BIRD: DISTILLING DECISION BOUNDARIES INTO RATIONALES FOR MLLM ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting general-purpose multimodal large language models (MLLMs) to specialized domains requires learning domain-specific decision criteria, which often hinge on subtle visual distinctions between otherwise plausible answers. Rationale augmentation aims to expose such evidence through additional observations or inter-sample comparisons, yet a visually valid cue is not necessarily decision-relevant: it may describe how samples differ without changing the model's relative preference between competing answers. We therefore introduce **BIRD**, a self-improving **B**oundary-**I**nformed **R**ationale **D**istillation framework that uses model-specific confusions to locate unresolved local decision boundaries and distills the evidence that resolves these confusions into rationales. For each sample, BIRD retrieves candidate neighbors from the target MLLM's own representation space and selects the most confusable one according to its answer preferences. It then generates answer-blind candidate evidence from their visual differences and functionally verifies which evidence most effectively strengthens the model's preference for the correct answer while avoiding inappropriate transfer across the pair. The verified evidence is then distilled into a single-sample rationale for standard supervised fine-tuning. Experiments on medical and chart VQA show that BIRD outperforms competing rationale-augmentation methods across two target MLLMs, while further analyses demonstrate clearer separation of confusable answers and stronger gains from model-matched supervision. Anonymous code: link.

1 INTRODUCTION

General-purpose multimodal large language models (MLLMs) possess broad visual-language capabilities (Li et al., 2023b; Liu et al., 2023; Dai et al., 2023), but specialized domains often hinge on distinctions that are subtle, domain-dependent, and decisive for the final prediction (Li et al., 2023a; Tu et al., 2024; Masry et al., 2024). In medical image analysis, for example, visually similar findings may indicate different diagnoses because of subtle differences in morphology or spatial distribution; in scientific chart understanding, a small trend reversal or relative change may alter the correct conclusion. At their core, specialized domains differ in the criteria that determine which visual distinctions are decision-defining. Rationales provide a natural way to express such criteria by making decision-relevant visual evidence explicit, thereby offering richer supervision than answer labels alone (Park et al., 2018; Kayser et al., 2021; Sammani et al., 2022).

Recent methods have therefore sought to improve domain-specialized MLLMs by augmenting rationale supervision (Carbune et al., 2024; Zhu et al., 2024; Zhang et al., 2025; Wang et al., 2025), where additional evidence is derived either from individual samples or from relations between samples (Figure 1(a)). The latter is especially useful when visually similar samples lead to different answers, as their comparison can expose subtle distinctions that are difficult to identify from either sample alone, motivating recent inter-sample approaches (Zou & Yin, 2025; Xiong et al., 2026). However, a visible difference is not necessarily relevant to the model's decision—a cue may accurately distinguish two samples while leaving their relative answer preferences unchanged. Such a cue adds descriptive detail to the rationale without providing evidence that resolves the model's "confusion". Effective rationale augmentation should therefore identify the evidence that changes the competition between plausible answers, increasing the model's preference for the correct answer while suppressing its confusable alternatives. Such "confusion" reveals where the model's local decision boundaries remain unresolved.

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.

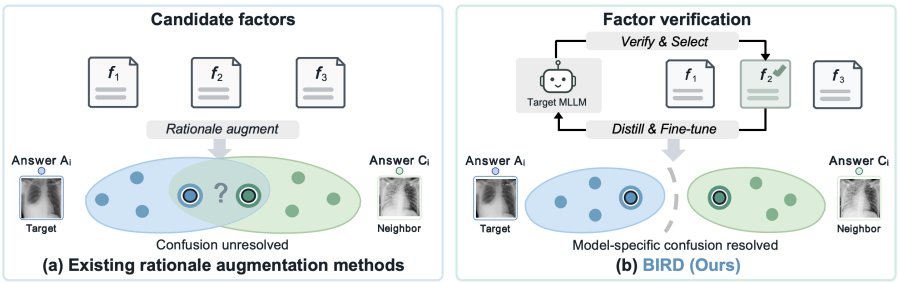


Figure 1: Existing rationale augmentation methods leave confusable answers unresolved. BIRD verifies candidate factors with the target MLLM and distills the selected factor to resolve model-specific confusion.

Equally importantly, these unresolved decision boundaries are model-specific. Different MLLMs organize samples into different visual neighborhoods, confuse different alternatives, and rely on different cues to resolve those confusions. A pair that lies near an unresolved boundary for one model may be readily distinguished by another; even for the same pair, the evidence needed to resolve the confusion may differ across models. Existing model-agnostic rationale augmentation methods (Chen et al., 2024; Xiong et al., 2025; Wu et al., 2025) cannot consistently target the specific weaknesses of a given MLLM. This motivates a self-improving formulation in which the target MLLM identifies its own confusions, locates the corresponding unresolved decision boundaries, and converts the evidence needed to resolve them into rationale supervision (Figure 1(b)).

Building on this perspective, we propose **BIRD**, a self-improving **B**oundary-**I**nformed **R**ationale **D**istillation framework that converts model-specific confusions into rationale supervision. For each training sample, BIRD retrieves semantically compatible samples with different answers from the target model's visual space and uses its answer preferences to identify the most confusable neighbor. The model then performs an answer-blind comparison to generate question-relevant visual differences as candidate evidence. Rather than treating these differences as equally useful, BIRD evaluates how each candidate changes the model's preference across the pair and selects the one that most effectively resolves the confusion while preserving the distinction between the two sides. The verified evidence is then distilled into a single-sample rationale, allowing the target MLLM to turn its own unresolved decision boundaries into supervision for adaptation. Experiments on medical and chart VQA show that BIRD achieves the strongest average performance across two target MLLMs, outperforming competing rationale-augmentation methods on in-domain benchmarks. Further analyses reveal clearer separation between confusable answers, boundary-specific effects of the selected factors, and stronger gains from model-matched supervision. Our contributions are threefold:

- *Conceptualizing rationale augmentation as model-specific boundary resolution.* We recast rationale augmentation from enriching sample descriptions to resolving the target MLLM's unresolved decision boundaries, turning rationale construction into model-conditioned evidence selection based on whether evidence resolves the model's own confusions.
- *Self-improving distillation of model-specific boundary evidence.* We introduce BIRD, which discovers model-specific confusable neighbors, evaluates answer-blind candidate evidence by own-side gain and cross-boundary transfer, and distills only verified evidence into single-sample rationales for standard supervised fine-tuning.
- *Demonstrating the advantage of model-matched boundary supervision.* Across two target MLLMs and medical and chart VQA, BIRD achieves the best overall average on most model–benchmark combinations, with analyses showing sharper separation of confusable answers, boundary-specific evidence effects, and stronger gains from model-matched supervision.

2 Related Work

Self-Improving Multimodal Reasoning. Building on STaR's perspective (Zelikman et al., 2022), recent work has explored improving multimodal reasoning by turning model-generated solutions into supervision. R^3V iteratively bootstraps positive and negative multimodal reasoning traces and

learns to refine or select rationales through self-reflection (Cheng et al., 2025). M-STAR systematically studies self-evolving multimodal training through the choice of training objective, reward model, and prompt variation, and introduces adaptive balancing to alleviate performance saturation (Liu et al., 2024). VC-STaR exploits visually similar VQA pairs to help VLMs identify visual discrepancies, thereby converting the model's contrastive ability into improved reasoning supervision (Pan et al., 2026). Concurrent work further improves self-training by explicitly verifying perceptual grounding (Sharma et al., 2026). Despite these advances, existing self-improving methods mainly bootstrap better reasoning from model-generated trajectories or feedback, but do not explicitly target the model's own unresolved decision boundaries.

Inter-Sample Supervision. Building on complementary and counterfactual VQA, inter-sample supervision has been used to explicitly model the relationship among factual, original, and counterfactual samples (Liang et al., 2020). Hard-negative mining further exploits confusable instances to improve cross-modal discrimination in medical VQA, while counterfactual image substitution has been used to measure and strengthen visual reliance (Zafar et al., 2026). More recently, counterfactual-style retrieval has been used to select causally informative demonstrations for visual in-context learning (Xiong et al., 2026). However, these methods don't test whether a difference actually changes the target model's preference between competing answers.

3 Pilot Study: Probing Model-Specific Boundaries

Before introducing BIRD, we conduct a lightweight pilot study with Qwen3.5-9B (Team, 2026) and InternVL3-8B (Zhu et al., 2025) to examine two questions: whether local confusion is specific to the target MLLM, and whether visual differences between similar samples necessarily provide useful boundary evidence.

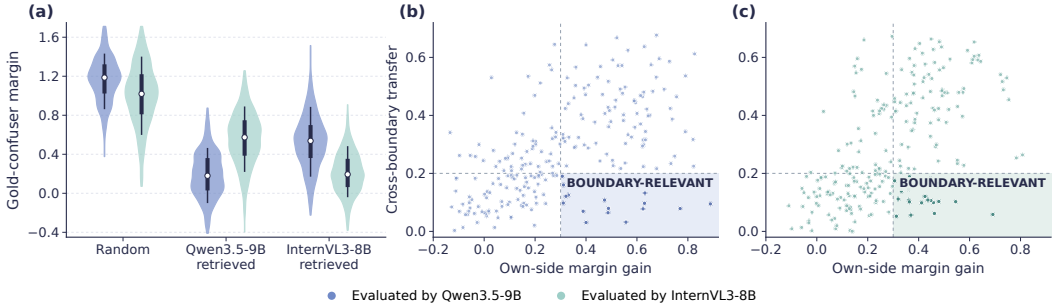


Figure 2: Pilot study of model-specific boundaries. (a) Model-retrieved neighbors yield lower gold–confuser margins than random pairs, with each model most confused by neighbors retrieved from its own representation space. (b,c) Candidate factors exhibit heterogeneous decision effects, only a small subset falls in the shaded boundary-relevant region.

Hard near neighbors are model-specific. For each anchor sample i, we retrieve its nearest different-answer neighbor using the target MLLM's own visual representation, and compare it with a random different-answer sample from the same compatible pool. Pair difficulty is measured by the gold–confuser margin:

$$m_M(i,j) = \ell_M(A_i \mid I_i, Q_i) - \ell_M(A_j \mid I_i, Q_i), \tag{1}$$

where ℓ_M is the length-normalized answer log-likelihood, and a smaller margin indicates stronger confusion. We repeat the retrieval independently with two MLLMs and evaluate all pairs under both models. As shown in Figure 2(a), nearest different-answer neighbors are consistently harder than random compatible pairs. Moreover, each model is most strongly confused by neighbors retrieved from its own representation space, suggesting that unresolved local boundaries are model-specific rather than fixed properties of the dataset.

Pairwise differences are not necessarily boundary evidence. For each hard pair, we hide the answers and ask the corresponding target MLLM to propose question-relevant visual differences. For each candidate factor e, we measure its *own-side gain*, i.e., the increase in the correct-versus-confuser margin on the image where it was observed, and its *cross-boundary transfer*, i.e., whether

the same evidence also pushes the opposite-side image toward the anchor's answer:

$$\Delta_{\text{own}} = m_M(I_i, e; A_i, A_j, Q_i) - m_M(I_i, \varnothing; A_i, A_j, Q_i), \tag{2}$$

$$\Delta_{\text{cross}} = \Big[m_M(I_j, e; A_i, A_j, Q_i) - m_M(I_j, \varnothing; A_i, A_j, Q_i) \Big]_+ . \tag{3}$$

We conduct this analysis separately for the hard pairs discovered by each target MLLM. Figures 2(b) and (c) show the candidate evidence distributions for Qwen3.5-9B and InternVL3-8B, respectively. Both models exhibit substantial, yet model-specific, heterogeneity: many visible differences have little effect on the decision, while others increase the desired margin but transfer similarly across the paired sample. For each model, only a subset exhibits the desired behavior of high own-side gain and low cross-boundary transfer.

Implication. These observations motivate two design principles: The target MLLM should discover its own confusable neighbors, and pairwise differences should be treated only as candidate evidence until their boundary relevance is functionally verified. BIRD follows these principles by discovering model-specific confusers, verifying discriminative factors, and distilling the validated evidence into sample-specific rationales.

4 BIRD: Boundary-Informed Rationale Distillation

4.1 Overview

The pipeline of BIRD is shown in Figure 3. Given a domain-specific training set

$$\mathcal{D}_0 = \{(I_i, Q_i, R_i^0, A_i)\}_{i=1}^N \tag{4}$$

and a target MLLM M_θ, BIRD improves the training rationales while keeping the images, questions, answers, and sample size unchanged. For each sample, BIRD retrieves model-specific neighbors, identifies the alternative most confusable to the target model, proposes and verifies boundary evidence, and distills the verified evidence into the rationale:

$$\mathcal{D}^* = \{(I_i, Q_i, R_i^*, A_i)\}_{i=1}^N. \tag{5}$$

The target MLLM's visual representations define the candidate neighborhoods, its answer preferences identify confusable alternatives and score candidate factors, and the selected factor is distilled into supervision for improving the same model.

4.2 Model-Specific Neighbor Retrieval

Our pilot study shows difficult local neighbors are model-specific. We therefore retrieve them using the target MLLM's own visual representation. Let $h_{i,t}^{\text{vis}}$ denote the projected visual tokens of I_i:

$$v_i = \text{Norm}\left(\frac{1}{T_i} \sum_{t=1}^{T_i} h_{i,t}^{\text{vis}} \right). \tag{6}$$

For each sample, we first restrict retrieval to semantically relevant questions using question templates when available or question similarity otherwise, and then retrieve the top-P samples by $\cos(v_i, v_j)$, while using GPT-5 to check whether these samples are performing the same task. Thus, the target MLLM itself defines the local neighborhoods in which BIRD searches for unresolved confusions.

We further remove QA pairs that cannot form a meaningful comparison with the target question. When question formulations differ, a neighbor is retained only if its known QA state can be unambiguously expressed as an alternative answer to the target question; otherwise it is discarded.

4.3 Confuser Selection

Among the retrieved neighbors, we follow the energy-based view (Liu et al., 2026) and identify the alternative that is most confusable to the target model by quantifying the model's preference,

$$E_\theta(a \mid I, Q) = -\frac{1}{|a|} \sum_{t=1}^{|a|} \log p_\theta(a_t \mid I, Q, a_{<t}), \tag{7}$$

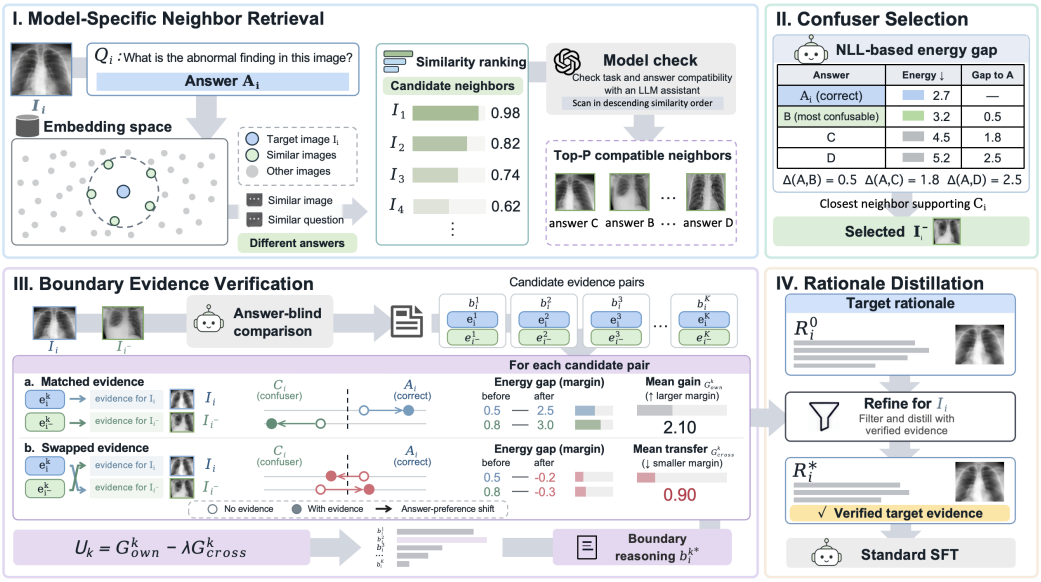


Figure 3: Overview of BIRD. The target MLLM retrieves model-specific neighbors, selects its most confusable alternative, verifies candidate boundary evidence with energy-based gain and cross-boundary transfer, and distills the validated evidence into a refined rationale for standard SFT.

where lower energy indicates stronger model preference.

Let $\mathcal{A}_i^-$ be the alternative answers provided by the retrieved neighbors in the answer space of Q_i. We select

$$C_i = \arg\min_{c \in \mathcal{A}_i^-} E_\theta(c \mid I_i, Q_i). \tag{8}$$

Hence, the confuser is determined directly by the target MLLM's current energy landscape. Among neighbors corresponding to C_i, we choose the visually closest one as the witness image I_i^-. The resulting pair (I_i, I_i^-) probes a local boundary that is difficult for the current model.

4.4 Boundary Evidence Verification

A confusable pair only suggests where a boundary may lie, and not every visible difference is decision-relevant. We therefore hide both answers and ask the target MLLM to propose K question-relevant visual factors. Finally, we get:

$$b_i^k = (e_i^k, e_{i^-}^k), \qquad k = 1, \ldots, K, \tag{9}$$

where e_i^k and $e_{i^-}^k$ describe corresponding observations on the two images. K is set to 3 by default. Candidates that explicitly reveal an answer are discarded.

We verify each factor through its effect on the target model's energy gap:

$$g_\theta(I, e; A, C, Q) = E_\theta(C \mid I, Q, e) - E_\theta(A \mid I, Q, e), \tag{10}$$

where a larger value indicates stronger preference for A over C.

Its own-side gain is

$$\Delta_t^k = g_\theta(I_i, e_i^k; A_i, C_i, Q_i) - g_\theta(I_i, \varnothing; A_i, C_i, Q_i), \tag{11}$$

$$\Delta_n^k = g_\theta(I_i^-, e_{i^-}^k; C_i, A_i, Q_i) - g_\theta(I_i^-, \varnothing; C_i, A_i, Q_i), \tag{12}$$

with

$$G_{\text{own}}^k = \frac{\Delta_t^k + \Delta_n^k}{2}. \tag{13}$$

We then swap the evidence across the pair to measure whether it transfers to the wrong side:

$$L^k_{t\to n} = \Big[g_\theta(I_i^-, e_i^k; A_i, C_i, Q_i) - g_\theta(I_i^-, \varnothing; A_i, C_i, Q_i)\Big]_+, \tag{14}$$

$$L^k_{n\to t} = \Big[g_\theta(I_i, e_{i^-}^k; C_i, A_i, Q_i) - g_\theta(I_i, \varnothing; C_i, A_i, Q_i)\Big]_+, \tag{15}$$

and

$$G^k_{\text{cross}} = \frac{L^k_{t\to n} + L^k_{n\to t}}{2}. \tag{16}$$

We define the *boundary utility* as

$$U_k = G^k_{\text{own}} - \lambda G^k_{\text{cross}}. \tag{17}$$

where λ is the cross-boundary penalty. A high-utility factor strengthens the appropriate decision on its own side while avoiding the same effect across the boundary. We select $k^* = \arg\max_k U_k$. Because both proposal and verification are performed by the target MLLM, the selected evidence specifically addresses its current decision weakness.

4.5 Rationale Distillation

Finally, BIRD converts the verified evidence into supervision for improving the same target model. If the best factor has sufficient boundary utility, we refine the original rationale as

$$R_i^* = \text{Refine}\left(R_i^0; I_i, Q_i, A_i, C_i, e_i^{k^*}, e_{i^-}^{k^*}\right). \tag{18}$$

The refined rationale preserves valid information in R_i^0, incorporates the verified target-side evidence, and introduces no unsupported visual observations. Neighbor evidence is used only as contrastive context.

If no factor reaches the utility threshold τ_U, we retain $R_i^* = R_i^0$. We then perform standard rationale-supervised fine-tuning on $\mathcal{D}^*$. No additional contrastive objective, preference optimization, or reinforcement learning is required, and no retrieval or paired image is used at training or inference.

5 Experiments and Analysis

5.1 Experimental Setup

Models and benchmarks. We evaluate BIRD with two target MLLMs, Qwen3.5-9B and InternVL3-8B, under two independent domain-adaptation settings. For chart and plot reasoning, we adapt each model on the ChartQA-X (Hegde et al., 2025) training set and evaluate on ChartQA (Masry et al., 2022), ChartBench (Xu et al., 2023), and ChartQAPro (Masry et al., 2025), which cover diverse chart-understanding and numerical reasoning tasks. For medical reasoning, we use OpenMedReason (Baghbanzadeh et al., 2026) for adaptation and evaluate on SLAKE (Liu et al., 2021), PathVQA (He et al., 2021), and MedXpertQA (Zuo et al., 2025), spanning general medical VQA, pathology-focused reasoning, and challenging medical visual question answering.

Baselines. We compare against the untuned base model and Original SFT, which directly fine-tunes on the original domain data without rationale enhancement. We further include representative rationale-enhancement methods: Reflective Instruction Tuning (Reflective IT) (Zhang et al., 2024), which augments rationale supervision without self-improvement; STaR, R^{3}V, and Self-Rationale Calibration (SRC) (Wu et al., 2025), which improve supervision using the target model's own reasoning or responses; and VC-STaR, which additionally exploits inter-sample visual contrast. These methods provide comparisons across both self-improving and non-self-improving settings, with and without inter-sample supervision.

Training protocol. All methods use the same source data and target MLLM within each model–domain setting. We train for 10 epochs on 8 NVIDIA A800 GPUs with AdamW (lr $= 2 \times 10^{-5}$, global batch size $= 128$), using a cosine learning-rate schedule with a 3% warmup ratio. For BIRD, we set the cross-boundary penalty to $\lambda = 2.0$ and the utility threshold to $\tau_U = 0.2$, so that only candidates with boundary utility above 0.2 are distilled into the training rationales. K and P are both set to 3 by default. Full details are provided in the Appendix.

Table 1: Main results across two target MLLMs and two domain-adaptation settings. Higher is better. "Self" indicates self-improvement, while "Inter" denotes inter-sample supervision. Arrows indicate absolute changes relative to the corresponding base model without SFT.

			Chart / Plot VQA			Medical VQA			
Method	**Self**	**Inter**	**ChartQA**	**ChartBench**	**ChartQAPro**	**SLAKE**	**PathVQA**	**MedXpertQA**	**Avg.**
Qwen3.5-9B									
Base (no SFT)			85.67	67.42	41.35	73.23	47.12	27.18	57.00
Original SFT			88.36 ↑2.69	68.20 ↑0.78	49.02 ↑7.67	75.49 ↑2.26	49.53 ↑2.41	27.84 ↑0.66	59.74 ↑2.74
Reflective IT			88.48 ↑2.81	68.17 ↑0.75	50.33 ↑8.98	75.81 ↑2.58	49.35 ↑2.23	27.92 ↑0.74	60.01 ↑3.01
STaR	✓		87.64 ↑1.97	67.99 ↑0.57	49.26 ↑7.91	75.19 ↑1.96	49.40 ↑2.28	27.83 ↑0.65	59.55 ↑2.55
R^3V	✓		88.26 ↑2.59	68.12 ↑0.70	47.39 ↑6.04	75.25 ↑2.02	49.78 ↑2.66	26.98 ↓0.20	59.30 ↑2.30
SRC	✓		83.62 ↓2.05	67.15 ↓0.27	46.67 ↑5.32	74.91 ↑1.68	48.66 ↑1.54	26.65 ↓0.53	57.94 ↑0.94
VC-STaR	✓	✓	89.45 ↑3.78	69.20 ↑1.78	51.38 ↑10.03	75.96 ↑2.73	50.29 ↑3.17	28.10 ↑0.92	60.73 ↑3.73
BIRD	✓	✓	89.84 ↑4.17	71.55 ↑4.13	52.96 ↑11.61	77.92 ↑4.69	51.49 ↑4.37	28.06 ↑0.88	61.97 ↑4.97
InternVL3-8B									
Base (no SFT)			82.72	65.07	37.78	72.83	48.64	22.39	54.91
Original SFT			85.31 ↑2.59	66.55 ↑1.48	47.26 ↑9.48	74.13 ↑1.30	50.08 ↑1.44	23.43 ↑1.04	57.79 ↑2.88
Reflective IT			85.44 ↑2.72	66.71 ↑1.64	48.11 ↑10.33	74.56 ↑1.73	50.21 ↑1.57	23.51 ↑1.12	58.09 ↑3.18
STaR	✓		84.92 ↑2.20	66.31 ↑1.24	47.58 ↑9.80	74.02 ↑1.19	50.03 ↑1.39	23.36 ↑0.97	57.70 ↑2.79
R^3V	✓		85.18 ↑2.46	66.44 ↑1.37	45.91 ↑8.13	74.11 ↑1.28	50.34 ↑1.70	22.81 ↑0.42	57.47 ↑2.56
SRC	✓		81.47 ↓1.25	65.12 ↑0.05	44.73 ↑6.95	73.62 ↑0.79	49.21 ↑0.57	22.46 ↑0.07	56.10 ↑1.19
VC-STaR	✓	✓	86.38 ↑3.66	67.53 ↑2.46	49.61 ↑11.83	74.81 ↑1.98	50.86 ↑2.22	23.74 ↑1.35	58.82 ↑3.91
BIRD	✓	✓	87.06 ↑4.34	69.42 ↑4.35	51.28 ↑13.50	76.31 ↑3.48	51.97 ↑3.33	23.69 ↑1.30	59.96 ↑5.05

5.2 Overall Domain Adaptation Performance

Table 1 reports the adaptation performance across the two target MLLMs and domains. BIRD achieves the highest overall averages, improving the base models by 4.97 and 5.05 points for both models, respectively. It further surpasses the strongest competing rationale-augmentation method, VC-STaR. These gains suggest that verified boundary-specific evidence provides more effective adaptation supervision.

5.3 Decision Boundary Sharpening

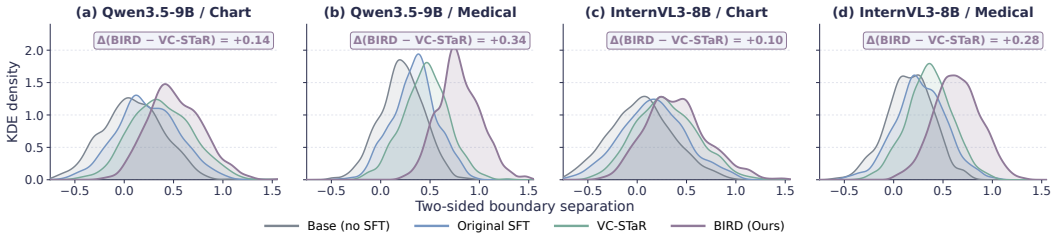


Figure 4: Held-out hard-pair distributions of two-sided boundary separation for base, Original SFT, VC-STaR, and BIRD. A consistent rightward shift for BIRD indicates that it more effectively separates the gold answer from the model's original confuser.

We further examine whether BIRD sharpens the decision boundaries targeted during adaptation. For each target MLLM, we construct fixed hard pairs from the held-out benchmark splits using Kimi K3 (Team et al., 2026). For each anchor, we retrieve question-compatible samples within the same benchmark, identify its lowest-energy alternative as the confuser, and select the closest valid neighbor supporting that alternative. We sample 50 hard pairs per benchmark, yielding 150 pairs for each model–domain setting, and use the same pairs to evaluate all methods.

For a pair (I_i, I_i^-) with competing answers (A_i, C_i), we measure two-sided boundary separation as:

$$S_i = \frac{1}{2}\big[E(C_i \mid I_i, Q_i) - E(A_i \mid I_i, Q_i) + E(A_i \mid I_i^-, Q_i) - E(C_i \mid I_i^-, Q_i)\big]. \tag{19}$$

Larger S_i indicates a clearer boundary. Figure 4 compares the distributions of S_i for the base model, Original SFT, the strongest prior baseline, and BIRD across both models and domains. A consistent rightward shift for BIRD indicates that it more effectively separates the gold answer from the model's original confuser.

5.4 Model-Specific Self-Improvement

BIRD is self-improving in that each target MLLM discovers and resolves its own decision ambiguities. We examine whether the resulting supervision is indeed model-specific from both the evidence and downstream perspectives.

Cross-model evidence utility. As shown in Figure 5, on the same held-out anchors, Qwen3.5-9B and InternVL3-8B independently run BIRD to discover their boundary pairs and selected evidence. For anchors where both models obtain valid evidence, we re-evaluate the factor discovered by each model using both target MLLMs, while keeping its discovered pair fixed. We then compare the mean boundary utility across evidence sources. If the supervision is model-specific, evidence discovered by a model should yield higher utility when evaluated by that same model.

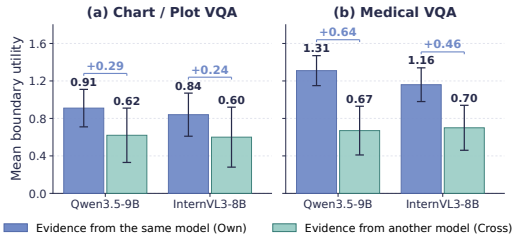


Figure 5: Cross-model boundary utility. Evidence discovered by each target MLLM achieves higher utility when evaluated by the same model, demonstrating model-specific boundary supervision.

Cross-model supervision transfer. We further test whether this specificity translates into downstream adaptation. As shown in Table 2, for each domain, we use the shared training datasets. Each target MLLM is fine-tuned using rationales produced either by itself or by the other model. Better performance with self-generated BIRD supervision would provide direct evidence that BIRD converts model-specific weaknesses into supervision that is particularly useful for improving the model itself.

Table 2: Cross-model transfer of BIRD supervision. "Own" denotes rationales constructed by the target MLLM itself, while "Cross" uses rationales constructed by the other MLLM.

		Chart / Plot VQA			Medical VQA			
Target Model	**BIRD Source**	**ChartQA**	**ChartBench**	**ChartQAPro**	**SLAKE**	**PathVQA**	**MedXpertQA**	**Avg.**
Qwen3.5-9B	**Qwen3.5-9B (Own)**	**89.84**	**71.55**	**52.96**	**77.92**	**51.49**	**28.06**	**61.97**
	InternVL3-8B (Cross)	88.05 ↓1.79	69.73 ↓1.82	51.43 ↓1.53	76.42 ↓1.50	49.57 ↓1.92	26.67 ↓1.39	60.31 ↓1.66
InternVL3-8B	**InternVL3-8B (Own)**	**87.06**	**69.42**	**51.28**	**76.31**	**51.97**	**23.69**	**59.96**
	Qwen3.5-9B (Cross)	85.45 ↓1.61	67.88 ↓1.54	49.14 ↓2.14	74.56 ↓1.75	49.98 ↓1.99	22.34 ↓1.35	58.23 ↓1.73

5.5 Boundary Evidence Analysis

Functional behavior of selected evidence. We examine whether BIRD selects factors with the intended boundary behavior. Figure 6 plots all candidate factors in the $(G_{\text{own}}, G_{\text{cross}})$ space for both target MLLMs and domains, with BIRD-selected factors highlighted. Across all four settings, the selected evidence concentrates in the high-own-side-gain and low-cross-boundary-transfer region, indicating that boundary utility filters generic pairwise differences into boundary-specific evidence.

Effect of boundary-utility selection. We further compare random candidate selection, selecting the factor with the largest G_{own}, and selecting by the full boundary utility U, while keeping the candidate pool and number of updated rationales fixed. As shown in Table 3, selection by U consistently performs best, showing that maximizing own-side gain alone is insufficient; controlling cross-boundary transfer provides additional benefit.

5.6 Ablation Study

We ablate two components as shown in Table 4.

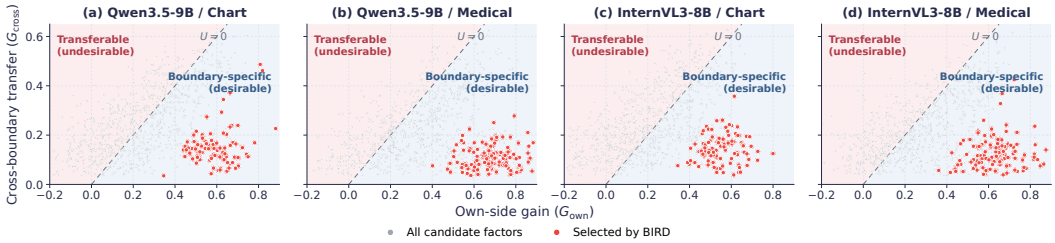


Figure 6: Functional distribution of candidate evidence across models and domains. BIRD-selected factors concentrate in the high-own-side-gain, low-cross-boundary-transfer region.

Table 3: Effect of different boundary-evidence selection strategies. Changes are relative to random selection. All variants use the same candidate pool and update the same number of rationales.

	Chart / Plot VQA			Medical VQA			
Selection	**ChartQA**	**ChartBench**	**ChartQAPro**	**SLAKE**	**PathVQA**	**MedXpertQA**	**Avg.**
Qwen3.5-9B							
Random	88.71	68.82	50.21	76.12	50.04	27.74	60.27
Max G_{own}	89.31 ↑0.60	69.94 ↑1.12	51.64 ↑1.43	76.91 ↑0.79	50.72 ↑0.68	27.95 ↑0.21	61.08 ↑0.81
Max U	**89.84** ↑1.13	**71.55** ↑2.73	**52.96** ↑2.75	**77.92** ↑1.80	**51.49** ↑1.45	**28.06** ↑0.32	**61.97** ↑1.70
InternVL3-8B							
Random	85.73	66.92	48.52	74.62	50.41	23.24	58.24
Max G_{own}	86.34 ↑0.61	68.06 ↑1.14	49.97 ↑1.45	75.38 ↑0.76	51.05 ↑0.64	23.47 ↑0.23	59.04 ↑0.80
Max U	**87.06** ↑1.33	**69.42** ↑2.50	**51.28** ↑2.76	**76.31** ↑1.69	**51.97** ↑1.56	**23.69** ↑0.45	**59.96** ↑1.72

Effect of Confuser Selection. We remove explicit confuser selection by directly using the alternative associated with the nearest valid neighbor, rather than selecting the lowest-energy alternative of the target MLLM. As shown in Table 4, this variant consistently degrades performance, indicating that confuser selection better targets the model's unresolved ambiguity.

Effect of Rationale Distillation. We remove rationale distillation by appending the verified target evidence to the original rationale without target-conditioned refinement. The resulting performance drop in Table 4 demonstrates that rationale distillation more effectively converts pairwise boundary evidence into sample-specific supervision.

Table 4: Ablation of confuser selection and rationale distillation. Arrows indicate absolute performance drops relative to full BIRD.

	Chart / Plot VQA			Medical VQA			
Variant	**ChartQA**	**ChartBench**	**ChartQAPro**	**SLAKE**	**PathVQA**	**MedXpertQA**	**Avg.**
Qwen3.5-9B							
w/o Confuser Selection	89.50 ↓0.34	69.45 ↓2.10	51.55 ↓1.41	76.18 ↓1.74	50.50 ↓0.99	27.92 ↓0.14	60.88 ↓1.09
w/o Rationale Distillation	89.66 ↓0.18	70.88 ↓0.67	52.43 ↓0.53	77.31 ↓0.61	51.12 ↓0.37	27.97 ↓0.09	61.59 ↓0.38
BIRD	**89.84**	**71.55**	**52.96**	**77.92**	**51.49**	**28.06**	**61.97**
InternVL3-8B							
w/o Confuser Selection	86.45 ↓0.61	67.80 ↓1.62	49.88 ↓1.40	75.02 ↓1.29	51.02 ↓0.95	23.05 ↓0.64	58.99 ↓0.97
w/o Rationale Distillation	86.78 ↓0.28	68.85 ↓0.57	50.82 ↓0.46	75.82 ↓0.49	51.55 ↓0.42	23.60 ↓0.09	59.60 ↓0.36
BIRD	**87.06**	**69.42**	**51.28**	**76.31**	**51.97**	**23.69**	**59.96**

6 Conclusion

We present BIRD, a self-improving framework that turns the target MLLM's own confusions into rationale supervision. BIRD retrieves model-specific candidate neighbors, selects the alternative the model finds most confusable, and generates answer-blind candidate factors from the resulting pair. It retains factors that improve answer separation on their corresponding samples without transferring across the pair, and distills the verified target-side evidence into a single-sample rationale. Experiments on medical and chart VQA show that BIRD achieves the strongest average performance across both target MLLMs. Further analyses demonstrate greater separation on held-out hard pairs and stronger gains from model-matched supervision. Overall, our findings establish decision boundaries as a useful source of self-discovered supervision and offer a promising direction for adapting MLLMs to specialized domains.

AI USE STATEMENT

AI-assisted tools were used to improve the language and clarity of portions of this manuscript, generate synthetic datasets, and implement parts of the code under the direct guidance and supervision of the authors. All AI-assisted code was subsequently reviewed and validated by two authors, who examined its implementation logic and verified its correctness and consistency with the intended methodology.

ETHICS STATEMENT

This work uses publicly released research datasets and introduces no new human-subject data collection. Because our experiments include medical VQA, the resulting models and rationales should not be interpreted as clinically reliable or used for diagnosis or treatment. Model-generated rationales may inherit biases or contain unsupported statements, and therefore require appropriate human oversight in high-stakes applications.

REPRODUCIBILITY STATEMENT

We provide dataset and training details, hyperparameters, the complete BIRD algorithm, and all prompt templates in the Appendix. Evaluation protocols and construction statistics are also reported. The implementation and configurations required to reproduce our experiments are available in the anonymous code repository linked in the paper.

REFERENCES

Negin Baghbanzadeh, Pritam Sarkar, Michael Colacci, Abeer Badawi, Adibvafa Fallahpour, Arash Afkanpour, Leonid Sigal, Ali Etemad, and Elham Dolatabadi. Openmedreason: Scientific reasoning supervision for medical vision-language models. *arXiv preprint arXiv:2606.12169*, 2026.

Victor Carbune, Hassan Mansoor, Fangyu Liu, Rahul Aralikatte, Gilles Baechler, Jindong Chen, and Abhanshu Sharma. Chart-based reasoning: Transferring capabilities from LLMs to VLMs. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 989–1004, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.62. URL https://aclanthology.org/2024.findings-naacl.62/.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4156–4172, 2024.

Kanzhi Cheng, Li YanTao, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. Vision-language models can self-improve reasoning via reflection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8876–8892, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 708–718, 2021.

Shamanthak Hegde, Pooyan Fazli, and Hasti Seifi. Chartqa-x: Generating explanations for visual chart reasoning, 2025. URL https://arxiv.org/abs/2504.13275.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-ViL: A dataset and benchmark for natural language explanations in vision-language tasks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1244–1254. IEEE, 2021.

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. *Advances in neural information processing systems*, 36: 28541–28564, 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 *International conference on machine learning*, pp. 19730–19742. PmLR, 2023b.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 3285–3292, 2020.

Anglin Liu, Ruichao Chen, Yi Lu, Hongxia Xu, and Jintai Chen. Med-scout: Curing mllms' geometric blindness in medical perception via geometry-aware rl post-training. *arXiv preprint arXiv:2601.23220*, 2026.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Wei Liu, Junlong Li, Xiwen Zhang, Fan Zhou, Yu Cheng, and Junxian He. Diving into self-evolving training for multimodal reasoning. *arXiv preprint arXiv:2412.17451*, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. Chartinstruct: Instruction tuning for chart comprehension and reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10387–10409, 2024.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. ChartQApro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151, 2025.

Zhiyu Pan, Yizheng Wu, Jiashen Hua, Junyi Feng, Shaotian Yan, Bing Deng, Zhiguo Cao, and Jieping Ye. Through the lens of contrast: Self-improving visual reasoning in VLMs. *arXiv preprint arXiv:2603.02556*, 2026.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8779–8788. IEEE, 2018.

Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. NLX-GPT: A model for natural language explanations in vision and vision-language tasks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8322–8332. IEEE, 2022.

Sourabh Sharma, Sonam Gupta, et al. Improving reasoning in vision-language models via perception verified self-training. *arXiv preprint arXiv:2606.22158*, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Yuan Wang, Jiaxiang Liu, Shujian Gao, Bin Feng, Zhihang Tang, Xiaotang Gai, Jian Wu, and Zuozhu Liu. V2T-CoT: From Vision to Text Chain-of-Thought for Medical Reasoning and Diagnosis . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15964. Springer Nature Switzerland, September 2025.

Yuanchen Wu, Ke Yan, Shouhong Ding, Ziyin Zhou, and Xiaoqiang Li. Towards rationale-answer alignment of LVLMs via self-rationale calibration. *arXiv preprint arXiv:2509.13919*, 2025.

Feng Xiong, Hongling Xu, Yifei Wang, Runxi Cheng, Yong Wang, and Xiangxiang Chu. HS-STAR: Hierarchical sampling for self-taught reasoners via difficulty estimation and budget reallocation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5539–5555, 2025.

Guangzhi Xiong, Sanchit Sinha, Zhenghao He, and Aidong Zhang. Retrieving counterfactuals improves visual in-context learning. *arXiv preprint arXiv:2603.16737*, 2026.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. ChartBench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023.

Anas Zafar, Leema Krishna Murali, Siddhant Bharadwaj, Ashish Vashist, and Jia Wu. Do medical vision language models actually see? A counterfactual grounding framework and hard-negative contrastive training for visually-reliant medical VLMs. *arXiv preprint arXiv:2607.03647*, 2026.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. STAR: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, pp. 196–213. Springer, 2024.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. 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. 1631–1662, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.82. URL https://aclanthology.org/2025.acl-long.82/.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT: A scalable lightweight infrastructure for fine-tuning, 2024. URL https://arxiv.org/abs/2408.05517.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Wang Zhu, Alekh Agarwal, Mandar Joshi, Robin Jia, Jesse Thomason, and Kristina Toutanova. Efficient end-to-end visual document understanding with rationale distillation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8401–8424, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.465. URL https://aclanthology.org/2024.naacl-long.465/.

Yuanhao Zou and Zhaozheng Yin. Alignment, mining and fusion: Representation alignment with hard negative mining and selective knowledge fusion for medical visual question answering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29623–29633. IEEE, 2025.

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.

APPENDIX

A EXPERIMENTAL DETAILS

A.1 DATASETS

We consider two domain-adaptation settings, chart/plot VQA and medical VQA. ChartQA-X and OpenMedReason serve as the source datasets for domain adaptation, while the remaining datasets are used as held-out benchmarks.

Chart / Plot Domain. **ChartQA-X** extends ChartQA with natural-language explanations for chart question–answer pairs. Its training split contains 28,299 examples with model-generated explanations. We use this split as the source data for chart-domain adaptation.

ChartQA contains real-world charts paired with human-authored questions and questions generated from chart summaries. We evaluate on its official test split of 2,500 examples.

ChartBench substantially broadens the visual diversity of chart reasoning, covering nine major chart types and 42 fine-grained categories. Its training corpus contains approximately 66.6K charts and 599.6K QA pairs, while the held-out benchmark contains 2,100 charts and 18,900 QA pairs. Many charts omit explicit data-point annotations, requiring reasoning directly from visual elements such as axes, legends, colors, and graphical marks.

ChartQAPro targets more diverse and challenging real-world chart understanding. It contains 1,341 charts collected from 157 sources and 1,948 questions, spanning conventional charts as well as infographics and dashboards. The questions cover multiple formats, including factoid, multiple-choice, conversational, hypothetical, and unanswerable cases.

Medical Domain. **OpenMedReason** is a medical VQA dataset constructed from figures in open-access biomedical literature. Each example pairs an image with a multiple-choice question, an answer, and an image-grounded reasoning trace. The released training split contains 150,246 examples, which we use as the source data for medical-domain adaptation.

SLAKE is a physician-annotated bilingual medical VQA dataset built from 642 radiology images, including CT, MRI, and X-ray scans. The complete bilingual dataset contains 14,028 QA pairs. Following common English-language evaluation settings, we use its English subset, consisting of 4,919 training, 1,053 validation, and 1,061 test questions.

PathVQA focuses on pathology images and contains approximately 32.8K QA pairs over 4,998 images. It includes both open-ended questions and closed-ended yes/no questions covering visual properties such as location, appearance, shape, and color. The standard split used in our experiments contains 19,755 training, 6,279 validation, and 6,761 test QA pairs.

MedXpertQA is designed to evaluate expert-level medical understanding and reasoning, and we use its multimodal subset. It spans 17 medical specialties and 11 body systems and combines clinical context with one or more medical images. The released multimodal subset contains 5 development examples and 2,000 test questions, which constitute the benchmark used in our experiments.

A.2 FINE-TUNING DETAILS

We perform full-parameter supervised fine-tuning. Unless otherwise specified, the vision encoder, multimodal alignment modules, and language model are all trainable. All methods use the same source training data and target MLLM within each model–domain setting, and the same SFT configuration is used whenever applicable.

Table 5 reports the details. Our training and inference are both built on the ms-swift framework (Zhao et al., 2024). We train each model for 10 epochs on 8 NVIDIA A800 GPUs using AdamW with a learning rate of 2×10^{-5}, weight decay of 0.1, and $(\beta_1, \beta_2) = (0.9, 0.95)$. We use a cosine learning-rate schedule with a 3% warmup ratio and BF16 precision. The per-device batch size is 4 with 4 gradient-accumulation steps, resulting in a global batch size of 128. The maximum sequence length is 1024. Training uses DeepSpeed ZeRO-3 and FlashAttention.

Table 5: Training hyperparameters used for all model–domain settings.

Hyperparameter	**Value**
Fine-tuning type	Full-parameter SFT
Training epochs	10
Optimizer	AdamW
Learning rate	2×10^{-5}
Weight decay	0.1
Adam (β_1, β_2)	$(0.9, 0.95)$
LR schedule	Cosine
Warmup ratio	3%
Precision	BF16
Per-device batch size	4
Gradient accumulation	4
Global batch size	128
Maximum sequence length	1024
Gradient checkpointing	Yes
Distributed training	DeepSpeed ZeRO-3
Attention implementation	FlashAttention
Training hardware	8 NVIDIA A800 GPUs

B BIRD Construction Details

This section provides implementation details omitted from the main text. Algorithm 1 summarizes the complete offline construction procedure, followed by the prompt templates used in BIRD.

B.1 End-to-End Algorithm

We clearly show the pseudo code of BIRD, as shown in Algorithm 1.

B.2 Question-Compatible Neighbor Pool Construction

For datasets with recurring question templates, we group samples by their normalized template before visual retrieval. Otherwise, we use Prompt P1 to verify that two questions concern the same underlying visual property and to map the candidate QA state into the answer space of the target question. Candidates with ambiguous mappings or with the same mapped answer as the target are discarded. Top-P visual retrieval is then performed within this compatible pool using the target MLLM representation defined in the main text.

B.3 Answer-Blind Candidate Evidence Generation

Given the selected target–witness pair, Prompt P2 receives only the two images and the target question; the gold and confuser answers are withheld. The target MLLM produces K paired observations $(e_i^k, e_{i^-}^k)$, each describing the same question-relevant visual factor on the two sides. Candidates that explicitly reveal an answer are removed before scoring. No candidate is treated as valid boundary evidence until it passes the functional verification step.

B.4 Boundary Utility Computation

We compute G_{own}, G_{cross}, and U exactly as defined in Eqs. (11)–(17) of the main text. Answer energies are obtained by teacher-forcing each candidate answer and averaging its token-level negative log-likelihood. For a given pair, the evidence-free energy gaps are cached once and reused across candidate factors. Matched and swapped evaluations use the same input format; only the inserted evidence string is changed. The factor with the largest U is retained, and rationale refinement is performed only if its utility exceeds τ_U.

Algorithm 1 Boundary-Informed Rationale Distillation (BIRD)

Require: Training set $\mathcal{D}_0$; target MLLM M_θ; retrieval size P; number of candidate factors K; cross-boundary penalty λ; utility threshold τ_U
Ensure: Boundary-informed training set $\mathcal{D}^*$

Extract and cache visual representations $\{\mathbf{v}_i\}_{i=1}^N$
$\mathcal{D}^* \leftarrow \emptyset$
for each $(I_i, Q_i, R_i^0, A_i) \in \mathcal{D}_0$ **do**
  $R_i^* \leftarrow R_i^0$
  Construct a question-compatible pool $\mathcal{P}_i$
  Map each valid candidate answer into the answer space of Q_i
  Retrieve the top-P visual neighbors $\mathcal{N}_i \subseteq \mathcal{P}_i$
  if $\mathcal{N}_i \neq \emptyset$ **then**
    Form the alternative-answer set $\mathcal{A}_i^-$ from $\mathcal{N}_i$
    $C_i \leftarrow \arg\min_{c \in \mathcal{A}_i^-} E_\theta(c \mid I_i, Q_i)$
    Select the closest neighbor supporting C_i as the witness image I_i^-
    Use M_θ to generate K answer-blind factor pairs

$$\mathcal{B}_i = \{b_i^k = (e_i^k, e_{i^-}^k)\}_{k=1}^K$$

    if $\mathcal{B}_i \neq \emptyset$ **then**
      for each $b_i^k \in \mathcal{B}_i$ **do**
        Compute $G_{i,\text{own}}^k$ and $G_{i,\text{cross}}^k$
        $U_i^k \leftarrow G_{i,\text{own}}^k - \lambda G_{i,\text{cross}}^k$
      end for
      $k^* \leftarrow \arg\max_{k:\, b_i^k \in \mathcal{B}_i} U_i^k$
      if $U_i^{k^*} \geq \tau_U$ **then**
        $R_i^* \leftarrow \text{Refine}\Big(R_i^0; I_i, Q_i, A_i, C_i, e_i^{k^*}, e_{i^-}^{k^*}\Big)$
      end if
    end if
  end if
  $\mathcal{D}^* \leftarrow \mathcal{D}^* \cup \{(I_i, Q_i, R_i^*, A_i)\}$
end for
return $\mathcal{D}^*$

B.5 Rationale Distillation

For a verified factor, Prompt P3 refines the original rationale using the target-side observation, while the witness-side observation is provided only as contrastive context. The output must remain a rationale for the target image alone: valid content from the original rationale is preserved, whereas unsupported details and explicit references to the neighboring sample or the BIRD construction process are excluded. Samples below the utility threshold retain their original rationales.

B.6 Full Prompt Templates

Text enclosed in braces denotes an instance-specific field.

Prompt P1: Question-Compatible Neighbor Validation

[t]

Task. Determine whether the candidate QA sample can serve as a valid contrastive neighbor for the target question.

Target question:
{target_question}
Target answer:
{target_answer}
Candidate question:
{candidate_question}

Candidate answer:
{candidate_answer}

Criteria.
A candidate is valid only if:

1. both questions concern the same underlying visual property;
2. the candidate QA state can be unambiguously expressed as an answer to the target question; and
3. the mapped answer differs from the target answer.

Do not judge visual similarity.

Output. Return `Decision: VALID` or `Decision: INVALID`. If VALID, additionally return `Mapped answer: <answer>`.

Prompt P2: Answer-Blind Candidate Evidence Generation

Task. Compare the two images under the given question and identify visual differences that could distinguish them.

Question:
{question}

Constraints.

1. Each factor must contain corresponding observations for both images.
2. Include only differences relevant to the given question.
3. Do not use wording that explicitly reveals an answer.

Generate up to {num_factors} distinct factors.

Output. Use the following format:

```
Factor 1
Image A: <observation in the target image>
Image B: <corresponding observation in the witness
image>

Factor 2
Image A: ...
Image B: ...
```

Prompt P3: Boundary-Informed Rationale Distillation

Task. Refine the original rationale using the verified discriminative factor.

Question:
{question}
Correct answer:
{target_answer}
Confusable alternative:
{confuser}
Original rationale:
{original_rationale}
Verified target-side observation:
{target_evidence}

Corresponding neighbor-side observation:
{neighbor_evidence}

Requirements.

1. Preserve correct and useful content from the original rationale.
2. Incorporate the verified target-side observation.
3. Do not mention the neighboring sample.
4. Remove unsupported details from the original rationale and do not introduce new visual observations that are not supported by the target image.

Output. Return only the refined rationale for the target image.

C Additional Analysis

C.1 Per-Benchmark Boundary Sharpening

The main text reports boundary separation aggregated within each model–domain setting. Here, we further break down the analysis by benchmark to examine whether the observed boundary sharpening is consistent across individual evaluation sets.

We use the same fixed hard pairs constructed using Kimi K3 as in Sec. 5.3, with 50 pairs sampled from each benchmark. All methods are evaluated on exactly the same pairs. Figure 7 reports the distribution of the two-sided boundary separation S_i for Base, Original SFT, VC-STaR, and BIRD on each benchmark separately.

Across both target MLLMs, BIRD improves the mean separation in 11 of the 12 model–benchmark combinations and is nearly tied with VC-STaR on InternVL3-8B/ChartQAPro. The improvements are generally larger on the medical benchmarks. These results show that the domain-level boundary sharpening reported in the main text is not driven by a particular benchmark.

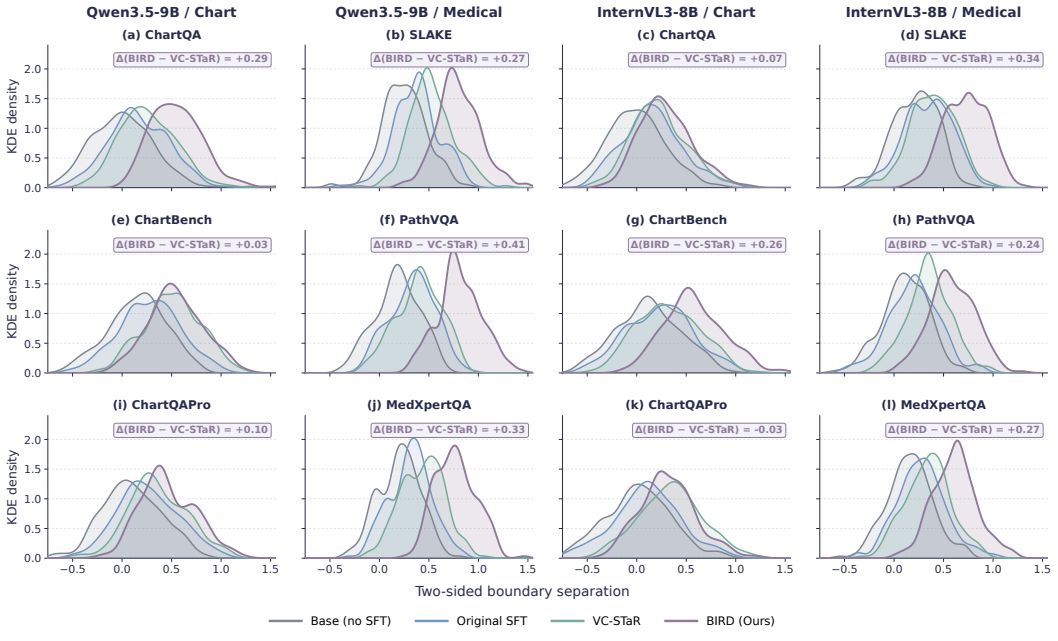


Figure 7: Per-benchmark analysis of decision-boundary sharpening. Each panel shows the distribution of two-sided boundary separation S_i on 50 fixed held-out hard pairs. Larger values indicate clearer separation between the gold answer and the pre-adaptation confuser.

C.2 Rationale Update Statistics

BIRD updates a training rationale only when a valid confusable pair can be formed, candidate evidence survives answer-leakage filtering, and the best factor satisfies $U^{k^*} \geq \tau$. Table 6 summarizes the resulting construction statistics for each target model–domain setting.

We report three coverage statistics: *pair coverage*, the fraction of training samples for which a valid question-compatible different-answer witness is obtained; *candidate coverage*, the fraction for which at least one answer-blind candidate remains after filtering; and *update rate*, the fraction whose rationale is ultimately refined. For updated samples, we additionally report the mean own-side gain, cross-boundary transfer, and boundary utility of the selected factor.

Table 6: Statistics of BIRD rationale construction. Pair and candidate coverage, and update rate are percentages over the source training set. G_{own}, G_{cross}, and U are averaged over samples whose rationales are updated.

Model	Domain	Pair Cov. (%)	Candidate Cov. (%)	Update Rate (%)	G_{own}	G_{cross}	U
Qwen3.5-9B	Chart / Plot	93.8	89.7	70.9	1.25	0.17	0.91
	Medical	88.4	84.2	67.8	1.61	0.15	1.31
InternVL3-8B	Chart / Plot	92.9	88.6	68.6	1.14	0.15	0.84
	Medical	87.6	82.9	65.7	1.46	0.15	1.16

BIRD forms valid comparison pairs for most training samples, while candidate filtering and boundary-utility thresholding make rationale updates more selective. With the default threshold $\tau = 0.2$, 65.7%–70.9% of the training rationales are refined. The selected factors consistently exhibit substantially larger own-side gains than cross-boundary transfer under the default penalty $\lambda = 2.0$.

C.3 Rationale Length Analysis

A possible concern is that BIRD may improve adaptation simply by producing longer rationales and thus providing more supervision tokens during fine-tuning. To test this, we compare the rationale-length distributions of VC-STaR and BIRD for both target MLLMs in Figure 8. The distributions substantially overlap and fall within similar overall length ranges, indicating comparable rationale budgets. Notably, BIRD is not shifted toward longer rationales: its average rationale length is 159.8 tokens versus 184.6 for VC-STaR on Qwen3.5-9B, and 119.6 versus 143.8 on InternVL3-8B. The density profiles show a consistent pattern: BIRD retains a broad distribution comparable to VC-STaR but places slightly more mass in the shorter-length region, despite differences in the models' absolute rationale lengths. Thus, BIRD's additional supervision does not arise from increased verbosity or more rationale tokens. Instead, it changes the content of supervision by selectively incorporating verified boundary-relevant evidence within a comparable, and on average smaller, rationale budget. Therefore, BIRD's advantage cannot be readily attributed to a larger amount of textual supervision.

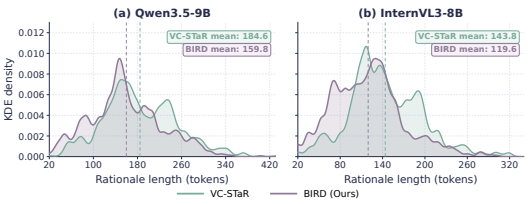


Figure 8: Rationale-length distributions of VC-STaR and BIRD for (a) Qwen3.5-9B and (b) InternVL3-8B. This indicates that BIRD does not obtain richer supervision simply by increasing rationale verbosity.

D Sensitivity Analysis

We examine the sensitivity of BIRD to four construction hyperparameters in Figure 9: the retrieval size P, cross-boundary penalty λ, utility threshold τ_U, and number of candidate factors K. We vary one hyperparameter at a time while keeping the others fixed at their default values. For each model–domain setting, performance is averaged over the three corresponding benchmarks. Performance

varies by less than 0.4 percentage points within every sweep, demonstrating that BIRD is generally robust to these hyperparameters. Moreover, the selected default consistently achieves the highest score across both target MLLMs and domains.

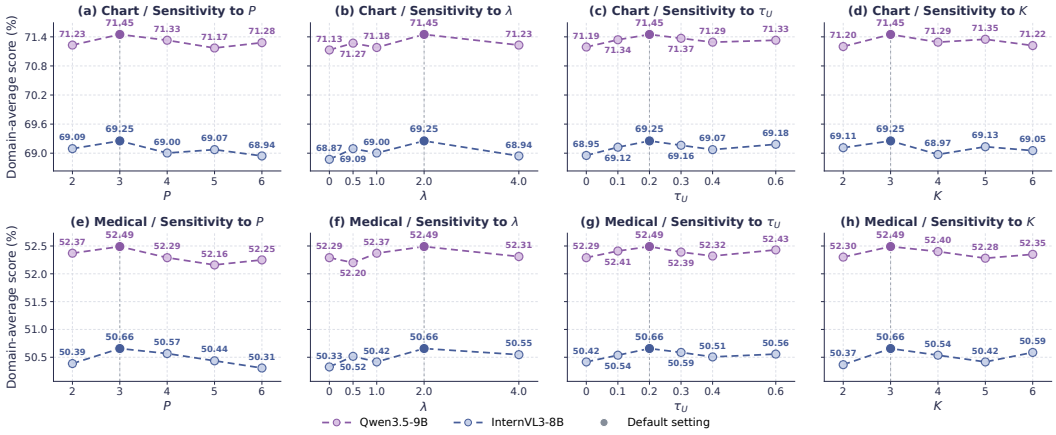


Figure 9: Sensitivity of BIRD to the retrieval size P, cross-boundary penalty λ, utility threshold τ_U, and number of candidate factors K. The top and bottom rows report results on the chart and medical domains, respectively. Scores are averaged over the three benchmarks in each domain. Filled markers indicate the default settings ($P = 3$, $\lambda = 2.0$, $\tau_U = 0.2$, and $K = 3$), which consistently achieve the strongest performance across both target MLLMs and domains.

D.1 Sensitivity to the Retrieval Size P

The retrieval size P controls how many visually similar, question-compatible neighbors are retained before BIRD identifies the alternative most confusable to the target MLLM. We vary $P \in 2, 3, 4, 5, 6$ while keeping the other settings fixed at their default values.

Increasing P from 2 to 3 improves performance in every model–domain setting. Further enlarging the retrieval pool provides no consistent benefit and generally leads to small declines. This suggests that a compact neighborhood already captures the most informative unresolved confusions, whereas a larger pool may introduce more distant or less relevant alternatives. We therefore set $P = 3$.

D.2 Sensitivity to the Cross-Boundary Penalty λ

The coefficient λ controls how strongly BIRD penalizes cross-boundary transfer when evaluating a candidate factor. We vary $\lambda \in 0, 0.5, 1.0, 2.0, 4.0$ while keeping the other settings fixed at their default values.

Compared with $\lambda = 0$, introducing a moderate cross-boundary penalty generally improves performance, with $\lambda = 2.0$ achieving the highest score in every model–domain setting. Increasing the penalty further to $\lambda = 4.0$ consistently reduces performance. These results confirm the importance of controlling cross-boundary transfer, while suggesting that an overly strong penalty may suppress factors that provide useful own-side gains despite limited transfer across the boundary.

D.3 Sensitivity to the Utility Threshold τ_U

The threshold τ_U determines whether the highest-utility factor is sufficiently boundary-relevant to trigger rationale refinement. We vary $\tau_U \in 0, 0.1, 0.2, 0.3, 0.4, 0.6$ while keeping the other settings fixed at their default values.

Performance improves as τ_U increases from 0 to 0.2, indicating that filtering weak candidate factors benefits rationale refinement. The default $\tau_U = 0.2$ achieves the highest score across all model–domain settings, while both lower and higher thresholds produce only modest decreases. This pattern reflects a balance between admitting weak evidence and rejecting potentially useful rationale updates.

D.4 Sensitivity to the Number of Candidate Factors K

The number K controls how many answer-blind factor pairs are generated from each target–witness pair before answer-leakage filtering and boundary-utility evaluation. We vary $K \in 2, 3, 4, 5, 6$ while keeping the other settings fixed at their default values.

Increasing K from 2 to 3 improves performance across both target MLLMs and domains. Beyond $K = 3$, the scores fluctuate slightly but remain consistently below the default. This suggests that a small candidate set is sufficient to cover the principal distinctions between paired samples, whereas additional factors are more likely to be redundant or less informative.

E Case Studies

We provide four case studies to illustrate how BIRD converts model-specific confusions into boundary-informed rationale supervision. Figures 10 and 11 present complete examples from chart and medical VQA, respectively. Figure 12 explains why a factor with a large target-side gain may still fail to capture the relevant decision boundary. Figure 13 further shows that both the confusable alternative and the evidence needed to resolve it depend on the target MLLM.

E.1 Chart/Plot VQA Examples

Figure 10 shows a chart question asking which series has the higher average across four quarters. The target sample has answer B, whereas the retrieved witness has answer A. Although the two charts share similar colors, scales, and overall structures, the balance between the early quarters and Q4 reverses the correct answer.

BIRD generates three answer-blind paired factors. Peak height (b1) and end-point rise (b3) both describe visible patterns, but these patterns occur on both sides of the pair and therefore do not explain the answer reversal. Their cross-boundary transfer reduces their utilities to 0.12 and 0.08, respectively. In contrast, the across-quarter balance (b2) captures the decisive difference. In the target, A's cumulative deficit of 45 over Q1–Q3 exceeds its gain of 25 in Q4, leaving B ahead overall. In the witness, the early deficit is only 15, while the Q4 gain is 35, reversing the result. This factor achieves $G_{\text{own}} = 0.90$ and $G_{\text{cross}} = 0.04$, giving the highest utility of 0.82. BIRD therefore distills this cumulative comparison, rather than the visually salient but non-decisive Q4 peak, into the target rationale.

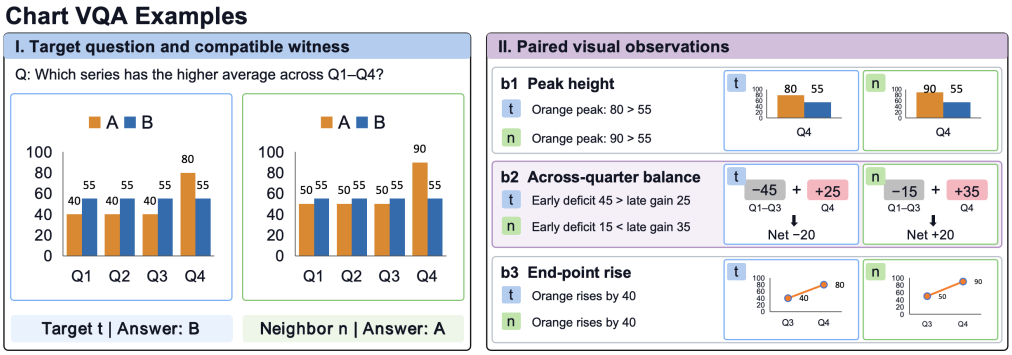
E.2 Medical VQA Examples

Figure 11 presents a chest-radiograph question about pneumothorax. The target is negative, while the retrieved witness is positive. BIRD proposes paired observations concerning regional density, the pleural boundary, and regional extent.

Regional density (b1) produces a relatively large own-side gain of 0.85. However, its substantial cross-boundary transfer of 0.30 reduces its utility to 0.25, indicating that differences in opacity and lucency alone do not reliably resolve the model's confusion. Regional extent (b3) is even less effective, with a utility of 0.08. The pleural-boundary factor (b2), by contrast, identifies markings extending toward the chest wall in the negative target and a retracted lung edge bordering peripheral lucency in the positive witness. It yields the largest own-side gain ($G_{\text{own}} = 1.10$), limited cross-boundary transfer ($G_{\text{cross}} = 0.07$), and the highest utility ($U = 0.96$). BIRD consequently emphasizes the pleural-boundary distinction in the refined rationale instead of relying on broader appearance differences that may occur in both cases.

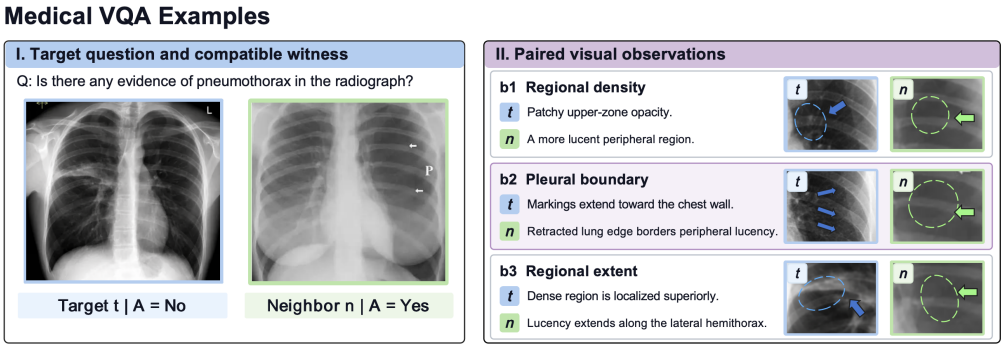
E.3 Why High-Gain Evidence Can Still Be Bad

Figure 12 isolates why own-side gain alone is insufficient for identifying boundary evidence. Both charts ask which category has the highest value, but the target is answered C and the witness is answered B. The partial-height factor (b1) appears useful when considering only the target: stating that C is taller than A and D increases the target model's correct-versus-confuser gap by $\Delta_t = 1.31$. However, the corresponding witness-side observation provides almost no support for the witness's



Candidate factor	$G_{own}\uparrow$	$G_{cross}\downarrow$	$U = G_{own} - 2G_{cross}$	Decision
b1 Peak height	0.60	0.24	0.12	× Reject
b2 Across-quarter balance	0.90	0.04	0.82	√ Select
b3 End-point rise	0.40	0.16	0.08	× Reject

Figure 10: Chart VQA example. BIRD compares a target answered B with a compatible witness answered A and evaluates three answer-blind paired factors. Across-quarter balance (b2) is selected because it provides high own-side gain and little cross-boundary transfer, whereas peak height (b1) and end-point rise (b3) describe patterns shared by both charts.



Candidate factor	$G_{own}\uparrow$	$G_{cross}\downarrow$	$U = G_{own} - 2G_{cross}$	Decision
b1 Regional density	0.85	0.30	0.25	× Reject
b2 Pleural boundary	1.10	0.07	0.96	√ Select
b3 Regional extent	0.50	0.21	0.08	× Reject

Figure 11: Medical VQA example. For a target without pneumothorax and a compatible witness with pneumothorax, BIRD selects the pleural-boundary factor (b2), which achieves $G_{\text{own}} = 1.10$, $G_{\text{cross}} = 0.07$, and $U = 0.96$. Regional density and extent are less boundary-specific and are therefore not selected.

correct answer ($\Delta_n = 0.09$). More importantly, these partial comparisons remain applicable when transferred across the pair because both B and C exceed A and D in both charts. The factor consequently has $G_{\text{cross}} = 0.56$, and its utility falls from $G_{\text{own}} = 0.70$ to $U = -0.42$.

Thus, this evidence raises confidence in the target answer without explaining why the alternative answer is correct on the neighboring sample. It strengthens a prediction but does not capture the boundary between B and C. By comparison, the pairwise-ordering factor (b2) directly states that C is slightly higher than B in the target and B is slightly higher than C in the witness. It benefits both samples under matched evaluation ($\Delta_t = 1.14$ and $\Delta_n = 1.06$) but has little effect when swapped ($G_{\text{cross}} = 0.05$), producing $U = 1.00$. This example illustrates why BIRD jointly rewards own-side gain and penalizes cross-boundary transfer instead of selecting evidence solely by its effect on the target.

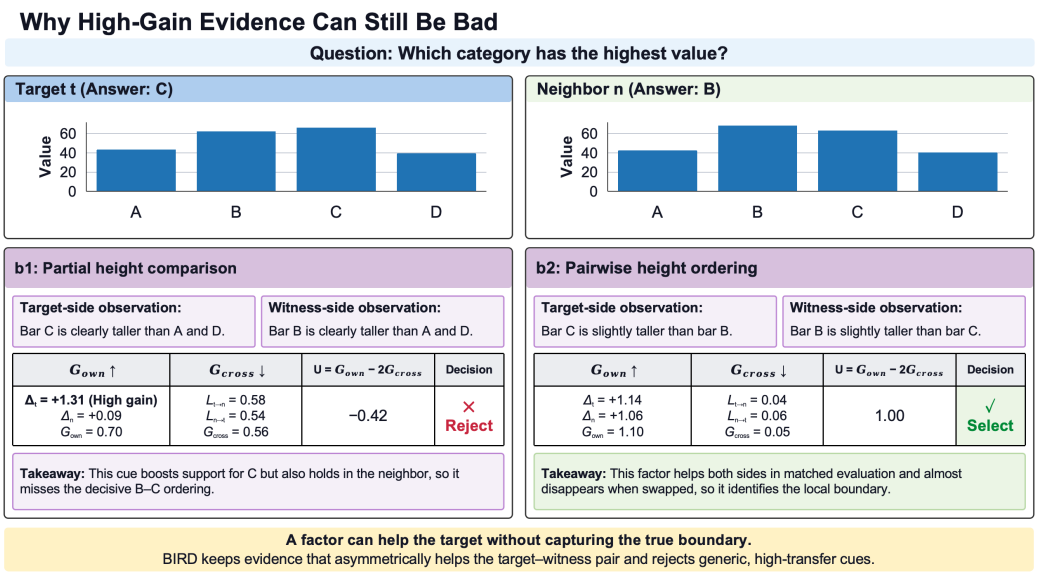


Figure 12: Why high-gain evidence can still be unsuitable. The partial-height factor (b1) strongly benefits the target but barely helps the witness and transfers across the pair, resulting in negative boundary utility. The pairwise B–C ordering (b2) supports the appropriate answer on both sides and nearly disappears when swapped, allowing it to identify the local decision boundary.

E.4 Model-Specific Example

Figure 13 demonstrates that the same sample can expose different unresolved boundaries for different MLLMs. The shared target asks for the Online sales value for East. Because the orange segment extends from 40 to 70, the correct answer is 30. Qwen3.5-9B instead prefers 40, with an energy gap of $g_\theta(30, 40) = -0.18$, reflecting confusion about the legend-to-segment mapping. Its retrieved witness preserves similar bars but reverses the legend, and the useful target observation therefore emphasizes that Online corresponds to the orange upper segment.

InternVL3-8B exhibits a different error on the same target. It prefers 70, with $g_\theta(30, 70) = -0.22$, treating the top of the stack as the value of the queried segment. Its witness has the same stack top but places the Online segment from zero to 70, exposing a confusion about segment span rather than legend binding. The useful observation consequently emphasizes that the target segment begins at the internal boundary of 40 and ends at 70.

The negative energy gaps show that each model initially favors its respective confuser over the gold answer. Moreover, when the evaluated pair and candidate factors are held fixed, the models still assign different boundary utilities. Qwen3.5-9B assigns higher utility to legend evidence (0.74 versus 0.26), whereas InternVL3-8B favors segment-span evidence (0.98 versus 0.28). Their distilled rationales accordingly emphasize different reasoning steps: legend mapping first for Qwen3.5-9B and segment subtraction first for InternVL3-8B. The example shows that model specificity arises in both where the unresolved boundary lies and which evidence is most effective for resolving it.

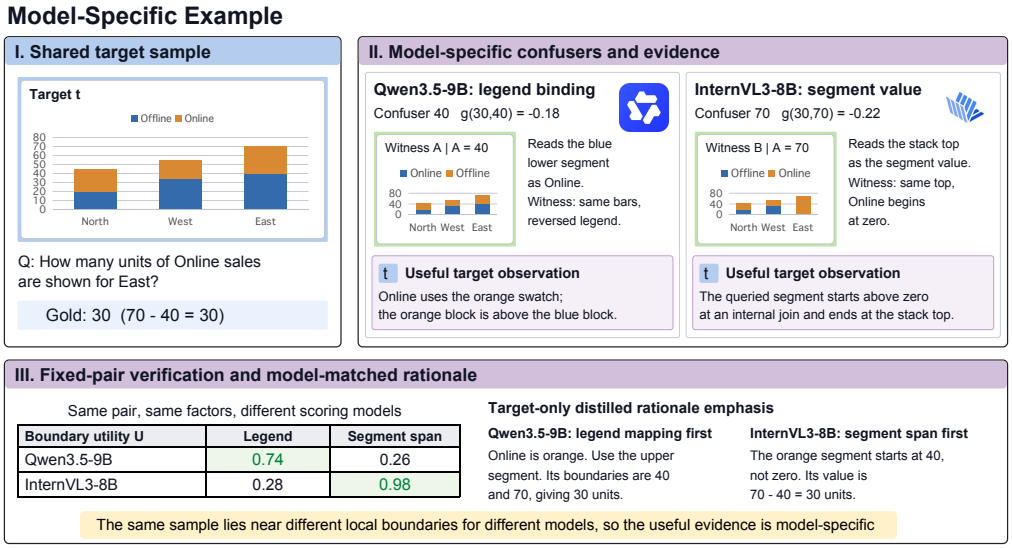


Figure 13: Model-specific boundary evidence for a shared target sample. Qwen3.5-9B confuses the gold answer 30 with 40 and assigns the highest utility to legend-binding evidence, whereas InternVL3-8B confuses it with 70 and favors segment-span evidence. Even under fixed-pair verification, the two models rank the same factors differently, leading to model-matched rationale supervision.

F Limitations and Perspectives

F.1 Limitations

BIRD has three main limitations. First, its boundary construction is local and pairwise. For each target sample, BIRD selects one confusable alternative and distills one verified factor. This design keeps the resulting supervision focused and interpretable, but some decisions may involve several plausible alternatives or require multiple complementary cues. A single target–witness pair may not fully capture such cases.

Second, boundary utility provides a functional rather than causal test of visual evidence. BIRD inserts a textual description of each candidate factor and measures how it changes the target MLLM's answer preferences. This reveals whether the factor helps distinguish the paired answers, but does not establish that the model has visually grounded the corresponding feature. The resulting scores may also be affected by answer phrasing, model calibration, and the wording of the candidate factor. Boundary utility should therefore be interpreted as a model-specific selection signal rather than causal evidence of visual grounding.

Finally, our experiments cover two target MLLMs and two specialized domains: chart/plot VQA and medical VQA. Although these settings differ substantially in their visual content and reasoning requirements, further evaluation is needed across additional model families, open-ended tasks, and specialized domains. In particular, improvements on medical VQA benchmarks should not be interpreted as establishing clinical reliability.

F.2 Perspectives

BIRD currently constructs boundary-informed rationales in a single offline round and then performs standard supervised fine-tuning. However, adaptation may resolve some confusions while exposing others. A natural extension is therefore to repeat neighbor retrieval, confuser selection, and evidence verification after each adaptation stage. This would allow the supervision to track the model's changing weaknesses while preserving single-sample training and inference.

BIRD also examines each sample against one confusable alternative and distills one factor. Future work could consider multiple competing answers and combinations of complementary factors when a decision cannot be resolved by a single visual distinction. The identified confusions could further guide a model-specific data curriculum that prioritizes samples near the model's remaining weaknesses. Evaluating these extensions across broader model families, open-ended tasks, and additional specialized domains would help establish the generality of boundary-informed supervision.