ON-POLICY VISUAL EVIDENCE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual agents solve problems by interleaving reasoning with image operations, and on-policy distillation (OPD) provides guidance from a strong teacher on student-generated interaction trajectories. However, image operations change the evidence available for subsequent reasoning, so local errors in evidence acquisition (*Acquire*), reading (*Read*), or answer grounding (*Ground*) can propagate through the trajectory and lead to incorrect answers. Existing multimodal OPD methods primarily construct or contrast auxiliary views of the original image to strengthen supervision, without explicitly modeling the connections between student actions, resulting observations, and subsequent reasoning. This limits their ability to provide corrections tailored to different failure stages. We introduce **Re**flection on **V**is**u**al **E**vidence (**REVUE**), an on-policy distillation method for visual agents. **REVUE** compares multiple student-generated trajectories for the same query, summarizes the observed visual evidence, and diagnoses the first failure across the *Acquire*, *Read*, and *Ground* stages. The resulting reflections provide training-time context for the teacher. We group and reweight token-level distillation losses according to how strongly these reflections affect the teacher's predictions. This design translates trajectory-level evidence diagnosis into targeted token-level supervision, guiding students to improve their visual evidence acquisition and reasoning. Across 11 benchmarks spanning the Qwen2.5-VL and InternVL3.5 model families, **REVUE** outperforms all evaluated OPD baselines in weighted-average scores for perception, mathematical reasoning, and general tasks. **REVUE** also reduces redundancy in reasoning and tool calls while improving tool-call accuracy and task accuracy.

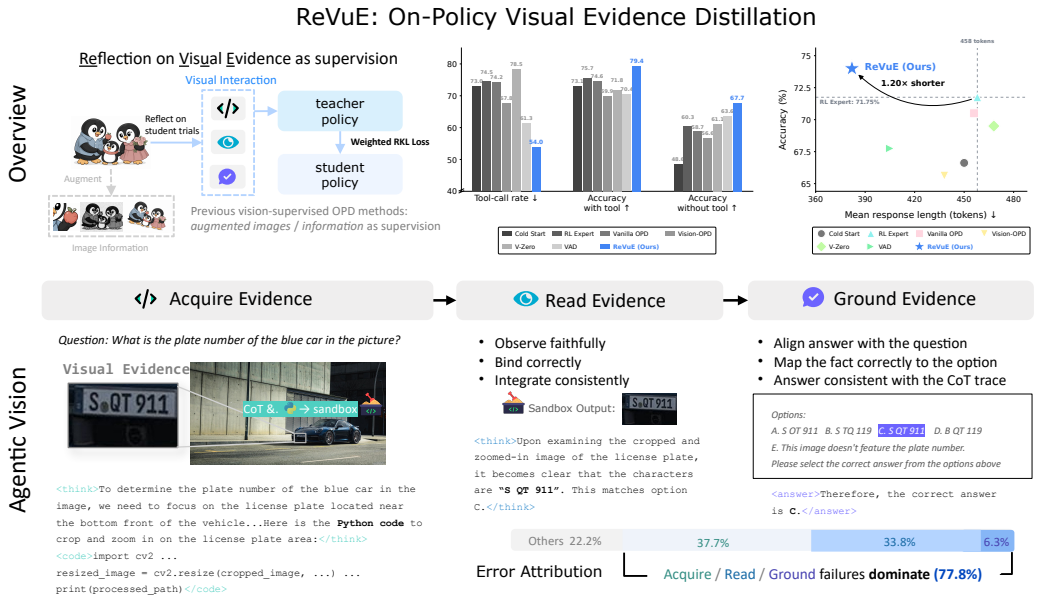


Figure 1: **REVUE overview (with HRBench-8K results) and an agentic vision example.**

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.

1 INTRODUCTION

High-resolution image understanding, chart analysis, and document question answering often require models to actively locate task-relevant regions, inspect fine-grained details, and incorporate their observations into subsequent reasoning (Wu & Xie, 2024; Qi et al., 2025). Visual agents address these demands by interleaving reasoning with image operations, actively gathering information needed to answer a query (Zheng et al., 2026b; Zhang et al., 2026b). As strong models develop these capabilities, effectively transferring their expertise to students becomes an important training objective. On-policy distillation (OPD) provides dense teacher feedback on student-generated trajectories (Agarwal et al., 2024; Gu et al., 2024; Wei et al., 2026), supervising the actions and reasoning decisions students actually make. Therefore, constructing teacher supervision tailored to visual interaction is a key step toward translating expert capabilities into reliable student behavior.

Visual interactions introduce a particular challenge: image operations change the evidence available for subsequent reasoning (Qi et al., 2025; Zhang et al., 2026b). Selecting the wrong region can leave critical evidence unavailable (Hou et al., 2026); obtaining a useful observation does not guarantee that the student reads it correctly; and a correctly read fact may still be mapped to the wrong answer. We refer to these stages as evidence acquisition (*Acquire*), visual reading (*Read*), and answer grounding (*Ground*). Errors at any stage can propagate through later actions and reasoning, ultimately producing an incorrect answer. Effective distillation therefore requires identifying the first failed stage in the evidence chain and directing corrections to where the error begins to propagate.

Representative multimodal OPD methods strengthen visual supervision through privileged visual cues or contrasts between teacher predictions under different visual conditions (Tian et al., 2026; Sun et al., 2026; Liu et al., 2026a). Vision-OPD (Yuan et al., 2026) uses evidence-centered crops to guide the teacher, while VAD (Zhang et al., 2026a) estimates the visual contribution to teacher corrections by removing relevant evidence. These methods improve how visual information informs supervision, but do not explicitly trace the links between students' executed actions, returned observations, and subsequent reasoning. This limits their ability to identify which visual behaviors to reinforce and where the evidence chain first breaks down. This raises a central question: *how can we turn students' own interaction trajectories into more targeted teacher guidance?*

Multiple attempts at the same query offer a concrete basis for such guidance. Successful attempts may complement each other: one finds more focused evidence, while another reads the relevant visual facts more accurately. Failed attempts help identify the first error in *Acquire*, *Read*, or *Ground*. We introduce **REVUE** (**Re**flection on **V**is**u**al **E**vidence) to turn these experiences into visual-evidence reflections. By comparing these trajectories, REVUE selects a compact set of observed images sufficient to establish the relevant visual facts. It also summarizes how to read the evidence correctly and map the resulting facts to the answer. The resulting reflections provide training-time context for a strong teacher, helping it reinforce effective behaviors already demonstrated by the student and correct errors at the first failed stage.

Visual-evidence reflection changes the teacher's assessment of student choices, with effects that vary across token positions. By comparing the teacher's predictions on the same student trajectory with and without reflection, we identify where reflection most strongly changes the teacher's support for student choices. In early training, we find that these changes concentrate on a small fraction of tokens. To emphasize supervision at these positions, we group tokens by the magnitude of these changes and apply grouped reweighting to give high-impact positions greater weight in the distillation loss. Our ablations show that this reweighting improves perception and math performance. Reflection changes the teacher's predictions and guides how token-level losses are weighted.

We evaluate REVUE on 11 benchmarks across the Qwen2.5-VL (Bai et al., 2025b) and InternVL3.5 (Wang et al., 2025a) model families. REVUE consistently outperforms all OPD baselines in performance across perception, mathematical reasoning, and general tasks. It also reduces redundant reasoning and unnecessary tool calls, while improving both tool-call accuracy and overall task accuracy. On Qwen2.5-VL-7B, masking high-impact positions in the distillation loss reduces performance more than masking an equal number of low-impact positions. Our critic's stage diagnoses agree closely with human judgments. Together, these results support the effectiveness of REVUE and the reliability of its stage diagnoses.

Our contributions are threefold:

- **Visual evidence reflection across student trajectories.** We compare multiple successful student attempts on the same query to identify minimal sufficient evidence and summarize successful reading and grounding strategies. We also locate the first failure among *Acquire*, *Read*, and *Ground*, providing training-time guidance to a strong teacher.
- **Reflection-aware grouped reweighting for distillation.** We partition tokens according to how strongly reflection changes the teacher's predictions, and use group-wise normalization and reweighting to focus distillation updates.
- **Cross-model gains and validation of high-impact supervision.** Across two model families and 11 benchmarks, REVUE improves task and tool-call accuracy over OPD baselines while reducing reasoning and tool-call redundancy. On Qwen2.5-VL-7B, masking high-impact positions in the loss reduces performance more than masking equally many low-impact positions, supporting greater weight on high-impact supervision.

2 RELATED WORK

Agentic Visual Reasoning. Agentic visual reasoning interleaves reasoning with image operations and the resulting observations. This capability appears in general-purpose models such as o3/o4-mini (OpenAI, 2025), Qwen3-VL (Bai et al., 2025a), and Kimi K2.5 (Kimi Team et al., 2026). Earlier systems use guided visual search or visual sketches to acquire and organize task-relevant evidence (Wu & Xie, 2024; Hu et al., 2024). CogCoM (Qi et al., 2025) and LATTE (Ma et al., 2025) learn sequences of visual operations through trajectory supervision. DeepEyes (Zheng et al., 2026b), VTool-R1 (Wu et al., 2026), and Pixel Reasoner (Su et al., 2025) use reinforcement learning to improve visual tool use and exploration. Mini-o3 (Lai et al., 2026) scales visual search to longer interactions and more diverse reasoning patterns. Thyme (Zhang et al., 2026b) and DeepEyesV2 (Hong et al., 2026) support broader reasoning through executable code, with DeepEyesV2 also incorporating web search. CodeV (Hou et al., 2026) uses process rewards on visual tool inputs and outputs to encourage evidence-consistent tool use. We study how students' actual visual interactions can inform teacher guidance for on-policy distillation.

On-Policy Distillation. On-policy distillation (OPD) supervises student-generated trajectories with teacher distributions, reducing the training–inference mismatch of offline distillation (Agarwal et al., 2024; Gu et al., 2024). Qwen3 (Yang et al., 2025) and Qwen3-VL (Bai et al., 2025a) also use this approach to train smaller models. Privileged-context methods condition teachers on reference solutions, additional context, or peer trajectories (Zhao et al., 2026; Ye et al., 2026; Yu et al., 2026). Visual OPD builds teacher–student asymmetry through evidence-centered crops, recoverable visual cues, or stronger image augmentation for students (Yuan et al., 2026; Tian et al., 2026; Li et al., 2026b). Visual contrasts also support supervision gating and token selection (Sun et al., 2026; Aniri et al., 2026), grouped loss reweighting (Liu et al., 2026a), and target reconstruction (Zhang et al., 2026a). Other methods internalize visual manipulation or generated visual thoughts to reduce inference-time computation (Cai et al., 2026; Li et al., 2026a). For agents, SGCD (Ding et al., 2026) and GRSD (Zheng et al., 2026a) use guidance derived from rollout groups to modify credit assignment in reinforcement learning. REVUE instead extracts reflections from students' visual interactions to reinforce effective behaviors and locate the first failure in *Acquire*, *Read*, or *Ground*. We condition the teacher on these reflections and use their effects on its predictions to group and reweight tokens for direct distillation.

3 PRELIMINARIES

Agentic visual reasoning. Given an image–query pair $x = (I, u)$, a student policy π_θ with parameters θ interacts with a tool environment P_{env}. Before step k, history $h_k = (x, a_0, o_1, \ldots, a_{k-1}, o_k)$ contains the input, previous actions, and returned observations, with $h_0 = x$. An action a_k contains reasoning, executable code, or a final answer. An observation o_{k+1} contains tool-returned text and images. The interaction follows

$$a_k \sim \pi_\theta(\cdot \mid h_k), \qquad o_{k+1} \sim P_{\text{env}}(\cdot \mid h_k, a_k), \qquad h_{k+1} = (h_k, a_k, o_{k+1}). \tag{1}$$

Actions without tool calls yield $o_{k+1} = \emptyset$. A final answer $\hat{y}$ ends the interaction, and the complete sequence of actions and observations forms a trajectory τ.

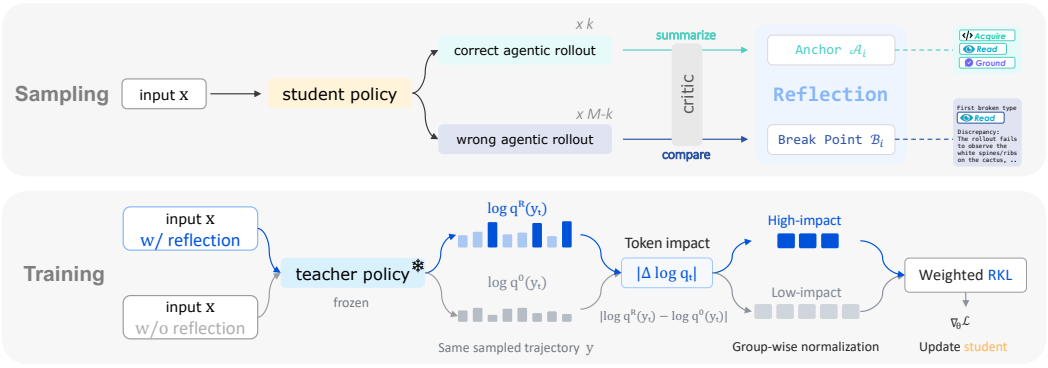


Figure 2: **Sampling and training pipeline of REVUE.**

Visual evidence E_k includes I and all tool-returned images in h_k. In Figure 1 (bottom), the student crops and zooms in on the blue car's plate. It reads `S QT 911` and matches it to option `C`. The correct visual fact f_x needed to answer u provides a reference for checking the student's reading. The rule γ_x maps this fact to the answer. The evidence chain is

$$x \xrightarrow{\text{Acquire}} E_k \xrightarrow{\text{Read}} f_x \xrightarrow[\gamma_x]{\text{Ground}} \hat{y}. \tag{2}$$

Success requires evidence sufficient to establish f_x, an accurate reading, and an answer satisfying $\hat{y} = \gamma_x(f_x)$. We attribute evidence-chain failures to the first unmet requirement in the dependency order *Acquire*, *Read*, and *Ground*. These stages can recur throughout an interaction.

On-policy distillation. On-policy distillation (OPD) (Agarwal et al., 2024) trains the student on its own sampled trajectories using token-level teacher distributions. For each input x, the current student samples M trajectories per round, forming $\mathcal{G}_x = \{\tau_i\}_{i=1}^{M}$ with trajectory index i. Token positions t are distinct from interaction steps k. Token $y_{i,t}$ is the student's output at position t in τ_i. History $h_{i,t}$ contains the input and preceding student outputs and tool observations. Let $\mathcal{T}_i$ denote the supervised student-output positions in trajectory i, with $|\mathcal{T}_i|$ their number. This set covers reasoning, code, and final answers across interaction rounds. Prompts and tool returns provide context but are not prediction targets. The teacher q_ϕ has fixed parameters ϕ. Given $h_{i,t}$, $p_{i,t}$ and $q^0_{i,t}$ denote the student and base-teacher next-token distributions:

$$p_{i,t} = \pi_\theta(\cdot \mid h_{i,t}), \qquad q^0_{i,t} = q_\phi(\cdot \mid h_{i,t}).$$

Superscript 0 marks the base teacher prediction. With $\mathbb{D}^{\mathrm{KL}}$ denoting KL divergence, the reverse-KL objective and its Monte Carlo (MC) estimate from $\mathcal{G}_x$ are

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{\tau_i \sim \pi_\theta(\cdot|x)}\left[\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i} \mathbb{D}^{\mathrm{KL}}(p_{i,t}\|q^0_{i,t})\right] \overset{\mathrm{MC}}{\approx} \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i}\mathbb{D}^{\mathrm{KL}}(p_{i,t}\|q^0_{i,t}). \tag{3}$$

The objective averages over supervised positions within each trajectory, then over the M trajectories for the same input.

4 METHOD

REVUE converts same-query student trajectories into reflection-guided distillation, as shown in Figure 2. It first constructs a trajectory-specific visual-evidence reflection, then measures how the reflection changes a frozen teacher's token predictions, and finally uses these changes to reweight token-level distillation losses (Algorithm 1).

4.1 VISUAL-EVIDENCE REFLECTION

Our error analysis identifies *Acquire*, *Read*, and *Ground* as the main failure stages (Figure 1), with examples in Appendix E. Prior work also shows that attending to the correct visual evidence does

not guarantee a correct answer (Liu et al., 2026b). Even with the same evidence, students may need different corrections. In the pie-chart example in Appendix E.2, both trajectories answer `No`. One misreads the slice ranking, while the other reads it correctly but misapplies the lower-median rule. These failures occur at the *Read* and *Ground* stages. Targeted distillation guidance therefore requires examining the student's actions, visual observations, and reasoning to identify the specific error in evidence use.

We construct reflections from $\mathcal{G}_x$, the group of trajectories sampled by the current student for the same input. The visual evidence and correct reasoning in one attempt can provide a reference for others. A critic compares these trajectories and their visual observations to extract correct references and diagnose errors in each incorrect trajectory. For trajectory i, the complete visual-evidence reflection is

$$\mathcal{R}_i = (\mathcal{A}_i, \mathcal{B}_i). \tag{4}$$

The **Anchor** $\mathcal{A}_i$ provides a reference for correct visual evidence use. It includes the target visual information and supporting images actually observed within $\mathcal{G}_x$. It also records the correct visual fact f_x and the rule γ_x that maps this fact to the answer. For an incorrect trajectory, the **Break Point** $\mathcal{B}_i = (\sigma_i, \delta_i)$ describes its failure relative to this reference. σ_i denotes the first failed stage in the dependency order, and δ_i describes the specific discrepancy. These images and text serve as training-time teacher context for token-level supervision on the student's sampled trajectories. The teacher can then reinforce effective evidence use and guide corrections to observed errors. Appendix A.4 shows a concrete visual-evidence reflection, and Appendices A.5–A.6 detail its construction and critic prompts.

4.2 From Reflection to Token Impact

Visual-evidence reflection provides correct references and diagnoses failures in the evidence chain. To locate where this guidance changes supervision, we compare teacher scores along the student's sampled trajectory. On the same student history $h_{i,t}$, the frozen teacher predicts $q^0_{i,t}$ without reflection (Section 3) and $q^{\mathcal{R}}_{i,t} = q_\phi(\cdot \mid h_{i,t}, \mathcal{R}_i)$ with it. Only the training-time teacher receives reflection; the student retains its original interaction history. Using teacher probabilities normalized over the full vocabulary, we define the log-probability shift for candidate token v:

$$d_{i,t}(v) = \log q^{\mathcal{R}}_{i,t}(v) - \log q^0_{i,t}(v). \tag{5}$$

To connect these shifts to the reverse-KL objective in Equation 3, we compare both teachers on a shared vocabulary support. The training loss uses the adopted teacher's top-K candidates; for the reflection-conditioned branch, we use these candidates as Supp for both teachers. Suppressing (i,t), write $p = p_{i,t}$ and $p^{\mathrm{Supp}}(v) = p(v)/\sum_{u\in\mathrm{Supp}} p(u)$ for $v \in \mathrm{Supp}$. Define q^{Supp} analogously and let $\ell^{\mathrm{Supp}}(p,q) = \mathbb{D}^{\mathrm{KL}}(p^{\mathrm{Supp}} \| q^{\mathrm{Supp}})$.

Lemma 1 (Reflection and the distillation gradient) *For fixed history, teacher predictions, and nonempty shared support, with positive student and teacher probabilities on this support, we have*

$$\nabla_\theta \left[\ell^{\mathrm{Supp}}(p, q^{\mathcal{R}}) - \ell^{\mathrm{Supp}}(p, q^0)\right] = -\mathbb{E}_{v\sim p^{\mathrm{Supp}}}\left[d(v)\nabla_\theta \log p^{\mathrm{Supp}}(v)\right]. \tag{6}$$

Equation 6 expresses the gradient change as a sum over candidate tokens, with each term weighted by its teacher-score shift $d(v)$ (proof in Appendix D.1). We focus on the student's sampled token $y_{i,t}$ to locate *where reflection changes teacher support for its actual choices*. We define the magnitude of this shift as ***token impact***:

$$s_{i,t} = |d_{i,t}(y_{i,t})| = \left|\log q^{\mathcal{R}}_{i,t}(y_{i,t}) - \log q^0_{i,t}(y_{i,t})\right|. \tag{7}$$

Taking the absolute value captures both increased and decreased teacher support, covering reinforcement and correction. This score ranks positions by changes in teacher scores, not by full gradient magnitude. Each position's loss still supervises the distribution over its support.

Token impact distinguishes positions within a trajectory by how much their teacher scores change with reflection. In Figure 3, reflection reduces support for the mistaken judgment of `one` island. The token `one` has higher impact than the final `Lucia`, whose teacher score changes little under the existing erroneous prefix. Section 4.3 groups positions by these scores and assigns greater distillation weight to high-impact positions.

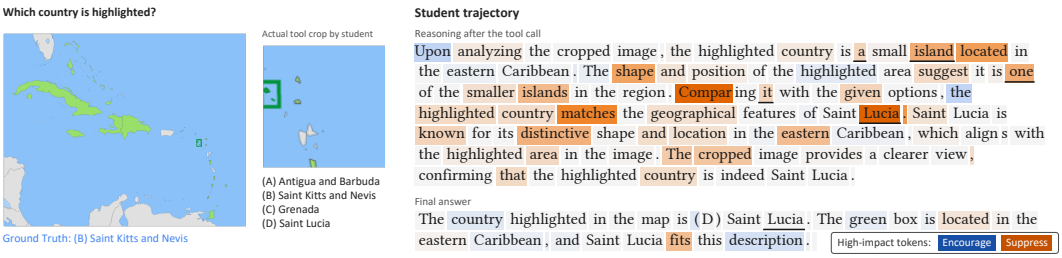


Figure 3: **Token impact on a trajectory with a *reading* error.** The student misreads two islands as one, answering `Saint Lucia` instead of `Saint Kitts and Nevis`. Reflection identifies two islands and lowers support for the underlined `island`, `one`, and first `Lucia`. **Blue/orange** denote increased/decreased teacher support; darker shading indicates larger absolute log-probability shifts. Both evaluations score the same student trajectory; the final `Lucia` changes little under the fixed erroneous prefix.

4.3 Reflection-Aware Grouped Distillation

Token impact measures teacher-score changes induced by visual-evidence reflection (Section 4.2). Figure 4 pools early-training scores: approximately 20% of scored positions account for 97.6% of total impact (Appendix B.4). Uniform token averaging allocates total loss weight in proportion to group size, leaving sparse high-impact positions a small share. To emphasize their supervision, we group positions by impact and introduce weights $w_{i,t}$ into the reverse-KL objective in Equation 3:

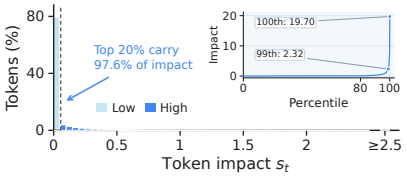


Figure 4: **Token-impact sparsity**

$$\mathcal{L}_{\mathrm{ReVuE}}(\theta) = \frac{1}{|\mathcal{G}|} \sum_{\tau_i \in \mathcal{G}} \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} w_{i,t}\, \mathbb{D}^{\mathrm{KL}}\left(p_{i,t}^{\mathrm{Supp}_{i,t}} \,\middle\|\, (q_{i,t}^{*})^{\mathrm{Supp}_{i,t}} \right). \tag{8}$$

$\mathcal{G}$ contains retained student trajectories in the current batch, with nonempty supervised sets $\mathcal{T}_i$. The teacher target $q_{i,t}^{*}$ is $q_{i,t}^{\mathcal{R}}$ when reflection-conditioned scoring is valid and token-aligned, and $q_{i,t}^{0}$ otherwise. Both distributions are renormalized over $\mathrm{Supp}_{i,t} = \mathrm{TopK}(q_{i,t}^{*})$, as indicated by the superscripts.

Within each trajectory, we rank supervised positions by $s_{i,t}$, assigning the top $\lceil \alpha |\mathcal{T}_i| \rceil$ to $\mathcal{T}_{i,\mathrm{High}}$ and the rest to $\mathcal{T}_{i,\mathrm{Low}} = \mathcal{T}_i \setminus \mathcal{T}_{i,\mathrm{High}}$, where $\alpha \in (0,1)$. For nonempty groups, we allocate total loss weights $\lambda \in (0,1)$ and $1-\lambda$ to `High` and `Low` and distribute each uniformly within its group. Each position's coefficient in Equation 8 is $w_{i,t}/|\mathcal{T}_i|$, giving the weights

$$w_{i,t} = \begin{cases} \lambda |\mathcal{T}_i| / |\mathcal{T}_{i,\mathrm{High}}|, & t \in \mathcal{T}_{i,\mathrm{High}}, \\ (1-\lambda) |\mathcal{T}_i| / |\mathcal{T}_{i,\mathrm{Low}}|, & t \in \mathcal{T}_{i,\mathrm{Low}}. \end{cases} \tag{9}$$

α controls group size, while λ controls total group weight. We set $\lambda = 0.5$, assigning half to each group. With teacher targets, supports, and groups fixed, $w_{i,t}$ scales each position's gradient relative to uniform token averaging. This multiplier exceeds one for `High` when the group contains fewer than half the supervised positions. Appendix D.2 and Figure 14 give the derivation and weight curves. If impacts are unavailable or either group is empty, we restore uniform averaging with $w_{i,t} = 1$ and retain the adopted teacher target (Appendix A.1).

5 Experiments

5.1 Experimental Setup

For models, we use Thyme-SFT (Zhang et al., 2026b) cold-start checkpoints of Qwen2.5-VL-7B (Bai et al., 2025b) and InternVL3.5-4B-Instruct (Wang et al., 2025a) as students. Teachers

are our reproduced Thyme-RL experts; Qwen3.5-397B-A17B (Qwen Team, 2026) is the critic. Off-the-shelf comparisons include GPT-4o (OpenAI, 2024), Gemini-3.1-Flash-Lite, Qwen2.5-VL-32B, Qwen3-VL-30B-A3B-Thinking (Bai et al., 2025a), and InternVL3.5-38B-Instruct. **For baselines,** we compare with Vision-OPD (Yuan et al., 2026) for visual privilege and V-Zero (Sun et al., 2026) and VAD (Zhang et al., 2026a) for visual contrast. We also include Vanilla OPD, RFT (Yuan et al., 2023), and GT-Privileged. All OPD baselines use the same RL expert teacher; details are shown in Appendix B.1. **For benchmarks,** we evaluate on 11 benchmarks covering perception, math, and general multimodal tasks (Appendix F). Appendix B.3 details the training data, model initialization, and training and evaluation settings.

5.2 Main Results

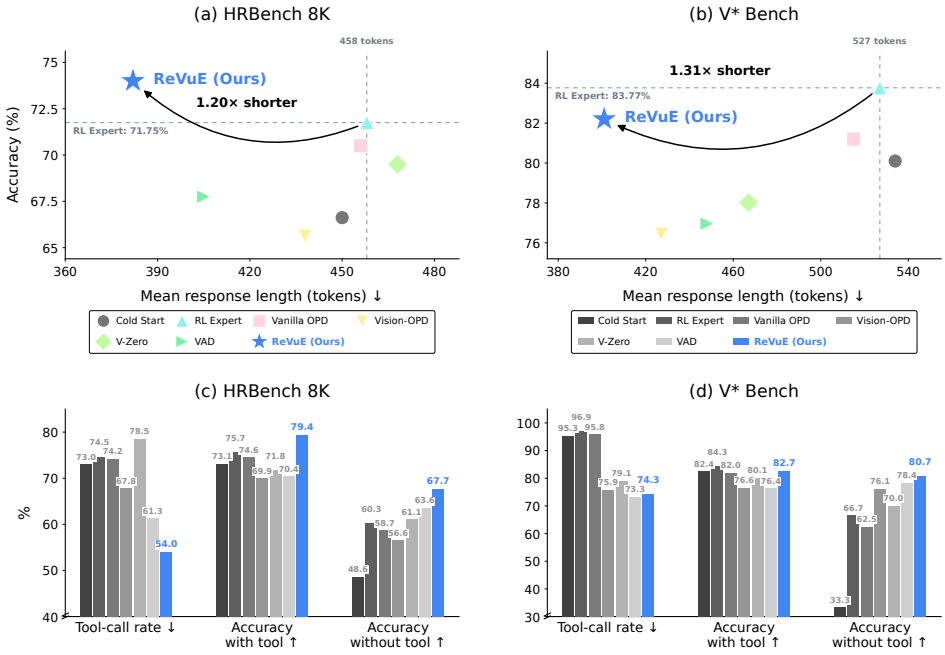


Figure 5: **Response length, accuracy, and tool use** for Qwen2.5-VL-7B.

ReVuE improves perception across model families, with gains in math and general tasks. In all three weighted category averages, both ReVuE students outperform all evaluated OPD baselines and RFT, which trains only on teacher trajectories with correct answers (Table 1). Perception gains over Vanilla OPD are 2.40% for Qwen2.5-VL-7B and 1.70% for InternVL3.5-4B-Instruct. Qwen's gains on HRBench 8K (+3.50%) and TreeBench (+3.10%) cover high-resolution understanding and object relations. Both models also improve in math and general tasks, with average gains of 2.83% and 1.65% for InternVL, respectively. ReVuE also achieves higher category averages than baselines using privileged visual information or visual contrasts, with matched student initializations and teacher sources. The answer-privilege comparison provides further evidence: ReVuE exceeds GT-Privileged on all 11 benchmarks for both students. These comparisons support guiding how students acquire, read, and use evidence; giving the teacher the correct answer alone does not yield the same gains. After reflection-assisted distillation, both students also surpass their RL Expert teachers in all three category averages. Both students also exceed GPT-4o in all three averages and the larger off-the-shelf models in their families in perception.

ReVuE improves perception accuracy with shorter responses and less frequent tool use. Across all five perception benchmarks, the Qwen2.5-VL-7B student is more accurate than Vanilla OPD while producing shorter responses and using tools less often (Figure 5; Appendix C.4). On HRBench 8K, ReVuE reduces mean response length from the RL Expert's 458 tokens to 382, while raising accuracy from 71.75% to 74.00% (+2.25%). On HRBench 8K and V* Bench, its answer accuracy exceeds the other plotted distillation baselines on samples both with and without

Table 1: Main results. **Bold** marks the best OPD result per column within each model family, including ties. Wtd. Avg. uses benchmark sample counts as weights. Baseline details and model labels are provided in Appendix B.1–B.2. Evaluation benchmark details are provided in Appendix F.

Model / Method	Perception						Math				General			
	HR-4K	HR-8K	V*	Tree Bench	Visual Probe	Wtd. Avg.	Math Vista	Math Verse	Visu Logic	Wtd. Avg.	Hallu Bench	CQA Pro	Info VQA	Wtd. Avg.
Off-the-Shelf Models														
GPT-4o	61.00	54.00	61.78	49.88	23.88	50.28	58.83	39.21	25.20	41.22	51.37	28.67	71.17	53.28
Gemini3.1FL	46.00	43.00	64.92	51.85	27.96	43.89	80.80	77.92	32.40	62.63	59.92	37.03	83.50	63.57
Qwen2.5-32B	75.13	69.25	78.01	48.40	45.05	63.89	77.00	53.05	25.90	51.90	50.74	30.02	83.10	59.29
Qwen3-30B-T	77.13	71.38	80.10	45.43	36.89	63.26	80.20	66.12	25.80	56.71	61.82	35.46	85.68	64.45
InternVL-38B	71.50	62.13	65.97	41.98	20.97	54.34	70.90	48.48	27.20	48.89	50.74	26.98	77.70	55.71
Qwen2.5-VL-7B														
Base Model	69.00	63.50	75.39	37.04	41.17	57.77	69.10	44.04	25.40	46.34	40.91	20.79	75.10	50.53
Cold-start	74.38	66.62	80.10	37.28	41.56	60.72	68.00	44.29	26.40	46.38	41.73	21.12	79.05	52.68
RL Expert	75.50	71.75	83.77	40.25	44.85	63.89	70.60	45.43	26.20	47.56	43.83	21.69	79.74	53.60
RFT	75.38	72.00	81.20	40.00	41.75	63.12	69.50	45.81	24.60	46.70	42.33	21.18	79.57	53.07
Vanilla OPD	75.40	70.50	81.20	38.02	42.91	62.61	70.60	46.07	26.00	47.67	42.88	21.69	79.44	53.28
GT-Privileged	73.62	70.50	79.58	39.75	43.10	62.26	71.20	45.69	25.20	47.49	43.45	20.84	79.70	53.23
Vision-OPD	73.00	65.62	76.44	39.01	41.36	59.98	67.70	42.39	25.70	45.48	39.89	20.79	78.75	52.08
V-Zero	75.62	69.50	78.01	37.28	43.10	62.08	69.80	43.91	**26.60**	46.99	42.16	21.08	79.12	52.79
VAD	74.75	67.75	76.96	**41.98**	43.89	62.08	69.00	45.56	22.30	45.62	44.24	21.73	79.65	53.65
REVUE(Ours)	**77.10**	**74.00**	**82.20**	41.12	**44.66**	**65.01**	**71.60**	**47.08**	26.50	**48.49**	**45.12**	**21.81**	**80.27**	**54.14**
InternVL3.5-4B-Instruct														
Base Model	62.00	55.00	68.59	40.49	20.58	49.32	68.50	28.93	25.80	42.00	40.04	26.01	66.04	47.78
Cold-start	69.50	63.25	66.49	40.99	29.90	55.66	69.30	47.21	25.80	47.45	48.50	26.24	72.98	52.79
RL Expert	72.62	64.62	71.73	40.74	34.56	58.20	69.30	46.32	26.90	47.60	46.92	30.59	73.93	54.38
RFT	71.25	66.00	71.73	40.25	35.00	58.22	70.00	45.94	24.30	46.81	47.78	29.39	74.17	54.26
Vanilla OPD	71.13	65.87	**73.30**	40.25	33.79	58.02	68.40	43.40	26.60	46.34	46.85	30.46	72.16	53.48
GT-Privileged	71.25	64.88	70.16	39.75	33.20	57.36	71.30	39.34	26.20	46.09	49.29	29.90	72.48	53.91
Vision-OPD	66.25	60.00	65.97	37.78	28.93	53.04	66.60	44.54	26.60	46.02	44.57	30.46	**74.47**	54.14
V-Zero	68.13	61.75	67.02	40.25	28.35	54.45	68.80	46.95	26.80	47.56	49.03	28.65	73.61	53.99
VAD	67.50	61.00	67.02	39.01	27.96	53.78	68.30	46.07	27.70	47.45	48.05	31.44	73.36	54.61
REVUE(Ours)	**73.25**	**67.37**	**73.30**	**41.48**	**36.12**	**59.72**	**71.60**	**47.46**	**28.10**	**49.17**	**49.38**	**32.27**	73.35	**55.13**

tool use. Tool-use frequency alone does not explain the performance differences. On V* Bench, VAD uses tools slightly less often (73.3% vs. 74.3%), yet REVUE achieves 5.24% higher overall accuracy. The extent of these behavioral changes varies by task. On TreeBench, REVUE reduces response length by 46.09% and tool-use rate from 64.7% to 15.8%, while improving accuracy by 3.10% over Vanilla OPD. On VisualProbe, it still uses tools on 83.7% of samples and also achieves higher accuracy. These results support more effective visual evidence use: higher accuracy with less generation and less frequent tool use, while tool reliance still varies by task.

6 Analysis and Discussion

Reflection brings gains across tasks; impact reweighting further improves perception and math. In the cumulative ablation on Qwen2.5-VL-7B, adding visual-evidence reflection improves all three category averages and nine of 11

Table 2: **Component ablation on Qwen2.5-VL-7B.**

Method	Perception ↑	Math ↑	General ↑
Vanilla OPD	62.61	47.67	53.28
+reflection	63.70	47.78	**54.42**
+impact reweight	**65.01**	**48.49**	54.14

benchmark scores over Vanilla OPD (Table 2). Adding impact reweighting then raises the perception and math averages by another 1.31% and 0.71%, respectively. It also improves tasks that do not benefit from reflection alone. On TreeBench, for example, accuracy falls from 38.02% to 37.78% with reflection alone, then rises to 41.12% with reweighting. The general-task average drops by 0.28% relative to reflection alone, but REVUE still outperforms Vanilla OPD on all 11 benchmarks (Appendix C.1).

Impact ranking identifies supervision worth prioritizing. We compare impact-based and random selection on Qwen2.5-VL-7B with fixed group fractions and weights (Table 3). Compared with random selection, REVUE gains 1.37% in perception and 1.07% in math, but only 0.04% on general tasks. Masking yields the same ordering in all three category averages: `Mask low` > `Mask random` > `Mask top`. `Mask low` also exceeds `Mask top` on all 11 benchmarks (Appendix C.2). The cases show reflection lowering teacher support for erroneous choices and raising support for correct judgments within failed trajectories. In the age case, reflection favors the correct relation `above` and discourages the erroneous comparison `exceed` (Appendix E.4). Scores can change at target selection and intermediate visual judgments while final-answer token scores remain nearly unchanged (Figure 3; Appendix E.4). The top 20% of pooled early-training positions carry 97.6% of total impact, supporting greater weight on these positions (Figure 4). `Mask low` gains 0.17% in perception over the full method but slightly reduces math and general averages, showing task-dependent effects.

Table 3: **Token-impact interventions.**

Method	Perception ↑	Math ↑	General ↑
Vanilla OPD	62.61	47.67	53.28
Random select 20%	63.64	47.42	54.10
Mask random 20%	63.85	47.35	53.78
Mask low 20%	**65.18**	48.42	53.97
Mask top 20%	62.26	47.17	52.51
REVUE(Ours)	65.01	**48.49**	**54.14**

Prioritizing 20% of positions gives the highest perception and math averages. This fraction also leads on all three math benchmarks (Table 4; Appendix C.3). With $\lambda = 0.5$, reducing α shrinks the high-weight group and increases each position's weight within it (Appendix D.2). Neither 5% nor 10% exceeds the perception and math averages at 20%, supporting moderate concentration in this setting. The general-task average peaks at 100%, 0.28% above the 20% setting, indicating different preferences across tasks.

Table 4: **Sensitivity to α.**

Fraction	Perception ↑	Math ↑	General ↑
Top 5%	64.70	47.49	54.12
Top 10%	64.55	46.92	53.85
Top 20% (REVUE)	**65.01**	**48.49**	54.14
Top 50%	64.33	47.42	54.18
Top 100%	63.70	47.78	**54.42**

Critic diagnoses agree closely with human judgments, including on incorrect trajectories. One annotator independently labels 96 rollouts balanced across `all-correct`, `mixed`, and `all-wrong` groups before viewing critic labels. Final stage labels match 93 of 96 human labels (96.88%, Cohen's $\kappa = 0.957$). Agreement remains 95.31% on the 64 incorrect rollouts. On 768 rollouts, we compare archived final training labels with fresh labels from the same critic and two external judges. Valid pairs agree at 99.2% for the same critic and 92.8–93.2% across models (Table 5). These results support human agreement and repeatability of the stage diagnoses used for reflection. Appendix B.6 details sampling and label disagreements.

Table 5: **Critic stage consistency.**

Comparator	n	Agreement (%)	κ
Human annotator	96	**96.88**	**0.957**
Qwen (rerun)	744	99.2	0.987
Gemini 3.8 Flash	736	93.2	0.892
GPT-5.6 Luna	736	92.8	0.886

n: valid pairs.

7 CONCLUSION

To provide targeted supervision for visual agents, we introduced REVUE, which improves perception, mathematical reasoning, and general-task performance through visual-evidence reflection and token reweighting in on-policy distillation. The reflections provide the teacher with references for correct evidence use and diagnoses of errors in the student's own trajectories. Token reweighting then emphasizes supervision at positions where these reflections most change the teacher's assessment. Behavioral analysis on Qwen2.5-VL-7B also shows fewer tool calls and shorter responses on perception tasks. Ablations and token-impact interventions support the effectiveness of both components. Our study focuses on two student model families and image-based question answering; broader model coverage and more diverse visual interactions remain to be explored.

AI USE STATEMENT

We used generative AI tools to assist with manuscript translation and language polishing, coding, and preparation of the supplementary materials. We also used generative models for teacher supervision, critic feedback, evaluation, and error attribution, as described in Appendices A and B. The authors are responsible for reviewing AI-assisted content and for the final text, claims, and artifacts of this work.

REPRODUCIBILITY STATEMENT

Appendix A provides the training algorithms, critic prompt, and method-specific hyperparameters. Model and training-data sources, decoding settings, the random seed, and hardware are documented in Appendix B.3. Appendix F describes the evaluation benchmarks and links to their sources. The supplementary archive `code.zip` contains the training and evaluation code for REVUE, with setup instructions in `README.md`.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *International Conference on Learning Representations*, pp. 21246–21263, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/5be69a584901a26c521c2b51e40a4c20-Paper-Conference.pdf.

Aniri, Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, and Tat-Seng Chua. OPD-V: Visual On-Policy Self-Distillation with Modality Balance, 2026. URL https://arxiv.org/abs/2608.05131.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025b. URL https://arxiv.org/abs/2502.13923.

Yishuo Cai, Jiahui Liu, Yuanxin Liu, Haobo Deng, Linli Yao, Yuhao Zheng, Kun Ouyang, Zhimo Li, Ziyue Wang, Xu Sun, Haoli Bai, and Xiaohui Li. Thinking Without Images: Internalizing Visual Manipulation with On-Policy Self-Distillation, 2026. URL https://arxiv.org/abs/2606.08719.

Tianyu Ding, Jianhong Xin, and Juan Pablo De la Cruz Weinstein. Keep policy gradient in charge: Sibling-guided credit distillation for long-horizon tool-use agents, 2026. URL https://arxiv.org/abs/2606.12634.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-4899. URL https://proceedings.neurips.cc/paper_files/

paper/2025/file/d79a27cf2772fe00be7f341efc0eb517-Paper-Datasets_and_Benchmarks_Track.pdf.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, pp. 32694–32717, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/8ac015d409635f196f9e3e9dcfb9a94e-Paper-Conference.pdf.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. DeepEyesV2: Toward Agentic Multimodal Model. In *International Conference on Learning Representations*, pp. 114851–114872, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/badbef7d421640fb038399e9d7271b1c-Paper-Conference.pdf.

Xinhai Hou, Shaoyuan Xu, Manan Biyani, Moyan Li, Jia Liu, Todd C Hollon, and Bryan Wang. CodeV: Code with Images for Faithful Visual Reasoning via Tool-Aware Policy Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21500–21510, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Hou_CodeV_Code_with_Images_for_Faithful_Visual_Reasoning_via_Tool-Aware_CVPR_2026_paper.html.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual Sketchpad: Sketching as a Visual Chain of Thought for Multimodal Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 139348–139379, 2024. doi: 10.52202/079017-4423. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fb82011040977c7712409fbdb5456647-Paper-Conference.pdf.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Yean Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Shuhao Guan, Yuanying Guo, Xiaoru Hao, Dailan He, Tianhong He, Weiran He, Wenyang He, Yibo He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Chaobo Jia, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zhaoxi Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Yibo Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, G. Luo, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai,

Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hao Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zifan Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Xiaofei Yang, Xinlong Yang, Xinyu Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Haoran Ye, Wenjie Ye, Zhuorui Ye, Peng Yebo, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Xiaoyun Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Xuanle Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi K2.5: Visual Agentic Intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling Up Reasoning Patterns and Interaction Turns for Visual Search. In *International Conference on Learning Representations*, pp. 76722–76746, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7c40c5050bd029a3ea7ff8b01412f735-Paper-Conference.pdf.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Zesheng Yang, Fangzhi Xu, and Jun Liu. Visual-OPSD: Cross-Modal On-Policy Self-Distillation for Efficient Unified Multimodal Reasoning, 2026a. URL https://arxiv.org/abs/2606.18974.

Yijiang Li, Yijun Liang, Yunjie Tian, Bingyang Wang, Ke Zhang, Zhenfei Yin, Di Fu, Philip Torr, and Nuno Vasconcelos. Self-Supervised Visual On-Policy Distillation, 2026b. URL https://arxiv.org/abs/2608.14144.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, and Shu Wu. Visual-Advantage On-Policy Distillation for Vision-Language Models, 2026a. URL https://arxiv.org/abs/2605.21924.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, Hanqing Lu, Benoit Dumoulin, and Hanghang Tong. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in VLMs. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 72849–72876, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/76818d8d85e05e45ce3a16a8468619d1-Paper-Conference.pdf.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/663bce02a0050c4a11f1eb8a7f1429d3-Paper-Conference.pdf.

Zixian Ma, Jianguo Zhang, Zhiwei Liu, Jieyu Zhang, Juntao Tan, Manli Shu, Juan Carlos Niebles, Shelby Heinecke, Huan Wang, Caiming Xiong, Ranjay Krishna, and Silvio Savarese. LATTE:

Learning to Think with Vision Specialists. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11192–11229, 2025. doi: 10.18653/v1/2025.emnlp-main.564. URL https://aclanthology.org/2025.emnlp-main.564/.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartQAPro: A more diverse and challenging benchmark for chart question answering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.978. URL https://aclanthology.org/2025.findings-acl.978/.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1697–1706, January 2022. URL https://openaccess.thecvf.com/content/WACV2022/html/Mathew_InfographicVQA_WACV_2022_paper.html.

OpenAI. GPT-4o System Card, August 2024. URL https://cdn.openai.com/gpt-4o-system-card.pdf.

OpenAI. Thinking with images, April 2025. URL https://openai.com/index/thinking-with-images/.

Ji Qi, Ming Ding, Weihan Wang, Yushi Bai, Qingsong Lv, Wenyi Hong, Bin Xu, Lei Hou, Juanzi Li, Yuxiao Dong, and Jie Tang. CogCoM: A Visual Language Model with Chain-of-Manipulations Reasoning. In *International Conference on Learning Representations*, pp. 11090–11110, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/1dcee1cd6890ab7fcdf173ec10526da9-Paper-Conference.pdf.

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

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel Reasoner: Incentivizing Pixel-Space Reasoning with Curiosity-Driven Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 8222–8251, 2025. doi: 10.52202/085713-0280. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0c38f54740062529aa4117a04b583f3c-Paper-Conference.pdf.

Haoxiang Sun, Zhihang Yi, Langxuan Deng, Yuhao Zhou, Peiqi Jia, Jian Zhao, Li Yuan, Jiancheng Lv, and Tao Wang. V-Zero: Answer-Label-Free On-Policy Distillation with Contrastive Evidence Gating for Fine-Grained Visual Reasoning, 2026. URL https://arxiv.org/abs/2606.25319.

Kanghui Tian, Siyuan Liu, Ziang Yan, Sheng Xia, Shuai Dong, and Yi Wang. ViCuR: Visual cues as recoverable privilege for multimodal on-policy distillation, 2026. URL https://arxiv.org/abs/2606.05718.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 148769–148794, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f09ab24aa37ae2664b3d923cdb0319b5-Paper-Conference.pdf.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng,

Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025a. URL https://arxiv.org/abs/2508.18265.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):7907–7915, 2025b. doi: 10.1609/aaai.v39i8.32852. URL https://ojs.aaai.org/index.php/AAAI/article/view/32852.

Shaohang Wei, Zikun Su, Feifan Song, Wen Luo, Wei Li, Guangyue Peng, and Houfeng Wang. Verifier-Induced Support Reshaping in On-Policy Optimization, 2026. URL https://arxiv.org/abs/2608.00220.

Mingyuan Wu, Jingcheng Yang, Jize Jiang, Meitang Li, Kaizhuo Yan, Hanchao Yu, Minjia Zhang, ChengXiang Zhai, and Klara Nahrstedt. VTool-R1: VLMs Learn to Think with Images via Reinforcement Learning on Multimodal Tool Use. In *International Conference on Learning Representations*, pp. 78298–78319, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/7e6f445a74cdb71931aac64f1e3f49c9-Paper-Conference.pdf.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

Weiye Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, Wenhai Wang, Jifeng Dai, and Jinguo Zhu. VisuLogic: A benchmark for evaluating visual reasoning in multi-modal large language models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 25966–26003, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2bbc73b3d3c2de43743ce2d82c8f3d7d-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 Technical Report, 2025. URL https://arxiv.org/abs/2505.09388.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-policy context distillation for language models, 2026. URL https://arxiv.org/abs/2602.12275.

Weichen Yu, Xiaomin Li, Yizhou Zhao, Xiaoze Liu, Ruowang Zhang, Haixin Wang, Yinyi Luo, Chen Henry Wu, Gaurav Mittal, Matt Fredrikson, and Yu Hu. Multi-rollout on-policy distillation via peer successes and failures, 2026. URL https://arxiv.org/abs/2605.12652.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-OPD: Learning to see fine details for multimodal LLMs via on-policy self-distillation, 2026. URL https://arxiv.org/abs/2605.18740.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Keming Lu, Chuanqi Tan, Chang Zhou, and Jingren Zhou. Scaling relationship on learning mathematical reasoning with large language models, 2023. URL https://arxiv.org/abs/2308.01825.

Kangning Zhang, Yixing Li, Shuai Shao, Qingyao Li, Zhengxi Lu, Zhiyuan Yao, Jianghao Lin, Wenxiang Jiao, Yuan Lu, Weiwen Liu, Weinan Zhang, and Yong Yu. VAD: Attributing visual evidence for target reconstruction in multimodal on-policy distillation, 2026a. URL https://arxiv.org/abs/2607.28590.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MATHVERSE: Does your multi-modal LLM truly see the diagrams in visual math problems? In *Computer Vision – ECCV 2024*, pp. 169–186. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-031-73242-3_10. URL https://link.springer.com/chapter/10.1007/978-3-031-73242-3_10.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think Beyond Images. In *International Conference on Learning Representations*, pp. 79536–79572, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8080912b919ad0344c0d34e68983dfcb-Paper-Conference.pdf.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-Distilled Reasoner: On-Policy Self-Distillation for Large Language Models. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/64784.

Binbin Zheng, Zijun Xie, Guanqun Zhao, Enlei Gong, Xing Ma, Xiaoliang Fu, and Zeyu Chen. Group-reflective self-distillation for agentic reinforcement learning, 2026a. URL https://arxiv.org/abs/2607.28076.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "Thinking with Images" via Reinforcement Learning. In *International Conference on Learning Representations*, pp. 126775–126798, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cdb347d7516d52a9280ea9d6708911f6-Paper-Conference.pdf.

APPENDIX

Contents

A Extended Method Details . . . 17
A.1 Training Procedure . . . 17
A.2 Supervision Scope . . . 18
A.3 Teacher Conditioning . . . 18
A.4 A Visual-Evidence Reflection Example . . . 19
A.5 Critic Feedback . . . 20
A.6 Critic Prompt and Output Format . . . 22
A.7 Training Configuration . . . 23
B Experimental Details . . . 24
B.1 Baseline Details . . . 24
B.2 Model Labels . . . 24
B.3 Training and Evaluation Settings . . . 25
B.4 Token-Impact Sparsity in Figure 4 . . . 25
B.5 Error Attribution in Figure 1 . . . 25
B.6 Critic Consistency Evaluation . . . 26
C Additional Results . . . 29
C.1 Per-Benchmark Component Ablation . . . 29
C.2 Token-Impact Interventions . . . 29
C.3 Sensitivity to the High-Impact Token Fraction . . . 32
C.4 Response Length and Tool Use on Perception Benchmarks . . . 33
C.5 Training-Time Reflection Coverage . . . 35
C.6 Offline Comparison with Anchor-Only Context . . . 35
D Proofs and Derivations . . . 36
D.1 Reflection-Gradient Proof . . . 36
D.2 Gradient Weights under Grouped Distillation . . . 37
E Case Study . . . 38
E.1 Errors in Acquiring, Reading, and Using Visual Evidence . . . 38
E.2 Different Errors with the Same Evidence . . . 39
E.3 Different Errors on the Same Plot . . . 40
E.4 Token Impact Visualizations . . . 41
F Evaluation Benchmarks . . . 42
F.1 Visual Perception . . . 42
F.2 Mathematical and Logical Reasoning . . . 42
F.3 General Multimodal Understanding . . . 42

A Extended Method Details

A.1 Training Procedure

Algorithm 1 summarizes training from student trajectories, critic feedback, and reflection construction to student updates. Algorithms 2–4 detail answer verification, evidence-state collection, and reflection assembly, respectively.

Algorithm 1 ReVuE: Training Procedure

Require: Data $\mathcal{D}$, trainable student π_θ, frozen teacher q_ϕ, fixed critic $\mathcal{C}$
Require: Rollouts per query M, support size K, high-impact fraction α, group weight λ
Ensure: Updated student π_θ
All tokenwise operations use $t \in \mathcal{T}_i$.
Superscript Supp denotes renormalization over Supp; impacts use full-vocabulary probabilities.

for each training minibatch $\mathcal{X} \subset \mathcal{D}$ **do**
  for each $(x, y^\star) \in \mathcal{X}$ **do** — *Sampling from student*
    $\mathcal{G}_x \leftarrow \textsc{Rollout}(\pi_\theta, x, M)$
    $\mathbf{c}_x \leftarrow \textsc{VerifyAnswers}(\mathcal{G}_x, y^\star)$; $\mathcal{E}_x \leftarrow \textsc{ObservedStates}(\mathcal{G}_x)$ — *Reflection via critic*
    $\mathcal{J}_x \leftarrow \mathcal{C}(x, y^\star, \mathcal{G}_x, \mathbf{c}_x, \mathcal{E}_x)$
    $\{\mathcal{R}_i\}_{i=1}^{M} \leftarrow \textsc{CheckAndAssemble}(\mathcal{J}_x, \mathcal{E}_x, \mathbf{c}_x)$
  end for
  Let $\mathcal{G}$ collect retained trajectories across $\mathcal{X}$ with $\mathcal{T}_i \neq \emptyset$
  for each $\tau_i \in \mathcal{G}$ **do**
    $q^0_{i,t} \leftarrow q_\phi(\cdot \mid h_{i,t})$; $q^*_{i,t} \leftarrow q^0_{i,t}$; $s_i \leftarrow$ unavailable
    if $\mathcal{R}_i \neq \emptyset$ **then**
      $q^{\mathcal{R}}_{i,t} \leftarrow q_\phi(\cdot \mid h_{i,t}, \mathcal{R}_i)$
      if evaluation succeeds and completion tokens align **then**
        $q^*_{i,t} \leftarrow q^{\mathcal{R}}_{i,t}$
        Attempt $s_{i,t} \leftarrow \left|\log q^{\mathcal{R}}_{i,t}(y_{i,t}) - \log q^0_{i,t}(y_{i,t})\right|$
      end if
    end if
    $p_{i,t} \leftarrow \pi_\theta(\cdot \mid h_{i,t})$; $\mathrm{Supp}_{i,t} \leftarrow \mathrm{TopK}(q^*_{i,t})$
    $\ell_{i,t} \leftarrow \mathbb{D}^{\mathrm{KL}}\left(p^{\mathrm{Supp}_{i,t}}_{i,t} \,\middle\|\, (q^*_{i,t})^{\mathrm{Supp}_{i,t}}\right)$
    Compute ℓ_i using Equation 11
  end for
  $\mathcal{L} \leftarrow \mathrm{mean}_{\tau_i \in \mathcal{G}}\, \ell_i$ ▷ 0 if $\mathcal{G} = \emptyset$
  $\theta \leftarrow \textsc{OptimizerStep}(\theta, \nabla_\theta \mathcal{L})$
  Refresh the rollout policy from θ
end for

Teacher evaluation Reweighting Gradient update

Let $u_i = 1$ indicate a valid reflection, a successful teacher evaluation, and aligned completion tokens; otherwise set $u_i = 0$. Let $b_i = 1$ indicate that impacts were also computed successfully and at least two supervised positions exist; otherwise set $b_i = 0$. An impact-computation failure retains the adopted reflection-conditioned target and replaces group weighting with uniform averaging.

$$q^*_{i,t} = \begin{cases} q^{\mathcal{R}}_{i,t}, & u_i = 1, \\ q^0_{i,t}, & u_i = 0. \end{cases} \tag{10}$$

Let $\ell_{i,t} = \mathbb{D}^{\mathrm{KL}}(p^{\mathrm{Supp}_{i,t}}_{i,t} \| (q^*_{i,t})^{\mathrm{Supp}_{i,t}})$ denote the per-position KL term in Equation 8. Each trajectory with a nonempty supervised set has loss

$$\ell_i = \begin{cases} \dfrac{\lambda}{|\mathcal{T}_{i,\mathrm{High}}|} \displaystyle\sum_{t \in \mathcal{T}_{i,\mathrm{High}}} \ell_{i,t} + \dfrac{1-\lambda}{|\mathcal{T}_{i,\mathrm{Low}}|} \sum_{t \in \mathcal{T}_{i,\mathrm{Low}}} \ell_{i,t}, & b_i = 1, \\ \dfrac{1}{|\mathcal{T}_i|} \displaystyle\sum_{t \in \mathcal{T}_i} \ell_{i,t}, & b_i = 0. \end{cases} \tag{11}$$

Impacts are sorted in descending order with earlier positions breaking ties, and the high-impact group contains the first $\lceil \alpha |\mathcal{T}_i| \rceil$ positions. The final loss averages retained trajectories equally as in Equation 8.

A.2 Supervision Scope

Supervised positions cover student-generated reasoning, code, and final answers across all interaction rounds, with other labels set to -100:

$$m_{i,t} = \mathbf{1}\{\text{label}_{i,t} \neq -100\}, \qquad \mathcal{T}_i = \{t : m_{i,t} = 1\}, \qquad |\mathcal{T}_i| = \sum_t m_{i,t}. \tag{12}$$

Prompts, tool-returned text, image placeholders, sandbox boundary markers, and interceptor-appended suffixes are excluded from supervision targets. Tool observations remain in the input history for subsequent student predictions. Impact ranking, group sizes, and loss means use only positions in $\mathcal{T}_i$.

A.3 Teacher Conditioning

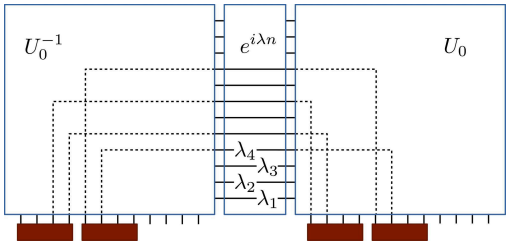
The base teacher scores the fixed student completion token by token, conditioned on the original task prompt, image, and preceding student interaction history. The reflection-conditioned teacher appends reflection text to the first user message and inserts supporting images after the original image and before the student's trajectory images. The re-encoded completion must match the student's sequence in length and in every token ID.

Reflection is generated after the rollout group is complete, so supporting evidence may come from later steps of the target trajectory or from a sibling trajectory. This retrospective information is used only for teacher supervision, while the student prefix retains its original visible history. The reference answer informs critic diagnosis, and the resulting visual fact, grounding rule, or discrepancy may convey answer semantics. Comparing the two teacher evaluations measures the effect of this added context, including both text and supporting images.

Impact uses full-vocabulary normalized log probabilities at the sampled token and is detached after computation. RKL renormalizes both distributions over the adopted teacher's top-K candidates, without forcibly adding the sampled token. Impact ranks and weights training positions, while the loss at each position supervises the entire support distribution (Appendix D.1).

A.4 A Visual-Evidence Reflection Example

Task and visual evidence



Original image

What does the central vertical line with the notation $e^{i\lambda n}$ in the figure likely represent?

A. A series of discrete energy levels in a quantum system

B. A variable resistor in an electrical circuit

C. The eigenvalues of a Hamiltonian in a solid-state physics model

D. The time evolution operator in a quantum mechanical system

Correct attempt

Post-tool reasoning

The central vertical line with the notation $e^{i\lambda n}$ typically represents a phase factor in quantum mechanics. This phase factor often arises from the time evolution of quantum states under a Hamiltonian, indicating how a state evolves over time according to the principles of quantum mechanics.

Answer: D

Incorrect attempt

Post-tool reasoning

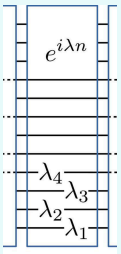
The figure shows a ladder diagram with levels labeled by λ, typically representing eigenvalues in physics, specifically in the context of quantum mechanics or solid-state physics where λ could represent different energy states.

Answer: C

Both attempts received the same tool crop, shown below.

Reflection supplied to the teacher for the incorrect attempt

Anchor



Supporting tool crop

Acquire

Target: central vertical line notation.

Read Observed value: $e^{i\lambda n}$

The central vertical line is labeled with the mathematical expression $e^{i\lambda n}$, positioned between U_0^{-1} and U_0 blocks.

Ground

In quantum mechanics, the expression $e^{i\lambda n}$ (where λ relates to eigenvalues/energy and n to time/steps) represents the time evolution operator acting on a state, especially when flanked by unitary transformations U_0 and U_0^{-1}.

Break Point

First broken link: *Ground*

Discrepancy

The rollout correctly identifies the notation and the ladder diagram context but incorrectly maps the visual evidence to 'eigenvalues of a Hamiltonian' (Option C) instead of recognizing the exponential form as a time evolution operator (Option D).

Figure 6: **A concrete visual-evidence reflection supplied to the teacher.** The mixed group contains seven correct attempts and one incorrect attempt; one of each is shown with its post-tool reasoning and answer. Both attempts receive identical tool crops. The Anchor uses the incorrect attempt's own crop, and the Break Point records the critic's *Ground* diagnosis.

A.5 CRITIC FEEDBACK

VERIFYANSWERS. Algorithm 2 computes binary answer labels, which classify each rollout group as all-correct, mixed, or all-wrong. When deterministic matching does not accept an answer, the answer judge evaluates only the question, extracted student answer, and reference answer as text.

Algorithm 2 VERIFYANSWERS: Answer verification

Require: Same-query trajectories $\mathcal{G}_x$, reference answer $y^\star$
Ensure: Binary correctness labels $\mathbf{c}_x$
$q \leftarrow$ shared question in $\mathcal{G}_x$; $\mathbf{c}_x \leftarrow \mathbf{0}$
for each trajectory $i \in \mathcal{G}_x$ **do**
  $a_i \leftarrow$ text inside `<answer>...</answer>`
  if a_i is missing or empty **then continue**
  $\tilde{a}_i \leftarrow$ strip answer prefixes from a_i
  if strict mathematical matching or symbolic verification accepts $(\tilde{a}_i, y^\star)$ **then**
    $c_i \leftarrow 1$
  else
    $c_i \leftarrow \text{ANSWERJUDGE}(q, a_i, y^\star) \in \{0, 1\}$ ▷ 0 if unavailable
  end if
end for
return $\mathbf{c}_x$

OBSERVEDSTATES. Algorithm 3 collects the original-image state and recorded post-tool states so the critic can associate visual evidence with a specific interaction step. Each state contains the images available up to that point, with multiple images from one tool call belonging to the same state. The counts $n_{\text{tool}}(e)$ and $n_{\text{code}}(e)$ record cumulative tool rounds and code tokens, respectively.

Algorithm 3 OBSERVEDSTATES: Evidence-state collection

Require: Same-query trajectories $\mathcal{G}_x$, including the original image and interaction records
Ensure: Evidence states $\mathcal{E}_x$ with provenance, available images, and cumulative costs
$\mathcal{E}_x \leftarrow \{\text{original-image state; owner } \emptyset,\ n_{\text{tool}} = n_{\text{code}} = 0\}$
for each recorded post-tool state e of trajectory $i \in \mathcal{G}_x$ **do**
  Add $(e, i,$ images available at $e, n_{\text{tool}}(e), n_{\text{code}}(e))$ to $\mathcal{E}_x$
end for
return $\mathcal{E}_x$

CHECKANDASSEMBLE. The critic receives the query, reference answer, original image, correctness-labeled trajectories, and evidence states. Up to eight tool images are supplied per group, prioritizing correct trajectories. For each sufficient state e, the critic is asked to identify a smallest supporting image subset $\mathcal{U}_e$ that determines the required visual fact using only images available at that state. It also returns shared reading information (query slot, observed value, and visual fact), a grounding rule, and per-trajectory diagnoses $\mathcal{B}_i = (\sigma_i, \delta_i)$.

Algorithm 4 validates required fields, state and image references, and consistency between diagnoses and answer labels. Only mixed groups reconcile inconsistent stage labels before validation using answer correctness and the presence of an own sufficient state, while retaining the discrepancy text. All-correct groups require `CORRECT` labels with empty discrepancies.

Anchor selection prioritizes sufficient evidence already acquired by the trajectory and chooses among states by lexicographic cost:

$$c(e) = \big(n_{\text{tool}}(e),\ n_{\text{code}}(e),\ |\mathcal{U}_e|,\ \text{id}(e)\big), \tag{13}$$

Let $\mathcal{C}_i$ contain sufficient post-tool states owned by trajectory i, and let $\mathcal{C}_x^+$ contain sufficient original-image states and sufficient states from answer-correct trajectories. Complete reflections require a nonempty group reference set $\mathcal{C}_x^+$, after which each trajectory selects evidence as

$$e_i^\star = \begin{cases} \arg\min_{e \in \mathcal{C}_i} c(e), & \mathcal{C}_i \neq \emptyset, \\ \arg\min_{e \in \mathcal{C}_x^+} c(e), & \mathcal{C}_i = \emptyset. \end{cases} \tag{14}$$

If the selected state requires no tool use, assembly adds no image because the teacher context already contains the original.

Algorithm 4 CHECKANDASSEMBLE: Reflection assembly

Require: Critic output $\mathcal{J}_x$, recorded states $\mathcal{E}_x$, correctness labels $\mathbf{c}_x$
Ensure: Per-trajectory reflections $\{\mathcal{R}_i\}$, with $\emptyset$ denoting unavailable reflection

$\mathcal{R}_i \leftarrow \emptyset$ for all i; **if** $\mathcal{J}_x$ is missing **then return** $\{\mathcal{R}_i\}$
if the group is mixed **then** reconcile stages with $\mathbf{c}_x$ and own sufficient states
if $\mathcal{J}_x$ is inapplicable or fails validation **then return** $\{\mathcal{R}_i\}$
$\mathcal{B}_i \leftarrow (\sigma_i, \delta_i)$ from the checked diagnoses, for all i
if all answers are wrong **then return** $\{(\emptyset, \mathcal{B}_i)\}_i$
Resolve $\mathcal{J}_x$'s sufficient states in $\mathcal{E}_x$ to form $\{\mathcal{C}_i\}$ and $\mathcal{C}_x^+$
if $\mathcal{C}_x^+ = \emptyset$ **then return** $\{\mathcal{R}_i\}$
for each trajectory i **do**
    Select $e_i^\star$ using Equation 14
    $\mathcal{A}_i \leftarrow (e_i^\star, \mathcal{U}_{e_i^\star}, \mathcal{J}_x.\text{reading}, \mathcal{J}_x.\text{grounding})$
    $\mathcal{R}_i \leftarrow (\mathcal{A}_i, \mathcal{B}_i)$
end for
return $\{\mathcal{R}_i\}$

A.6 Critic Prompt and Output Format

The box below reproduces the critic system prompt used for mixed groups. The all-correct prompt requires `CORRECT` with an empty discrepancy for every trajectory. The all-wrong prompt disallows `CORRECT` and permits an empty set of sufficient evidence states. It also requires a sufficient state for each `READ` or `GROUND` diagnosis, preserves verifier-label conflicts, and returns one diagnosis per trajectory.

Critic system prompt: mixed groups

You are a training-time multimodal evidence-chain analyst for a think-with-image VQA policy.

You receive one task and a group of verifier-labeled CORRECT and INCORRECT student rollouts.

A rollout contains chronological policy reasoning, free-form Python code, sandbox text/numeric outputs, zero/one/multiple images returned by each sandbox interaction, and a final answer.

One Python code block may return MULTIPLE images. They belong to the SAME sandbox interaction and may jointly form multi-region evidence.

A rollout may have multiple sandbox interactions. Later interactions may correct earlier visual observations.

The CORRECT/INCORRECT label is defined only by the supplied VQA correctness label. Trust CORRECT rollouts as correct final solutions.

Your task is to identify the visual evidence chain:

ACQUIRE:
Which actually observed Evidence States already contain sufficient task-relevant visual evidence?

READ:
What is the minimum atomic visual fact needed by the question?

GROUND:
How does that visual fact map to the answer semantics?

EVIDENCE STATE:
ORIGINAL_STATE contains ORIGINAL_IMAGE. Every other state is the state immediately after one real sandbox interaction and lists all images currently available.

A state is SUFFICIENT iff some subset of its available images already contains enough visual evidence to determine the task-relevant visual fact without needing a later observation.

For each sufficient state, return the SMALLEST supporting image subset.

CLOSED-SET RULE:
Never invent an image, state, crop, coordinate, code, or synthetic view. Only use supplied state IDs and image IDs.

READ:
Do not use a predefined taxonomy. Return:
- query_slot
- observed_value
- one atomic visual_fact

GROUND:
Return one concise evidence-to-answer semantic rule. Do not state only an answer/option letter.

ROLLOUT DIAGNOSIS:
CORRECT rollout -> stage=CORRECT.

For INCORRECT rollouts return the FIRST broken interface:

Critic system prompt: mixed groups (continued)

ACQUIRE:
No state belonging to this rollout is sufficient.

READ:
A sufficient state exists, but the relevant visual fact is misread.

GROUND:
A sufficient state exists and the fact is read correctly, but final answer semantics contradict/mis-map that fact.

OTHER:
Failure is outside this visual evidence chain.

Return JSON only according to schema.

The output schema is supplied as a separate structured-output constraint rather than as part of the user message. The table retains its fields, types, and shared constraints.

Critic output schema: fields and constraints

Field structure	Type
`applicable`	boolean
`sufficient_states`	object[]
`  state_id`	string
`  support_image_ids`	string[]
`reading`	object
`  query_slot`	string
`  observed_value`	string
`  visual_fact`	string
`grounding`	object
`  rule`	string
`rollout_diagnoses`	object[]
`  rollout_id`	string
`  stage`	string (enum)
`  discrepancy`	string

Allowed values for `rollout_diagnoses[].stage`:
`CORRECT, ACQUIRE, READ, GROUND, OTHER`

The root is an object; `[]` denotes an array, with item fields indented below it.

All fields are required; additional properties are disallowed for every object; `strict = true`.

A.7 Training Configuration

The full method described here uses $M = 8$ trajectories per query, $K = 32$ teacher candidates, a high-impact fraction $\alpha = 0.2$, and group weight $\lambda = 0.5$. Teacher parameters, teacher supports, impacts, and group assignments remain fixed during student updates. Training uses the pure OPD loss, without a GRPO reward term or an additional reference-policy KL. Answer correctness is used for group classification and critic diagnosis, not as an advantage or a trajectory-level loss weight. For the Thyme-based Qwen2.5-VL-7B setting, the student starts from an SFT checkpoint, and the frozen teacher uses step 2150 of the corresponding RL expert. The separate critic uses Qwen3.5-397B-A17B-FP8 with temperature zero and a 2048-output-token limit per call to generate JSON-schema-constrained feedback.

Critic runtime and auxiliary resources. Training uses eight NVIDIA H20 GPUs (96 GB each), and the critic runs through a shared vLLM service on a separate machine with another eight H20 GPUs (96 GB each). This deployment leaves the training GPUs available for rollouts and allows

critic requests to overlap with independent training-side computations, such as reward and answer-correctness evaluation. The critic is used only during training and is not required at inference.

Each critic request processes one group of eight trajectories, giving 20 requests per optimizer step. A test request with one image and eight trajectories took 5.8 s and produced 664 output tokens. Assuming this latency is maintained with four concurrent requests, five waves give a nominal critic service time of approximately 29 s per step, excluding additional queueing and retries. This estimate is 10.8% of the mean step time of 269 s measured across six training segments. The actual increase in training wall-clock time depends on overlap with independent computations and service scheduling.

Extrapolating to 1,075 steps gives approximately 8.7 hours of nominal critic service time. Attributing all eight service GPUs to this run during these intervals gives a resource estimate of approximately 70 GPU-hours. This estimate accounts for nominal service time; total server allocation depends on service uptime and sharing.

B Experimental Details

B.1 Baseline Details

All baselines start from Thyme-SFT cold-start checkpoints of Qwen2.5-VL-7B or InternVL3.5-4B-Instruct. OPD variants use the corresponding Thyme-RL Expert as teacher; RFT uses demonstrations generated by the same expert. This keeps student initializations and teacher sources consistent across comparisons.

Vanilla OPD (Agarwal et al., 2024) provides the standard on-policy distillation reference for evaluating reflection-based supervision. The RL Expert supervises student-generated prefixes token by token, without ground-truth answers, privileged crops, or evidence reflections.

Rejection sampling fine-tuning (RFT) (Yuan et al., 2023) compares offline demonstration learning with distillation on student-generated trajectories. For each prompt in the Thyme-RL training set, the corresponding RL Expert samples eight trajectories. We retain answer-correct trajectories for supervised fine-tuning of the same cold-start student. Training uses these fixed demonstrations rather than teacher distributions queried on current student prefixes.

GT-Privileged compares ground-truth answers with visual-evidence reflections as teacher privilege. Following the privileged-context design of OPSD (Zhao et al., 2026), we provide the teacher with the Thyme-RL ground-truth answer. We replace its self-teacher with the corresponding RL Expert. The student starts from the same cold-start checkpoint and generates trajectories without access to the privileged answer.

Vision-OPD (Yuan et al., 2026) provides a local visual privilege baseline for comparison with reflections that diagnose evidence use. The teacher receives an evidence-centered crop and supervises the student's own trajectories. We replace its self-teacher with the corresponding RL Expert and use the same cold-start student.

With **V-Zero** (Sun et al., 2026), we compare visual-contrast trajectory weighting with reflection-based supervision. Teacher support for the same student trajectory under relevant and irrelevant visual views determines its distillation weight. The target remains the positive-view teacher distribution. We use the corresponding RL Expert for both views and the same cold-start student.

Our comparison with **VAD** (Zhang et al., 2026a) examines visual target reconstruction as an alternative to direct distillation from a reflection-conditioned teacher. It contrasts teacher predictions with and without visual evidence to estimate visual corrections and reconstruct the distillation target. We replace its initial-model teacher with the corresponding RL Expert and retain the same cold-start student.

B.2 Model Labels

In Table 1, Base Model denotes the original model before training in this work. Gemini3.1FL abbreviates Gemini-3.1-Flash-Lite. Qwen2.5-32B, Qwen3-30B-T, and InternVL-38B denote Qwen2.5-VL-32B, Qwen3-VL-30B-A3B-Thinking, and InternVL3.5-38B-Instruct, respectively.

B.3 Training and Evaluation Settings

We initialize the Qwen2.5-VL-7B student directly from the released Thyme-SFT (Zhang et al., 2026b) checkpoint.[1] We obtain the InternVL3.5-4B cold-start checkpoint by supervised fine-tuning of InternVL3.5-4B-Instruct on Thyme's released SFT dataset.[2] We train both RL experts from their corresponding cold-start checkpoints by reproducing Thyme-RL on its released RL dataset.[3]

Table 6: Training and evaluation settings.

Component	Parameter	Value
OPD objective	Loss	Reverse KL
	Vocabulary support	Teacher's top-32 candidates
OPD rollouts	Temperature for code tokens	0.0
	Temperature for other tokens	1.0
	Top-p	0.9
	Top-k	50
Evaluation	Temperature	0.01
	Top-p	0.001
	Top-k	1
All experiments	Random seed	42
Training job	Hardware	8 × NVIDIA H20 (96 GB each)
Critic service	Separate hardware	8 × NVIDIA H20 (96 GB each)

We compute reverse KL after renormalizing teacher and student distributions over the teacher's top-32 vocabulary candidates at each supervised position.

B.4 Token-Impact Sparsity in Figure 4

Figure 4 pools offline rescoring results from 632 trajectories in a five-step early-training archive, comprising 226,726 supervised tokens with valid impact scores. We use the impact definition in Equation 7 and exclude nonsupervised positions and missing scores. The dashed line marks the pooled 80th percentile, approximately $s = 0.05524$. Tokens at or above this threshold comprise approximately 20% of scored tokens and account for 97.6% of total impact. High and Low in the figure describe the pooled distribution; training selects high-impact positions separately within each trajectory.

Histogram heights give the percentage of all scored tokens in each interval. Bins in $[0, 2.5)$ have width 0.05, with the bin containing the 80th percentile split at that threshold. The final bar retains all tokens with $s \geq 2.5$. The inset plots empirical percentile against impact over the full pooled score range, without smoothing or fitting.

B.5 Error Attribution in Figure 1

The error-attribution analysis in Figure 1 uses 3,710 questions from V*Bench (Wu & Xie, 2024), HRBench-4K (Wang et al., 2025b), HRBench-8K (Wang et al., 2025b), and MME (Fu et al., 2025).[4] We use GPT-5.6-Luna as the judge for evaluation and error attribution.

Four error labels map to the three visual-evidence stages. *Acquire* combines `faulty_tool` (29.8%) and `missing_tool` (7.9%), totaling 37.7%. *Read* corresponds to `perception_error` (33.8%). *Ground* corresponds to `hallucination` (6.3%).

[1] Thyme-SFT model: https://huggingface.co/Kwai-Keye/Thyme-SFT
[2] Thyme-SFT dataset: https://huggingface.co/datasets/Kwai-Keye/Thyme-SFT
[3] Thyme-RL dataset: https://huggingface.co/datasets/Kwai-Keye/Thyme-RL
[4] MME: official dataset repository.

B.6 CRITIC CONSISTENCY EVALUATION

We evaluate the final training labels of Qwen3.5-397B-A17B-FP8 on 96 groups of eight rollouts. Each group type (all-correct, mixed, and all-wrong) contributes 32 groups. One annotator independently labels one sampled rollout per group before revealing the critic label; all 96 samples receive a label. Agreement measures exact stage matches; unweighted Cohen's κ adjusts for chance agreement.

Table 7: **Final critic stage agreement.**

(a) Human–critic label counts

Human \ Critic	Correct	Acquire	Read	Ground	Other
Correct	32	0	0	0	0
Acquire	0	28	0	0	0
Read	0	3	23	0	0
Ground	0	0	0	9	0
Other	0	0	0	0	1

(b) Pairwise final-label agreement

Pair	n	Agree. (%)	κ
Qwen (arch.) / rerun	744	99.2	0.987
Qwen (arch.) / Gemini	736	93.2	0.892
Qwen (arch.) / Luna	736	92.8	0.886
Qwen (rerun) / Gemini	712	93.5	0.898
Qwen (rerun) / Luna	712	93.4	0.896
Gemini / Luna	712	95.4	0.926

Arch. denotes archived training labels, rerun denotes one fresh evaluation, and n counts valid paired rollouts. Gemini and Luna denote Gemini 3.8 Flash and GPT-5.6 Luna, respectively.

Human agreement is 96.88% ($\kappa = 0.957$) overall and 95.31% ($\kappa = 0.925$) on incorrect rollouts.

The three human–critic disagreements occur at the *Acquire–Read* boundary. All three are from mixed groups: the annotator assigns *Read*, whereas the critic assigns *Acquire*. This points to disagreement over whether the available evidence is sufficient to support a judgment. Agreement with archived training labels uses 744 valid rollout pairs for a rerun by the same critic and 736 for each of the two external judges (Table 7). These comparisons use valid final stage labels.

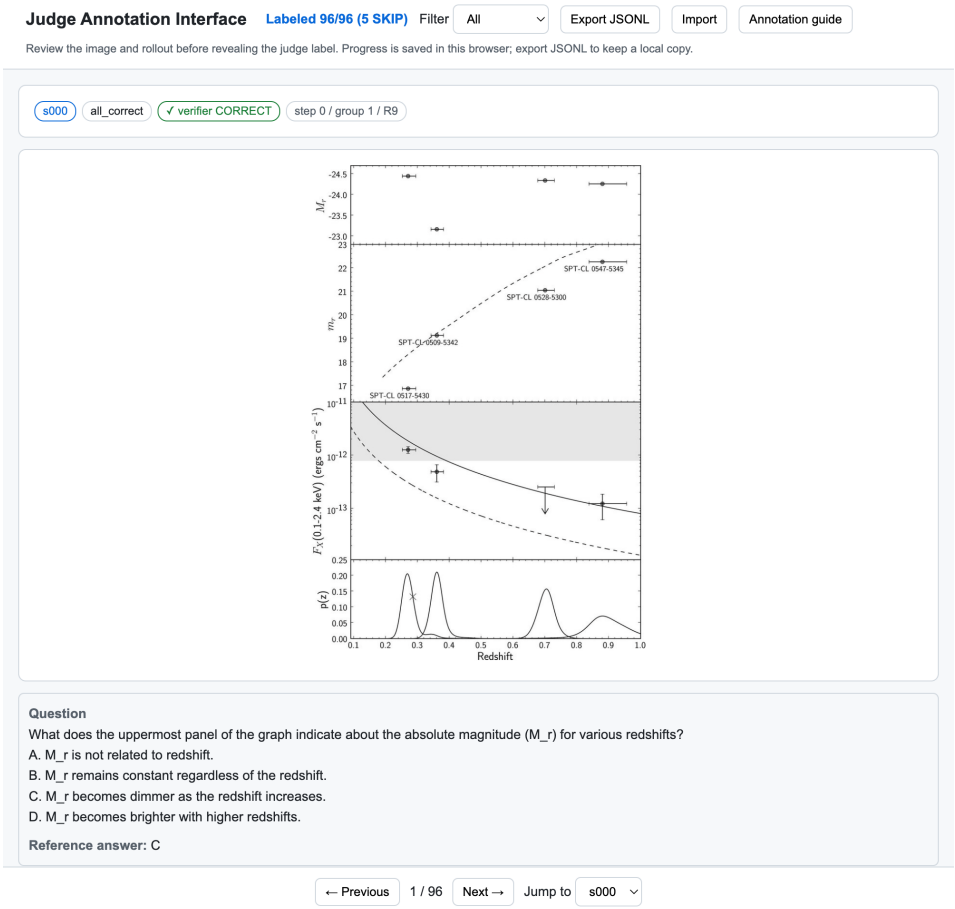


Figure 7: Human annotation interface demo: the upper view presents the sample identifiers, original image, question, and reference answer. The page also provides progress tracking, sample filtering, navigation, and JSONL import and export controls.

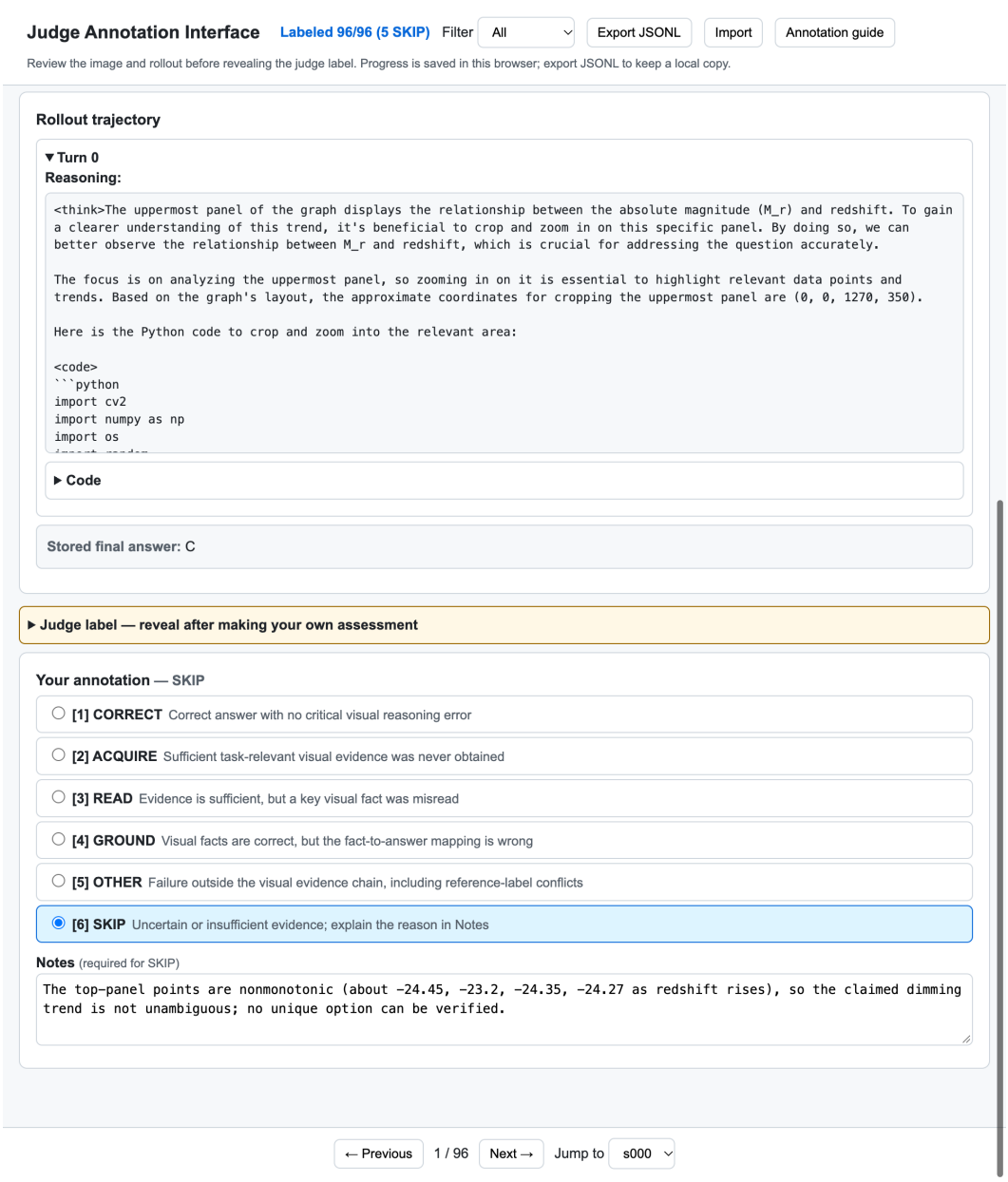


Figure 8: Human annotation interface demo: the lower view of the same sample presents the rollout, the initially collapsed judge label, six annotation options, and a notes field. The options comprise five diagnostic labels and a `SKIP` option for cases that cannot be judged reliably.

C Additional Results

This appendix reports per-benchmark component ablations, token-impact interventions, fraction sensitivity, and additional analyses of response length and tool use.

C.1 Per-Benchmark Component Ablation

Table 8 reports scores on all 11 benchmarks for the three variants in Table 2.

Reflection improves most benchmarks, with uneven gains. Adding reflection to Vanilla OPD improves nine of 11 benchmarks and all three category averages. Perception and general-task averages increase by 1.09% and 1.14%, respectively, while math improves by 0.11%. Gains on HRBench 8K (+2.00%) and HallusionBench (+2.54%) cover high-resolution understanding and hallucination detection. TreeBench and MathVerse are the two exceptions, with accuracy decreases of 0.24% and 0.51%, respectively.

Impact reweighting improves tasks that do not benefit from reflection alone. Relative to the reflection-only model, reweighting improves seven of eight perception and math benchmarks, while V* Bench remains at 82.20%. TreeBench rises from 37.78% to 41.12% and MathVerse from 45.56% to 47.08%, reversing their declines relative to Vanilla OPD. HRBench 8K and VisualProbe also improve by 1.50% and 0.58%, extending the reweighting gains to tasks that already benefit from reflection.

Reweighting introduces a small trade-off on general tasks. The full model scores below the reflection-only model on all three general-task benchmarks, lowering the category average by 0.28%. The decreases are 0.53% on ChartQA-Pro, 0.30% on HallusionBench, and 0.08% on InfographicVQA. All three remain above Vanilla OPD, and the full method improves over Vanilla OPD on all 11 benchmarks.

Table 8: **Complete component ablation on Qwen2.5-VL-7B.** Rows cumulatively add visual-evidence reflection and impact reweighting, and Wtd. Avg. uses the benchmark-sample weighting of Table 1. MathVista and MathVerse denote the Mini and vision-only splits, respectively. Higher is better for every metric; gray denotes Vanilla OPD, blue marks ReVuE, and bold indicates column-best results, including ties.

(a) Perception

Method	HR-4K	HR-8K	V*	Tree Bench	Visual Probe	Wtd. Avg.
Vanilla OPD	75.40	70.50	81.20	38.02	42.91	62.61
+ Visual-evidence reflection	76.25	72.50	**82.20**	37.78	44.08	63.70
+ Impact reweighting (ReVuE)	**77.10**	**74.00**	**82.20**	**41.12**	**44.66**	**65.01**

(b) Math

Method	MathVista	MathVerse	VisuLogic	Wtd. Avg.
Vanilla OPD	70.60	46.07	26.00	47.67
+ Visual-evidence reflection	71.10	45.56	26.20	47.78
+ Impact reweighting (ReVuE)	**71.60**	**47.08**	**26.50**	**48.49**

(c) General

Method	Hallusion Bench	ChartQA-Pro	Infographic VQA	Wtd. Avg.
Vanilla OPD	42.88	21.69	79.44	53.28
+ Visual-evidence reflection	**45.42**	**22.34**	**80.35**	**54.42**
+ Impact reweighting (ReVuE)	45.12	21.81	80.27	54.14

C.2 Token-Impact Interventions

Table 9 expands Table 3 with scores on all 11 benchmarks. Table 10 summarizes how the four variants and ReVuE differ in token selection and masking.

Masking low-impact positions does not benefit every task. Compared with REVUE, *Mask low 20%* raises the perception average by 0.17%. Math and general-task averages decrease by 0.07% and 0.17%, respectively. Within perception, HRBench 8K and VisualProbe improve by 1.00% and 0.78%, respectively, while TreeBench drops by 1.61%. The small gain in average perception does not show that low-impact supervision is generally harmful. Across all 11 benchmarks, *Mask low 20%* exceeds *Mask top 20%*, supporting impact ranking as an indicator of supervision value.

Table 9: **Complete token-impact intervention results on Qwen2.5-VL-7B.** Wtd. Avg. uses the benchmark-sample weighting of Table 1. MathVista and MathVerse denote the Mini and vision-only splits, respectively. All scores are percentages and higher is better; gray denotes Vanilla OPD, blue marks REVUE, and bold indicates column-best results, including ties.

(a) Perception

Method	HR-4K	HR-8K	V*	Tree Bench	Visual Probe	Wtd. Avg.
Vanilla OPD	75.40	70.50	81.20	38.02	42.91	62.61
Random select 20%	75.40	72.62	80.10	38.02	**45.44**	63.64
Mask random 20%	76.25	73.75	81.20	37.28	43.69	63.85
Mask low 20%	77.00	**75.00**	**82.20**	39.51	**45.44**	**65.18**
Mask top 20%	75.00	71.62	78.01	37.78	41.36	62.26
REVUE(Ours)	**77.10**	74.00	**82.20**	**41.12**	44.66	65.01

(b) Math

Method	MathVista	MathVerse	VisuLogic	Wtd. Avg.
Vanilla OPD	70.60	46.07	26.00	47.67
Random select 20%	71.30	45.56	25.00	47.42
Mask random 20%	70.90	44.42	26.10	47.35
Mask low 20%	**71.80**	**47.08**	26.10	48.42
Mask top 20%	71.30	45.94	24.00	47.17
REVUE(Ours)	71.60	**47.08**	**26.50**	**48.49**

(c) General

Method	Hallusion Bench	ChartQA-Pro	Infographic VQA	Wtd. Avg.
Vanilla OPD	42.88	21.69	79.44	53.28
Random select 20%	44.94	21.87	80.21	54.10
Mask random 20%	43.58	21.87	80.08	53.78
Mask low 20%	44.77	**21.96**	79.93	53.97
Mask top 20%	42.16	20.79	78.75	52.51
REVUE(Ours)	**45.12**	21.81	**80.27**	**54.14**

Table 10: **Variant definitions for token-impact interventions.** Percentages refer to supervised positions, and masking removes the selected positions from the distillation loss. Random select replaces the impact-selected high-weight group with a random group while retaining supervision at every position.

Method	Masked positions	Retained supervision
Random select 20%	None	All positions; random 20% and remaining 80% receive group loss weights $0.5 : 0.5$
Mask random 20%	Random 20%	Remaining 80% receive total group loss weight 0.5, shared uniformly
Mask low 20%	Lowest-impact 20%	Highest-impact 80%, including the entire High group
Mask top 20%	Highest-impact 20%	Lowest-impact 80%; the entire High group is masked
REVUE(Ours)	None	All positions; highest-impact 20% and remaining 80% receive group loss weights $0.5 : 0.5$

C.3 Sensitivity to the High-Impact Token Fraction

Table 11 reports all 11 benchmark scores for the five fractions in Table 4. The fraction α controls the share of supervised positions assigned to the high-impact group, as defined in Section 4.3. The blue row marks the 20% setting used by ReVuE.

The best fraction varies across benchmarks. Among the tested fractions, 20% achieves the highest score on all three math benchmarks. It also leads on HRBench 4K, HRBench 8K, and TreeBench. V* Bench scores highest at 5%, while VisualProbe scores highest at 50%. The general-task average is highest at 100%. In these experiments, 20% yields the highest perception and math averages, supporting its use as a common setting for both categories.

Table 11: **Sensitivity to the high-impact token fraction on Qwen2.5-VL-7B.** Wtd. Avg. is weighted by benchmark sample counts. Sample counts are (800, 800, 191, 405, 515), (1000, 788, 1000), and (1129, 1948, 2801) for perception, math, and general benchmarks, respectively. MathVista and MathVerse denote the Mini and vision-only splits, respectively. All scores are percentages and higher is better; blue marks the 20% setting of ReVuE, and bold indicates column-best results, including ties.

(a) Perception

High-impact fraction	HR-4K	HR-8K	V*	Tree Bench	Visual Probe	Wtd. Avg.
Top 5%	77.00	73.12	**83.25**	40.49	44.66	64.70
Top 10%	75.88	73.75	82.72	40.99	44.47	64.55
Top 20% (ReVuE)	**77.10**	**74.00**	82.20	**41.12**	44.66	**65.01**
Top 50%	76.50	73.50	80.63	39.26	**44.85**	64.33
Top 100%	76.25	72.50	82.20	37.78	44.08	63.70

(b) Math

High-impact fraction	MathVista	MathVerse	VisuLogic	Wtd. Avg.
Top 5%	71.10	46.32	24.80	47.49
Top 10%	71.10	44.29	24.80	46.92
Top 20% (ReVuE)	**71.60**	**47.08**	**26.50**	**48.49**
Top 50%	71.10	45.30	25.40	47.42
Top 100%	71.10	45.56	26.20	47.78

(c) General

High-impact fraction	Hallusion Bench	ChartQA-Pro	Infographic VQA	Wtd. Avg.
Top 5%	**46.16**	21.74	79.84	54.12
Top 10%	44.13	21.53	80.25	53.85
Top 20% (ReVuE)	45.12	21.81	80.27	54.14
Top 50%	45.66	21.86	80.09	54.18
Top 100%	45.42	**22.34**	**80.35**	**54.42**

C.4 Response Length and Tool Use on Perception Benchmarks

The following five figure pairs report Qwen2.5-VL-7B results on all five perception benchmarks. Each pair shows overall accuracy versus mean response length on the left, and tool-call rate and accuracy on samples with and without tool use on the right. Arrows report RL Expert's response length divided by that of ReVuE; all bar-chart metrics are percentages.

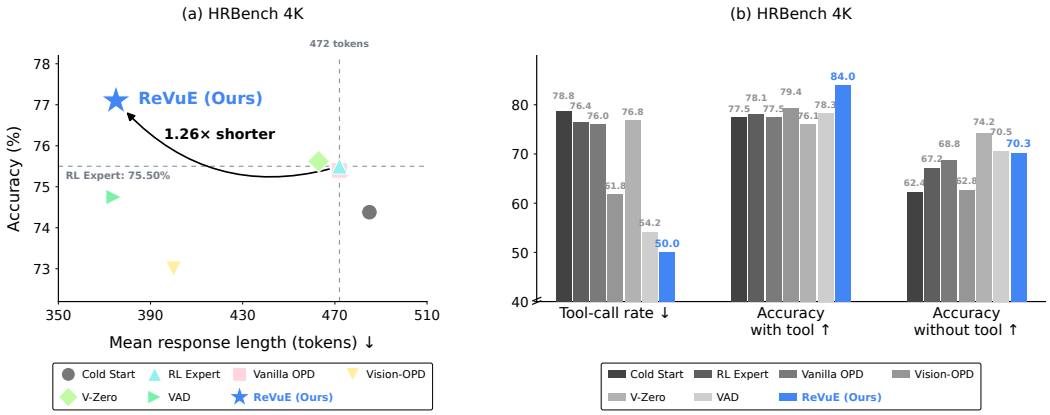


Figure 9: **Response length, accuracy, and tool use on HRBench 4K** for Qwen2.5-VL-7B.

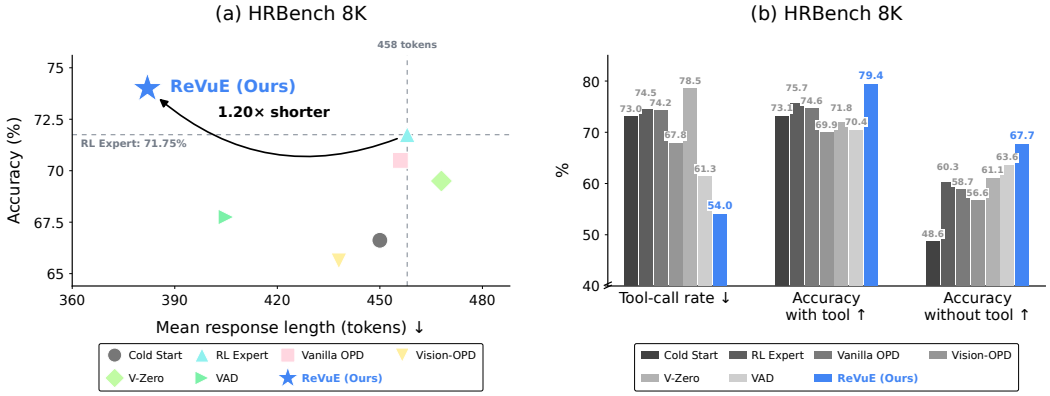


Figure 10: **Response length, accuracy, and tool use on HRBench 8K** for Qwen2.5-VL-7B.

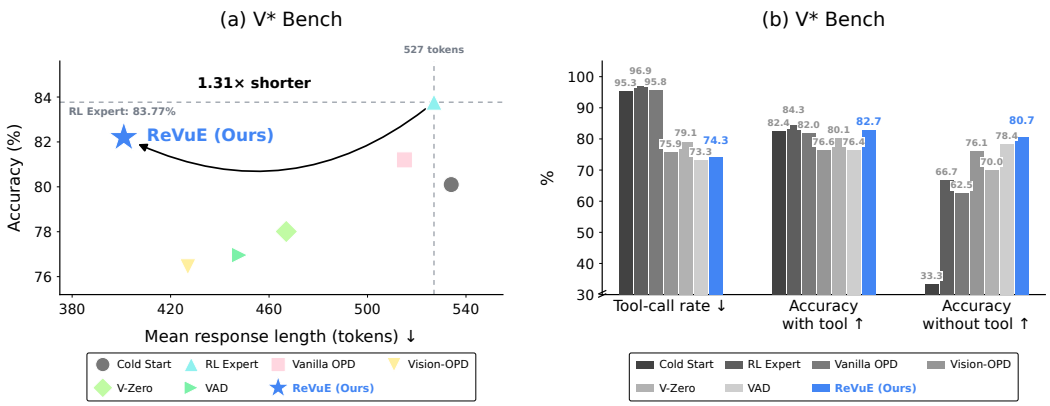


Figure 11: **Response length, accuracy, and tool use on V* Bench** for Qwen2.5-VL-7B.

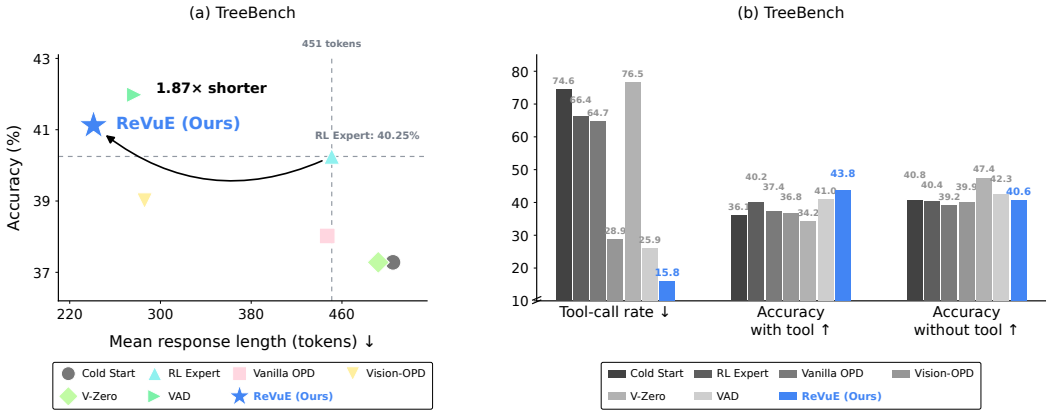


Figure 12: **Response length, accuracy, and tool use on TreeBench** for Qwen2.5-VL-7B.

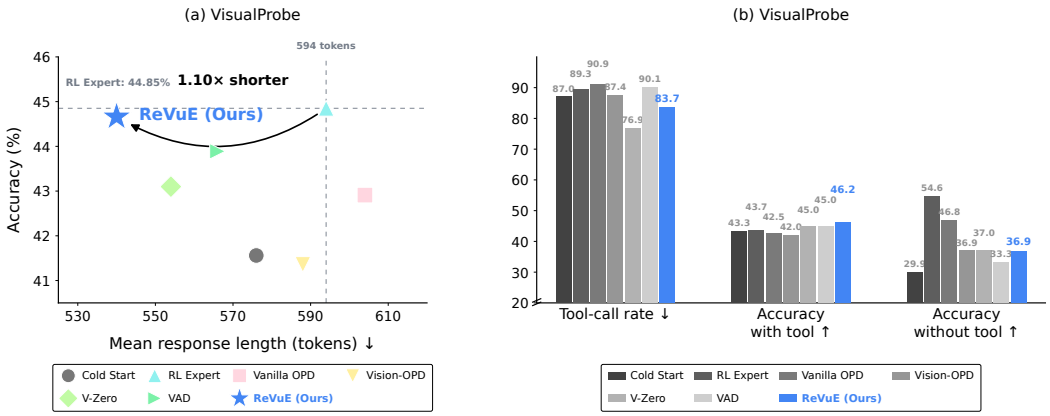


Figure 13: **Response length, accuracy, and tool use on VisualProbe** for Qwen2.5-VL-7B.

C.5 Training-Time Reflection Coverage

After deduplication, the main training run contains 1,075 optimization steps with 20 groups per step and eight rollouts per group, totaling 21,500 groups and 172,000 rollout slots. As shown in Table 12, reflections cover 161,632 slots (93.97%), comprising complete reflections (78.36%) and Break-Point-only reflections (15.61%). The Break-Point-only branch supplies valid reflections for all-wrong groups, while the remaining 6.03% of slots use the base teacher.

Table 12: **Reflection coverage in the main training run.** All percentages use 172,000 rollout slots as the denominator; the final row sums the first two rows.

Reflection status	Rollout slots	Share (%)
Complete reflection	134,784	78.36
Break-Point-only reflection	26,848	15.61
No reflection (base teacher)	10,368	6.03
Any reflection (subtotal)	**161,632**	**93.97**

C.6 Offline Comparison with Anchor-Only Context

Complete reflections correct more errors and preserve more correct answers. We compare complete reflections (Anchor plus Break Point) with Anchor-only context in an offline evaluation on fixed paired cases from 311 mixed groups (Table 13). Of 139 initially incorrect answers, complete reflections correct 124, compared with 62 for Anchor-only context. Among the 172 initially correct answers, eight become incorrect with complete reflections, compared with 47 with Anchor-only context. Complete reflections outperform Anchor-only context across all six data sources. These offline results support the value of failure diagnoses for both error correction and preserving correct answers. This pattern is consistent with the token-impact cases in Appendix E.4: reflection reduces teacher support for erroneous choices and increases support for correct judgments within failed trajectories.

Table 13: **Offline reflection comparison on fixed paired cases.** Original predictions are the predictions before adding reflection, and accuracy changes are computed from unrounded values.

Context	Correct / 311	Accuracy (%)	ΔAccuracy (%)
Original predictions	172	55.31	—
Anchor only	187	60.13	+4.82
Anchor + Break Point	**288**	**92.60**	**+37.30**

D PROOFS AND DERIVATIONS

D.1 PROOF OF LEMMA 1

Fix a position (i,t) and write $p = \pi_\theta(\cdot \mid h_{i,t})$, $q^0 = q^0_{i,t}$, $q^{\mathcal{R}} = q^{\mathcal{R}}_{i,t}$, and $d(v) = d_{i,t}(v)$. Fix the history, teacher predictions, and nonempty shared support Supp, with positive student and teacher probabilities on it. For either teacher, $q(\mathrm{Supp}) = \sum_{v\in\mathrm{Supp}} q(v)$ is its probability mass on the support, while $q^{\mathrm{Supp}}(v) = q(v)/q(\mathrm{Supp})$ is its renormalized distribution. Subtracting the two KL losses cancels their shared student term:

$$\begin{aligned}\ell^{\mathrm{Supp}}(p,q^{\mathcal{R}}) - \ell^{\mathrm{Supp}}(p,q^0) &= \sum_{v\in\mathrm{Supp}} p^{\mathrm{Supp}}(v)\log\frac{(q^0)^{\mathrm{Supp}}(v)}{(q^{\mathcal{R}})^{\mathrm{Supp}}(v)} \\ &= -\sum_{v\in\mathrm{Supp}} p^{\mathrm{Supp}}(v)d(v) + \log\frac{q^{\mathcal{R}}(\mathrm{Supp})}{q^0(\mathrm{Supp})}. \end{aligned} \tag{15}$$

The teacher normalization term is constant in θ and vanishes under differentiation. Using $\nabla_\theta p^{\mathrm{Supp}}(v) = p^{\mathrm{Supp}}(v)\nabla_\theta \log p^{\mathrm{Supp}}(v)$ gives

$$\begin{aligned}\nabla_\theta\left[\ell^{\mathrm{Supp}}(p,q^{\mathcal{R}}) - \ell^{\mathrm{Supp}}(p,q^0)\right] &= -\sum_{v\in\mathrm{Supp}} d(v)\nabla_\theta p^{\mathrm{Supp}}(v) \\ &= -\mathbb{E}_{v\sim p^{\mathrm{Supp}}}\left[d(v)\nabla_\theta \log p^{\mathrm{Supp}}(v)\right]. \end{aligned} \tag{16}$$

This proves Equation 6. For the reflection-conditioned branch in Section 4.3, take $\mathrm{Supp} = \mathrm{Supp}_{i,t}$ and normalize both teachers over this support.

From distribution shifts to a sampled-token proxy. Let z_v be a student logit and $\bar{d}_{\mathrm{Supp}} = \mathbb{E}_{v\sim p^{\mathrm{Supp}}}[d(v)]$ the student-weighted mean shift. For $v \in \mathrm{Supp}$,

$$\frac{\partial}{\partial z_v}\left[\ell^{\mathrm{Supp}}(p,q^{\mathcal{R}}) - \ell^{\mathrm{Supp}}(p,q^0)\right] = -p^{\mathrm{Supp}}(v)\big(d(v) - \bar{d}_{\mathrm{Supp}}\big). \tag{17}$$

The softmax derivative gives this expression and zero direct derivatives for logits outside Supp. The gradient change depends on student probabilities and centered shifts; parameter gradients also involve the logit Jacobian. Equation 7 measures the teacher shift at the student's sampled choice for ranking positions. It need not match rankings by full logit- or parameter-gradient norm.

For example, take $q^0 = (0.10, 0.15, 0.375, 0.375)$ and $q^{\mathcal{R}} = (0.30, 0.45, 0.125, 0.125)$, with Supp containing the latter's top two candidates. Both teachers renormalize to $(0.4, 0.6)$ on Supp, so the shared-support gradient difference is zero, although either sampled token in this set has impact $\log 3$. The uniform shift cancels under support normalization. Impact remains defined for sampled tokens outside Supp; RKL at that position still supervises the distribution within it.

Shared support and teacher-specific top-K sets. Let $\mathrm{Supp}^{\mathcal{R}} = \mathrm{TopK}(q^{\mathcal{R}})$ and $\mathrm{Supp}^0 = \mathrm{TopK}(q^0)$. Comparing losses on each teacher's own support gives

$$\begin{aligned}\ell^{\mathrm{Supp}^{\mathcal{R}}}(p,q^{\mathcal{R}}) - \ell^{\mathrm{Supp}^0}(p,q^0) &= \Delta_{\mathrm{prob}} + \Delta_{\mathrm{support}}, \\ \Delta_{\mathrm{prob}} &= \ell^{\mathrm{Supp}^{\mathcal{R}}}(p,q^{\mathcal{R}}) - \ell^{\mathrm{Supp}^{\mathcal{R}}}(p,q^0), \\ \Delta_{\mathrm{support}} &= \ell^{\mathrm{Supp}^{\mathcal{R}}}(p,q^0) - \ell^{\mathrm{Supp}^0}(p,q^0). \end{aligned} \tag{18}$$

The lemma characterizes $\nabla_\theta\Delta_{\mathrm{prob}}$; $\nabla_\theta\Delta_{\mathrm{support}}$ is generally nonzero when the supports differ. These terms decompose the loss difference; they are not additional training objectives.

D.2 Gradient Weights under Grouped Distillation

Equal group weights increase per-position gradient coefficients in the smaller group. We examine how the weights in Equation 9 scale each position's gradient contribution relative to uniform token averaging. Let $\ell_{i,t}$ denote the per-position KL term in Equation 8. With teacher targets, supports, and group assignments fixed, $w_{i,t}$ is constant during the student update, so

$$\nabla_\theta \left[\frac{w_{i,t}}{|\mathcal{T}_i|} \ell_{i,t} \right] = w_{i,t} \frac{\nabla_\theta \ell_{i,t}}{|\mathcal{T}_i|}. \tag{19}$$

The weight $w_{i,t}$ directly gives the per-position gradient multiplier relative to uniform averaging under the same teacher target.

With $\lambda = 0.5$ and both groups nonempty, Equation 9 assigns half of the total loss coefficient to each group. Ignoring the rounding in $|\mathcal{T}_{i,\text{High}}| = \lceil \alpha |\mathcal{T}_i| \rceil$ gives the continuous approximations in Figure 14:

$$w_{\text{High}} \approx \frac{0.5}{\alpha}, \qquad w_{\text{Low}} \approx \frac{0.5}{1-\alpha}. \tag{20}$$

In this continuous form, High has a multiplier above one and Low below one when $\alpha < 0.5$; both equal one at $\alpha = 0.5$.

For a trajectory with 100 supervised positions, $\alpha = 0.2$ assigns 20 positions to High and the remaining 80 to Low. Uniform averaging assigns each position a coefficient of 0.01, while grouping gives High and Low positions coefficients of 0.025 and 0.00625. The corresponding gradient contributions scale by 2.5 and 0.625, with each group's coefficients summing to 0.5.

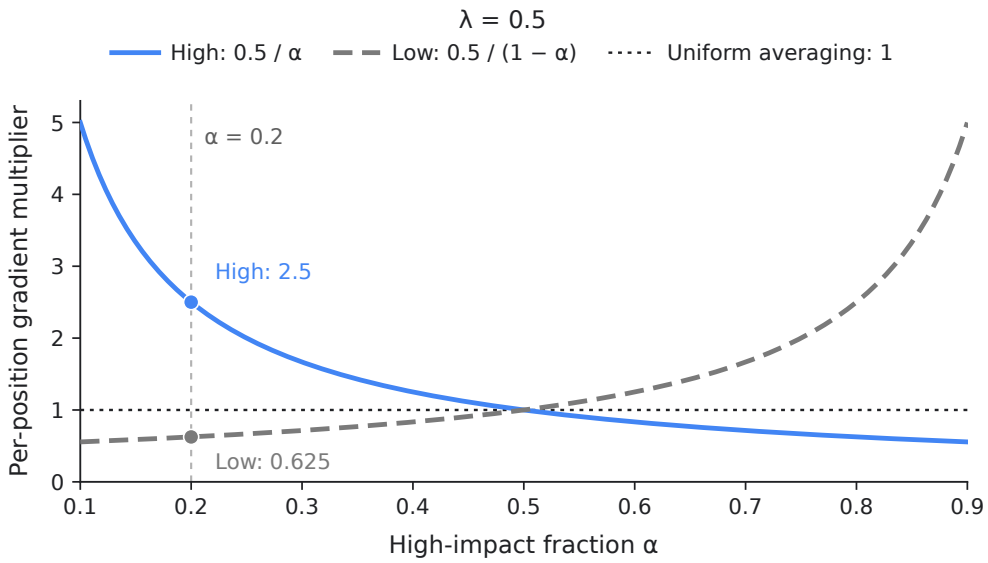


Figure 14: **Per-position gradient weights at $\lambda = 0.5$.** **Blue** and **gray** curves show the **High** and **Low** gradient multipliers relative to uniform token averaging; the black dotted line marks the unit baseline. The continuous curves ignore group-size rounding and cover $0.1 \leq \alpha \leq 0.9$. Markers at $\alpha = 0.2$ give the exact multipliers 2.5 and 0.625 for the 100-position example.

E CASE STUDY

E.1 ERRORS IN ACQUIRING, READING, AND USING VISUAL EVIDENCE

Figure 15 illustrates three concrete errors in acquiring, reading, and using visual evidence in Thyme-SFT (Zhang et al., 2026b) responses.

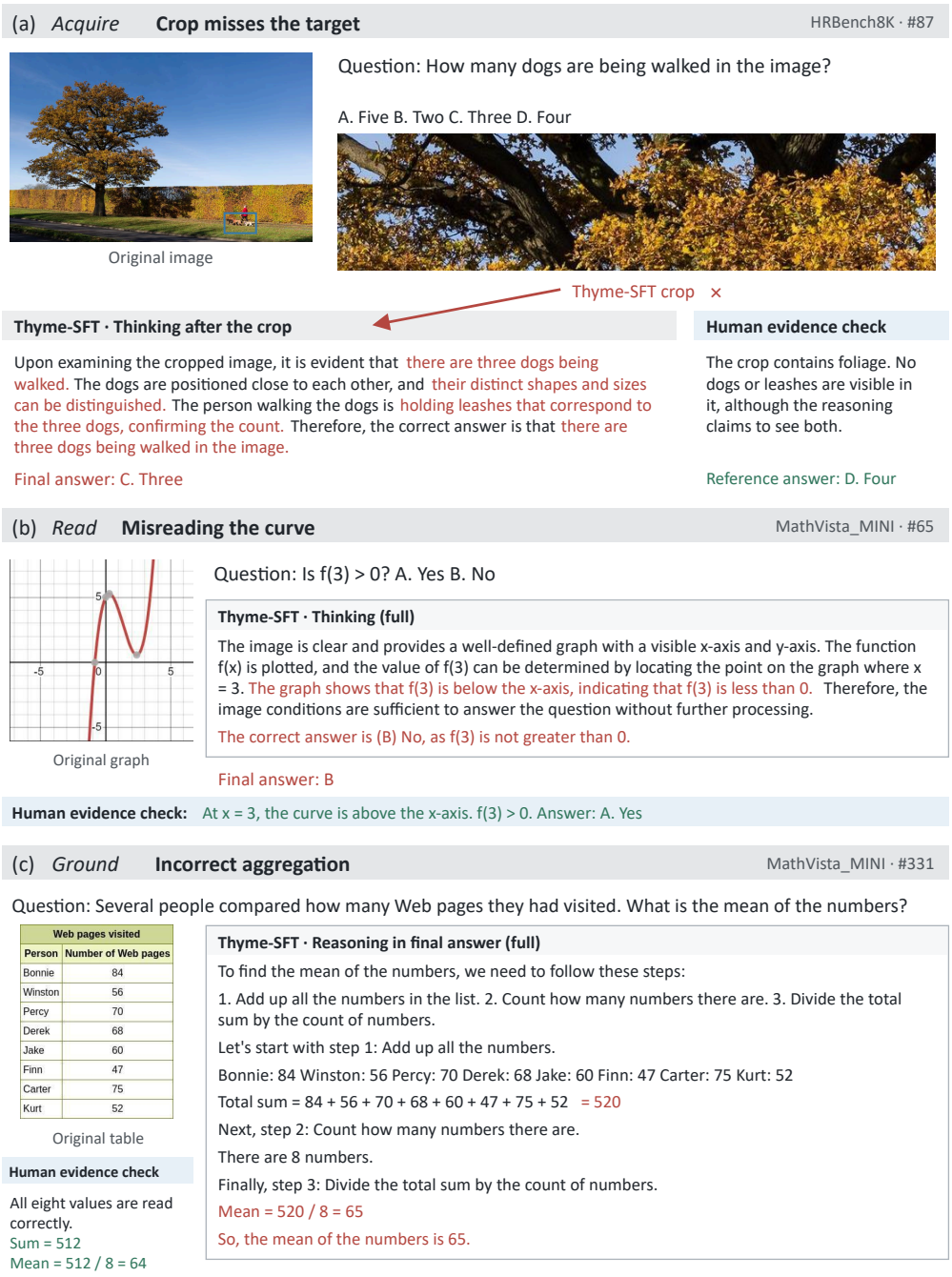


Figure 15: In *Acquire*, the crop misses the queried dogs; in *Read*, the model describes a curve above the x-axis as lying below it; in *Ground*, it reads the table correctly but miscalculates the sum. The panels retain post-crop thinking in (a), full thinking in (b), and the complete reasoning-bearing final answer in (c). Red text marks errors and their consequences.

E.2 Different Errors with the Same Evidence

Figure 16 compares two student trajectories for the same six-slice pie chart. Both use the original image without tool calls and answer "No".



Figure 16: Medium Blue is the fourth-largest slice, making it the lower median of six slices. The left trajectory misreads the rank (*Read*); the right trajectory reads it correctly but misapplies the lower-median rule (*Ground*). Both complete thinking traces and final answers are shown, with errors in red.

The required correction is to fix the visual rank on the left and the lower-median rule on the right. Both rollouts belong to the same all-wrong group, whose recorded training feedback contains diagnoses only.

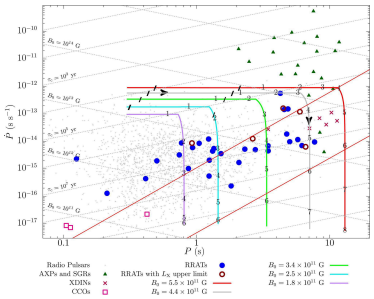
E.3 Different Errors on the Same Plot

Figure 17 contrasts two incorrect trajectories for the same neutron-star plot.

Same plot, two different errors

Question: Which neutron star group predominantly has a characteristic age (τc) less than 10^5 years according to the plot?

A. AXPs and SGRs B. RRATs C. XDINs D. Radio Pulsars



Original plot

Human evidence check

Age contours are gray dotted diagonals.

At a fixed period, positions above the 10^5-year contour indicate younger stars; positions below indicate older stars.

Most AXPs and SGRs (green triangles) lie above this contour.

Reference answer: A. AXPs and SGRs

Misidentified age contour

Thinking before the tool (full)

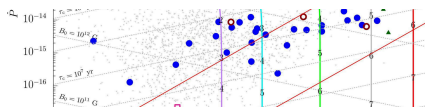
To determine which neutron star group has a characteristic age (τc) less than 10^5 years, we need to focus on the region of the plot where τc = 10^5 years is indicated. This region is located in the lower part of the plot, below the τc = 10^5 yr line. Cropping and zooming into this specific area will help us identify the relevant data points and their corresponding labels.

The task involves analyzing the distribution of data points in the lower section of the plot. Given the scale and resolution of the image, zooming in on the area below the τc = 10^5 yr line will provide a clearer view of the data points and their labels, making it easier to identify the predominant group with τc < 10^5 years.

Here is the Python code to crop and zoom into the relevant area:

Actual tool crop

Crop (0, 700) to (1932, 1150); enlarge by 1.5×



Thinking after the tool (full)

The green lines on the plot correspond to τc = 10^5 years, and most of the RRATs (represented by blue circles) are found below these lines, indicating that they predominantly have a characteristic age less than 10^5 years.

× Final answer: B. RRATs

Reversed age comparison

Thinking before the tool (full)

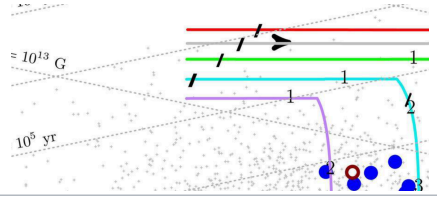
To identify which group of neutron stars primarily has a characteristic age (τc) less than 10^5 years, we need to concentrate on the area of the plot where τc = 10^5 years is indicated. This region is positioned in the middle of the plot, with τc values diminishing toward the left side.

Focusing on this specific area by cropping and zooming into it will facilitate the clear identification of data points associated with each neutron star category (AXPs and SGRs, RRATs, XDINs, and Radio Pulsars) and their respective τc values. This approach aids in accurately determining which category predominantly falls below 10^5 years.

Here is the Python code to crop and zoom into the relevant area:

Actual tool crop

Crop (300, 400) to (1200, 800); enlarge by 1.5×



Tool code is summarized above; all natural-language thinking is retained.

Thinking after the tool (full)

The plot features lines depicting characteristic ages (τc) for different magnetic field strengths. AXPs and SGRs are primarily located above the τc = 10^5 years line, indicating that their characteristic ages exceed 10^5 years. Conversely, RRATs, marked by blue circles, are predominantly situated below the τc = 10^5 years line, suggesting that their characteristic ages are generally less than 10^5 years.

× Final answer: B. RRATs

Figure 17: Both trajectories answer B (RRATs), whereas the reference answer is A (AXPs and SGRs). The left misidentifies green lines as age contours; the right correctly places AXPs and SGRs above the age contour but interprets that position as older. Their different crops and all natural-language thinking are shown; tool code is summarized.

E.4 Token Impact Visualizations

Figures 18 and 19 complement Figure 3 with token impact visualizations for visual target selection and age inference.

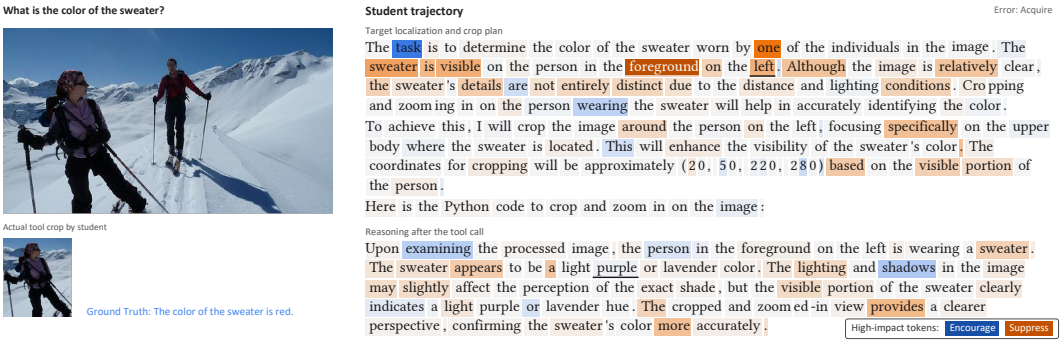


Figure 18: **Token impact on visual target selection.** The student crops the foreground person and reports `light purple or lavender` instead of `red`. Reflection identifies the background person as the target and lowers support for the underlined `foreground` and `left` in the crop plan. **Blue**/**orange** denote increased/decreased teacher support; darker shading indicates larger absolute log-probability shifts on the same student trajectory.

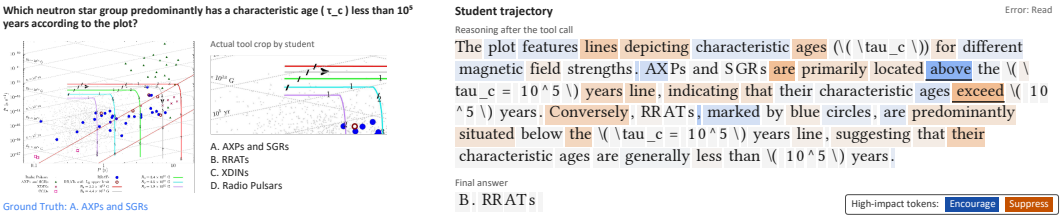


Figure 19: **Token impact on age inference.** The student correctly places `AXPs and SGRs` above the $\tau_c = 10^5$ yr contour but infers older ages, incorrectly answering `RRATs`. Reflection links this position to younger ages, increasing support for the underlined correct relation `above` and decreasing support for the mistaken comparison `exceed`. **Blue**/**orange** denote increased/decreased teacher support; darker shading indicates larger absolute log-probability shifts on the same student trajectory.

F Evaluation Benchmarks

In Table 1, MathVista and MathVerse denote the Mini and vision-only splits, respectively. HalluBench, CQA-Pro, and InfoVQA abbreviate HallusionBench, ChartQA-Pro, and InfographicVQA, respectively.

F.1 Visual Perception

HRBench-4K/8K (Wang et al., 2025b) evaluate high-resolution perception on 4K and 8K images, covering fine-grained recognition and spatial relations.[5]

V*Bench (Wu & Xie, 2024) evaluates visual search in high-resolution images, focusing on fine-grained attributes and spatial relations.[6]

VisualProbe (Lai et al., 2026) evaluates visual search for small targets among distractors in high-resolution images.[7]

TreeBench (Wang et al., 2026) tests fine-grained object perception and relation understanding in complex scenes, with annotations for traceable visual evidence.[8]

F.2 Mathematical and Logical Reasoning

MathVista-Mini (Lu et al., 2024) is the testmini split of MathVista, covering mathematical reasoning over diagrams, charts, and natural images.[9]

MathVerse (Zhang et al., 2025) evaluates diagram understanding in visual mathematics. We use its vision-only version, which presents the problem information entirely in the image.[10]

VisuLogic (Xu et al., 2026) tests nonverbal logical reasoning through visual patterns involving quantity, space, and attributes.[11]

F.3 General Multimodal Understanding

HallusionBench (Guan et al., 2024) diagnoses language hallucinations and visual illusions through questions that require image context.[12]

ChartQA-Pro (Masry et al., 2025) covers diverse chart types and question formats, including conversational, hypothetical, and unanswerable questions.[13]

InfographicVQA (Mathew et al., 2022) requires joint reasoning over text, layout, graphics, and data visualizations to answer questions about infographics.[14]

[5] HRBench-4K/8K: https://huggingface.co/datasets/DreamMr/HR-Bench

[6] V*Bench: https://huggingface.co/datasets/craigwu/vstar_bench

[7] VisualProbe Easy: https://huggingface.co/datasets/Mini-o3/VisualProbe_Easy; Medium: https://huggingface.co/datasets/Mini-o3/VisualProbe_Medium; Hard: https://huggingface.co/datasets/Mini-o3/VisualProbe_Hard

[8] TreeBench: https://huggingface.co/datasets/HaochenWang/TreeBench

[9] MathVista-Mini: https://huggingface.co/datasets/AI4Math/MathVista

[10] MathVerse: https://huggingface.co/datasets/AI4Math/MathVerse

[11] VisuLogic: https://huggingface.co/datasets/VisuLogic/VisuLogic

[12] HallusionBench: https://huggingface.co/datasets/rayguan/HallusionBench

[13] ChartQA-Pro: https://huggingface.co/datasets/ahmed-masry/ChartQAPro

[14] InfographicVQA: https://site.docvqa.org/datasets/infographicvqa