA as a whole. All questions on both benchmarks are five-way multiple choice, so the accuracy floor is 20% by chance.

Backbones. We align four VLM backbones spanning different scales and pretraining lineages: Qwen3-VL at the 8B and 4B dense checkpoints (Bai et al., 2025); the clinically-adapted MedGemma

at the 4B checkpoint (Sellergren et al., 2025); and GLM at the 9B checkpoint (GLM-4.1V-9B-Thinking) (Hong et al., 2025). All four are fine-tuned with LoRA rather than full-parameter updates, so that the alignment *recipe* (SFT epochs, RL or not, LoRA rank) is the object of study, with backbone held fixed within each comparison.

Ablation grid. For each backbone, we sweep SFT epochs $\in \{1, 2, 4, 6, 8\}$ crossed with LoRA rank $\in \{16, 64, 128\}$ (Hu et al., 2022), yielding 15 SFT configurations per backbone and 60 in total. For the RL stage, we take the SFT checkpoints at epoch 1 and epoch 8, each under LoRA rank 64 and rank 128, as four starting points, and continue training from each with GRPO (Shao et al., 2024; DeepSeek-AI, 2025) for 20, 40, 80, 160, and 200 steps, using a separate LoRA rank of 16 or 64 for the RL update. The full RL grid is eight SFT-to-GRPO runs per backbone at five step counts each. Section 2.3 reports the three runs completed at submission time, all on Qwen3-VL-8B, and we mark its findings as preliminary accordingly.

2.2 Supervised Fine-tuning of VLM

The SFT ablation covers four backbones: Qwen3-VL-8B, Qwen3-VL-4B, MedGemma, and GLM. Figure 2 plots ID and OOD accuracy against epoch for every backbone and LoRA rank. Each backbone has fifteen SFT configurations (three LoRA ranks times five epochs).

Finding 1

SFT's in-domain gains are large and near-universal, but its out-of-domain gains shrink as the backbone's zero-shot OOD accuracy rises, and for the strongest OOD baseline SFT is net harmful.

In three backbones, SFT configurations improve ID performance almost universally across different backbones: 15/15 configurations beat the ID baseline for Qwen3-VL-4B and GLM, and 12/15 for Qwen3-VL-8B. The pattern is inverted in MedGemma; however, where only 7/15 SFT configurations beat the ID baseline, and eight configurations fall *below* zero-shot, namely all three ranks at epochs 4 and 8 as well as ranks 16 and 128 at epoch 6. OOD behaves differently: the gain over baseline is largest where the baseline is weakest. MedGemma beats its OOD baseline in 15/15 of SFT configurations, Qwen3-VL-4B 15/15, and GLM 13/15, but Qwen3-VL-8B only 3/15, and its post-SFT OOD accuracy averages 25.41, below zero-shot. The net effect is convergence in performance. Across the three backbones that do not collapse, zero-shot ID spans 7.91 points (32.78 to 40.69) but best-configuration ID spans only 4.03 (45.14 to 49.17), and epoch-averaged OOD spans 1.51 (23.90 to 25.41) against a 7.16-point baseline spread. When it is successful, SFT narrows the backbones' starting differences.

However, the debate over what SFT teaches models may also apply here, particularly in the case of MedGemma. (Chu et al., 2025; Ren et al., 2026). Ren et al. (2026) argue that weaker models copy surface-level verbosity instead of learning transferable patterns. MedGemma's zero-shot accuracy is roughly half of every other backbone's, and it is the only backbone that SFT repeatedly pushes below its own zero-shot ID score.

Finding 2

A mid-training dip in ID accuracy, typically at epoch 4, appears in every backbone and rank tested. Where accuracy resettles afterwards is backbone-specific.

All twelve backbone-rank ID series show a clear local dip. Ten dip at epoch 4; the other two dip earlier (Qwen3-VL-8B rank 16, epoch 2) or later (Qwen3-VL-4B rank 16, epoch 6). How a model recover; however, differs by backbone. Both Qwen backbones dip and then return close to, but usually below, their epoch-1 level: Qwen3-VL-4B rank 128 falls from 45.14 to 41.39 and recovers only to 42.86. GLM recovers past its starting point, with every rank's epoch-8 value at or above its epoch-2 value and rank 128 peaking at epoch 8 (49.03), its highest point overall. MedGemma does not recover: rank 64 rebounds at epoch 6 (17.08) then falls to 13.75, below its own epoch-4 dip (15.00), and rank 16 rebounds to 16.11 then falls back to 12.64.

The dip's consistent timing supports a simple account: one epoch is enough to shift output format toward the target domain, and further training briefly overfits the narrow training split before the

model resettles (Zhou et al., 2023; Ren et al., 2026; Fesser et al., 2026). Because the dip lands at the same point in all four backbones, it plausibly stems from the training data and schedule rather than from any single backbone. What the schedule does not explain is the direction of the resettling, which is higher for GLM, roughly unchanged for both Qwen models, and lower for MedGemma.

Finding 3

LoRA rank sets the height of the epoch-1 peak, but backbone identity predicts the instability. Within a model family, scale does not buy robustness.

At epoch 1, ID accuracy rises with rank in both Qwen backbones: 42.50/43.33/45.97 (ranks 16/64/128) for Qwen3-VL-8B and 40.42/43.06/45.14 for Qwen3-VL-4B. MedGemma breaks this order (rank 64 at 30.00 beats rank 128 at 27.64, with rank 16 lowest at 18.89), and GLM has no order to break, its epoch-1 values sitting within half a point of each other (46.39/46.11/46.53). Instability, measured as the gap between a rank's best and worst epoch, separates backbones far more cleanly than it separates ranks. Qwen3-VL-4B (2.63 to 3.75) and GLM (3.06 to 3.47) occupy one stable band, Qwen3-VL-8B (7.92 to 10.97) a second, and MedGemma (16.53 to 19.03) a third; the three bands do not overlap, whereas the three ranks within a backbone always fall inside one of them. Scale within a family does not help either: Qwen3-VL-4B beats Qwen3-VL-8B on ID in 11 of 15 configurations, losing three comparisons at epoch 1 but winning the majority at later epochs.

The rank-as-capacity reading of the epoch-1 ordering is consistent with the broader LoRA literature: higher rank adds trainable capacity (Hu et al., 2022), and lower-rank adapters trade target-task performance for better preservation of other capabilities (Biderman et al., 2024). The effect survives within a backbone, where rank 128 is the least stable rank in three of four backbones (MedGemma, whose rank 64 is least stable, is the exception). However, it cannot explain the cross-backbone differences: the same rank sweep barely moves GLM yet destabilizes MedGemma five times more than either Qwen backbone. What makes a backbone robust or fragile under LoRA fine-tuning lies outside the rank-as-capacity account.

Finding 4

LoRA rank affects OOD accuracy in some backbones and not others. Rank 128 gives the best OOD accuracy in 9 of 20 backbone-and-epoch settings, against 4 each for ranks 16 and 64.

Qwen3-VL-8B is the flattest case: its epoch-averaged OOD accuracy varies by only 0.48 points across ranks (25.20 to 25.68). The others spread more widely, at 1.86 points for Qwen3-VL-4B, 1.53 for MedGemma, and 2.24 for GLM. GLM is also the only backbone in which one rank dominates throughout: rank 128 matches or beats both other ranks at every epoch, and it is GLM's most stable rank as well (a 1.67-point range, versus 3.58 and 4.30). Elsewhere the winning rank changes from epoch to epoch, so the aggregate count overstates how systematic the effect is.

Our reading of Finding 4 draws on intrinsic dimensionality (Aghajanyan et al., 2021): if all three ranks already exceed the capacity a low-dimensional update needs, OOD accuracy should be flat across ranks. Qwen3-VL-8B fits this prediction; GLM does not. Its consistent rank-128 advantage suggests that adding LoRA capacity results in OOD generalization, not just overfitting ID.

2.3 Impact of Reinforcement Learning on VLM

We next ask how a reinforcement learning stage applied after SFT affects performance on clinical multi-image analysis as shown in Figure 3.

Finding 5

GRPO substantially lowers ID accuracy relative to the SFT checkpoint it starts from.

From the epoch 1 checkpoint (ID 45.97), ID falls immediately and stays low through step 200: 8.33–11.25 points below start at rank 16 (mean 36.36), 8.61–10.28 below at rank 64 (mean 36.64). From the weaker epoch 8 checkpoint (ID 41.94), rank 64 drops less but still substantially: 3.75–6.80 points below start (mean 36.53), 5.4 below by step 200.

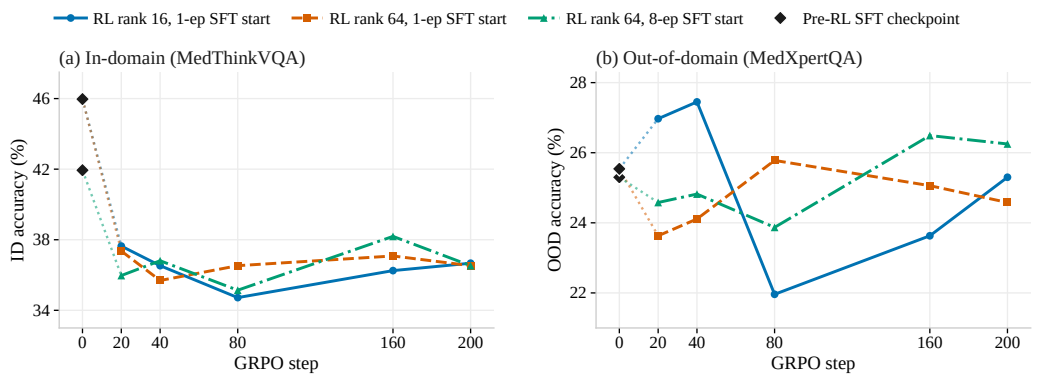


Figure 3: GRPO from a rank-128 SFT base: (a) ID accuracy (Finding 5); (b) OOD accuracy (Finding 6), both by GRPO step. Diamonds mark the pre-RL SFT checkpoint.

We hypothesize three potential reasons. First, our reward is purely outcome-based, which other work has shown to impact generalization while SFT stabilizes format (Chu et al., 2025). With no pressure to preserve that format, the policy is free to drift while chasing reward. Second, KL regularization toward the reference policy does not reliably prevent drift once the reward is imperfect (Kwa et al., 2024), consistent with the large, fast drop here. Third, an epoch 1 checkpoint is already a sharp, possibly fragile peak per Finding 3, so RL may perturb it more easily. The epoch 8 result supports this: its smaller drop matches prior work where RL only recovers out-of-domain performance from a strong SFT checkpoint rather than a weak one (Fesser et al., 2026).

Finding 6

GRPO's effect on OOD accuracy is inconsistent across the RL LoRA ranks.

At rank 16, OOD starts above the 25.54 baseline, peaks at step 40 (27.45), drops sharply at step 80 (21.96), then partially recovers to 25.30 by step 200. At rank 64 (same start), OOD starts below baseline, peaks at step 80 (25.78), then eases just under baseline to 24.58. The two ranks trace almost opposite trajectories; neither settles by step 200. The epoch 8, rank 64 run traces a third pattern: it dips to 23.87 at step 80 before climbing past its own lower baseline to 26.25 by step 200.

One candidate explanation is the SFT/RL capacity asymmetry from Finding 5: RL needs less adapter capacity than SFT to reach its ceiling (Fesser et al., 2026), which would put rank 16 and rank 64 on different sides of whatever capacity GRPO needs. The epoch 8 run's own third trajectory points to the SFT start mattering as well, not rank alone.

3 MULTI-MULTI: Decomposed Reasoning for Multi-image Multi-modal Clinical Analysis

Prior work shows VLMs fail in two ways on multi-image medical reasoning: confident diagnoses when images are absent (Asadi et al., 2026), and failure to ground and compose evidence when images are present (Yao et al., 2026). Both challenges are diagnosed at inference time and neither proposes a fix solution. We identify a third failure introduced *during* training. Our SFT-only models are accurate and largely hallucination-free in-domain, yet hallucinate substantially more out-of-domain; this asymmetry shows that caption-supervised SFT can teach a model to format reasoning around images without genuinely deriving it from them, a failure we call supervision-modality mismatch. Motivated by the insights drawn in Section 2, we propose a framework named MULTI-MULTI that addresses this with two components, both realized entirely through supervised fine-tuning: a decomposed reasoning chain (visual reliance inspectable at each step) and Image-Swap Supervision (that reliance enforced). Both components are realized through SFT-stage data construction alone, with no reinforcement learning or preference optimization anywhere in the pipeline. Figure 4 previews both components together before the detailed construction below.

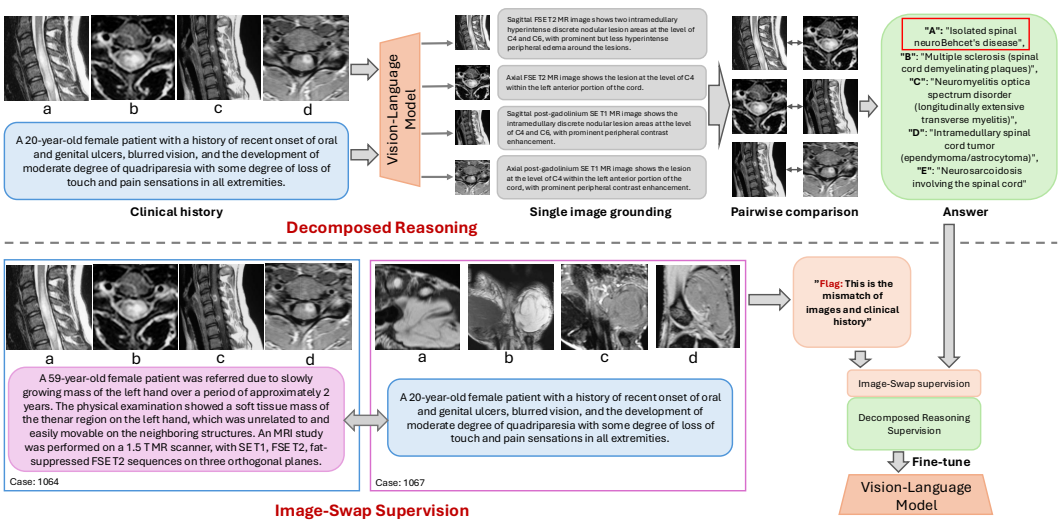


Figure 4: MULTI-MULTI's SFT pipeline. *Top:* the decomposed reasoning chain concatenates three checkable parts: one grounding statement per image, a sentence comparing the two images whose captions differ most, and a justification for the answer. *Bottom:* Image-Swap Supervision runs the same construction, but pairs the images with a clinical history from a different sample, so the target flags the mismatch instead of reasoning from it. Swapped and unswapped examples train together under one objective, with no separate loss term.

3.1 Decomposed Reasoning Chain

We decompose the SFT target y for an n-image case into three concatenated components:

$$y = \underbrace{g_1 \,\|\, \cdots \,\|\, g_n}_{\text{per-image grounding}} \;\|\; \underbrace{s_{i^*,j^*}}_{\text{composition}} \;\|\; \underbrace{a}_{\text{answer justification}} \tag{1}$$

so grounding and composition stay separate, individually checkable parts of the output rather than one opaque block. Each g_i rewrites image i's expert caption as "Image i shows {caption content}."

No dataset used here provides human cross-image comparisons, so we auto-construct the composition sentence s_{i^*,j^*} from existing captions. Define the pairwise caption-difference operator

$$\mathrm{diff}(c_i, c_j) \;=\; \mathrm{words}(c_i) \,\triangle\, \mathrm{words}(c_j), \tag{2}$$

the symmetric difference of the content-word sets of captions c_i and c_j. We select the most informative pair as

$$(i^*, j^*) \;=\; \underset{1 \le i < j \le n}{\arg\max} \,\big|\mathrm{diff}(c_i, c_j)\big|, \tag{3}$$

and instantiate a fixed template naming up to six terms from $\mathrm{diff}(c_{i^*}, c_{j^*})$: *"Comparing Image i^* and Image j^*, the findings differ with respect to {terms}, which is relevant when weighing the candidate diagnoses against each other."* This needs no new annotation and is necessary: without it, no training target ever contains a cross-image comparison, so per-image supervision alone gives the model no signal for what one looks like. This split converges independently with MedThinkVQA's own zero-shot decomposition, suggesting it reflects the task's natural structure.

3.2 Image-Swap Supervision (ISS)

Decomposition alone cannot guarantee content is image-derived: a model can fill every step fluently while drawing on clinical history text instead. ISS targets this: for a subset of cases we swap in another case's history h' while keeping the images $\{I_i\}_{i=1}^n$ unchanged, and reuse the same construction as Eq. 1 to build the target for the pairing $(\{I_i\}, h')$, so every step comes under counterfactual pressure, while the target explicitly flags the mismatch in the form "The clinical history provided does not correspond to…" Donors are constrained to different ICD chapters so a same-category swap cannot dilute the signal.

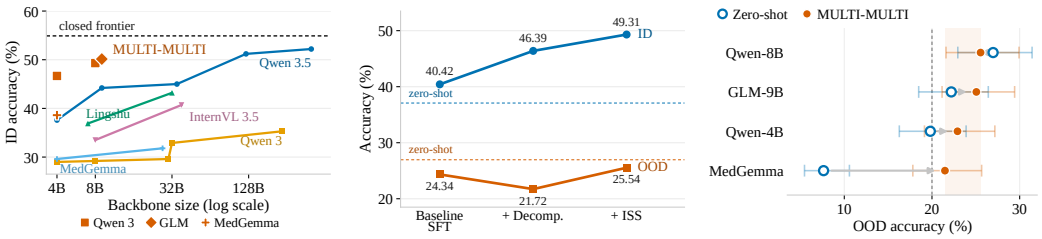


Figure 5: (a) ID accuracy against backbone size for MULTI-MULTI and published baselines; the dashed rule marks the lower edge of the closed-source band (54.9 to 57.2). (b) Component ablation on Qwen3-VL-8B, adding decomposed reasoning and then ISS to the SFT baseline; dashed rules mark the un-fine-tuned model. (c) OOD accuracy on MedXpertQA ($n = 419$).

3.3 Results

3.3.1 Comparison with the Baselines

Comparison with Baselines Figure 5(a) positions MULTI-MULTI within the broader published landscape: reported accuracy across five VLM families at various sizes (Qwen 3.5&3, Lingshu, InternVL 3.5, and MedGemma (Sellergren et al., 2025)), alongside a proprietary reference band.

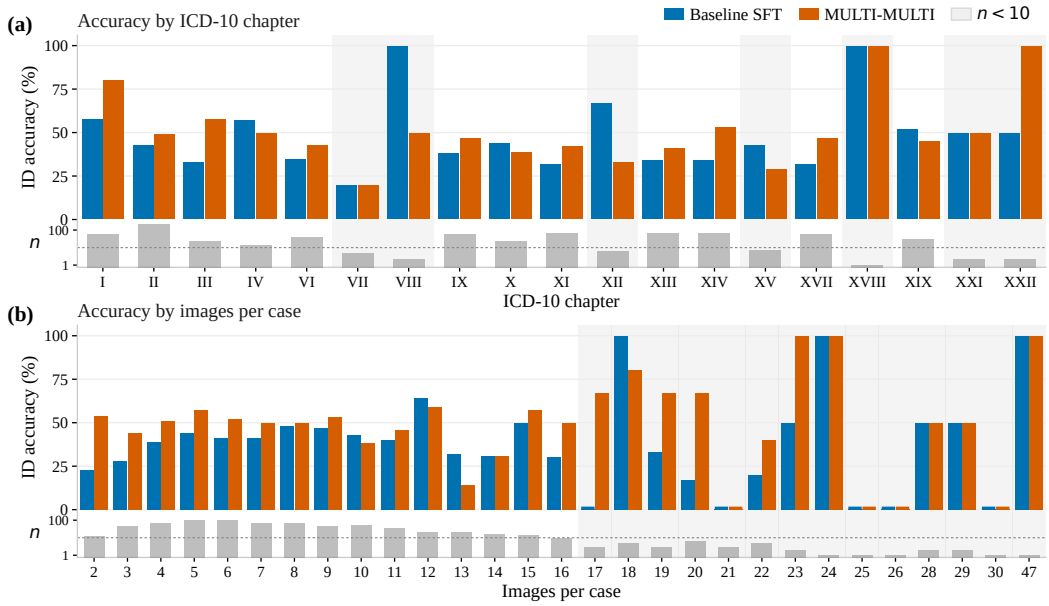


Figure 6: Where the in-domain gain comes from: MULTI-MULTI against the SFT baseline on Qwen3-VL-8B, ID (MedThinkVQA) test set. Each panel carries a sample-size strip (n, log scale), since equal-width columns hide unequal n. Shaded columns hold fewer than ten cases and are not individually informative; a stub at zero is a measured 0%, not a missing column. (a) By ICD-10 chapter: MULTI-MULTI leads in 10 of 19 chapters, ties in 3 and trails in 6, but the seven shaded chapters hold only 25 of 720 cases between them. Chapters V, XVI and XX have no test cases. (b) By images per case: the 14 shaded counts occupy half the panel yet hold 36 of 720 cases (5.0%), so volatility past 16 images reflects sample size, not behavior.

First, at 8B parameters MULTI-MULTI's 49.31% matches or exceeds every plotted Qwen 3, Lingshu, InternVL 3.5 and MedGemma checkpoint at any scale (up to 235B, 32B, 38B and 27B respectively), and every Qwen 3.5 checkpoint up to roughly 35B, trailing only Qwen 3.5's 122B and 397B points (51.2% and 52.2%). Second, this generalizes across backbones and scales: GLM-9B reaches 50.14% and MedGemma-4B 38.61%, while the 4B Qwen 3 backbone alone (46.67%) exceeds the

largest Qwen 3 baseline (235B, 35.3%). MedGemma isolates the effect of alignment with backbone and scale held fixed: 38.61% against 29.6% at the same 4B size, and above even its 27B baseline (31.8%). Third, the closed-source band remains 5.6 to 7.9 points ahead: Claude-4.6-Opus, Gemini-3-Pro and GPT-5.2, prompted zero-shot on the same cases, reach 54.9–57.2% (Yao et al., 2026). Out of domain the effect is convergence rather than gain: on MedXpertQA (Figure 5(c)) MULTI-MULTI compresses a 19.3-point spread across the four zero-shot backbones (7.6 to 27.0) into a 4.1-point band (21.5 to 25.5), raising MedGemma by 13.8 points and GLM by 2.9, while Qwen3-VL-8B, the only one already above that band, falls by 1.4.

Component ablation. To attribute the gain above to each piece individually, we compare three conditions on Qwen3-VL-8B: the SFT baseline; the decomposed reasoning chain alone, without ISS; and the full method. Figure 5(b) plots all three on ID/OOD accuracy. ID accuracy improves monotonically across all three conditions (40.42 $\rightarrow$ 46.39 $\rightarrow$ 49.31), consistent with decomposition changing what the model is asked to produce and ISS adding a further gain on top. OOD accuracy does not: it drops from 24.34 to 21.72 when the decomposed chain is added alone, before recovering past the baseline to 25.54 once ISS is added. This pattern shows each component doing distinct, necessary work: decomposition drives the ID gains by giving the model an explicit per-image structure to reason with, but that structure alone does not yet force the reasoning to be image-derived, so OOD accuracy briefly dips (the same format-matching-without-grounding gap Finding 1, Section 2.2, already shows for the SFT baseline). ISS is what closes that gap: forcing the model to detect a history-image mismatch makes genuine image-grounding necessary rather than optional, which is why OOD accuracy recovers past the baseline only once ISS is added on top of decomposition.

Per-chapter breakdown. Aggregate accuracy can hide whether a gain is broad or concentrated in a few categories. Figure 6(a) breaks MULTI-MULTI vs. the SFT baseline down by ICD-10 chapter, both on Qwen3-VL-8B. MULTI-MULTI leads in 10 of the 19 represented chapters, ties in 3 (Chapters VII, XVIII, XXI), and trails in 6 (Chapters IV, VIII, X, XII, XV, XIX); the three chapters not shown have zero ID test cases (Table 1, Section A.3, reports n per chapter for both splits). The largest gap in MULTI-MULTI's favor at a meaningful sample size is Chapter III (58.3 vs. 33.3, $n = 24$); Chapter XXII's larger-looking gap (100.0 vs. 50.0) and the largest gap against it, Chapter VIII (50.0 vs. 100.0), are both just $n = 2$ and read as anecdotal rather than a stable effect.

Robustness to image count. Section 3.1 flags that s_{i^*,j^*} references at most two of n images regardless of n, and commits to measuring whether that costs accuracy as n grows rather than leaving it unquantified. Figure 6(b) reports ID accuracy by images per case for MULTI-MULTI and the SFT baseline, both on Qwen3-VL-8B, one column per image count with at least one test case (count 27 has none). MULTI-MULTI comes out ahead broadly: across all 29 image counts it leads outright at 16, ties at 9, and trails at only 4. The advantage is largest and most consistent at low counts (2–6 images), where MULTI-MULTI leads at every one, by 16.5 points on average. It narrows through 7–14 images – MULTI-MULTI still leads at four of the eight counts against three trailing, though one large trailing count (13 images, -18.2) pulls the average to roughly flat (-0.8 points) – consistent with the concern above, since grounding only two of n images is a smaller share of the case at $n = 2$ than at $n = 12$. From 15 images on the numbers get noisier, but the lead reasserts itself: MULTI-MULTI wins 5 of the 6 counts from 15–20, by as much as 66.7 points (17 images), with only 18 images favoring the baseline (-20.0). Past 20 images the data cannot really speak either way: 8 of the 10 counts at 21 or above show identical accuracy for both models, a pattern more consistent with very few test cases per count than with any real effect.

4 Conclusion

A sweep over four backbones, SFT epochs, LoRA rank and an optional RL stage yields six findings. They expose a training-time failure: caption-supervised SFT rewards reasoning formatted around the images but not derived from them. MULTI-MULTI addresses it in the SFT target alone: each trace decomposes into per-image grounding, cross-image comparison and justification, and a quarter of cases carry another patient's history, a mismatch only the images reveal. Our 8B model matches the in-domain accuracy of zero-shot open models nearly fifty times larger, and narrows the gap to the leading proprietary models.

REFERENCES

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 23716–23736, 2022.

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Rajabalifardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage: The illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. Reports dense 2B/4B/8B/32B and MoE 30B-A3B/235B-A22B variants.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15016–15027, 2023.

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research (TMLR)*, 2024.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2024.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *International Conference on Machine Learning (ICML)*, 2025.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. Also published in Nature 645, 633–638 (2025), doi:10.1038/s41586-025-09422-z.

Lukas Fesser, Hanlin Zhang, Michelle M. Li, Eric Wang, Bryan Perozzi, Shekoofeh Azizi, Sham M. Kakade, and Marinka Zitnik. How post-training shapes biological reasoning models. *arXiv preprint arXiv:2606.16517*, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024. doi: 10.1007/978-3-031-73337-6_9. URL https://arxiv.org/abs/2404.12390.

Krzysztof J. Geras, Stacey Wolfson, Yiqiu Shen, Nan Wu, S. Gene Kim, Eric Kim, Laura Heacock, Ujas Parikh, Linda Moy, and Kyunghyun Cho. High-resolution breast cancer screening with multi-view deep convolutional neural networks. *arXiv preprint arXiv:1703.07047*, 2017.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, et al. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. GLM-V Team. NEEDS CONFIRMATION: which checkpoint (e.g., GLM-4.1V-9B-Thinking vs. GLM-4.5V) was actually used as the backbone.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2405.01483.

Thomas Kwa, Drake Thomas, and Adrià Garriga-Alonso. Catastrophic goodhart: Regularizing RLHF with KL divergence does not mitigate heavy-tailed reward misspecification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023b.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2312.07533.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

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

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021. doi: 10.1038/s41551-020-00682-w.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2408.02718.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2025. URL https://arxiv.org/abs/2502.19634.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Qihan Ren, Peng Wang, Ruikun Cai, Shuai Shao, Dadi Guo, Yuejin Xie, Yafu Li, Quanshi Zhang, Xia Hu, Jing Shao, and Dongrui Liu. Rethinking generalization in reasoning SFT: A conditional analysis on optimization, data, and model capability. *arXiv preprint arXiv:2604.06628*, 2026.

Jonathan Rubin, Deepan Sanghavi, Claire Zhao, Kathy Lee, Ashequl Qadir, and Minnan Xu-Wilson. Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. *arXiv preprint arXiv:1804.07839*, 2018.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. Google Research and Google DeepMind. NEEDS CONFIRMATION: which checkpoint (4B vs. 27B) was used as the backbone.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. Introduces Group Relative Policy Optimization (GRPO).

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578, 2024. URL https://arxiv.org/abs/2401.06209.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2406.09411.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*, 2023.

Zonghai Yao, Benlu Wang, Yifan Zhang, Junda Wang, Iris Xia, Zhipeng Tang, Shuo Han, Feiyun Ouyang, Zhichao Yang, Arman Cohan, and Hong Yu. Medical thinking with multiple images. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2604.16506. Introduces the MedThinkVQA benchmark.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.01936.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025. Accepted to ICML 2025.

A Appendix

A.1 Related Work

Vision-Language Models. Modern vision-language models (VLMs) align a pretrained visual encoder with a large language model (LLM). Contrastive pretraining on web-scale image-text pairs, popularized by CLIP (Radford et al., 2021), remains the dominant source of visual representations, and architectures such as BLIP-2 (Li et al., 2023b) and Flamingo (Alayrac et al., 2022) bridge such encoders to an LLM via a querying transformer or gated cross-attention, respectively. Visual instruction tuning (Liu et al., 2023; 2024) popularized the simpler projector-based recipe underlying most current VLMs, including InternVL (Chen et al., 2024). In every case, the alignment mechanism is designed and frozen around a nominally single-image input. Our work instead treats the number and relationship of input images as a first-class design axis of alignment itself, a question these architectures do not address by default.

Multi-Image Reasoning with Vision-Language Models. Giving VLMs more images does not translate into more understanding: benchmarks such as MuirBench (Wang et al., 2025) and MMIU (Meng et al., 2025) show even the strongest proprietary models degrade sharply on cross-image comparison, and BLINK (Fu et al., 2024) traces part of this to basic perceptual failures. Methods such as Mantis (Jiang et al., 2024) and VILA (Lin et al., 2024) respond by training explicitly on interleaved multi-image data, while a complementary line of work shows contrastively-aligned VLMs behave like bags-of-words that ignore relations (Yuksekgonul et al., 2023) and rely on encoders that cannot distinguish visually distinct images (Tong et al., 2024). These efforts target natural images and general-domain benchmarks, where images are only loosely related and errors are comparatively low-stakes; none is designed for or evaluated on the tighter, anatomically-grounded correspondences that clinical multi-image reading requires, which is the setting our work targets.

Multi-Image Reasoning in Clinical Applications. Clinical VLMs have largely inherited single-image alignment: LLaVA-Med (Li et al., 2023a) and BiomedCLIP (Zhang et al., 2023) adapt general-domain recipes to biomedicine one image at a time, and even RadFM (Wu et al., 2023), trained on 2D and 3D data at web scale, does not explicitly model how alignment should adapt to a variable number of clinically related images at inference time. This gap shows up directly in evaluation: MedThinkVQA (Yao et al., 2026), a benchmark averaging 6.6 images per case, finds that even the strongest closed models (Claude-4.6-Opus, Gemini-3-Pro, GPT-5.2) reach only 54.9–57.2% accuracy and open-source and smaller models fall well below 50%, with over 70% of errors traced to failures in reading and integrating evidence across images rather than in diagnostic reasoning itself. Pre-VLM architectures instead hardwire a fixed image structure into dedicated fusion layers – multi-view CNNs for mammography's four standard views (Geras et al., 2017), DualNet for frontal/lateral chest X-rays (Rubin et al., 2018), CLAM for whole-slide-image patches (Lu et al., 2021) – and closest to our setting, BioViL-T (Bannur et al., 2023) aligns a vision-language model across a current and prior chest X-ray but is specialized to exactly this one pair and modality. No existing work offers a comprehensive alignment mechanism analysis for the variable number and type of images a clinical VLM must jointly reason over.

A.2 Implementation Details

SFT implementation. The SFT target concatenates the per-image findings, the discussion, and the final answer. We use this target to train an SFT baseline for four backbones: Qwen3-VL-4B-Thinking, Qwen3-VL-8B-Thinking, MedGemma-4B-it, and GLM-4.1V-9B-Thinking.[1] We sweep LoRA rank over $\{16, 64, 128\}$ at 8 epochs, learning rate 2×10^{-5}, and a maximum output length of 4096 tokens.

RL implementation. We apply GRPO with reward $r = r_{\text{format}} + r_{\text{analysis}} + r_{\text{answer}}$, sampling temperature 1.0, learning rate 1×10^{-6}, KL coefficient 0.04, and the same maximum output length of 4096 tokens.

MULTI-MULTI implementation. The Multi-Multi method are described in algorithm 1. We fine-tune all four backbones with LoRA adapters rather than full-parameter tuning. Qwen3-VL-

[1] HuggingFace identifiers: `Qwen/Qwen3-VL-4B-Thinking`, `Qwen/Qwen3-VL-8B-Thinking`, `google/medgemma-4b-it`, `zai-org/GLM-4.1V-9B-Thinking`.

Algorithm 1 MULTI-MULTI: SFT Data Construction and Training

Require: Dataset $\mathcal{D} = \big\{\big(\{I_i^{(k)}\}_{i=1}^{n_k}, \{c_i^{(k)}\}_{i=1}^{n_k}, h^{(k)}, q^{(k)}, a^{(k)}\big)\big\}_{k=1}^{N}$
(images, expert captions, clinical history, question, ground-truth answer)
Require: ISS swap rate $\rho \in [0, 1]$, ICD-chapter map $\chi(\cdot)$
Ensure: Fine-tuned model parameters θ
Phase 1: Decomposed Target Construction (with Image-Swap Supervision)
$\mathcal{D}' \leftarrow \emptyset$
for $k = 1$ **to** N **do**
  for $i = 1$ **to** n_k **do**
    $g_i \leftarrow$ *"Image i shows* $\{c_i^{(k)}\}$*."* {per-image grounding}
  end for
  $(i^*, j^*) \leftarrow \arg\max_{1 \le i < j \le n_k} \big|\mathrm{diff}(c_i^{(k)}, c_j^{(k)})\big|$ {Eq. equation 3}
  $s \leftarrow$ composition template filled with terms from $\mathrm{diff}(c_{i^*}^{(k)}, c_{j^*}^{(k)})$ {Eq. equation 2}
  $\tilde{h} \leftarrow h^{(k)}$; swapped $\leftarrow$ **false**
  if $\mathrm{Bernoulli}(\rho) = 1$ **then** {Image-Swap Supervision, Sec. 3.2}
    sample donor case k' with $\chi(h^{(k')}) \neq \chi(h^{(k)})$
    $\tilde{h} \leftarrow h^{(k')}$; swapped $\leftarrow$ **true**
  end if
  $a \leftarrow$ synthesis over $(g_1, \ldots, g_{n_k}, s)$ + justification for $a^{(k)}$, flagging mismatch if swapped
  $y^{(k)} \leftarrow g_1 \| \cdots \| g_{n_k} \| s \| a$ {Eq. equation 1}
  $x^{(k)} \leftarrow \big(\{I_i^{(k)}\}_{i=1}^{n_k}, \tilde{h}, q^{(k)}\big)$
  $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(x^{(k)}, y^{(k)})\}$
end for
Phase 2: Supervised Fine-Tuning
$\theta \leftarrow \arg\min_\theta \ \mathbb{E}_{(x,y)\sim\mathcal{D}'}\big[-\log p_\theta(y \mid x)\big]$ {plain SFT; no RL or preference optimization}
return θ =0

8B-Thinking is the primary backbone, used for the hyperparameter ablations; the other three test whether the method generalizes beyond this one backbone. For Qwen3-VL-8B-Thinking, we use LoRA rank 64 and learning rate 2×10^{-5}. By default, ISS (Section 3.2) swaps in randomly mismatched images for 25% of training cases, keeping the clinical history and question fixed.

Compute and Training Resources All experiments are conducted on the Stanford Sherlock computing cluster, using NVIDIA H100 GPUs with 80GB of memory. SFT and RL training each use 2 GPUs; inference uses 1 GPU.

A.3 Dataset Statistics

Section 2.1 reports the headline sample counts: 7347 MedThinkVQA training cases, 720 MedThinkVQA test (ID) cases, and 419 MedXpertQA cases filtered to its multi-image subset (OOD). MedThinkVQA cases average 6.6 images each (Section 1); MedXpertQA spans 17 specialties (Section 1).

In-domain and Out-Of-Domain Datasets Comparison Our in-domain and out-of-domain benchmarks are built from disjoint source pools by independent research groups, so no item-level overlap is possible between them. MedThinkVQA, our in-domain benchmark, is adapted from Eurorad, the European Society of Radiology's peer-reviewed teaching-case database, and consists of real clinical cases submitted by radiologists across major subspecialties, curated into 8,067 cases and filtered so that each case requires genuine cross-image reasoning rather than being solvable from text alone (Yao et al., 2026). MedXpertQA, our out-of-domain benchmark, instead draws on board-exam-style questions covering 17 of the 25 American Board of Medical Specialties member boards, with its multimodal track's clinical content curated through the authors' own data synthesis and multi-round expert review pipeline, designed specifically to mitigate data leakage risk (Zuo et al., 2025). Beyond this difference in origin and content type, the two benchmarks also differ sharply in structure: MedThinkVQA cases average 6.62 images each, and MedThinkVQA's own related-work comparison notes that MedXpertQA's multimodal questions average well under two images

each (at most 1.43), consistent with the two benchmarks being built around fundamentally different notions of a case rather than sharing or re-deriving material from one another (Yao et al., 2026). Taken together, the disjoint source institutions, independently reviewed construction pipelines, and distinct image-density profiles support treating in-domain and out-of-domain evaluation in this work as drawing on non-overlapping underlying data.

Per-ICD-10-chapter breakdown. Adopted from the statistics of samples from each chapter in supplementary of MedThink Yao et al. (2026), we summarized them into Table 1. Table 1 gives the case count and share of each MedThinkVQA split, train and ID test, by ICD-10 chapter. The same chapters underlying the per-chapter accuracy comparison in Figure 5(c). Chapters V, XVI, and XX have zero test cases and are therefore absent from that figure; several of the remaining 19 have very few (Chapter XVIII: 1; Chapters VIII, XXI, XXII: 2 each), so per-chapter accuracy there should be read as anecdotal rather than a stable effect.

Ch.	ICD-10 Chapter	Train n (%)	Test n (%)
I	Certain infectious and parasitic diseases	364 (5.0%)	55 (7.6%)
II	Neoplasms	1934 (26.3%)	202 (28.1%)
III	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	142 (1.9%)	24 (3.3%)
IV	Endocrine, nutritional and metabolic diseases	217 (3.0%)	14 (1.9%)
V	Mental and behavioural disorders	2 (0.0%)	0 (0.0%)
VI	Diseases of the nervous system	359 (4.9%)	37 (5.1%)
VII	Diseases of the eye and adnexa	23 (0.3%)	5 (0.7%)
VIII	Diseases of the ear and mastoid process	22 (0.3%)	2 (0.3%)
IX	Diseases of the circulatory system	759 (10.3%)	58 (8.1%)
X	Diseases of the respiratory system	222 (3.0%)	23 (3.2%)
XI	Diseases of the digestive system	892 (12.1%)	66 (9.2%)
XII	Diseases of the skin and subcutaneous tissue	43 (0.6%)	6 (0.8%)
XIII	Diseases of the musculoskeletal system and connective tissue	524 (7.1%)	61 (8.5%)
XIV	Diseases of the genitourinary system	344 (4.7%)	64 (8.9%)
XV	Pregnancy, childbirth and the puerperium	35 (0.5%)	7 (1.0%)
XVI	Certain conditions originating in the perinatal period	25 (0.3%)	0 (0.0%)
XVII	Congenital malformations, deformations and chromosomal abnormalities	793 (10.8%)	60 (8.3%)
XVIII	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	36 (0.5%)	1 (0.1%)
XIX	Injury, poisoning and certain other consequences of external causes	572 (7.8%)	31 (4.3%)
XX	External causes of morbidity and mortality	3 (0.0%)	0 (0.0%)
XXI	Factors influencing health status and contact with health services	18 (0.2%)	2 (0.3%)
XXII	Codes for special purposes	18 (0.2%)	2 (0.3%)
	Total	7347	720

Table 1: Case counts per ICD-10 chapter in MedThinkVQA's training split and ID test split (Section 2.1). Percentages are of each split's own total. Chapters V, XVI, and XX have zero test cases and are omitted from Figure 6(a); several of the remaining 19 chapters have very few test cases, so per-chapter accuracy there is not statistically meaningful on its own.

A.4 Prompts, Scoring, and Supervision Templates

This subsection gives the exact prompt text, scoring rules and target-construction templates behind every number in the paper. Two artefacts are reproduced verbatim: the prompt that defines the evaluation task (Figure 7) and the template that defines the supervision signal (Figure 8). Everything else in this subsection specifies how outputs produced under the first are scored, and how targets conforming to the second are built.

Inference. All evaluation is single-pass and deterministic: greedy decoding, one sample per item, no self-consistency, no retries, and a maximum output length of 4096 tokens shared by every condition. Images are supplied in the order the benchmark stores them, and that order fixes the indices $1, \ldots, n$

Images per Case (n)	Train Cases	Test Cases
1	121	0
2	475	13
3	877	46
4	1151	72
5	1009	98
6	854	100
7	682	70
8	538	66
9	422	47
10	311	53
11	209	35
12	169	22
13	120	22
14	92	16
15	79	14
16	54	10
17	40	3
18	21	5
19	24	3
20	24	6
21	17	3
22	16	5
23	5	2
24	6	1
25	6	1
26	5	1
27	3	0
28	5	2
29	3	2
30	1	1
31	1	0
33	1	0
34	1	0
38	2	0
39	1	0
44	1	0
47	0	1
49	1	0
Total	**7347**	**720**

Table 2: Number of MedThinkVQA cases by images per case (n), train and test splits. Every n with at least one case in either split is listed; a 0 means that split has no cases at that n. This is the per-count sample size underlying the per-image-count ID accuracy comparison in Figure 6(b) (test split) – $n = 27$ has no test cases, which is why that figure has no column for it.

Images per Case (n)	Test Cases
2	236
3	59
4	22
5	78
6	24
Total	**419**

Table 3: Number of MedXpertQA (test split) cases with at least two images, by images per case (n); single-image cases, if any exist in this split, are not shown here. MedXpertQA is the out-of-domain (OOD) evaluation set referenced throughout this paper (the OOD accuracy columns in Tables 4–7).

LoRA Rank	Epochs	ID Acc. (%)	OOD Acc. (%)
N/A	N/A	37.08	26.97
128	1	45.97	25.54
128	2	39.31	26.73
128	4	35.00	24.11
128	6	39.72	24.34
128	8	41.94	25.30
64	1	43.33	27.21
64	2	40.42	24.34
64	4	35.83	25.06
64	6	41.53	23.63
64	8	43.75	28.16
16	1	42.50	27.21
16	2	32.78	25.06
16	4	37.94	25.30
16	6	38.89	23.87
16	8	39.03	25.30

Table 4: Baseline SFT hyperparameter sweep for Qwen3-VL-8B (LoRA rank $\times$ epochs), ID accuracy on MedThinkVQA and OOD accuracy on MedXpert. First row (LoRA Rank = Epochs = N/A) is the zero-shot, pre-SFT reference point.

LoRA Rank	Epochs	ID Acc. (%)	OOD Acc. (%)
N/A	N/A	32.78	19.81
128	1	45.14	25.30
128	2	43.19	23.87
128	4	41.39	23.63
128	6	41.67	26.25
128	8	42.86	24.82
64	1	43.06	22.20
64	2	41.25	22.43
64	4	40.56	22.91
64	6	43.19	23.63
64	8	43.06	23.39
16	1	40.42	24.34
16	2	43.19	24.11
16	4	42.08	24.82
16	6	39.58	24.11
16	8	40.28	22.67

Table 5: Baseline SFT hyperparameter sweep for Qwen3-VL-4B (LoRA rank $\times$ epochs), ID accuracy on MedThinkVQA and OOD accuracy on MedXpert. First row (LoRA Rank = Epochs = N/A) is the zero-shot, pre-SFT reference point.

LoRA Rank	Epochs	ID Acc. (%)	OOD Acc. (%)
N/A	N/A	16.39	7.64
128	1	27.64	22.20
128	2	24.58	19.57
128	4	10.14	15.04
128	6	14.17	11.69
128	8	16.25	10.02
64	1	30.00	22.20
64	2	32.78	22.20
64	4	15.00	15.51
64	6	17.08	13.13
64	8	13.75	10.02
16	1	18.89	19.33
16	2	26.25	20.53
16	4	9.72	12.41
16	6	16.11	11.46
16	8	12.64	11.69

Table 6: Baseline SFT hyperparameter sweep for MedGemma (LoRA rank $\times$ epochs), ID accuracy on MedThinkVQA and OOD accuracy on MedXpert. First row (LoRA Rank = Epochs = N/A) is the zero-shot, pre-SFT reference point.

LoRA Rank	Epochs	ID Acc. (%)	OOD Acc. (%)
N/A	N/A	40.69	22.2
128	1	46.53	26.01
128	2	48.33	25.78
128	4	45.56	27.45
128	6	48.06	26.25
128	8	49.03	26.49
64	1	46.11	25.54
64	2	49.03	25.06
64	4	46.81	24.34
64	6	49.17	21.96
64	8	48.75	23.87
16	1	46.39	24.11
16	2	47.64	25.78
16	4	45	24.82
16	6	47.64	21.48
16	8	48.06	25.06

Table 7: Baseline SFT hyperparameter sweep for GLM (LoRA rank $\times$ epochs), ID accuracy on MedThinkVQA and OOD accuracy on MedXpert. First row (LoRA Rank = Epochs = N/A) is the zero-shot, pre-SFT reference point.

used by both the prompt and the target template, so *Image 3* means the same study to the model, to the target, and to a reader of the trace. The same system prompt is used for zero-shot, SFT-baseline and MULTI-MULTI evaluation, so that differences between conditions are attributable to training rather than to prompting.

The prompt in Figure 7 asks for per-image inspection, cross-image comparison and synthesis, in that order. It is worth noting what this implies for the baseline results in Section 2.2: the decomposition that MULTI-MULTI supplies through supervision is already requested of every model in the prompt. A baseline that fabricates findings is therefore not failing for want of instruction, and the instruction *"Do not invent clinical history, measurements, mechanisms of injury, treatments, or outcomes that are not given to you"* is present in every evaluation whose hallucination rate we report.

Answer extraction and format compliance. An answer is scored only when the response ends in a well-formed `<answer>X</answer>` tag with $X \in \{\texttt{A}, \ldots, \texttt{E}\}$. Both benchmarks are five-way multiple choice, so the chance floor is 20%. Because a model that answers correctly but does not emit the tag is indistinguishable under this rule from one that answers wrongly, we report the *format-compliance rate*, the fraction of items yielding a parseable tag, alongside accuracy for every configuration (Tables 4–7). This matters for interpreting the weaker backbones: a zero-shot accuracy at or below the chance floor is a statement about format compliance at least as much as about clinical ability, and should not be read as the latter alone. Responses that emit no tag, or a tag after which further text follows, are counted as incorrect rather than discarded, so accuracy is always computed over the full 720 in-domain and 419 out-of-domain items.

The SFT target. Figure 8 shows the target y for an n-image case exactly as it is written into the training data, including the `<think>...</think>` wrapper and the terminal answer tag. The three segments instantiate $y = g_1 \| \cdots \| g_n \| s_{i^*,j^*} \| a$ from Eq. 1. Their boundaries are the unit of inspection: because grounding, composition and justification occupy fixed, separately identifiable positions, a trace can be checked one segment at a time, and a failure can be attributed to a segment rather than to the output as a whole.

Per-image grounding. Each g_i rewrites image i's expert caption into the fixed frame *"Image i shows {caption content}."* No content is added, removed or paraphrased; the rewrite supplies only the index and the frame. The template is deliberately minimal, since its purpose is to bind a claim to an image rather than to change what the claim says.

Composition. No dataset used here provides human cross-image comparisons, so s_{i^*,j^*} is constructed from the captions themselves. The pair (i^*, j^*) maximises $|\text{diff}(c_i, c_j)|$, the size of the symmetric difference of the content-word sets (Eq. 2, Eq. 3), and up to six of its terms instantiate *"Comparing Image i^* and Image j^*, the findings differ with respect to {terms}, which is relevant when weighing the candidate diagnoses against each other."* Two properties of this construction are worth stating plainly, because both are visible in the outputs. The selection criterion is lexical, not clinical, so the chosen pair is the most differently worded rather than the most diagnostically informative; and the terms are surface content words, so degenerate sets containing stems or generic nouns are possible when captions are terse or stylistically uneven. Section A.5 shows an item where both effects occur together, and Section **??** discusses the consequences.

Image-Swap Supervision. For a Bernoulli-sampled fraction $\rho = 0.25$ of training cases, the case's own clinical history h is replaced by the history h' of a donor case drawn from a different ICD-10 chapter, with the images $\{I_i\}_{i=1}^n$ and the question left unchanged (Algorithm 1). The target for the swapped pairing is built by the same three templates above, with the justification a opening on the mismatch in the fixed form *"The clinical history provided does not correspond to…"* rather than reasoning as though h' were correct. The chapter constraint is a coarse proxy: it makes a same-category swap impossible, but it neither guarantees that a cross-chapter history is detectable from the images nor that a same-chapter one would not have been. Swapped and unswapped examples are mixed in a single training set and optimised under the same objective; no separate loss term, reward or preference signal is attached to the swapped subset.

System Prompt

```
You are an expert radiologist. You will be given one or more
    medical images and a clinical question with multiple-choice
    options.

Before selecting your answer, think carefully through the following
    steps:
1. Examine each image and note the key radiological findings.
2. Compare images to identify differences, temporal changes, or
    complementary evidence across modalities or time points.
3. Synthesise all findings into a unified diagnostic impression and
    explain why the other options are less likely.

Base your reasoning only on what is visible in the images and
    stated in the question. Do not invent clinical history,
    measurements, mechanisms of injury, treatments, or outcomes
    that are not given to you.

End your response with exactly:
<answer>X</answer>
where X is a single option letter (A, B, C, D, or E). No text is
    permitted after the answer tag.
```

Figure 7: System prompt used for zero-shot baseline evaluation.

A.5 Qualitative Examples

Figure 9 places four outputs side by side for three out-of-domain cases from the MedXpertQA MM test split (items MM-0, MM-34 and MM-18, top to bottom): the SFT baseline and MULTI-MULTI, each on the Qwen3-VL and GLM backbones. This figure is about the trace rather than the answer. Accuracy is established in Section 3.3.1; what is at issue here is which claims a trace makes, which of them the case actually supports, and whether a reader can tell the two apart. The three cases were chosen to span the answer outcomes for exactly that reason: MULTI-MULTI is correct and the baseline is not on MM-0, both are wrong on MM-34, and MULTI-MULTI alone is wrong on MM-18.

```
<think>
# per-image grounding:  g_1 ‖ ··· ‖ g_n
Image 1 shows {caption content of image 1}.
Image 2 shows {caption content of image 2}.
...
Image n shows {caption content of image n}.

# composition:  s_{i*,j*}
Comparing Image i* and Image j*, the findings differ with respect to
{terms from diff(c_{i*}, c_{j*})}, which is relevant when weighing the candidate
diagnoses against each other.

# answer justification:  a
{synthesis paragraph integrating all images}
This is most consistent with option X:  {option text}.

</think>
<answer>X</answer>
```

Figure 8: SFT target template for an n-image case, shown exactly as wrapped in the generated data. The three segments inside <think>...</think> instantiate $y = g_1 \| \cdots \| g_n \| s_{i^*,j^*} \| a$ from Eq. 1, with comparison terms drawn from $\mathrm{diff}(\cdot,\cdot)$ in Eq. 2. Braced spans are filled per case; the remaining text is fixed template language.

Holding the answer aside is the point, because the property we are examining varies independently of it.

MM-0: an operative narrative written from nothing. The prompt gives a 26-year-old man with a closed elbow injury after a fall, two radiographs, and the statement that surgery used a sequential approach. It does not name the fracture, the procedure, or the order of steps.

The GLM baseline supplies seven case-level assertions the prompt does not contain, inside a block labelled <imaging_findings>. It states that the radiographs *"show a radial head fracture"*, that the patient *"undergoes open reduction internal fixation (ORIF), which includes radial head resection"*, two procedures that exclude one another, and that the injury was *"associated with a lateral condylar fracture"*, which contradicts its own preceding sentence. It adds a capsular plication, a medial collateral reconstruction with *"autologous semitendinosus tendon grafts"*, and a follow-up, *"resulted in good functional outcome"*. Nothing in the block distinguishes what was read from the images from what was written into them. The Qwen3-VL baseline fails the same way with different content, opening *"Radiographs demonstrate a displaced olecranon fracture dislocation with associated coronal plane instability"* and filling its discussion with generic anatomy under *Background* and *Take Home Message* headings inherited from the training corpus.

MULTI-MULTI on GLM opens with per-image grounding, *"Image 1 shows AP view demonstrating radial head dislocation. Image 2 shows Lateral view demonstrating radial head dislocation"*, then the composition step, *"Comparing the two images, there are no additional or contrasting findings"*. Three claims, each attached to a named image, each checkable against that image on its own. Its justification then reconstructs a surgical sequence it was not given, so the trace is not free of invention. But the invention is confined to the justification component a, it does not include a self-contradictory procedure, a second fracture, or an outcome, and it is no longer presented as an imaging finding. Four invented assertions replace seven, and the three that carry the visual premise are separated out where a clinician can check them.

MM-34: a fabricated premise that excludes the right answer. The prompt gives a 37-year-old woman with acute quadriparesis, sagittal and axial T2 sequences of the cervical and upper thoracic cord, and a normal concurrent brain study. It supplies no history, no post-contrast series, and no follow-up.

The GLM baseline supplies about ten assertions the case does not contain, and two of them are not merely unsupported but impossible. It reports that *"there were no associated abnormalities at the level of the conus medullaris"*, a region outside the imaged field entirely, and that *"the absence*

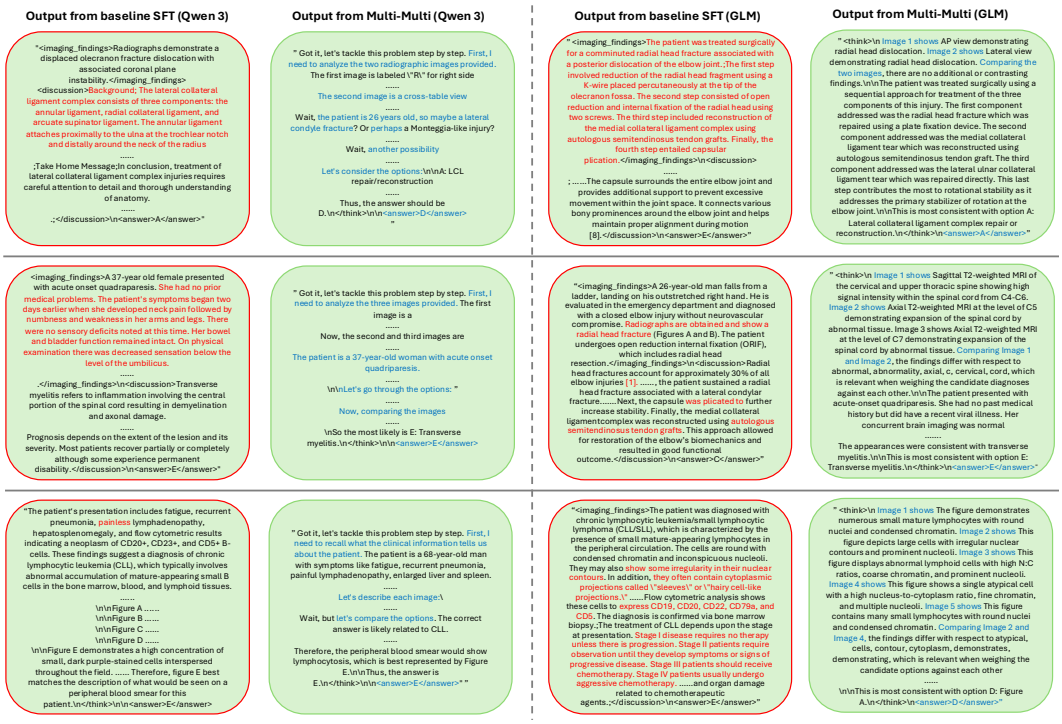


Figure 9: Model outputs on three out-of-domain cases from the MedXpertQA MM test split (MM-0, MM-34 and MM-18, top to bottom), for the SFT baseline and MULTI-MULTI on each of the Qwen3-VL and GLM backbones. Highlighted spans assert content the case does not contain, or contradicts. The baseline places such content inside a block labelled `<imaging_findings>`, where an invented operative sequence, a fabricated viral prodrome, a report of a region outside the imaged field and a morphology recited without reference to any figure are indistinguishable from what was actually read. MULTI-MULTI opens instead with per-image grounding $g_1, \ldots, g_n$ and the composition sentence s_{i^*,j^*}, so every visual claim is attached to a named image and can be checked on its own; invention is reduced but not removed, and what remains sits in the answer justification a. Reference answers are A, C and E. The three cases span the answer outcomes deliberately, because the property shown here varies independently of correctness. Section A.5 analyses each case.

of enhancement excluded tumours and abscesses", although only T2 sequences were provided and enhancement cannot be assessed on them. It invents a *"recent viral illness"*, the classic prodrome favouring an inflammatory myelopathy, and an absent neurological history. It then reaches the reference diagnosis and discards it, *"Spinal cord infarction was also considered but the distribution of the abnormality was too extensive for this possibility"*, and closes with an outcome, *"the patient made a good recovery over several weeks without treatment"*. Both the exclusion criterion and the prodrome are load-bearing: remove them and the stated reasoning no longer supports the answer it gives. This is the clearest case in our sample that hallucination is not a cosmetic property of a trace. The register says where it comes from, past tense throughout, a differential adjudicated retrospectively, a follow-up: the model answers a prospective question by writing the genre its supervision was drawn from, and generates the history, contrast findings and outcome that genre requires.

MULTI-MULTI on GLM reaches the same wrong answer, and its grounding step is where it goes wrong, *"Image 2 shows Axial T2-weighted MRI at the level of C5 demonstrating expansion of the spinal cord by abnormal tissue"*, a description of an infiltrative process rather than the anterior grey matter pattern of infarction. That is the useful part: the error is attributable to a named component and to a named image, and it is stated in a form an expert can contradict. Two invented assertions remain, both in the justification and both also present in the baseline, against roughly ten in the baseline, and the impossible ones are gone. The two traces are also mutually incompatible, one placing the lesion from C4 to T5 and the other from C4 to C6, which answer accuracy records as a single shared error.

MM-18: a purely visual task, attempted and not attempted. The prompt gives a 68-year-old man with a CD20+, CD23+, CD5+ B-cell neoplasm and five candidate smears. Answering requires reading the figures.

The GLM baseline does not read them. Its imaging block recites CLL morphology from memory, asserts that the cells *"often contain cytoplasmic projections called 'sleeves' or 'hairy cell-like projections'"*, a feature of hairy cell leukaemia and not of CLL, omits the smudge cells that actually distinguish a CLL smear, restates the flow panel as CD19, CD20, CD22, CD79a and CD5 while dropping the CD23 the prompt supplies, and never names a single figure. It answers correctly without having attempted the task the question asks.

MULTI-MULTI on GLM describes all five, and two of its descriptions are appropriate for CLL: Image 1 as *"numerous small mature lymphocytes with round nuclei and condensed chromatin"* and Image 5 as *"many small lymphocytes with round nuclei and condensed chromatin"*. It then selects the wrong one of that pair. The composition step explains why: choosing the most lexically distant caption pair puts Images 2 and 4 under comparison, neither of which is a candidate, and the terms it surfaces, *"atypical, cells, contour, cytoplasm, demonstrates, demonstrating"*, are stems and generic nouns rather than a clinical distinction. The comparison the case turns on, Image 1 against Image 5, is never posed. This is the limitation named in Section **??** occurring in practice, and it localises: grounding succeeded, composition failed, and it failed because $\arg\max |\mathrm{diff}(c_i, c_j)|$ optimises lexical distance rather than diagnostic relevance.

Taken together. Across the three cases the baseline traces place invented case facts inside a block labelled as imaging, so no part of them is separable for checking: an operative sequence that never happened, a prodrome that decides a differential, a region outside the field of view, a morphology recited without looking. MULTI-MULTI's traces do not eliminate invention, and we do not claim they do; what they change is where it can occur and whether it can be found. Every visual claim is attached to a named image and stands alone, the count of unsupported assertions falls substantially on both cases where we can compare them directly, the assertions that are impossible rather than merely unsupported disappear, and what invention remains sits in the justification step, legible as inference. On MM-18 and MM-34 this is what allows a wrong answer to be attributed to a specific component rather than to the model as a whole. It is also why decomposition alone is not the method: Section 3.2 introduces ISS precisely because an inspectable structure does not by itself force the content filling it to be image-derived.