VISTA: INTERNALIZING COLLECTIVE VISUAL EXPERIENCE VIA ON-POLICY DISTILLATION FOR ACTIVE MULTIMODAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Active multimodal agents use visual tools to acquire task-relevant evidence while reasoning. Although reinforcement learning samples multiple interaction trajectories per input, outcome-based objectives primarily use the group to estimate scalar advantages, leaving complementary visual discoveries underused. We introduce VISTA, which internalizes collective visual experience through on-policy distillation by turning observations from same-input rollouts into shared supervision. *Collective visual experience distillation* (CVED) organizes these observations with their interaction context and aligns them with individual decisions, while *heterogeneity-aware policy improvement* (HAPI) reinforces successful trajectories and provides experience-guided distillation for unsuccessful attempts. An experience-conditioned teacher evaluates the student's sampled response prefixes, allowing discoveries from one trajectory to guide learning in another without replacing the student's original history or generating new target trajectories. The trained agent retains its visual tools and acts using its own interaction history. VISTA achieves the strongest average performance among the evaluated active multimodal agents of comparable size and consistently outperforms same-backbone training baselines across fine-grained perception and general reasoning tasks, demonstrating the value of collective experience for active multimodal learning.

1 INTRODUCTION

Multimodal agents are moving beyond answering questions from a fixed image toward actively acquiring the evidence required to solve them. They can inspect local regions with crop and zoom tools, transform returned views, and chain visual actions with reasoning or search (Wu & Xie, 2023; Hu et al., 2024; Shen et al., 2024). This setting makes visual interaction part of the policy. Success depends not only on what the model knows, but also on whether it acquires useful evidence and incorporates it into subsequent decisions. Reinforcement learning has therefore become a natural training paradigm for active multimodal agents. Recent systems sample multiple tool-using trajectories for each image–question pair and use verifiable outcomes to reinforce effective behavior (Zheng et al., 2025; Wu et al., 2025; Wang et al., 2025a; Liu et al., 2025; Yang et al., 2026b; Shao et al., 2024).

However, current outcome-based objectives use a rollout group primarily as a source of scalar comparisons. Once relative advantages have been computed, the visual discoveries made by one trajectory are not directly available to the others: a trajectory that finds a decisive region can teach its peers only through a reward difference. This creates an *experience-utilization gap*. The group may collectively contain the evidence needed to solve a task, including useful observations from partial attempts, yet most of this task-specific experience disappears after the policy update. On-policy distillation offers a natural way to narrow this gap by evaluating a student's response under an enhanced context and thereby providing denser token-level supervision (Penaloza et al., 2026; Ye et al., 2026; Yang et al., 2026a). Visual approaches derive supervision from privileged regional views, generated visual thoughts, or contrasts between visual inputs (Yuan et al., 2026; Cai et al., 2026; Li et al., 2026; Liang et al., 2026).

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.

Applying visual OPD to active multimodal agents nevertheless leaves two problems unresolved. The first concerns the source of privileged supervision. Existing approaches enrich teacher predictions through privileged context or visual contrast, but leave open how to exploit complementary evidence discovered across the agent's own same-input interaction rollouts. Turning these discoveries into supervision requires organizing their interaction context and aligning them with the student's current decision. The second problem concerns the allocation of supervision across heterogeneous trajectories. A failed attempt may contain a useful observation despite reaching an incorrect answer, whereas a successful attempt may benefit more from reinforcing its behavior than from matching a teacher's preferences. Final success therefore does not fully characterize the value of a trajectory's experience, and uniform distillation does not distinguish behavioral reinforcement from corrective guidance. Effective learning therefore requires both a shared representation of the group's experience and a policy-improvement strategy that respects the heterogeneous roles of its trajectories.

These observations motivate VISTA. Our first principle is to distill the shared evidence produced by the rollout group rather than treating trajectories as independent reward samples. We implement this principle with *collective visual experience distillation* (CVED), which organizes the group's visual artifacts and interaction context into training-only privilege and aligns that context with each student's original history. Our second principle is to improve the policy according to the heterogeneous roles of different trajectories. We implement it with *heterogeneity-aware policy improvement* (HAPI): successful trajectories receive reward-aligned RLVR updates, while unsuccessful trajectories receive OPSD guidance from the experience-conditioned teacher. The teacher evaluates the exact response prefix sampled by the student rather than generating a replacement trajectory. The student remains conditioned only on its own interaction history, and the rollout group, teacher, and collective context are removed at deployment. VISTA therefore transfers useful discoveries across trajectories during training while preserving the original active-agent interface at inference time.

Our contributions are:

- **Mechanism.** We identify an experience-utilization gap in active multimodal learning, where outcome-based supervision underuses complementary visual discoveries and overlooks the distinction between trajectory success and informational value.
- **Methodology.** We introduce VISTA, combining CVED for decision-aligned experience distillation with HAPI for reinforcing successful trajectories and guiding unsuccessful ones through collective visual experience.
- **Performance.** VISTA leads the evaluated active multimodal agents of comparable size in average visual-reasoning accuracy and improves over same-backbone training baselines across fine-grained perception and general reasoning tasks.

2 Related Work

RLVR for multimodal agents. Reinforcement learning with verifiable rewards enables multimodal agents to acquire evidence through visual tools rather than reason only over a fixed image (Zheng et al., 2025; Wang et al., 2025a; Liu et al., 2025). VTool-R1 and Thyme extend this paradigm to executable image operations, while DeepEyesV2 develops broader multimodal tool use (Wu et al., 2025; Zhang et al., 2025a; Hong et al., 2025). Further work improves reasoning–action alignment and scales interaction depth (Yang et al., 2026b; Lai et al., 2025). These studies primarily improve how agents explore and learn from task feedback. VISTA addresses the complementary problem of reusing what exploration discovers: visual observations from same-input rollouts become shared supervision for the agent's decisions.

OPD for multimodal agents. On-policy distillation trains on student-generated responses, and self-distillation can obtain stronger supervision by conditioning the same model on additional information (Agarwal et al., 2023; Zhao et al., 2026). Vision-OPD and Imagine-OPD exploit regional evidence, while Visual-OPSD transfers the benefit of generated visual thoughts (Yuan et al., 2026; Cai et al., 2026; Li et al., 2026). VA-OPD, V-Zero, and VCSD use visual dependence or contrastive evidence to refine the distillation signal (Liu et al., 2026; Sun et al., 2026; Liang et al., 2026). Context- and skill-conditioned approaches also distill trajectory experience for language models and agents (Ye et al., 2026; Yang et al., 2026a). VISTA focuses on visual experience acquired collectively

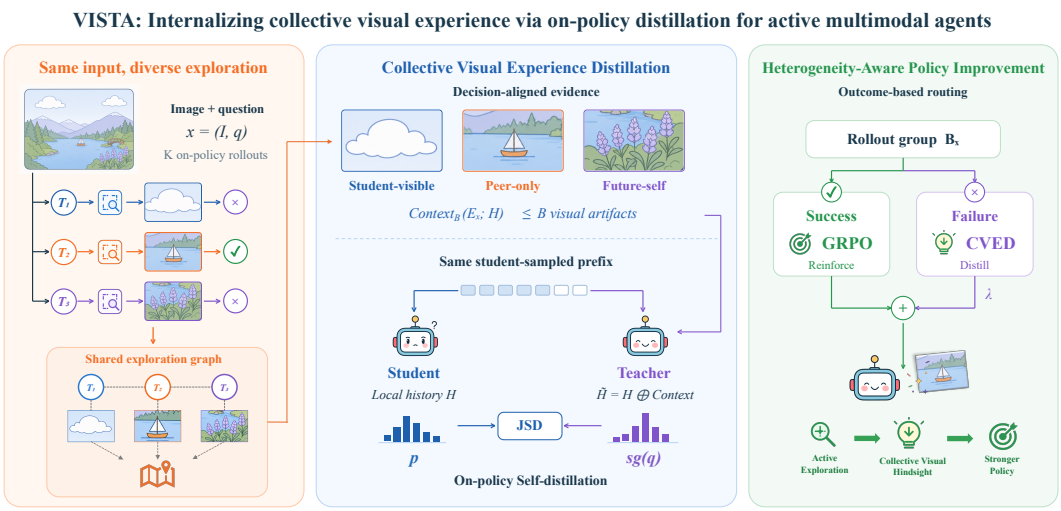

Figure 1: VISTA combines experience transfer with heterogeneous policy improvement. The complete rollout group supplies experience for CVED, while HAPI allocates reinforcement and distillation according to trajectory outcomes. The teacher scores the student's sampled responses without generating new training trajectories.

through active interaction: CVED aligns this evidence with student decisions, and HAPI distinguishes reinforcement of successful behavior from distillation on unsuccessful attempts. Appendix F provides a detailed discussion.

3 VISTA

VISTA turns the visual experience of an on-policy rollout group into supervision for an individual agent. Its two components address complementary aspects of this learning problem. CVED constructs a decision-aligned teacher from the group's interaction experience, providing distributional guidance on the student's own responses. HAPI determines how this guidance is used, reinforcing successful trajectories with RLVR and applying collective-experience distillation to unsuccessful ones.

3.1 Problem Formulation

Given an image–question pair $x = (I, q)$, an active multimodal policy alternates between generating responses and observing tool results. At decision t, the interaction history H_t contains the original input and all preceding responses and tool observations. The student policy π_θ samples a response y_t, which can invoke a visual tool or provide a final answer. A tool result becomes part of the next history. For each input, we collect

$$\mathcal{B}_x = \{\tau_k\}_{k=1}^{K}, \qquad \tau_k \sim \pi_{\theta_{\text{old}}}(\cdot \mid x), \tag{1}$$

where θ_{old} denotes the policy used for rollout collection. Each trajectory receives a verified task reward and a terminal status indicating success, an incorrect answer, a protocol-invalid action, or an incomplete attempt. We use these outcomes to choose the learning objective, while retaining observations from the full group as potential sources of experience. The resulting training problem is to improve the policy using what the group discovered, without changing the histories on which its responses were sampled.

3.2 Collective Visual Experience Distillation

Collective experience. A rollout group provides several attempts to acquire evidence about the same input. Their value is not limited to the final answers. A trajectory may reveal a small visual detail that its peers missed, or obtain a useful observation before making a subsequent reasoning error. CVED compiles these discoveries into a task-local experience object $\mathcal{E}_x = \mathcal{C}(\mathcal{B}_x)$. It associates each

visual artifact with the decision that produced it and preserves its position within the originating interaction. Relations between artifacts retain how successive crops were derived from one another and which regions they cover. Repeated views can share a visual representation without losing their separate trajectory occurrences. This organization lets the teacher interpret a visual discovery in context rather than treating the group as an unordered collection of images. Trajectory outcomes provide feedback for interpreting this experience, while both successful and unsuccessful attempts remain eligible evidence sources.

Decision-aligned context. The experience is shared across the group, but its relevance depends on the decision being supervised. For a response generated from history H, the teacher is conditioned on an augmented history

$$\widetilde{H} = H \oplus \mathrm{Context}_B(\mathcal{E}_x; H), \tag{2}$$

where $\mathrm{Context}_B$ selects and encodes collective experience under a budget of B visual artifacts, and $\oplus$ denotes context augmentation. Alignment distinguishes artifacts already observed by the student from those available through peers or later in its own trajectory. Thus, the teacher can use hindsight without confusing it with evidence available when the student acted.

Collective experience therefore provides a broader evidence base for supervising decisions made under a partial interaction history. By conditioning the teacher on this experience, CVED allows discoveries from one trajectory to inform learning in another, including useful observations from attempts that ultimately fail. This cross-trajectory transfer expands the information available for supervision while keeping policy learning grounded in the student's own behavior. Teacher-context construction is detailed in Appendix B.2.

On-policy knowledge transfer. Let y be a student-generated response and $y_{<n}$ its prefix before token n. The student π_θ and the experience-conditioned teacher π_{T} evaluate this same prefix using H and $\widetilde{H}$, respectively. For a subset $\mathcal{S}$ of the rollout group, the on-policy distillation objective is

$$\mathcal{L}_{\mathrm{CVED}}(\theta; \mathcal{S}) = \mathbb{E}_{(H, y_{<n}) \sim \mathcal{S}} \Big[D\Big(\pi_\theta(\cdot \mid H, y_{<n}),\, \pi_{\mathrm{T}}(\cdot \mid \widetilde{H}, y_{<n}) \Big) \Big], \tag{3}$$

where the expectation denotes the empirical average over generated response-token prefixes in $\mathcal{S}$, excluding prompt tokens, tool observations, and padding. The teacher distribution is held fixed during student optimization. Its context is drawn from the full rollout group even when only a subset receives distillation. Thus, the teacher supplies distributional supervision on student-visited prefixes rather than generating replacement trajectories. The additional experience changes how the teacher evaluates the student's behavior, and the resulting distributional signal trains the policy under its ordinary local history.

3.3 Heterogeneity-Aware Policy Improvement

Trajectories within a rollout group differ in both behavioral quality and informational value. An unsuccessful attempt may discover useful visual evidence even when its subsequent decisions lead to an incorrect answer, whereas a successful attempt provides behavior supported by a verified outcome. These differences motivate distinct forms of policy improvement. While CVED makes experience available across the group, HAPI determines how each trajectory learns from it.

HAPI reinforces successful trajectories through outcome-based policy optimization and applies experience-guided distillation to unsuccessful ones. For successful trajectories, the verified reward supports strengthening behavior that already solves the task. For unsuccessful trajectories, the experience-conditioned teacher provides distributional guidance informed by discoveries beyond the agent's local history. This guidance covers the full sequence of generated responses, allowing collective experience to inform intermediate decisions as well as the final answer. Both types of trajectory remain sources of collective experience, irrespective of the objective applied to their behavior.

Let $\mathcal{B}_x^-$ denote the unsuccessful trajectories in the group, including invalid and incomplete attempts. Writing $\mathcal{L}_{\mathrm{GRPO}}^{\mathrm{succ}}$ for the GRPO objective restricted to successful trajectories, the overall objective is

$$\mathcal{L}_{\mathrm{VISTA}}(\theta) = \mathbb{E}_{x, \mathcal{B}_x} \Big[\mathcal{L}_{\mathrm{GRPO}}^{\mathrm{succ}}(\theta; \mathcal{B}_x) + \lambda \mathcal{L}_{\mathrm{CVED}}(\theta; \mathcal{B}_x^-) \Big]. \tag{4}$$

Here λ controls the strength of experience distillation. Relative advantages are computed from the full group's rewards, while the CVED teacher draws on experience contributed by all trajectories. The two objectives provide complementary directions for learning: reinforcement strengthens successful behavior, and distillation guides improvement where the current policy has not yet succeeded. Together, CVED and HAPI couple shared experience with differentiated supervision to improve the active interaction policy.

We use Jensen–Shannon divergence for distribution matching; teacher updates and optimization details are provided in Appendix B. At inference, the student continues to use visual tools under its own interaction history, without the collective context or teacher.

4 Experiments

Our evaluation examines the effectiveness of collective visual experience and the roles of its two components. We first compare VISTA with existing agents, examine generalization across task families under different training methods, and then isolate the contributions of CVED and HAPI. Further analyses examine teacher updates and distribution matching, generalization across backbones, and the effects of experience scale and distillation strength.

4.1 Experimental Setup

Models and training. We use Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct (Bai et al., 2025) as the default backbones and train on the Vision-OPD-6K split (Yuan et al., 2026). Training inputs contain the original full image without target-box overlays and the question without added spatial-restriction hints; the student acquires additional observations through visual tools. Our default configuration uses $K = 8$ rollouts per input, distillation weight $\lambda = 0.1$, and an artifact budget of $B = 4$ for teacher-context construction. Answer correctness provides the task reward, and the policy may answer directly without invoking a tool.

Benchmarks and metrics. The main evaluation covers V*Bench (Wu & Xie, 2023), ZoomBench (Wei et al., 2026), HR-Bench-4K/8K (Wang et al., 2025b), and the English and Chinese splits of MME-RealWorld (Zhang et al., 2025b). To assess whether improvements extend beyond fine-grained perception, the training-method comparison also includes MMStar, MMBench, and MathVista (Chen et al., 2024; Liu et al., 2024; Lu et al., 2024). Component ablations and parameter analyses use V*Bench and both HR-Bench variants unless otherwise specified. We report accuracy (%) and the unweighted mean over the benchmarks in each table. Bold and underlined scores indicate the best and second-best results, respectively.

Baselines and comparison protocol. We compare against the base checkpoint, GRPO (Shao et al., 2024), and OPSD, VCSD, Vision-OPD, and Imagine-OPD (Zhao et al., 2026; Liang et al., 2026; Yuan et al., 2026; Cai et al., 2026). Within each backbone, controlled comparisons share the starting checkpoint, training data, optimizer and schedule, with matched visual tools and interaction limits. Evaluation is deterministic, and all experimental results are averaged over three seeds. All reported results are obtained from our own experiments, including the broader comparison with thinking-with-images agents and general multimodal models; larger and proprietary models serve as capability references. Appendix C provides the reference-model descriptions and further protocol details.

4.2 Main Results

Consistent improvements across visual tasks. VISTA improves on its base checkpoints across all six benchmarks at both scales (Table 1), with the largest gains on ZoomBench. The improvements on both HR-Bench variants and both language splits of MME-RealWorld show that the benefit extends beyond a single visual-search benchmark. Both variants also outperform the compared thinking-with-images baselines throughout the suite.

The trained 4B model exceeds the base 8B checkpoint across this comparison, suggesting that better use of interaction experience can complement increases in model capacity. Although the trained 8B

model generally performs better, the consistent gains at both scales support collective experience as a useful source of supervision rather than a capacity-specific remedy.

Table 1: Performance comparison on visual perception and reasoning benchmarks. We report accuracy (%). Δ denotes the gain over the corresponding Qwen3-VL base model.

Model	V* Bench	ZoomBench	HR-Bench-4K	HR-Bench-8K	MME-RW-EN	MME-RW-CN	Average
Closed-source models							
GPT-5.2	79.06	50.89	81.12	78.38	<u>72.60</u>	68.80	71.81
GPT-5.4	76.96	52.66	**84.00**	77.88	**74.20**	**70.93**	72.77
Gemini-2.5-Pro	78.01	49.82	80.63	77.50	71.29	69.11	71.06
Gemini-3-Flash	78.53	50.06	80.87	78.00	71.35	69.38	71.37
Open-source models							
Qwen3-VL-4B	81.68	44.97	78.50	76.25	63.27	62.92	67.93
Qwen3-VL-8B	84.82	42.96	79.63	75.25	63.19	64.61	68.41
MiMo-VL-RL-7B	83.25	45.68	73.50	69.38	62.73	55.89	65.07
MiniCPM-V-4.5-9B	70.68	42.60	69.63	61.50	62.65	61.64	61.45
GLM-4.6V-106B	86.91	50.06	82.13	78.88	65.57	65.62	71.53
Kimi-K2.6-1T	<u>88.48</u>	53.14	81.88	78.00	69.22	66.13	72.81
Thinking-with-images models							
Thyme-7B	82.20	45.09	77.00	72.00	64.80	64.59	67.61
DeepEyes-7B	85.86	46.51	75.13	72.63	64.10	64.09	68.05
DeepEyesV2-7B	81.68	44.97	77.88	73.75	64.90	65.07	68.04
Pixel-Reasoner-7B	84.29	44.73	76.50	72.00	64.40	64.12	67.67
Our models							
VISTA-4B	<u>88.48</u>	<u>53.85</u>	83.50	<u>79.50</u>	68.24	69.29	<u>73.81</u>
Δ (vs. Qwen3-VL-4B)	+6.80	+8.88	+5.00	+3.25	+4.97	+6.37	+5.88
VISTA-8B	**89.53**	**54.43**	<u>83.75</u>	**79.88**	68.19	<u>69.55</u>	**74.22**
Δ (vs. Qwen3-VL-8B)	+4.71	+11.47	+4.12	+4.63	+5.00	+4.94	+5.81

4.3 Cross-Task Generalization

Generalization beyond fine-grained perception. Table 2 pairs fine-grained visual tasks with MMStar, MMBench, and MathVista to examine the breadth of the learned capabilities under fixed backbones. VISTA leads both task groups at 4B and 8B, with improvements in fine-grained perception accompanied by stronger general reasoning rather than a trade-off between them. The result on MathVista further extends this pattern to mathematical reasoning in visual contexts.

Consistent gains across training methods and scales. GRPO and the competing distillation methods also improve over the base checkpoints, but their relative strengths vary across tasks. In contrast, VISTA maintains its advantage on every evaluated benchmark at both scales. This consistency supports cross-task generalization within the evaluated suite, rather than an average gain driven by one task family. The ablations below distinguish the contributions of experience construction and supervision allocation.

4.4 Ablation Study

We isolate experience construction and supervision allocation on Qwen3-VL-4B. The CVED ablation fixes HAPI and varies the teacher's privileged context, whereas the HAPI ablation fixes CVED and varies the learning rule.

CVED: the content and organization of experience. Table 3 compares CVED with reference-answer supervision (Answer-only), annotated evidence crops (Oracle ROI), and trajectory-ordered observations and interaction records (Unstructured). All variants use the same training set, and CVED and Unstructured share the selected observations and context budget. CVED performs best across all three tasks, followed by Oracle ROI, Unstructured, and Answer-only.

The ordering suggests that the teacher benefits from visual evidence beyond knowing the correct endpoint, while the strength of Oracle ROI highlights the value of relevant, localized information. CVED improves further by retaining interaction context and aligning evidence with the student's

Table 2: Comparison of training methods on fine-grained visual tasks and general reasoning tasks. We report accuracy (%) for Qwen3-VL-4B and Qwen3-VL-8B.

Methods	Fine-Grained Visual Tasks			General Reasoning Tasks			Average
	V* Bench	HR-Bench-4K	HR-Bench-8K	MMStar	MMBench	MathVista	
Qwen3-VL-4B							
Base	81.68	78.50	76.25	68.73	81.22	73.70	76.68
GRPO	84.12	80.08	77.50	70.87	82.48	75.10	78.36
OPSD	85.34	81.25	78.25	71.53	82.91	75.60	79.15
VCSD	83.77	80.38	76.88	69.20	83.08	74.55	77.98
Vision-OPD	85.34	81.25	77.38	70.55	82.72	75.65	78.82
Imagine-OPD	86.39	82.00	77.80	71.47	82.95	75.90	79.42
VISTA	**88.48**	**83.50**	**79.50**	**73.07**	**83.52**	**76.30**	**80.73**
Qwen3-VL-8B							
Base	84.82	79.63	75.25	70.67	82.68	76.50	78.26
GRPO	86.91	82.13	78.25	73.07	83.62	77.80	80.30
OPSD	87.96	82.75	79.00	74.20	83.91	78.10	80.99
VCSD	87.43	83.25	79.13	74.35	83.18	77.65	80.83
Vision-OPD	88.48	82.88	78.63	74.65	83.82	78.35	81.14
Imagine-OPD	87.96	83.05	78.88	75.15	84.12	78.50	81.28
VISTA	**89.53**	**83.75**	**79.88**	**76.67**	**84.36**	**78.80**	**82.17**

decision. Its advantage over Unstructured is particularly informative because the observations and budget are unchanged: organizing the evidence, rather than simply supplying more images, makes the supervision more effective. Together with the Oracle ROI comparison, this suggests that localization and interaction context serve complementary roles: useful supervision depends on both the evidence presented and its relation to the decision being learned.

HAPI: allocating complementary learning signals. Table 4 compares GRPO-only, CVED-only, Uniform hybrid, and HAPI with the same rollout-group size and optimization budget. Uniform hybrid applies both objectives to every trajectory, whereas HAPI reinforces successful trajectories and distills unsuccessful ones. All distillation variants share the same CVED context construction.

Uniform hybrid outperforms either objective alone, and HAPI improves further across all three tasks. Thus, reinforcement and distillation are complementary, but their allocation also matters. Since Uniform hybrid already receives both signals, HAPI's advantage supports matching supervision to trajectory outcomes rather than merely adding a loss term. Meanwhile, unsuccessful attempts can still contribute useful observations to CVED, separating their informational value from how their behavior is optimized. Thus, the routing rule changes how trajectories learn without restricting the shared evidence pool to successful attempts, preserving potentially useful discoveries from failures.

Table 3: CVED ablation: privileged supervision under fixed HAPI routing.

Setting	V*	HR-4K	HR-8K	Avg.
Answer-only	86.56	81.30	77.80	81.89
Unstructured	86.91	82.13	78.25	82.43
Oracle ROI	87.96	82.80	79.00	83.25
CVED	**88.48**	**83.50**	**79.50**	**83.83**

Table 4: HAPI ablation: learning rules under fixed experience construction.

Setting	V*	HR-4K	HR-8K	Avg.
GRPO-only	84.12	80.08	77.50	80.57
CVED-only	86.21	81.50	78.13	81.95
Uniform hybrid	87.43	82.38	78.80	82.87
HAPI	**88.48**	**83.50**	**79.50**	**83.83**

4.5 Training Design Analysis

We next examine two supporting choices within the fixed CVED–HAPI framework: how the teacher is updated and how its distributions are matched. Both comparisons use Qwen3-VL-4B and the six-benchmark suite from Table 1.

Teacher update strategy. Freezing the teacher at initialization retains much of the performance of EMA (Table 5). EMA gives the stronger average, especially on V*Bench and ZoomBench, while the frozen teacher performs better on HR-Bench-8K and MME-RW-CN. We therefore adopt EMA as an aggregate preference rather than a uniformly better choice for every task.

Both teachers receive newly collected experience, so fixed weights do not imply a fixed teaching signal. The competitive frozen-teacher result suggests that context enrichment can provide useful supervision in its own right, with parameter updates offering an additional, task-dependent benefit.

Table 5: Teacher-update ablation on Qwen3-VL-4B.

Teacher	V*Bench	ZoomBench	HR-Bench-4K	HR-Bench-8K	MME-RW-EN	MME-RW-CN	Avg.
Frozen initial	87.43	51.95	83.38	**80.75**	68.19	**69.51**	73.54
EMA (ours)	**88.48**	**53.85**	**83.50**	79.50	**68.24**	69.29	**73.81**

Divergence objective. Table 6 compares forward KL, $D_{\mathrm{KL}}(\pi_{\mathrm{T}} \| \pi_\theta)$, reverse KL, $D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{T}})$, and JSD with an equally weighted mixture. All variants match distributions on student-generated prefixes with the same CVED context construction, HAPI routing, and teacher updates.

JSD and reverse KL outperform forward KL across the suite. JSD gives the best average, while reverse KL is stronger on ZoomBench and HR-Bench-8K. Since all three variants improve over the base model, collective-experience learning is not tied to a particular divergence. The task-dependent ordering supports JSD as an effective overall choice without implying that it is optimal for every benchmark.

Table 6: Divergence-objective comparison on Qwen3-VL-4B with fixed CVED and HAPI.

Objective	V*Bench	ZoomBench	HR-Bench-4K	HR-Bench-8K	MME-RW-EN	MME-RW-CN	Avg.
Forward KL	86.91	50.89	82.25	77.80	67.10	68.10	72.18
Reverse KL	87.96	**53.96**	82.95	**79.75**	67.80	69.05	73.58
JSD	**88.48**	53.85	**83.50**	79.50	**68.24**	**69.29**	**73.81**

4.6 Backbone Generalization

We repeat the base, GRPO, OPSD, and VISTA comparisons on Qwen3.5-4B and Qwen3.5-9B (Qwen Team, 2026) using the same protocol within each backbone (Table 7). VISTA leads on all three tasks for both backbones. Its advantage over OPSD is concentrated on V*Bench at 4B, where the high-resolution results are close, but extends more clearly across the suite at 9B.

On Qwen3.5-9B, OPSD is stronger than GRPO on V*Bench but weaker on the high-resolution tasks, whereas VISTA improves over both throughout. The combined supervision therefore remains useful across the tested backbones and across tasks that favor different baseline objectives, although the size of the benefit depends on the model and task.

Table 7: Backbone generalization on Qwen3.5-4B and Qwen3.5-9B. All methods use the same training and evaluation protocol within each backbone.

Qwen3.5-4B					**Qwen3.5-9B**				
Method	V*	HR-4K	HR-8K	Avg.	Method	V*	HR-4K	HR-8K	Avg.
Base	82.20	82.75	78.38	81.11	Base	81.68	84.50	79.88	82.02
GRPO	83.25	82.88	78.75	81.63	GRPO	84.29	85.75	82.38	84.14
OPSD	84.29	83.00	79.00	82.10	OPSD	87.96	84.25	81.25	84.49
VISTA	**88.48**	**83.13**	**79.13**	**83.58**	**VISTA**	**92.15**	**87.25**	**84.38**	**87.93**

4.7 Scalability

Increasing the rollout-group size $K \in \{4, 6, 8, 12, 16\}$ improves or preserves accuracy on all three benchmarks, with the strongest response on V*Bench (Figure 2(a)). Gains continue beyond the

default group size but diminish toward the largest tested groups, indicating that additional attempts remain useful with decreasing returns.

Because the artifact budget is fixed, larger groups expand the candidate experience pool without requiring the teacher to process every new observation. Together with the budget sweep below, this favors collecting broadly and conditioning selectively. The sweep also increases rollout volume, so it measures scaling with additional experience rather than a fixed-compute allocation. In particular, the different trends for K and B suggest that the supply of candidate evidence and the teacher's capacity to use it should be treated as separate design choices.

4.8 Hyperparameter Sensitivity

Having varied the experience pool, we examine how strongly that experience should supervise the student and how much should enter each teacher context. We vary λ and B independently on Qwen3-VL-4B, keeping the other settings at their defaults.

Distillation strength. Figure 2(b) varies $\lambda \in \{0, 0.05, 0.1, 0.5\}$ under fixed HAPI routing; setting $\lambda = 0$ retains only successful-trajectory reinforcement in Equation 4. Every tested nonzero weight improves all three tasks over this zero-distillation setting, while the shared peak at $\lambda = 0.1$ favors a moderate distillation strength. The decline at the largest weight shows that increasing emphasis on the teacher signal is not uniformly beneficial, although the gains across several nonzero settings indicate that the benefit is not confined to one coefficient. The common peak across tasks also supports a shared default rather than selecting a separate weight for each benchmark.

Visual artifact budget. Figure 2(c) shows that an intermediate budget is preferable to the smallest or largest tested context: HR-4K favors $B = 4$, while V*Bench and HR-8K favor $B = 6$. Longer teacher contexts may make decision-relevant evidence harder to integrate, while processing additional artifacts increases computation and can prolong training. We retain $B = 4$ as a compact default. In contrast to the continued gains from larger rollout groups, this non-monotonic trend supports expanding the experience pool while selecting a concise context for each decision.

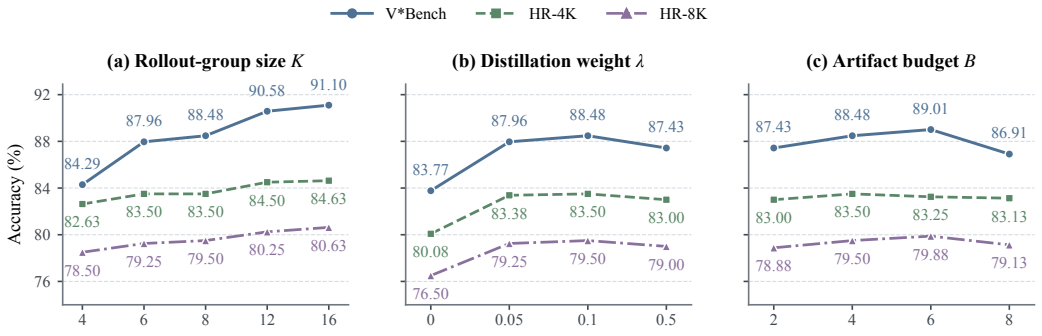

Figure 2: Sensitivity of VISTA on Qwen3-VL-4B to (a) rollout-group size K, (b) distillation weight λ, and (c) visual artifact budget B. Points report V*Bench, HR-4K, and HR-8K accuracy. Tested settings are equally spaced along each horizontal axis.

5 Conclusion

We reframed grouped active-agent training as an experience-internalization problem. Instead of using on-policy rollouts only to estimate scalar advantages, VISTA constructs collective visual experience through CVED and uses HAPI to combine successful-behavior reinforcement with experience-guided distillation on unsuccessful trajectories. The deployed agent remains a single ordinary tool-using policy. This view makes experience sources, interaction structure, and outcome feedback independently testable components of privileged supervision. More broadly, it suggests that the transient population created by on-policy training can serve as a renewable source of experience for the policy itself. Appendix G discusses limitations and directions for future work.

AI USE STATEMENT

Large language models were used solely to assist with language polishing, including correcting grammar, refining phrasing, and improving the clarity and readability of the manuscript. The authors reviewed and revised the suggested edits to ensure that they accurately conveyed the intended technical meaning. All experimental results and implementation details reported in the paper were checked and confirmed by the authors. The authors take full responsibility for the accuracy and integrity of the final manuscript, including its technical claims and conclusions.

REPRODUCIBILITY STATEMENT

Section 3 specifies the learning problem, CVED, HAPI, and the training objective. Appendix B describes experience construction and teacher supervision. Section 4 describes the evaluation protocol and experimental comparisons. Appendix B.4 provides the student and teacher prompt templates, and Appendix B.2 specifies the information supplied to the teacher. We provide the code in the supplementary material to facilitate reproducibility.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. *arXiv preprint arXiv:2306.13649*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025.

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. *arXiv preprint arXiv:2606.08719*, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv preprint arXiv:2403.20330*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, et al. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Tulsee Doshi. Gemini 3 Flash: Frontier Intelligence Built for Speed, 2025. URL https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. DeepEyesV2: Toward Agentic Multimodal Model. *arXiv preprint arXiv:2511.05271*, 2025.

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. *arXiv preprint arXiv:2406.09403*, 2024.

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. *arXiv preprint arXiv:2509.07969*, 2025.

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. *arXiv preprint arXiv:2606.18974*, 2026.

Yijun Liang, Yunjie Tian, Yijiang Li, Yuqi Jia, Furong Huang, Tianyi Zhou, and Di Fu. Visual Contrastive Self-Distillation. *arXiv preprint arXiv:2607.21556*, 2026.

Jianhua Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

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. *arXiv preprint arXiv:2605.21924*, 2026.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is Your Multi-modal Model an All-around Player? In *European Conference on Computer Vision*, 2024.

Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual Agentic Reinforcement Fine-Tuning. *arXiv preprint arXiv:2505.14246*, 2025.

LLM-Core-Team Xiaomi. MiMo-VL Technical Report. *arXiv preprint arXiv:2506.03569*, 2025.

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 *International Conference on Learning Representations*, 2024.

Moonshot AI. Kimi K2.6, 2026. URL https://huggingface.co/moonshotai/Kimi-K2.6. Official model card.

OpenAI. Introducing GPT-5.2, 2025. URL https://openai.com/index/introducing-gpt-5-2/.

OpenAI. Introducing GPT-5.4, 2026. URL https://openai.com/index/introducing-gpt-5-4/.

Emiliano Penaloza, Dheeraj Vattikonda, Nicolas Gontier, Alexandre Lacoste, Laurent Charlin, and Massimo Caccia. Privileged Information Distillation for Language Models. *arXiv preprint arXiv:2602.04942*, 2026.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. ZoomEye: Enhancing Multimodal LLMs with Human-Like Zooming Capabilities through Tree-Based Image Exploration. *arXiv preprint arXiv:2411.16044*, 2024.

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. *arXiv preprint arXiv:2606.25319*, 2026.

Antti Tarvainen and Harri Valpola. Mean Teachers Are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel Reasoner: Incentivizing Pixel-Space Reasoning with Curiosity-Driven Reinforcement Learning. *arXiv preprint arXiv:2505.15966*, 2025a.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025b. doi: 10.1609/aaai.v39i8.32852.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without Zooming: Region-to-Image Distillation for Fine-Grained Multimodal Perception. *arXiv preprint arXiv:2602.11858*, 2026.

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. *arXiv preprint arXiv:2505.19255*, 2025.

Penghao Wu and Saining Xie. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. *arXiv preprint arXiv:2312.14135*, 2023.

Shuo Yang, Jinyang Wu, Zhengxi Lu, Yuhao Shen, Fan Zhang, Lang Feng, Shuai Zhang, Haoran Luo, Zheng Lian, Zhengqi Wen, and Jianhua Tao. OPID: On-Policy Skill Distillation for Agentic Reinforcement Learning. *arXiv preprint arXiv:2606.26790*, 2026a.

Wenhao Yang, Yu Xia, Jinlong Huang, Shiyin Lu, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, Yuchen Zhou, Xiaobo Xia, Yuanyu Wan, Lijun Zhang, and Tat-Seng Chua. Walk the Talk: Bridging the Reasoning-Action Gap for Thinking with Images via Multimodal Agentic Policy Optimization. *arXiv preprint arXiv:2604.06777*, 2026b.

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. On-Policy Context Distillation for Language Models. *arXiv preprint arXiv:2602.12275*, 2026.

Tianyu Yu, Zefan Wang, Chongyi Wang, et al. MiniCPM-V 4.5: Cooking Efficient MLLMs via Architecture, Data, and Training Recipe. *arXiv preprint arXiv:2509.18154*, 2025.

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. *arXiv preprint arXiv:2605.18740*, 2026.

Z.ai. GLM-4.6V, 2025. URL https://huggingface.co/zai-org/GLM-4.6V. Official model card.

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. *arXiv preprint arXiv:2508.11630*, 2025a.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? In *International Conference on Learning Representations*, 2025b.

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. *arXiv preprint arXiv:2601.18734*, 2026.

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. *arXiv preprint arXiv:2505.14362*, 2025.

A VISTA TRAINING ALGORITHM

Algorithm 1 summarizes the training procedure in Section 3. Each rollout group supplies both the rewards for relative advantages and the experience for CVED. HAPI determines which trajectories receive each loss; it does not filter the sources used to construct collective experience. Context projection follows Appendix B.1, and optimization follows Appendix B.3.

Algorithm 1 VISTA training with CVED and HAPI

Input: Initial student θ; training data and task verifier; rollout count K; artifact budget B; distillation weight λ; EMA decay ρ.
Output: Trained active multimodal policy π_θ.

Initialize teacher parameters $\bar{\theta} \leftarrow \theta$.
for each training minibatch $\mathcal{Q}$ **do**
  Set rollout parameters $\theta_{\text{old}} \leftarrow \theta$.
  for each input $x \in \mathcal{Q}$ **do**
    Collect K interactive trajectories $\mathcal{B}_x$ with $\pi_{\theta_{\text{old}}}$; retain original histories, response tokens, tool observations, and rollout log probabilities.
    Verify rewards and terminal statuses for every trajectory.
    Compute advantages $\{A_k\}_{k=1}^{K}$ from the full group's rewards.
    Compile $\mathcal{E}_x \leftarrow \mathcal{C}(\mathcal{B}_x)$ from all trajectories.
    Count response tokens $N \leftarrow \sum_k |\mathcal{T}_k|$ and $N^- \leftarrow \sum_{k:\, \tau_k \text{ unsuccessful}} |\mathcal{T}_k|$.
    Compute $\mathcal{L}_{\text{GRPO}}^{\text{succ}}$ using Equation 5, with the success mask and full-group denominator N.
    Initialize accumulated distillation divergence $S_x \leftarrow 0$.
    for each unsuccessful trajectory $\tau_k \in \mathcal{B}_x$ **do**
      for each response y sampled from history H in τ_k **do**
        Project and align context: $\widetilde{H} \leftarrow H \oplus \text{Context}_B(\mathcal{E}_x; H)$.
        for each generated response-token position n **do**
          $P_n \leftarrow \pi_\theta(\cdot \mid H, y_{<n})$.
          $Q_n \leftarrow \text{stopgrad}\Big[\pi_{\bar{\theta}}(\cdot \mid \widetilde{H}, y_{<n})\Big]$.
          $S_x \leftarrow S_x + \text{JSD}(P_n, Q_n)$.
        end for
      end for
    end for
    $\mathcal{L}_{\text{CVED}} \leftarrow S_x / \max(1, N^-)$;
    $\mathcal{L}_x \leftarrow \mathcal{L}_{\text{GRPO}}^{\text{succ}} + \lambda \mathcal{L}_{\text{CVED}}$.
  end for
  Update θ with AdamW on $|\mathcal{Q}|^{-1} \sum_{x \in \mathcal{Q}} \mathcal{L}_x$, holding $\bar{\theta}$ fixed.
  Update the teacher: $\bar{\theta} \leftarrow \rho\bar{\theta} + (1-\rho)\theta$.
end for

Unsuccessful trajectories include incorrect, protocol-invalid, and incomplete attempts. Both losses exclude prompt tokens, tool observations, and padding; an empty eligible subset contributes zero, including a zero-token group. The teacher context includes selected visual artifacts, interaction structure, and categorical outcome feedback, but excludes reference answers and generated final-answer text from the added experience. The teacher scores the student's unchanged response prefixes without generating targets. At inference, only π_θ and its visual tools are used, conditioned on the agent's own history.

B Implementation Details

B.1 Experience Representation and Projection

The compiler retains decision records and artifact occurrences before applying a teacher-context budget. Canonical artifact clusters share representative images while retaining each occurrence's source trajectory and producing decision. Nested crops are mapped to input-image coordinates. Exact duplicates are merged; near-duplicate observations are merged when their root-coordinate box IoU is at least 0.90 or their perceptual-hash Hamming distance is at most 4. The internal graph stores temporal, provenance, derivation, and spatial relations for teacher-context construction.

Experience is compiled once per input, with a separate projected view for each decision. Selection prioritizes student-visible and peer evidence, favors successful-peer observations, and deprioritizes future-only observations from the target trajectory. Visual ancestors are included within the artifact budget. Alignment marks student-visible, peer-only, and future-self evidence, while caching avoids repeated projection and image processing.

Budgeted selection. The budget counts distinct selected crop artifacts, including their required crop ancestors; the original input image is already present in the student's history. The projector starts with an empty selected set and greedily adds a candidate together with its unselected ancestors, rejecting additions that would exceed B. For each candidate, the decision-conditioned score combines student visibility, the fraction of source trajectories that are peers, the fraction of those peers that succeeded, spatial novelty, newly covered source trajectories, and inverse source frequency. Their coefficients are 6, 4, 2, 1.5, 0.5, and 0.5, respectively; future-only evidence from the target trajectory receives a penalty of 2. Spatial novelty is one minus the largest root-coordinate IoU with an already selected artifact, or one when the selected set is empty. The score is divided by the number of new artifacts required by the candidate's ancestor closure. Ties are resolved by greater spatial novelty, fewer added artifacts, and then the smaller artifact identifier. Selection stops when the budget is exhausted or no feasible addition remains. If all artifacts fit, the complete set is retained. Only decisions producing selected artifacts and relations whose endpoints remain visible are serialized into the projected view.

B.2 Teacher Information and Supervision

The teacher receives selected visual observations together with their interaction context and categorical outcome feedback. Both successful and unsuccessful trajectories contribute candidate experience. The added context excludes generated final answers from all trajectories and does not include externally supplied reference answers. Interaction records describe evidence-producing decisions and relations without appending terminal answer text.

Table 8: Additional teacher context used by CVED. Generated final answers and reference answers are not included.

Channel	Content	Main setting
Visual observations	selected images from the rollout group	enabled
Interaction context	producing decisions and artifact relations	enabled
Outcome feedback	trajectory status and failure category	enabled

Teacher instructions distinguish incorrect direct answers from errors after visual interaction, malformed actions, and unfinished attempts. For an unsuccessful trajectory, a short instruction identifies the failure category and directs the teacher's guidance toward the current decision. These additions affect only the teacher context; the student's original history and sampled response remain unchanged. All response turns in a non-success trajectory receive distillation supervision. Synthetic decision records align malformed or unfinished actions that lack ordinary experience events, without replacing student responses or generating new target tokens.

The outcome source is the task verifier. Reference answers are used for verification, while only categorical feedback is exposed to the teacher.

B.3 Distribution Matching and Teacher Updates

At each response prefix, we compute the Jensen–Shannon divergence between the student and teacher token distributions (Lin, 1991). Teacher probabilities are detached from the gradient.

The GRPO term in Equation 4 retains the clipped policy ratio and full-group relative advantages. Successful-trajectory contributions are normalized by the full group's generated response-token count, whereas CVED averages over generated response tokens in unsuccessful trajectories. A term is zero when no eligible trajectories are present.

The teacher is initialized from the student and updated by exponential moving averaging after student optimization (Tarvainen & Valpola, 2017).

Update order and token alignment. For each training input, all K trajectories are first collected with the rollout policy, retaining the original response tokens and their rollout log probabilities. The verifier determines terminal outcomes, and group-relative advantages are computed before any trajectory is masked out of the policy loss. The complete rollout group is then compiled into collective experience. For each unsuccessful response, the teacher scores the same generated token prefixes as the student, with the selected experience appended only to the teacher's input. Prompt tokens, tool outputs, and padding do not contribute to either learning objective. Gradients update only the student; the teacher remains fixed during the student update and is updated afterward. This order prevents either teacher-generated replacements or teacher-context tokens from becoming student training targets.

Masked policy objective. Let $\mathcal{T}_k$ contain all generated response-token positions in trajectory k, and let A_k denote its full-group relative advantage. For token position u, write $r_{k,u}(\theta)$ for the probability ratio between the current student and the rollout policy, evaluated under the unchanged student history. The successful-trajectory term in Equation 4 is

$$\mathcal{L}_{\mathrm{GRPO}}^{\mathrm{succ}} = -\frac{1}{N} \sum_{k:\ \tau_k \text{ succeeds}} \sum_{u \in \mathcal{T}_k} \min(r_{k,u} A_k,\ \mathrm{clip}(r_{k,u}, 1-\epsilon, 1+\epsilon) A_k)\,, \tag{5}$$

where $N = \sum_{k=1}^{K} |\mathcal{T}_k|$ includes response tokens from the entire group and ϵ is the policy-ratio clipping threshold. Thus, masking unsuccessful trajectories removes their policy-gradient contributions without recomputing advantages from the successful subset or renormalizing the policy loss by its size. CVED instead averages its token divergences over $N^- = \sum_{k:\ \tau_k \text{ unsuccessful}} |\mathcal{T}_k|$. The two group-level terms are combined with weight λ and averaged across training inputs. An empty successful or unsuccessful subset contributes zero to its corresponding term.

Teacher update and divergence. Writing $\bar{\theta}$ for teacher parameters, the update after a student optimizer step is $\bar{\theta} \leftarrow \rho\bar{\theta} + (1-\rho)\theta$, initialized with $\bar{\theta} = \theta$. No gradient is propagated through this update or through teacher probabilities. The default divergence is equal-mixture JSD: for student and teacher distributions P and Q, its mixture is $M = (P+Q)/2$ and its value is $[D_{\mathrm{KL}}(P\|M) + D_{\mathrm{KL}}(Q\|M)]/2$. The divergence ablation changes this distribution-matching objective while retaining CVED context construction and HAPI routing. The frozen-teacher ablation retains the initialized teacher throughout training instead of applying EMA.

Learning-rule controls. GRPO-only applies the policy objective to all trajectories and omits distillation. CVED-only applies experience-conditioned distillation to all trajectories and omits the policy objective. Uniform hybrid applies both objectives to every trajectory, whereas HAPI restricts reinforcement to successful trajectories and distillation to unsuccessful trajectories. All distillation controls construct their teacher context from the complete rollout group. In the λ sweep, routing and the token denominators of HAPI remain fixed: $\lambda = 0$ disables only the unsuccessful-trajectory distillation term, leaving Equation 5 unchanged.

B.4 Core Prompt Templates

This section specifies the prompt templates used by the student and the experience-conditioned teacher. Native image content and the tool schema are inserted during model-input construction.

Visual tool interface. The agent calls `image_zoom_in_tool` with a bounding box `bbox_2d`, a region `label`, and an image index `img_idx`. The box uses coordinates in $[0, 1000]$ relative to the selected image, and the zero-based image index selects the original image or a previously returned crop. The environment extracts the requested region and appends the resulting image to the interaction history. Nested crops retain their parent image and are mapped back to full-image coordinates for experience construction. These boxes are generated by the policy; they are not target annotations supplied with the training input.

Prompt layouts. The teacher receives visual experience, interaction context, and categorical outcome feedback, but no reference answer or peer final answer.

```
Your role is that of a research assistant specializing in visual
information. Answer questions about images by looking at them
closely and then using research tools. Please follow this
structured thinking process and show your work.

Start an iterative loop for each question:
- First, look closely: describe the image and identify what must
be found.
- Next, find information: use a tool to research the missing
information.
- Then, review the findings: analyze the result and decide on
the next action.

Continue until the research is complete. Put the final answer
within \boxed{} and make it contain only the answer itself. For
multiple-choice questions, return the selected option's letter
directly, for example \boxed{A}.
```

Figure 3: Student system prompt used for active visual rollouts.

Wrapper serialization. Figure 5 follows `append_teacher_context` and the view and alignment serializers in `ExperienceSerializer`. The teacher instruction is appended to the system message, whereas this wrapper is appended to the last message of the student's unchanged input history. The trajectory record repeats for every trajectory in the group. Decision and artifact records repeat for the selected view, followed by the retained relations and diagnostics. Braced variables are replaced with input-specific values. Fields ending in `_json` denote JSON-encoded strings or lists, including their quotation marks or brackets. A missing decision bounding box is serialized as `[]`, and `valid` is an integer boolean.

The `failure` field is appended only when a failure type is available. For a trajectory without an outcome, its record contains only `trajectory` and `decisions`. Each relation record groups one relation type, with `source>target` pairs separated by semicolons. Relations of type `member_of` are omitted, and other relations appear only when both endpoints are visible in the selected view. A diagnostic record appears only when it references at least one selected artifact. In the alignment block, a missing target decision is written as `unknown`. Each `<image>` marker is replaced by a native image content item before model input construction. The number of records depends on the input. Generated intent descriptions are sanitized and length-limited during serialization.

Final answers are not serialized into the main teacher context, and the student's original history and sampled response remain unchanged. The resulting augmentation is therefore available only during training and does not create an inference-time memory.

```
[COLLECTIVE_EXPERIENCE_TEACHER_INSTRUCTION]
You are the privileged visual teacher for the student's current
decision. COLLECTIVE_VISUAL_EXPERIENCE is a shared,
training-only view compiled from completed on-policy
trajectories for the same input. It contains visual artifacts,
action/provenance relations, categorical outcomes, and a
DECISION_ALIGNMENT block marking student-visible, peer-only, and
future-self evidence.

Use collective experience to evaluate the student's exact
current response. Do not generate or imitate a replacement
trajectory. Outcome association is not causal evidence: an
artifact from a successful trajectory may be incidental, and an
artifact from a failed trajectory may still be useful. Inspect
the images and reason about the original question yourself.

Do not reveal teacher-only context. If evidence is insufficient,
prefer one non-redundant crop that can resolve the remaining
uncertainty. Follow the native tool schema and make at most one
tool call in this turn.
[/COLLECTIVE_EXPERIENCE_TEACHER_INSTRUCTION]
```

Figure 4: Collective-experience teacher instruction. The teacher evaluates the student's unchanged response under training-only visual hindsight.

```
[COLLECTIVE_EXPERIENCE_TEACHER_CONTEXT]
The following shared experience view is teacher-only privilege:
[COLLECTIVE_VISUAL_EXPERIENCE]
task={task_id_json} policy_version={policy_version}
outcome_semantics=categorical trajectory hindsight; association
is not causal evidence
trajectory={trajectory_id_json} outcome={status}
valid={valid_integer} crops={crop_count}
decisions={decision_count} failure={failure_type}
decision={decision_id_json} kind={kind} tool={tool_name_or_none}
src={source_artifact_id_or_none}
produces={produced_artifact_id_or_none} bbox={bbox_list}
label={label_json} intent={sanitized_intent_json}
artifact={artifact_id} bbox={root_bbox_list}
seen={source_trajectory_ids_json} outcome={association_type}
ok={successful_trajectory_ids_json}
fail={failed_trajectory_ids_json}
unknown={unverified_trajectory_ids_json} <image>
relations={relation_type}:{semicolon_joined_source_target_pairs}
diagnostic={diagnostic_type} target={target_trajectory_id_json}
artifacts={visible_artifact_ids_list}
[/COLLECTIVE_VISUAL_EXPERIENCE]
[DECISION_ALIGNMENT]
target_trajectory={target_trajectory_id_json} target_step={step}
target_decision={target_decision_id_or_unknown}
student_visible={student_visible_artifact_ids_json}
peer_only={peer_only_artifact_ids_json}
self_future={self_future_artifact_ids_json}
[/DECISION_ALIGNMENT]
[/COLLECTIVE_EXPERIENCE_TEACHER_CONTEXT]
```

Figure 5: Teacher-context serialization with visual experience, interaction context, and outcome feedback. Braced variables are instantiated for each input.

C Additional Experimental Details

Models and training data. We use Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct as the default backbones (Bai et al., 2025). Within each backbone, trainable methods start from the same checkpoint and use the Vision-OPD-6K training split (Yuan et al., 2026), which contains 6,241 image–question–answer examples. The training inputs use the original full images without overlaid target boxes. Questions contain no added instructions that identify a target region or restrict attention to an annotated spatial region. The student therefore chooses where to inspect from the image and task question, rather than following a supplied localization hint. Reference answers are retained for reward verification, not appended to the student's input. Regional annotations are used only for training-time privileged supervision in the relevant baselines and the Oracle ROI ablation. The default configuration uses $K = 8$ trajectories per input, distillation weight $\lambda = 0.1$, and visual artifact budget $B = 4$. Controlled comparisons share the optimizer and training schedule, with matched rollout settings, visual tools, and interaction limits. Rewards are based on answer correctness, and tool use is not mandatory.

Training configuration. Table 9 summarizes the default configuration shared by Qwen3-VL-4B and Qwen3-VL-8B. Training lasts 65 optimizer steps, corresponding to approximately one epoch of input exposure: $65 \times 96/6{,}241 \approx 1$ equivalent data pass. The teacher is initialized from the student and updated after each student optimizer step.

Benchmarks and metrics. We evaluate on V*Bench (Wu & Xie, 2023), ZoomBench (Wei et al., 2026), HR-Bench-4K and HR-Bench-8K (Wang et al., 2025b), and the English and Chinese splits of MME-RealWorld (Zhang et al., 2025b). The training-method comparison pairs V*Bench and the two HR-Bench variants with MMStar (Chen et al., 2024), MMBench (Liu et al., 2024), and MathVista (Lu et al., 2024) to assess both fine-grained perception and general reasoning. Component ablations, backbone comparisons, and parameter sweeps use V*Bench, HR-Bench-4K, and HR-Bench-8K. We report accuracy (%), with Average or Avg. denoting the unweighted mean of the benchmarks in the corresponding table. Unless otherwise specified in the caption, bold and underlined values mark the best and second-best scores, including ties, with rankings computed separately within each backbone where applicable.

Baselines and protocol. We compare VISTA with the base checkpoint, GRPO (Shao et al., 2024), and the distillation methods OPSD (Zhao et al., 2026), VCSD (Liang et al., 2026), Vision-OPD (Yuan et al., 2026), and Imagine-OPD (Cai et al., 2026) at both model scales. All experiments report mean accuracy over three random seeds, with deterministic evaluation. All scores in the broader model comparison are obtained by our own evaluation of the open-source, thinking-with-images, and proprietary models on the benchmark splits specified above. Models in the 4B–9B range provide size-comparable references, while larger open-source and proprietary models provide additional capability references.

Scope of the controlled comparison. Table 2 compares training methods separately within each backbone, rather than attributing differences between unrelated released checkpoints to the learning objective. The base row evaluates the starting checkpoint without additional training. For the trained methods, the shared controls are the starting checkpoint, training examples, student input format, optimizer and schedule, visual-tool interface, and interaction limits. The method-specific supervision is the variable of interest: GRPO uses verified outcome rewards, whereas the OPD baselines retain their respective teacher-conditioning mechanisms and distillation objectives. Matching the student-side interface does not give a baseline access to CVED's collective context or apply HAPI routing to it. Training-schedule matching is distinct from equal computational cost, since constructing privileged context and evaluating a teacher introduce method-dependent overhead.

Privileged-context ablations. The CVED ablation keeps the training inputs and HAPI learning rule unchanged. Answer-only supplies the reference answer to the teacher in place of visual interaction experience; the reference answer is not appended to the student history. Oracle ROI supplies annotated region crops as teacher-side visual evidence, without drawing boxes on the student's full image or adding localization instructions to its question. Unstructured and CVED use the same selected observations and context budget. Unstructured presents observations and interaction

Table 9: Default VISTA training configuration for both backbone sizes. Equivalent data passes are computed as optimizer steps times input questions per batch divided by training-set size.

Parameter	Value
Optimization and training duration	
Optimizer	AdamW
Learning rate	1×10^{-6}
Weight decay	0.01
Optimizer steps	65
Equivalent data passes (epochs)	≈ 1
Input questions per batch	96
Rollouts per question (K)	8
Sampled trajectories per batch	768
Policy-ratio clipping threshold (ϵ)	0.2
Maximum gradient norm	1.0
Gradient checkpointing	Enabled
Collective experience and teacher	
Distillation weight (λ)	0.1
Visual artifact budget (B)	4
Distribution-matching objective	Equal-mixture JSD
Teacher initialization	Student checkpoint
Teacher update	EMA after each optimizer step
EMA decay (ρ)	0.99
Rollout sampling and interaction limits	
Temperature / top-p	1.0 / 1.0
Top-k truncation	Disabled
Maximum interaction turns / crop calls	5 / 4 per trajectory
Maximum generated tokens	1,024 per response
Student / teacher prompt limit	16,384 / 24,576 tokens
Maximum image pixel budget	12,845,056
Checkpointing and repeated runs	
Checkpoint saving	Every 10 optimizer steps and at training end
Random seeds	3

records in trajectory order, whereas CVED retains artifact relations and decision-relative alignment. Consequently, the privileged input changes on the teacher side without changing the image–question task solved by the student.

Parameter and teacher controls. The parameter analyses use Qwen3-VL-4B with $K = 8$, $\lambda = 0.1$, and $B = 4$ unless the corresponding parameter is varied. Changing K changes the number of sampled attempts per input, whereas changing B changes only the number of artifacts available in a projected teacher context. The λ sweep varies the coefficient of unsuccessful-trajectory distillation under fixed HAPI routing, including its literal zero-weight setting. The teacher-update comparison changes whether the initialized teacher remains frozen or follows the student through EMA; it does not change the teacher's information channels.

Reference models. The open-source references include MiMo-VL (LLM-Core-Team Xiaomi, 2025), MiniCPM-V 4.5 (Yu et al., 2025), GLM-4.6V (Z.ai, 2025), and Kimi K2.6 (Moonshot AI, 2026). Proprietary references comprise GPT-5.2 and GPT-5.4 (OpenAI, 2025; 2026), Gemini 2.5 Pro (Comanici et al., 2025), and Gemini 3 Flash (Doshi, 2025). The thinking-with-images baselines are Thyme, DeepEyes, DeepEyesV2, and Pixel Reasoner, discussed in Appendix F.

D Effect of Cross-Trajectory Visual Experience

Table 10 compares hindsight from the student's own trajectory with collective visual experience from the rollout group. Self-only hindsight restricts candidate visual observations and their interaction records to the target trajectory, including observations acquired later in that trajectory. Collective visual experience uses the full group's observations and interaction records, as in the default CVED configuration. Both settings retain the same group-level categorical outcome-feedback records, HAPI routing, teacher updates, and context-serialization format. The comparison uses Qwen3-VL-4B with $K = 8$, $\lambda = 0.1$, and the same artifact-budget cap $B = 4$, following the training and evaluation protocol in Appendix C. This contrast tests the added value of peer visual experience beyond the target trajectory's own hindsight under shared outcome feedback.

Table 10: Experience-source ablation on Qwen3-VL-4B. Accuracy (%) is averaged over three seeds; Avg. is the unweighted mean across the three benchmarks. The collective-experience row reproduces the default CVED results in Table 3. Bold and underlined values mark the highest and lowest scores in each column, respectively.

Experience source	V*	HR-4K	HR-8K	Avg.
Self-only hindsight	<u>85.34</u>	<u>82.25</u>	<u>77.50</u>	<u>81.70</u>
Collective visual experience	**88.48**	**83.50**	**79.50**	**83.83**

Collective visual experience improves average accuracy from 81.70% to 83.83%, a gain of 2.13 percentage points over self-only hindsight. The gains are positive on all three benchmarks: 3.14 points on V*Bench, 1.25 on HR-Bench-4K, and 2.00 on HR-Bench-8K. With categorical outcome feedback available in both settings, this comparison supports the value of peer visual observations and their interaction context beyond evidence acquired within the target trajectory.

E Disentangling HAPI Routing

Table 11 crosses the trajectory subsets used for reinforcement and distillation. The RL mask restricts reinforcement to successful trajectories, while the distillation mask restricts CVED to unsuccessful trajectories, including invalid and incomplete attempts. The two intermediate settings apply one mask at a time; HAPI applies both. All four settings construct CVED context from the complete rollout group and use Qwen3-VL-4B with $K = 8$, $\lambda = 0.1$, and $B = 4$. Relative advantages are computed from the full group, and the policy term retains the full-group response-token denominator N. The distillation term averages over response tokens in its eligible subset, following the learning-rule definitions in Appendix B.3. The teacher updates, optimizer, training schedule, and evaluation protocol are shared across settings. The crossed comparison tests the contribution of each routing mask and their combination under these loss-normalization conventions.

Table 11: Crossed routing ablation on Qwen3-VL-4B. We report three-seed mean accuracy (%) and the unweighted benchmark average. Uniform hybrid and HAPI reproduce Table 4. Bold and underlined values mark the highest and lowest scores in each column, respectively.

Setting	RL subset	Distillation subset	V*	HR-4K	HR-8K	Avg.
Uniform hybrid	All	All	87.43	82.38	78.80	82.87
RL mask only	Successful	All	<u>86.91</u>	82.50	<u>78.25</u>	<u>82.55</u>
Distillation mask only	All	Unsuccessful	87.43	<u>82.25</u>	78.92	82.87
HAPI	Successful	Unsuccessful	**88.48**	**83.50**	**79.50**	**83.83**

Neither mask alone improves average accuracy over the uniform hybrid. Restricting RL to successful trajectories lowers the average from 82.87% to 82.55%, while restricting distillation to unsuccessful trajectories leaves it essentially unchanged at 82.87%. Both individual masks produce mixed changes across benchmarks. Applying the two masks together yields the highest accuracy on every benchmark and raises the average to 83.83%, 0.96 percentage points above the uniform hybrid. This pattern supports the joint allocation of reinforcement and distillation under the stated normalization conventions, with the benefit of each routing choice depending on the other.

F Additional Related Work

F.1 RLVR for Multimodal Agents

Reinforcement learning with verifiable rewards provides a way to learn visual tool use from task outcomes rather than requiring demonstrations for every intermediate action. DeepEyes learns active perception through end-to-end reinforcement learning without a cold-start supervised fine-tuning stage (Zheng et al., 2025). VTool-R1 integrates Python-based visual editing into reinforcement fine-tuning, allowing models to learn multimodal reasoning chains under task-accuracy rewards (Wu et al., 2025). Visual-ARFT extends agentic reinforcement fine-tuning to image manipulation and web search (Liu et al., 2025). Thyme learns to generate executable code for image processing and computation (Zhang et al., 2025a), while DeepEyesV2 combines cold-start training with reinforcement learning for multimodal tool use (Hong et al., 2025). Together, these studies shift multimodal policy learning from reasoning over a fixed visual input toward acquiring new evidence during interaction.

Subsequent developments refine how visual actions are explored and evaluated. Pixel Reasoner combines instruction tuning with curiosity-driven reinforcement learning to encourage the use of pixel-space operations (Wang et al., 2025a). MAPO addresses the mismatch between textual reasoning and visual actions by incorporating description–observation alignment into advantage estimation (Yang et al., 2026b). Mini-o3 scales visual search through diverse reasoning patterns and longer tool interactions (Lai et al., 2025). Where these approaches improve exploration or reward-based credit assignment, VISTA focuses on transferring the observations produced by exploration. It treats a same-input rollout group as a source of collective visual experience and turns discoveries across trajectories into decision-aligned supervision for the policy.

F.2 OPD for Multimodal Agents

On-policy distillation provides teacher supervision along responses generated by the student, reducing the distribution mismatch associated with fixed teacher-generated sequences (Agarwal et al., 2023). On-policy self-distillation uses the same model as teacher and student with different contexts, allowing privileged information to provide a stronger teaching signal (Zhao et al., 2026; Penaloza et al., 2026). On-Policy Context Distillation demonstrates how a context-conditioned teacher can transfer experiential knowledge into a language model through this mechanism (Ye et al., 2026). For language agents, OPID extracts hierarchical skill guidance from completed on-policy trajectories (Yang et al., 2026a).

For multimodal learning, Vision-OPD matches a full-image student's next-token distributions to those of a crop-conditioned teacher along student-generated rollouts (Yuan et al., 2026). Imagine-OPD conditions the teacher on annotated regional evidence to supervise tool-free imagination trajectories (Cai et al., 2026). Visual-OPSD instead conditions the teacher on generated visual-thought traces and distills their reasoning benefit into a student that does not generate those visual thoughts at inference (Li et al., 2026). These approaches show how a richer visual context can improve a student without retaining that context at deployment.

Other work focuses on the selectivity of visual supervision. VA-OPD measures the teacher's dependence on fine-grained visual detail and uses this visual advantage for rollout reweighting and token-grouped distillation (Liu et al., 2026). V-Zero contrasts question-relevant regional evidence with a negative visual view to gate distillation on student-sampled trajectories (Sun et al., 2026). VCSD instead contrasts the original image with a content-erased control to construct visually grounded targets without privileged answers or evidence crops (Liang et al., 2026). VISTA addresses a different learning problem: internalizing visual experience discovered collectively through active interaction. CVED assembles this experience from same-input rollouts and aligns it with the student's actual decisions. HAPI uses trajectory outcomes to distinguish reward-based reinforcement of successful behavior from experience-guided distillation on failed attempts, rather than reweighting distillation solely by visual dependence or contrastive evidence.

G LIMITATIONS AND FUTURE WORK

VISTA offers several directions for further development. First, the value of collective experience depends on the diversity and relevance of the group's observations; adaptive exploration and evidence selection could improve supervision when rollouts are redundant or miss useful regions. Second, HAPI uses terminal outcomes to route supervision, motivating finer-grained feedback that distinguishes productive intermediate decisions from unsuccessful ones. Third, experience construction and teacher inference introduce additional training overhead; more compact experience representations and selective teacher queries could improve efficiency, with comparisons under matched total compute budgets clarifying the trade-offs. Finally, our evaluation focuses on image-question answering with crop-and-zoom interaction and Qwen-family backbones. Extending this framework to other model families, video, and broader tool-use environments would help assess its generality.