DISTILL THE TEACHER, CREDIT THE ACTION: WHAT ON-POLICY DISTILLATION TRANSFERS TO A VISUAL AGENT'S ACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) densely supervises student-generated trajectories, but for agents that act before answering, an action is distilled before execution reveals whether it was useful. For visual crop agents, we find that nearly all of OPD's gain comes from learning to read the acquired crop rather than where to look. Meanwhile, sampled crops shift toward the teacher's action distribution, which in our main setting is less aligned with the evidence. OPD uses each executed crop to train the reader but never credits the action for what that crop reveals—an *action-credit gap*. Execution-Conditioned Entropy Drop (Exec-ED) closes this gap without labels or changes to OPD: it credits each sampled action by how much its executed crop reduces the teacher's answer uncertainty, centered among sibling crops sampled for the same state. Across two teacher sizes and two benchmarks, Exec-ED improves crop acquisition while maintaining comparable end-to-end accuracy. In the main 9B setting, it raises ZoomBench crop precision from 55.0% under OPD to 57.4–58.8% (Base: 59.2%), and its crops are more useful to a held-out reader. This weak credit works in tandem with OPD's action KL: the KL anchors the actor near the teacher, while Exec-ED tilts its sampled actions toward more useful crops. More broadly, a teacher can show a student what to do, but only execution shows what worked; distillation for agents that act should learn from both.

1 INTRODUCTION

Multimodal agents increasingly answer questions about high-resolution images by acting on them: they crop or zoom into a region, observe the result, and answer from the acquired view (Wu & Xie, 2024; Zheng et al., 2026; Su et al., 2026; Wu et al., 2026). Small agents can be trained with reinforcement learning from task rewards (Su et al., 2026; Wu et al., 2026), while on-policy distillation (OPD) offers dense supervision by matching a frozen teacher on student-generated trajectories (Agarwal et al., 2024; Gu et al., 2024). Recent multimodal variants further focus this supervision on visually dependent or teachable tokens (Yuan et al., 2026; Liu et al., 2026; Wang et al., 2026). Yet acting creates a mismatch that token supervision alone does not resolve: answer tokens are supervised after the crop is available, whereas action tokens are supervised before the model can know whether the crop they produce will be useful. What, then, does OPD actually teach a visual agent to acquire?

We find that OPD teaches the student far more about how to read than where to look. With a Qwen3.5-9B teacher and Qwen3.5-2B crop agent, reader–actor swaps attribute 9.5 of OPD's 9.6-point gain in correct-answer probability on ZoomBench to the reader (Section 3). The actor instead shifts toward the teacher's wider, less evidence-aligned action distribution: the greedy crop moves toward the evidence, while the sampled crops—the very actions OPD trains on—move away from it. This asymmetry is structural. OPD matches actions to the teacher before execution, while its post-execution answer loss trains the reader on the resulting crop without crediting that crop back to the action that produced it. That the action loss is execution-blind follows directly from the objective; the learned consequence does not. Consistently, the action KL is essentially uncorrelated with crop quality, reweighting distilled tokens does not restore the sampled crops, and explicit execution-conditioned credit does. We call this the *action-credit gap*: OPD supervises every sampled action but never credits it by what its execution achieves.

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

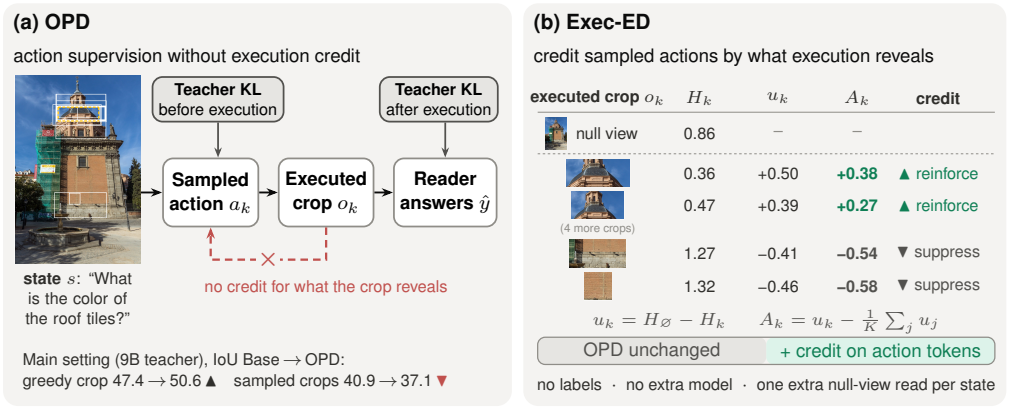

Figure 1: **Distill the teacher, credit the action.** (a) OPD supervises action tokens before execution and answer tokens after, but never credits a sampled action by what its executed crop reveals. In our main setting (9B teacher), the greedy crop improves while sampled crops worsen. (b) Exec-ED leaves OPD unchanged, computes the teacher's entropy drop for each executed crop relative to a null view, centers it across $K{=}8$ sibling crops, and applies the resulting credit only to action tokens.

Can this missing credit be supplied without task labels? A weak signal is already available from the teacher: its answer-entropy drop after reading a crop weakly ranks crops by evidence quality. **Execution-Conditioned Entropy Drop (Exec-ED)** turns this signal into credit (Fig. 1). It leaves OPD unchanged, measures each executed crop's entropy drop relative to an unzoomed null view, centers the drop among crops sampled for the same state, and applies the result only to the action tokens. Because OPD already evaluates the teacher on every executed crop, Exec-ED needs no labels, no additional model, and only one extra null-view read per state (Section 4).

Across 9B and 4B teachers, Exec-ED recovers much of the crop quality OPD loses while maintaining comparable task accuracy (Section 5). It improves ZoomBench crop precision and V* target coverage and makes the crops more useful to a held-out reader. The improvement persists when OPD and Exec-ED sample at lower temperatures, and reader–actor swaps locate the gain in the actor rather than the reader. Analysis shows how the two signals interact (Section 6): OPD's action KL anchors the actor, while Exec-ED acts as a standing execution-conditioned counterweight.

Our contributions are threefold: **(1)** we diagnose an action-credit gap in OPD: nearly all of its answer gain is reader-side, while its sampled actions shift toward the teacher and away from visual evidence; **(2)** we introduce Exec-ED, a minimal label-free residual that credits sampled actions using sibling-relative post-execution entropy drop and recovers much of OPD's lost crop quality at comparable accuracy; and **(3)** we show that distillation and execution credit are complementary: OPD anchors the actor, while execution credit tilts the anchored distribution toward more useful actions. Together, these results suggest viewing distillation for agents as two signals rather than one: what the teacher would do, and what the student's actions achieve.

2 Related Work

On-policy distillation for VLMs. Knowledge distillation transfers teacher predictions, generated sequences, and policies into smaller students (Hinton et al., 2015; Kim & Rush, 2016; Rusu et al., 2015). On-policy distillation instead queries the teacher on student-generated prefixes, reducing training–inference mismatch in the spirit of learner-state supervision such as DAgger (Ross et al., 2011; Agarwal et al., 2024; Gu et al., 2024). On these prefixes, Agarwal et al. (2024) match the student's full next-token distributions to the teacher's using divergences such as reverse KL, which we adopt here. For multimodal students, recent variants focus this supervision on visual information: VA-OPD reweights rollouts and token groups by the teacher's dependence on visual detail (Liu et al., 2026), Yoon et al. (2026) separate language-prior and visual-grounding gradients, and Xue et al. (2026) project teacher corrections onto locally realizable visual directions; TA-OPD instead selects teacher corrections within the student's predictive support (Wang et al., 2026). These methods refine

token targets, weights, or update directions; none of them, however, credits a sampled action by what its execution reveals, and what OPD transfers to the actions a student samples has not, to our knowledge, been examined.

Supervising executed visual actions. Outcome supervision has also been combined with OPD. VOLD adds GRPO to distillation, RG-OPD gates teacher supervision with verifier feedback, and OPDVR adjusts token rewards using trajectory correctness (Bousselham et al., 2026; Akhondzadeh et al., 2026; Lin et al., 2026). Outside OPD, Pixel Reasoner and VTool-R1 train visual operations with task rewards (Su et al., 2026; Wu et al., 2026), while Ma et al. (2026) show that crop-and-zoom RL gains need not come primarily from better tool use. More recent work assigns credit at finer granularity: Wang et al. (2025) reward individual search turns by their increase in gold-answer confidence, TACO probes the effect of each visual tool call on answer correctness (Feng et al., 2026), and CounterCredit tests whether an executed visual call was both needed and used (Peng et al., 2026). NTEP instead supervises necessary tool–evidence paths (Long et al., 2026). Privileged-view methods take a different route: Vision-OPD distills from crop-conditioned perception, while V-Zero contrasts relevant and negative regional views to gate token-level distillation (Yuan et al., 2026; Sun et al., 2026). Existing execution-level credit thus relies on task outcomes or supplied visual evidence; a crop the student executes has not itself been credited inside OPD without either.

Label-free rewards from model certainty. Entropy minimization has long provided learning signals without labels, from semi-supervised learning to test-time adaptation (Grandvalet & Bengio, 2004; Wang et al., 2020). Recent reasoning methods derive rewards from the model itself: Intuitor uses self-certainty, while EM-RL and RENT optimize negative entropy (Zhao et al., 2026; Agarwal et al., 2026; Prabhudesai et al., 2025). Self-consistency motivates another route: TTRL converts majority-voted answers into rewards on unlabeled problems (Wang et al., 2022; Zuo et al., 2026), and Hu et al. (2026) extend self-induced outcome structure to verifier-free turn-level credit in search agents. Closest to the signal studied here, Li et al. (2026) use entropy reduction to supervise tool use, but their rewards remain anchored to final-answer correctness. Exec-ED's utility belongs to this family, and its estimator is the sibling-centered REINFORCE of GRPO (Shao et al., 2024); what differs is where they act: on the action tokens inside OPD, with the frozen teacher reading each executed view, rather than in a stand-alone objective. Model certainty is not a correctness guarantee (Guo et al., 2017): Zhang et al. (2025) show that prolonged internal-feedback optimization can erase early gains or degrade instruction-tuned models. Reference-policy KL regularization is a standard stabilizer in language-model RL (Ouyang et al., 2022); Intuitor likewise reports sensitivity to this constraint (Zhao et al., 2026). Whether a label-free belief change can credit an executed visual action inside distillation is therefore a question of anchoring as much as of signal.

3 Diagnosing visual acquisition under on-policy distillation

We ask what OPD transfers to a visual agent's actor, what it misses, and what post-execution signal could supply the missing supervision. That OPD's action loss is blind to what an executed action later reveals follows from the objective; its consequence for the learned actor does not, and is what we measure.

3.1 Setup

Agent and objective. A policy p_θ acquires evidence before answering. In state s (image and question), it emits a crop action $a \sim p_\theta(\cdot \mid s)$ whose four coordinates are tokens on a 0–999 grid; executing a returns the cropped view $o = \mathcal{E}(s, a)$, after which the policy answers $y \sim p_\theta(\cdot \mid s, o)$. We call $p_\theta(a \mid s)$ the *actor*, $p_\theta(y \mid s, o)$ the *reader*, and use *acquisition* to refer to the usefulness of the acquired view for answering.

OPD (Agarwal et al., 2024) minimizes the full-vocabulary reverse KL to a frozen teacher p_T, averaged over the tokens of trajectories $\tau = (a, o, y)$ sampled by the student:

$$\mathcal{L}_{\mathrm{OPD}}(\theta) = \mathbb{E}_{s,\,\tau\sim p_{\bar\theta}}\left[\frac{1}{|\mathcal{A}\cup\mathcal{Y}|}\sum_{t\in\mathcal{A}\cup\mathcal{Y}} D_t\right], \qquad D_t = D_{\mathrm{KL}}\big(p_\theta(\cdot \mid h_t)\,\big\|\,p_T(\cdot \mid h_t)\big), \tag{1}$$

Table 1: Training signals of the diagnosed models (Eqs. (1)–(2)).

	OPD	w/o action KL	TA-OPD	VA-OPD	Exec-Acc	OPD+RLVR
Teacher KL on	$\mathcal{A} \cup \mathcal{Y}$	$\mathcal{Y}$	$\mathcal{A} \cup \mathcal{Y}$	$\mathcal{A} \cup \mathcal{Y}$	$\mathcal{A} \cup \mathcal{Y}$	$\mathcal{A} \cup \mathcal{Y}$
KL weights	uniform	uniform	top 10%	visual	uniform	uniform
Correctness credit on	–	–	–	–	$\mathcal{A}$	$\mathcal{A} \cup \mathcal{Y}$

where $\bar{\theta}$ is a copy of the current parameters that receives no gradient. Thus trajectories are drawn on-policy but held fixed, and gradients reach the student only through D_t. Here $\mathcal{A}$ and $\mathcal{Y}$ denote the action and answer tokens, respectively, and h_t is the prefix before token t. For answer tokens, h_t contains the executed crop o, so OPD also supervises how the student reads its own crop.

Diagnosed models. We diagnose a Qwen3.5-9B teacher and a Qwen3.5-2B student (Team, 2026). All models start from the initial student (Base), share one training recipe (Section 5.1), and differ only in their training signal (Table 1). Besides OPD and OPD without the action-token KL, *token-side* variants reweight the KL terms of Eq. (1): TA-OPD (Wang et al., 2026) keeps only tokens scored as most teachable, including action tokens in our adaptation, while VA-OPD (Liu et al., 2026) weights trajectories and tokens by the teacher's dependence on visual detail (Appendix B).

Label-based controls instead add to OPD a REINFORCE residual (Williams, 1992) over the K trajectories τ_k (*siblings*) sampled for each state:

$$\mathcal{L}_{\mathcal{C}}(\theta) = \mathcal{L}_{\mathrm{OPD}}(\theta) - \lambda\, \mathbb{E}_{s,\, \tau_{1:K} \sim p_{\bar{\theta}}} \left[\frac{1}{K} \sum_{k=1}^{K} A_k \sum_{t \in \mathcal{C}_k} \log p_\theta(\tau_{k,t} \mid h_{k,t}) \right], \tag{2}$$

where $A_k = r_k - \frac{1}{K}\sum_j r_j$ centers the correctness $r_k = \mathbf{1}[\hat{y}_k = y^\star]$ within the state. *Exec-Acc* credits only the action tokens ($\mathcal{C}_k = \mathcal{A}_k$), whereas *OPD+RLVR* credits both action and answer tokens ($\mathcal{C}_k = \mathcal{A}_k \cup \mathcal{Y}_k$), following trajectory-level verifiable-reward objectives combined with OPD (Bousselham et al., 2026; Lin et al., 2026). Both retain the OPD term.

Measuring acquisition. We measure acquisition on 1,000 random training states, whose annotated evidence boxes are never used for training, and on ZoomBench-MCQ (606 questions with evidence boxes; Wei et al., 2026), from two angles (Appendix D):

- *Sampled and greedy crops.* Crops sampled as during training (T=1, top-p 0.95) are the on-policy actions OPD trains on; the greedy crop takes the most likely token at every step.
- *Evidence geometry.* Without using any reader, IoU measures overlap with the evidence box, mean pairwise IoU between a state's sampled crops measures their spread, and precision (intersection over crop area) measures how tightly a crop targets the evidence.

3.2 OPD DISTILLS READING AND TRANSFERS THE TEACHER'S ACTION DISTRIBUTION

Where does OPD's gain come from? Measured by a reader's probability of answering correctly on ZoomBench, OPD gains 9.6 points. Reader–actor swaps with Base show that OPD's reader alone recovers 9.5 of them, whereas its actor alone contributes at most 0.9 (Table 9).

The actor nevertheless changes, and in two opposite ways (Fig. 2a). On the 1,000 diagnosis states, OPD moves the greedy crop toward the evidence (IoU $47.4 \rightarrow 50.6$) but its sampled crops away from it ($40.9 \rightarrow 37.1$). The samples also spread nearly as widely as the teacher's (mean pairwise IoU $0.53 \rightarrow 0.43$; teacher 0.40). The same split holds out of sample on ZoomBench: greedy IoU rises from 39.2 to 43.1, while sampled IoU falls from 33.4 to 31.1. This distinction matters because sampled crops are the on-policy actions OPD trains on and the views its reader learns to answer from.

OPD therefore improves the reader while shifting the actor toward the teacher's action distribution: in our 9B setting, its greedy crop improves, but its sampled crops get worse.

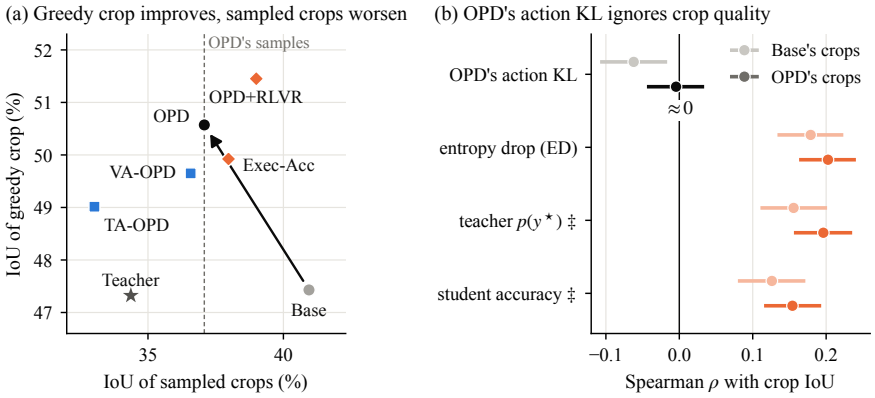

Figure 2: **Diagnosis of OPD** (Qwen3.5-9B→2B, 1,000 training states). (a) IoU of each model's greedy crop against the mean IoU of its sampled crops; the arrow marks the shift from Base to OPD, and the dashed line OPD's sampled-crop IoU. Blue: token-side variants; orange: label-based controls. (b) Within-state Spearman ρ between crop IoU and either OPD's action KL or a post-execution readout, for crops sampled by Base (lighter) and OPD (darker), with 95% bootstrap intervals; ‡ uses the task label.

3.3 OPD SUPERVISES EVERY SAMPLED ACTION BUT CREDITS NONE

Why do the sampled crops not improve? In Eq. (1), action tokens are supervised before the crop is executed, toward the teacher's prediction at each prefix, regardless of what the crop later reveals. Answer tokens are supervised after execution and therefore condition on the crop, but the sampled trajectory is held fixed: their loss trains the reader on that crop without sending credit back to the action that produced it. **OPD observes each crop's consequence when training the reader, but never credits that consequence to the sampled action.**

Consistently, OPD's action KL on a sampled crop—the student–teacher KL summed over its action tokens—is essentially uncorrelated with crop IoU ($|\rho| < 0.07$; Fig. 2b). The teacher's own sampled crops also overlap the evidence less than Base's (Fig. 2a), so imitation alone need not favor evidence-aligned actions.

Reweighting imitation does not supply the missing credit. TA-OPD and VA-OPD change the weighting of the KL but introduce no execution-dependent signal to the action. On the diagnosis states, their sampled crops remain as far from the evidence as OPD's, or farther (Fig. 2a). By contrast, explicitly crediting each sibling by the consequence of its execution does move the samples: the label-based controls Exec-Acc and OPD+RLVR shift them back toward the evidence while retaining most of OPD's greedy-crop gain.

What could supply this credit without labels? Post-execution readouts weakly rank a state's crops by IoU ($\rho = 0.13$–0.20; Fig. 2b). One requires no task label: the teacher's entropy drop (ED), the reduction in its answer entropy after reading the crop rather than the unzoomed image. Among the label-free readouts, it ranks crops best ($\rho = 0.18$–0.20, against 0.15–0.17 for the teacher's maximum option probability and 0.06–0.11 for its belief change; Appendix D). The label-dependent alternatives—the teacher's probability of the correct answer and the student's answer accuracy—show no clear ranking advantage over ED (Appendix D). The remaining question is whether such a weak, label-free signal is sufficient when credited directly to the sampled action on top of OPD.

4 EXECUTION-CONDITIONED ENTROPY DROP

Exec-ED closes the action-credit gap diagnosed in Section 3 by leaving OPD unchanged and adding a label-free post-execution credit to the action tokens.

Execution credit. Let $o_k = \mathcal{E}(s, a_k)$ be the executed view of sibling k, $u(s, o_k)$ a post-execution utility, and $A^u_k = u(s, o_k) - \frac{1}{K}\sum_j u(s, o_j)$ its within-state centered value. Replacing correctness

Table 2: Design choices of Exec-ED and alternatives considered.

Design choice	Alternative	Needs	Rationale
Utility: entropy drop	correctness, $p_T(y^\star)$	label	no better crop ranking (Sec. 3.3)
	privileged crop score	evidence box	outside the label-free setting
Reader: frozen teacher	current student	–	moving during training; no better empirically
Locus: action only	action and answer	–	reader already receives post-execution OPD supervision
Anchor: action KL	no action KL	–	unanchored actor collapses (Sec. 6.2)

in Eq. (2) by u and restricting the residual to action tokens gives

$$\mathcal{L}_{\text{Exec-}u}(\theta) = \mathcal{L}_{\text{OPD}}(\theta) - \lambda\, \mathbb{E}_{s,\, \tau_{1:K}\sim p_{\bar{\theta}}} \left[\frac{1}{K} \sum_{k=1}^{K} A_k^u \, \log p_\theta(a_k \mid s) \right], \tag{3}$$

where $\log p_\theta(a_k \mid s)$ sums over action tokens and A_k^u is held fixed. This is an unclipped, action-only REINFORCE estimator with a GRPO-like sibling baseline (Shao et al., 2024); in expectation, its descent direction is $\lambda \frac{K-1}{K} \nabla_\theta \mathbb{E}_{a\sim p_\theta(\cdot|s)} u(s, \mathcal{E}(s,a))$ (Proposition 1). Thus, Eq. (3) attaches post-execution credit to the actor while leaving OPD's reader supervision unchanged.

Entropy drop. We instantiate u with the label-free readout identified in Section 3.3. For a reader R with answer distribution $q_R(\cdot \mid s, o)$,

$$u_{\text{ED}}(s,o) = H\big(q_R(\cdot \mid s, \varnothing)\big) - H\big(q_R(\cdot \mid s, o)\big), \tag{4}$$

where $\varnothing$ is the null, unzoomed view. Equation (3) with $u = u_{\text{ED}}$ defines **Execution-Conditioned Entropy Drop (Exec-ED)**. Within-state centering keeps only the differences among sibling crops, matching the weak ranking signal observed in Section 3.3. For valid siblings, the null-view entropy cancels under centering, so the credit depends only on their relative post-execution entropy (Lemma 1); the null view matters only when malformed actions are present.

Design rationale. The estimator and entropy-based utility are standard (Section 2); the design lies in where they act inside OPD (Table 2). The credit is computed after execution because OPD's pre-execution action KL contains no crop consequence; entropy drop is used because it is label-free and ranks crops comparably to the label-dependent readouts we measure; the residual acts only on action tokens because the reader already receives post-execution supervision from OPD; and the original action KL is retained as an anchor, whose role is examined in Section 6.2.

Instantiation. We use the frozen teacher as R, with q_R its softmax over answer options at the first answer token. Because OPD already evaluates the teacher on every executed crop, Exec-ED adds only one null-view teacher read per state; with our implementation, which reads the teacher in a separate pass, an update takes 14% longer than OPD's (Appendix B). We use $\lambda \in \{0.1, 0.2\}$ for all teachers and benchmarks (Section 5.1). Malformed actions receive $u = 0$; Section 6.4 analyzes this choice and gives a label-free floor.

5 Experiments and Results

We test whether label-free execution credit repairs the acquisition degradation diagnosed in Section 3 while preserving task performance, across two teacher sizes and two benchmarks. We then repeat the complete 9B setting with a second training seed.

5.1 Experimental setup

Models and training. All students are Qwen3.5-2B, distilled from the Qwen3.5-9B teacher of Section 3, our main setting, or from a smaller Qwen3.5-4B teacher. The main comparisons are entirely label-free: Base, OPD, TA-OPD, VA-OPD, and Exec-ED with $\lambda \in \{0.1, 0.2\}$, fixed across

Table 3: **Main results** (%). Accuracy is end-to-end under the sampled-action protocol. Precision and Hit score crop quality; G is the held-out gain of a frozen Qwen3.5-27B reader. V* uses its 96 four-option attribute questions. Bold and underline mark the best and second best within each block. Table 10 gives paired intervals and Table 13 results on all 170 V* questions.

	ZoomBench-MCQ			V* Bench		
Method	Accuracy	Precision	G	Accuracy	Hit	G
Teacher Qwen3.5-9B (main setting)						
Base	44.99	**59.18**	6.39	59.64	57.55	5.39
OPD	53.69	55.04	6.64	72.14	61.20	5.19
TA-OPD	52.54	51.61	5.58	70.05	56.77	4.64
VA-OPD	54.99	54.02	6.84	<u>73.44</u>	61.72	4.84
Exec-ED (λ=0.1)	**55.38**	57.38	<u>7.49</u>	**74.74**	<u>65.36</u>	<u>5.96</u>
Exec-ED (λ=0.2)	<u>55.09</u>	<u>58.78</u>	**7.70**	**74.74**	**66.93**	**7.49**
Teacher Qwen3.5-4B						
Base	44.99	<u>59.18</u>	6.39	59.64	57.55	5.39
OPD	53.63	56.18	6.94	<u>75.13</u>	67.19	6.20
TA-OPD	49.79	54.57	5.87	72.79	61.98	5.45
VA-OPD	<u>53.69</u>	56.50	7.10	73.96	67.19	6.60
Exec-ED (λ=0.1)	**54.33**	57.65	<u>7.41</u>	<u>75.13</u>	**69.79**	<u>6.78</u>
Exec-ED (λ=0.2)	52.93	**60.07**	**8.00**	**75.78**	<u>68.23</u>	**7.74**

teachers and benchmarks. Label-based controls remain diagnostic and are reported only in Appendix E. Every model is trained from Base for one epoch of Vision-OPD-6K (6,240 states; 65 updates of 96 states with K=8 siblings), with shared data order, sampling seeds, and decoding (Appendix B).

Evaluation. We evaluate on ZoomBench-MCQ (606 four-option questions with evidence boxes) and V* Bench (Wu & Xie, 2024). Accuracy is end-to-end, averaging two answers from each of four sampled actions; malformed actions count as wrong. We use sampled actions as the primary protocol because they are the on-policy actions OPD trains on, and report greedy decoding and matched lower-temperature comparisons separately. Crop quality is measured without a reader by precision against ZoomBench evidence boxes and, for V*'s much smaller targets, by the share of crops containing at least half of every target (*hit*). We additionally measure held-out gain $G = p_R(y^\star \mid s, o) - p_R(y^\star \mid s, \varnothing)$, the increase in a frozen Qwen3.5-27B reader's probability of the correct answer from the crop over the null view. For V*, the main table uses the 96 four-option, single-target attribute questions that match the training setting; all 170 questions are reported by type in Appendix E. On the 68 relative-position questions, a single crop lowers G for every model, including Base, placing them outside the single-crop acquisition studied here. Our acquisition conclusions do not depend on this choice: on all 170 questions, both Exec-ED strengths still have the highest hit and G with either teacher (Table 13). Alternative treatments of malformed actions do not change the conclusions below, although excluding them lowers the recovered share of precision (Appendix E).

5.2 Exec-ED tightens the sampled crops that OPD loosens

Exec-ED consistently counteracts the crop-quality degradation introduced by OPD while maintaining comparable task accuracy. Across both teachers and both credit strengths, its sampled crops are more precise on ZoomBench and hit V* targets more often than those of every OPD-trained baseline (Table 3). With the 9B teacher, ZoomBench precision rises from 55.0 under OPD to 57.4 and 58.8, while V* hit rises from 61.2 to 65.4 and 66.9. With the 4B teacher, the corresponding changes are 56.2 to 57.7 and 60.1 in precision, and 67.2 to 69.8 and 68.2 in hit. TA-OPD and VA-OPD, which reweight distillation without execution-dependent credit, recover little or none of the degradation identified in Section 3.

Relative to Base, λ=0.1 recovers about half of the precision lost under OPD, while λ=0.2 recovers at least 90% with either teacher. In the 9B setting, λ=0.2 nearly returns to Base precision (58.8 vs. 59.2); with the 4B teacher it exceeds it (60.1 vs. 59.2). The same improvement appears under

an independent reader: Exec-ED has higher G than every baseline in all four teacher–benchmark blocks. Paired bootstrap intervals exclude zero for all eight ZoomBench precision comparisons against OPD and VA-OPD and for six of eight V* hit comparisons; support for G is strongest at λ=0.2, where six of eight intervals exclude zero (Table 10).

End-to-end accuracy remains comparable rather than driving these gains. At λ=0.1, Exec-ED matches or exceeds every baseline's accuracy in all four teacher–benchmark blocks. The stronger λ=0.2 improves acquisition further but, on ZoomBench with the 4B teacher, is 0.7 and 0.8 points below OPD and VA-OPD. Greedy behavior is not systematically degraded either: at λ=0.1, greedy accuracy is numerically above OPD's in all four blocks by 0.7–3.9 points (Table 15). Section 6.3 further shows that Exec-ED's crop advantage persists when both models sample at matched lower temperatures.

5.3 Replication with a second training seed

Table 4: **Replication with a second training seed** (seed 43; %). The complete 9B setting of Table 3 is retrained; Base is unchanged. Bold and underline mark the best and second best within each block.

	ZoomBench-MCQ			V* Bench		
Method	Accuracy	Precision	G	Accuracy	Hit	G
Base	44.99	<u>59.18</u>	6.39	59.64	57.55	5.39
OPD	53.53	53.17	6.47	72.14	58.85	4.82
TA-OPD	52.21	50.67	5.52	69.66	55.99	3.66
VA-OPD	**54.41**	54.92	6.95	**75.52**	62.76	5.71
Exec-ED (λ=0.1)	<u>54.25</u>	57.75	**7.67**	<u>73.44</u>	<u>64.32</u>	<u>5.88</u>
Exec-ED (λ=0.2)	53.92	**59.79**	<u>7.61</u>	**75.52**	**65.62**	**6.80**

Exec-ED's acquisition advantage replicates with seed 43 (Table 4). On ZoomBench, Exec-ED again raises precision over OPD from 53.17 to 57.75 and 59.79, and G from 6.47 to 7.67 and 7.61. On V*, hit rises from 58.85 to 64.32 and 65.62, and G from 4.82 to 5.88 and 6.80. Against OPD, seven of the eight precision, hit, and G differences exclude zero; both strengths also exceed VA-OPD in all three, although only the precision differences exclude zero (Table 10). Accuracy rankings vary across seeds: VA-OPD is highest on ZoomBench (54.41 vs. 54.25 for Exec-ED at λ=0.1), while VA-OPD and Exec-ED at λ=0.2 tie on V* (75.52). Thus the replicated effect is the one predicted by the diagnosis: improved acquisition, with comparable rather than uniformly improved end-to-end accuracy.

6 Analysis: how execution credit interacts with OPD

Section 3.3 found that entropy drop ranks crops only weakly, yet Section 5 showed a consistent improvement in acquisition. We now locate that improvement and examine how the execution credit interacts with OPD.

6.1 The gain is actor-side

Reader–actor swaps separate whether Exec-ED improves the acquired crop or merely trains a better reader. On ZoomBench, fixing OPD's reader and replacing only its actor with Exec-ED's raises the probability of a correct answer from 56.18% to 56.93% (+0.75 points), while fixing OPD's actor and replacing only its reader changes it by just +0.02 points (Table 9). Under Exec-ED's own reader, replacing OPD's actor with Exec-ED's similarly gives a +1.04-point gain.

Actor effects are small in absolute terms: OPD's actor adds at most 0.9 points over Base's under any reader, so these gains are substantial. Together with the higher held-out gain in Section 5.2, the swap localizes Exec-ED's gain to the acquired view rather than the reader. This complements Section 3.2: OPD's gain is predominantly reader-side, whereas Exec-ED acts on the actor, where the execution credit is applied (examples in Appendix G).

6.2 OPD ANCHORS THE EXECUTION CREDIT

Removing the action KL reveals the role of OPD's anchor. With Exec-ED but no action KL (its credit rescaled to its initial share of the gradient; Appendix B), the sampled metrics become superficially stronger (ZoomBench precision 66.0 and G 11.0), but the actor collapses: about 40% of sibling pairs become near-duplicates during training (Appendix F), and 23.7% of sampled actions equal the greedy action, versus at most 0.5% for the other models. The greedy crop exposes the cost of this collapse: its IoU falls to 34.6, below Base's 39.2 and anchored Exec-ED's 44.1. Narrowing alone also raises sampled scores—lowering Exec-ED's sampling temperature from 1.0 to 0.5 increases precision by 7.6 points and G by 2.2 without changing its greedy action (Appendix E). Thus, the high sampled scores without the KL do not indicate a better actor; OPD's action KL keeps the distribution anchored while execution credit reshapes it.

6.3 EXEC-ED ACTS AS A STANDING COUNTERWEIGHT

Fig. 3 returns to the diagnosis states of Fig. 2a: each arrow connects OPD to Exec-ED (λ=0.1) trained with the same seed, and the dashed line marks Base's sampled-crop IoU. Across both seeds, Exec-ED moves OPD's sampled crops back toward the evidence (IoU 37.1 $\rightarrow$ 39.3 and 35.5 $\rightarrow$ 38.5) while leaving the greedy crop nearly unchanged. The samples narrow only modestly, and the advantage persists when OPD and Exec-ED are evaluated at matched lower sampling temperatures from T=1.0 to 0.5 (Appendix E). Switching objectives after half of training shows that the effect lasts only while the credit is applied: replacing Exec-ED by OPD lowers the batch IoU within a few updates, and replacing OPD by Exec-ED restores it (Appendix F). Thus, Exec-ED acts as a standing counterweight to OPD: distillation anchors the actor toward the teacher, while execution credit favors more useful sampled actions within that distribution.

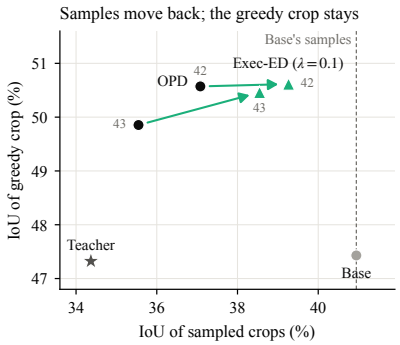

Figure 3: **Exec-ED counteracts OPD on sampled actions.** Across two seeds, sampled-crop IoU improves while greedy-crop IoU changes little.

6.4 STRENGTH AND VALIDITY

A stronger credit further tightens the crops: with the 9B teacher on ZoomBench, increasing λ from 0.1 to 0.2 and 0.4 raises precision from 57.4 to 58.8 and 61.1 and G from 7.5 to 7.7 and 8.4, while sampled accuracy remains within 1.4 points (Appendix E). One edge case is malformed actions: the default utility u=0 can give them positive centered credit when valid crops have negative entropy drop, yet Exec-ED leaves no more malformed actions than OPD, and a label-free floor that ranks them last would remove this case by construction (Proposition 2; Fig. 4); we keep the simpler default.

7 CONCLUSION

On-policy distillation teaches a visual agent how to read its acquired view but does not credit sampled actions by what their execution reveals, creating an *action-credit gap*. Exec-ED adds this missing signal with a label-free, sibling-centered post-execution credit applied only to action tokens, recovering much of the crop quality lost under OPD at comparable task accuracy. Its effect is actor-side and depends on OPD's action KL as an anchor: distillation constrains what the agent would do, while execution credit shifts its sampled actions toward what works. More broadly, execution-aware distillation can learn from both the teacher's behavior and the consequences of the student's own actions.

AI USE STATEMENT

Generative AI tools were used to assist with and polish writing, to support literature retrieval and discovery, and to provide suggestions during research execution. All retrieved literature and citations were manually verified by the authors for relevance, correctness, and potential hallucinations. The core research idea was conceived by the authors; AI-generated suggestions on possible approaches were treated only as references, with all research decisions made by the authors. Coding agents were also used to assist implementation, and their outputs were cross-audited and repeatedly checked by the authors for correctness. All AI-assisted content was reviewed by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

All authors have read and follow the ICLR Code of Ethics. Our study investigates what on-policy distillation transfers to crop-based visual agents and develops execution-level credit assignment for improving visual acquisition, using existing model checkpoints and benchmark datasets, including Vision-OPD-6K, ZoomBench-MCQ, and V^* Bench; we do not conduct a human-subject study or collect new user data. The work focuses on analyzing and improving how multimodal agents acquire visual evidence rather than on deployment in safety-critical settings. As with the pretrained models and datasets on which the experiments rely, the results may inherit biases, annotation artifacts, or other limitations from these resources. We additionally audit benchmark overlap with the training data and report the resulting exclusions and other methodological limitations in the paper and appendix.

REPRODUCIBILITY STATEMENT

Our diagnostic framework, Exec-ED method, and experimental design are described in Sections 3–5. Implementation and training details, including the training data, action representation, optimization objectives, hyperparameters, decoding settings, and random seeds, are provided in Appendix B. Appendix C specifies the benchmark construction, data-overlap checks, evaluation metrics, malformed-action handling, and statistical procedures; additional diagnostic protocols, complete results, and training-dynamics analyses are reported in Appendices D–F. Together, these sections document the experimental and analysis protocols used for the reported diagnoses, comparisons, and ablations.

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*, volume 2024, pp. 21246–21263, 2024.

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. *Advances in Neural Information Processing Systems*, 38:107150–107180, 2026.

Mohammad Sadegh Akhondzadeh, Vijay Lingam, Atula Tejaswi, Chanakya Ekbote, Sujay Sanghavi, and Aleksandar Bojchevski. Reward-gated on-policy distillation. *arXiv preprint arXiv:2607.04037*, 2026.

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. Vold: Reasoning transfer from llms to vision-language models via on-policy distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26209–26218, 2026.

Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics: Methodology and distribution*, pp. 569–593. Springer, 1992.

Mingkuan Feng, Jinyang Wu, Hao Gu, Fangrui Lv, Ruihan Jin, Chuyuan Zhang, Zhengqi Wen, and Jianhua Tao. Taco: Tool-augmented credit optimization for agentic tool use. *arXiv preprint arXiv:2606.30251*, 2026.

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Minillm: Knowledge distillation of large language models. In *International Conference on Learning Representations*, volume 2024, pp. 32694–32717, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Senkang Hu, Yong Dai, Xudong Han, Zhengru Fang, Yuzhi Zhao, Sam Tak Wu Kwong, and Yuguang Fang. Self-induced outcome potential: Turn-level credit assignment for agents without verifiers. *arXiv preprint arXiv:2605.04984*, 2026.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free!, 2019. URL https://openreview.net/forum?id=r1lgTGL5DE.

Zeping Li, Hongru Wang, Yiwen Zhao, Guanhua Chen, Yixia Li, Keyang Chen, Yixin Cao, Guangnan Ye, Hongfeng Chai, and Zhenfei Yin. Rethinking the role of entropy in optimizing tool-use behaviors for large language model agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27955–27967, 2026.

Wenze Lin, Jiale Zhao, Xitai Jiang, Songde Rao, Yining Li, Shenzhi Wang, Bingxiang He, and Gao Huang. On-policy distillation with verifiable reward. *arXiv preprint arXiv:2608.24696*, 2026.

Ruiqi Liu, Xiaolei Lv, Gengsheng Li, Ximo Zhu, Zhiheng Wang, Zhengbo Zhang, Junkai Chen, Zhiheng Li, Bo Li, Jun Gao, et al. Visual-advantage on-policy distillation for vision-language models. *arXiv preprint arXiv:2605.21924*, 2026.

Xingming Long, Yu Liu, Zhiwei Yang, Hanqi Feng, Shaojie Zhang, Barnabas Poczos, Chao Jiang, Zhenbo Luo, Lei Jiang, and Pei Fu. Making every tool call count: Necessary tool-evidence path rewards for agentic vision-language models. *arXiv preprint arXiv:2609.03493*, 2026.

Yan Ma, Weiyu Zhang, Tianle Li, Linge Du, Xuyang Shen, and Pengfei Liu. What does vision tool-use reinforcement learning really learn? disentangling tool-induced and intrinsic effects for crop-and-zoom. *arXiv preprint arXiv:2602.01334*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Kunyu Peng, Junming Liu, Ruiqi He, Qingzhuo Wang, Jianzhong Qi, and Xianhui Liu. When should a vlm look? paying only for visual calls that were needed and used, 2026. URL https://arxiv.org/abs/2609.22910.

Mihir Prabhudesai, Lili Chen, Alex Ippoliti, Katerina Fragkiadaki, Hao Liu, and Deepak Pathak. Maximizing confidence alone improves reasoning. *arXiv preprint arXiv:2505.22660*, 2025.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pp. 2564–2571. Ieee, 2011.

Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. *arXiv preprint arXiv:1511.06295*, 2015.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:8222–8251, 2026.

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.

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

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Guoqing Wang, Sunhao Dai, Guangze Ye, Zeyu Gan, Wei Yao, Yong Deng, Xiaofeng Wu, and Zhenzhe Ying. Information gain-based policy optimization: A simple and effective approach for multi-turn llm agents. *arXiv e-prints*, pp. arXiv–2510, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Yuanyi Wang, Su Lu, Yanggan Gu, Pengkai Wang, Yifan Yang, Zhaoyi Yan, Congkai Xie, Jianmin Wu, and Hongxia Yang. Not all disagreement is learnable: Token teachability in on-policy distillation. *arXiv preprint arXiv:2605.26844*, 2026.

Lai Wei, Liangbo He, Jun Lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, et al. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. *arXiv preprint arXiv:2602.11858*, 2026.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

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*, volume 2026, pp. 78298–78319, 2026.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

Leyan Xue, Feng Xiong, Mingjun Ma, and Changqing Zhang. Distill what the student can see: Fisher-projected on-policy distillation for vision-language models. *arXiv preprint arXiv:2608.01263*, 2026.

Hee Suk Yoon, Eunseop Yoon, Jaehyun Jang, SooHwan Eom, Ji Woo Hong, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D Yoo. Decomposed on-policy distillation for vision-language reasoning: Steering gradients for visual grounding. *arXiv preprint arXiv:2606.00564*, 2026.

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.

Yanzhi Zhang, Zhaoxi Zhang, Haoxiang Guan, Yilin Cheng, Yitong Duan, Chen Wang, Yue Wang, Shuxin Zheng, and Jiyan He. No free lunch: Rethinking internal feedback for llm reasoning. *arXiv preprint arXiv:2506.17219*, 2025.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. In *International Conference on Learning Representations*, volume 2026, pp. 2548–2581, 2026.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, and Chao Shen. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, Ermo Hua, Biqing Qi, et al. Ttrl: Test-time reinforcement learning. *Advances in Neural Information Processing Systems*, 38:131459–131483, 2026.

A PROPERTIES OF THE EXECUTION CREDIT

This section proves the properties of Eq. (3) used in the main text. Throughout, s is a state, its siblings $a_1, \dots, a_K$ are drawn independently from $p_{\bar{\theta}}(\cdot \mid s)$, where $\bar{\theta}$ equals θ but receives no gradient, so expectations below are under p_θ and gradients act only on the log-probabilities; u_k is the utility of sibling k, and the advantage

$$A_k^u = u_k - \frac{1}{K}\sum_{j=1}^{K} u_j$$

is held constant during differentiation.

A.1 THE CREDIT ASCENDS THE EXPECTED UTILITY

Proposition 1 (Expected credit gradient)**.** *Suppose that u_k depends on the siblings only through a_k and, given a_k, not on θ, and let $\bar{u}(a) = \mathbb{E}[u_k \mid a_k = a]$. Then*

$$\mathbb{E}\Big[\frac{1}{K}\sum_{k=1}^{K} A_k^u \,\nabla_\theta \log p_\theta(a_k \mid s)\Big] = \frac{K-1}{K}\,\nabla_\theta\, \mathbb{E}_{a\sim p_\theta(\cdot|s)}\big[\bar{u}(a)\big]. \tag{5}$$

Proof. Let $g_k = \nabla_\theta \log p_\theta(a_k \mid s)$. The score has zero mean:

$$\begin{aligned}
\mathbb{E}[g_k] &= \sum_a p_\theta(a \mid s)\,\nabla_\theta \log p_\theta(a \mid s) \\
&= \sum_a \nabla_\theta\, p_\theta(a \mid s) \\
&= \nabla_\theta \sum_a p_\theta(a \mid s) \\
&= \nabla_\theta\, 1 \;=\; 0.
\end{aligned}$$

For $j \neq k$, the siblings a_j and a_k are independent, so

$$\mathbb{E}[u_j\, g_k] = \mathbb{E}[u_j]\,\mathbb{E}[g_k] = 0.$$

Hence, for each k,

$$\begin{aligned}
\mathbb{E}\big[A_k^u\, g_k\big] &= \mathbb{E}[u_k\, g_k] - \frac{1}{K}\,\mathbb{E}[u_k\, g_k] - \frac{1}{K}\sum_{j\neq k}\mathbb{E}[u_j\, g_k] \\
&= \frac{K-1}{K}\,\mathbb{E}[u_k\, g_k] \\
&= \frac{K-1}{K}\,\mathbb{E}\big[\bar{u}(a_k)\, g_k\big] \\
&= \frac{K-1}{K}\sum_a p_\theta(a \mid s)\,\bar{u}(a)\,\nabla_\theta \log p_\theta(a \mid s) \\
&= \frac{K-1}{K}\,\nabla_\theta \sum_a p_\theta(a \mid s)\,\bar{u}(a) \\
&= \frac{K-1}{K}\,\nabla_\theta\, \mathbb{E}_{a\sim p_\theta(\cdot|s)}\big[\bar{u}(a)\big].
\end{aligned}$$

Averaging over $k = 1, \dots, K$ gives Eq. (5). □

Remarks. This is the score-function estimator of Williams (1992) with the state mean as baseline; a leave-one-out mean would remove the factor $\frac{K-1}{K}$ (Kool et al., 2019). The condition holds exactly for Exec-ED and Exec-p_T, whose reader is the frozen teacher and whose execution is deterministic, but not for the floor of Section 6.4, whose value for a malformed sibling depends on the other siblings. For Exec-ED$_S$ and Exec-Acc, u_k also depends on θ through the student as reader; as the credit reaches only the action tokens, Eq. (5) then gives the actor's gradient with the reader held fixed. Here p_θ is the grammar-restricted, top-p-truncated distribution that actions are sampled from (Appendix B), with its support held fixed.

A.2 THE NULL-VIEW TERM CANCELS AMONG VALID SIBLINGS

Lemma 1 (Cancellation of the null-view term)**.** *Write $H_\varnothing = H\big(q_R(\cdot \mid s, \varnothing)\big)$ and $H_k = H\big(q_R(\cdot \mid s, o_k)\big)$. If all K siblings of a state are valid, then*

$$A_k^{\mathrm{ED}} = \frac{1}{K}\sum_{j=1}^{K} H_j - H_k. \tag{6}$$

Proof. With $u_{\mathrm{ED}}(s, o_k) = H_\varnothing - H_k$ from Eq. (4),

$$\begin{aligned} A_k^{\mathrm{ED}} &= u_{\mathrm{ED}}(s, o_k) - \frac{1}{K}\sum_{j=1}^{K} u_{\mathrm{ED}}(s, o_j) \\ &= \big(H_\varnothing - H_k\big) - \frac{1}{K}\sum_{j=1}^{K}\big(H_\varnothing - H_j\big) \\ &= H_\varnothing - H_k - H_\varnothing + \frac{1}{K}\sum_{j=1}^{K} H_j \\ &= \frac{1}{K}\sum_{j=1}^{K} H_j - H_k. \end{aligned}$$

□

The credit thus ranks valid siblings by the reader's post-execution entropy alone; the null view matters only through malformed actions, which receive its level, $u = 0$.

A.3 WHERE A MALFORMED ACTION RANKS

Proposition 2 (Advantage of a malformed action)**.** *Let V be the set of valid siblings of a state, $n_V = |V| \geq 1$, $\bar{u}_V$ their mean utility, and let every malformed sibling receive the utility c. Then every malformed sibling m has the advantage*

$$A_m^u = \frac{n_V}{K}\big(c - \bar{u}_V\big). \tag{7}$$

In particular:

- *(i) if $u_j \geq c$ for all $j \in V$, as for the label-based utilities in $[0, 1]$ with $c = 0$, then $A_m^u \leq 0$;*
- *(ii) for the entropy drop with $c = 0$, $A_m^u > 0$ whenever $\bar{u}_V < 0$, that is, whenever the valid crops leave the reader less certain than the null view on average;*
- *(iii) the floor $c = \min\big(0, \min_{j\in V} u_{\mathrm{ED}}(s, o_j)\big)$ gives $A_m^u \leq 0$ in every state.*

Proof. The state mean splits into the valid and the malformed siblings:

$$\frac{1}{K}\sum_{j=1}^{K} u_j = \frac{1}{K}\big(n_V\,\bar{u}_V + (K - n_V)\,c\big).$$

Therefore

$$\begin{aligned} A_m^u &= c - \frac{1}{K}\big(n_V\,\bar{u}_V + (K - n_V)\,c\big) \\ &= c - \frac{K - n_V}{K}\,c - \frac{n_V}{K}\,\bar{u}_V \\ &= \frac{n_V}{K}\,c - \frac{n_V}{K}\,\bar{u}_V \\ &= \frac{n_V}{K}\big(c - \bar{u}_V\big). \end{aligned}$$

For (i), $u_j \geq c$ for all $j \in V$ gives $\bar{u}_V \geq c$, hence $A^u_m \leq 0$. For (ii), $c = 0$ gives

$$A^u_m = -\frac{n_V}{K}\,\bar{u}_V > 0 \quad \text{whenever} \quad \bar{u}_V < 0.$$

For (iii),

$$c \;\leq\; \min_{j \in V} u_{\mathrm{ED}}(s, o_j) \;\leq\; \bar{u}_V,$$

hence $A^u_m \leq 0$. □

If no sibling of a state is valid, all advantages are zero. Case (ii) is common: on the diagnosis states (Appendix D), 37% of Base's and 36% of OPD's valid crops leave the teacher less certain than the null view, and among the states with a malformed action, the valid crops' mean entropy drop is negative in 51% (Base) and 54% (OPD) (Fig. 4a). Even so, Exec-ED leaves no more malformed actions than OPD (Fig. 4b).

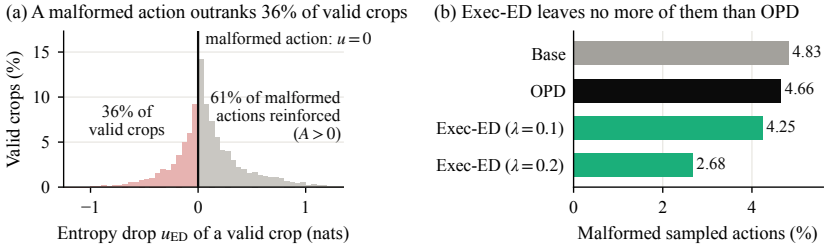

Figure 4: **Malformed actions under entropy-drop credit.** (a) A malformed action receives u=0, above 36% of OPD's valid sampled crops on the diagnosis states. (b) Malformed sampled actions on ZoomBench (9B teacher).

A.4 What each credit requires

Table 5 lists each execution credit of the paper with its utility, its reader, its use of the label, and its cost beyond OPD.

Table 5: Execution credits. q_T and q_S: teacher's and student's distributions over the answer options at the first answer token; $\hat{y}$: the student's sampled answer; $\varnothing$: null (unzoomed) view. Malformed actions receive $u = 0$.

Credit	Utility $u(s, o)$	Reader	Label	Tokens	Extra read per state
Exec-ED	$H(q_T(\cdot \mid s, \varnothing)) - H(q_T(\cdot \mid s, o))$	teacher	–	$\mathcal{A}$	teacher on $\varnothing$
Exec-ED$_S$	$H(q_S(\cdot \mid s, \varnothing)) - H(q_S(\cdot \mid s, o))$	student	–	$\mathcal{A}$	student on $\varnothing$
Exec-p_T	$q_T(y^\star \mid s, o)$	teacher	✓	$\mathcal{A}$	none
Exec-Acc	$\mathbf{1}[\hat{y} = y^\star]$	student	✓	$\mathcal{A}$	none
OPD+RLVR	$\mathbf{1}[\hat{y} = y^\star]$	student	✓	$\mathcal{A} \cup \mathcal{Y}$	none

B TRAINING DETAILS

Data, actions, and answers. We train on Vision-OPD-6K (Yuan et al., 2026), using 6,240 of its 6,241 multiple-choice questions. In each state, the agent sees an image and a question, emits one crop action `crop(x1=...,y1=...,x2=...,y2=...)` with three-digit coordinates on a 0–999 grid (27 tokens: 12 coordinate digits, 14 fixed syntax tokens, and an end token), receives the cropped view as an additional image, and answers with an option letter (two answer tokens). The dialogue uses the Qwen3.5 chat template with thinking disabled and two user turns: the first shows the image, the question, and its options and asks for one crop; the second shows the image with the cropped view and asks for one option letter, after which the end token is forced. The null view uses the same template without a crop, and the processor resizes the image and the crop to between 256^2 and 1024^2 pixels. No objective uses the training set's evidence boxes; they serve only the diagnosis (Appendix D) and the telemetry of Appendix F.

Executing an action. An action is malformed if its coordinates are out of order or degenerate ($x_1 \geq x_2$ or $y_1 \geq y_2$) or if the image processor rejects its aspect ratio; malformed actions are not executed and receive no answer during training. A valid action on an image of width W and height H selects the pixel box

$$b(a) = \big(\lfloor x_1 W/999 \rfloor,\ \lfloor y_1 H/999 \rfloor,\ \lceil x_2 W/999 \rceil,\ \lceil y_2 H/999 \rceil\big), \tag{8}$$

and the null view $\varnothing$ executes the full-image action $(000, 000, 999, 999)$.

Objective as implemented. With B states per update and K siblings per state, the loss is

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{B} \sum_{k=1}^{K} \sum_{t \in \mathcal{A}_{ik} \cup \mathcal{Y}_{ik}} D_{ik,t} \;-\; \frac{\lambda}{BK} \sum_{i=1}^{B} \sum_{k=1}^{K} A^{u}_{ik} \log p_\theta(a_{ik} \mid s_i), \tag{9}$$

where D is the reverse KL of Eq. (1), N the number of supervised tokens in the batch (27 action tokens per sibling and two answer tokens per executed sibling), and A^u_{ik} is centered over all K siblings of state i, malformed ones included. The action log-probability is that of the sampling distribution: syntax tokens are forced and contribute zero, and each coordinate digit is drawn from the digit tokens at temperature 1 after top-p truncation at 0.95. The teacher's belief $q_T(\cdot \mid s, o)$ is its softmax over the four option letters at the first answer position, at temperature 1; we read it in a separate forward pass whose values match those of OPD's own pass (all 47,706 readings of the main run agree exactly).

Hyperparameters. All methods share the settings of Table 6, with identical data order and sampling seeds.

Table 6: Shared training settings.

Teacher → student	Qwen3.5-9B → Qwen3.5-2B (main)
Second setting	Qwen3.5-4B → Qwen3.5-2B
Data	Vision-OPD-6K, one epoch (6,240 states)
States B / siblings K per update	96 / 8 (768 trajectories)
Updates	65
Optimizer	AdamW (β_1=0.9, β_2=0.95), weight decay 0.01
Learning rate	2×10^{-6}, linear warm-up over 10 updates, then constant
Gradient clipping	1.0 (global norm)
Precision	BF16 compute, FP32 master weights and optimizer states
Action decoding	temperature 1.0, top-p 0.95, no top-k
Answer decoding	temperature 0.7, top-p 0.8, top-k 20
Seed	42 for sibling sampling (43 in Section 5.3); data order fixed

Credit weights. The weight $\lambda = 0.1$ was set once for a correctness-based action credit in preliminary experiments and is reused, without tuning on the evaluation benchmarks, by Exec-ED, Exec-ED$_S$, VA-OPD + Exec-ED, Exec-Acc, and OPD+RLVR; Exec-ED is also run at $\lambda = 0.2$ with both teachers and at $\lambda = 0.4$ with the 9B teacher. On the first batch, which is identical for all runs,

the ratio $\rho_1 = \lambda \|g_{\text{credit}}\| / \|g_{\text{OPD}}\|$ between the gradient norms of the two terms of Eq. (9) is 0.0152 for Exec-ED at $\lambda = 0.1$, matching the correctness credit (0.0151); Exec-p_T, whose credit gradient is smaller, uses $\lambda = 0.13$ to match it, whereas Exec-ED$_S$, whose credit gradient is 1.55 times Exec-ED's, is not rescaled and is thus the stronger credit. Without the action-token KL, OPD supervises 1,490 instead of 22,226 tokens on the first batch and $\|g_{\text{OPD}}\|$ rises from 164.06 to 1,986.21, so Exec-ED w/o action KL uses $\lambda' = 0.1 \times 1{,}986.21/164.06 = 1.2107$ to keep ρ_1. This matching holds only at the first update: over training, the ratio stays near 0.25 at $\lambda = 0.1$ and 0.5 at $\lambda = 0.2$ with the anchor, but exceeds 1 from update 17 on without it (1.3 to 1.8 over later windows): besides restraining the actor directly, the anchor keeps the credit's share of the update small (Section 6.2).

Token-side baselines. TA-OPD (Wang et al., 2026) scores each token by the KL from teacher to student restricted to the union of both models' top-16 tokens, multiplied by the teacher's probability mass on the student's top-16 tokens; scores are normalized by their 5th and 95th percentiles, and only the top 10% of all supervised tokens, action tokens included, keep their KL ($\epsilon = 10^{-8}$, no entropy term; the 10% is our setting). VA-OPD (Liu et al., 2026) keeps its original settings, which we did not tune: a token's visual advantage measures how the teacher's prediction changes when every image side is downscaled to 1/10 and upscaled back; within a trajectory, the 20% of positions with the highest visual advantage and the remaining positions each receive half of the token weight; and trajectories are weighted by a softmax (temperature 1) over their standardized mean visual advantage, taken over the $K = 8$ siblings of a state instead of four.

Compute. Each run uses 8 NVIDIA H200 or A100 GPUs with 12 states per GPU; on H200, the median update takes 130.5 s for OPD and 149.2 s for Exec-ED, whose teacher reads run in a separate forward pass. All models are evaluated on the same hardware.

C Evaluation Protocol

Benchmarks. ZoomBench (Wei et al., 2026) has 845 test questions, 621 of them multiple choice; before any evaluation, we fixed our set to the 606 four-option questions (582 source images), which match the training format. Comparing all 845 images with all 6,241 training images by exact and perceptual hashes, DINOv2 features (Oquab et al., 2023), and ORB matching with RANSAC (Rublee et al., 2011; Fischler & Bolles, 1981), followed by manual inspection, found no shared source image. V* Bench (Wu & Xie, 2024) has 191 questions on high-resolution SA-1B images, each with annotated boxes of its target objects: 115 ask about an object's attributes (107 with four options, 8 with two) and 76 about the relative position of two objects (two options); we use the official questions and shuffled options (Hugging Face `craigwu/vstar_bench`, revision `d9ae62c`). As Vision-OPD-6K also draws on SA-1B, we compared all 191 images with the training images: exact hashes of the decoded pixels match 21 of them, and perceptual hashes (pHash, dHash) and DINOv2 nearest neighbors, followed by manual inspection, find no further duplicate. The remaining 170 questions (96 four-option and 6 two-option attribute questions, 68 relative-position questions) are evaluated with the prompt, decoding, and sampling seeds of ZoomBench; two-option questions keep the A–D answer format, and an answer outside their options counts as wrong. The main text reports the 96 four-option questions, which share the training format of one target and four options; on relative-position questions, a single crop lowers the reader's probability of the correct answer for every model (Table 13).

Accuracy. Each state is answered from four sampled actions, with two sampled answers per executed action, and from the greedy action, the argmax over the tokens that the action grammar allows, with four sampled answers; decoding follows Table 6, with sampling seeds shared by all models. Sampled and greedy accuracy average the correctness of these eight and four answers, and then over states; a malformed action counts as wrong.

Malformed actions. Nearly all malformed actions at evaluation have out-of-order or degenerate coordinates. Accuracy uses the unrepaired actions; evidence geometry and reader-based measures execute their repair, which clamps the coordinates to $[0, 999]$, sorts each pair, and widens a degenerate pair by one grid unit, and the rare actions that remain unexecutable count as zero. Appendix E adds two stricter treatments: *Zero* scores a malformed action as zero in every crop metric, with zero reader probability and hence a gain of $-p_R(y^\star \mid s, \varnothing)$, and keeps it in the denominator; *Excluded* averages over valid actions only, each weighted equally, so that states with more valid actions weigh more.

Evidence geometry. For a crop with pixel box b (Eq. (8)) and an evidence box e, $\mathrm{IoU} = |b \cap e|/|b \cup e|$, $\mathrm{Recall} = |b \cap e|/|e|$, and $\mathrm{Precision} = |b \cap e|/|b|$, with areas in pixels; IoU and recall are averaged over the four sampled crops of a state and then over states, and precision over executed crops. The target boxes of V* Bench cover a median 0.06% of the image, so there a crop *hits* if at least half of every target box lies inside it; hit is the share of sampled crops that hit, recall the mean covered fraction of the target boxes, and Area | hit the image share of the crops that hit.

Readers. A reader's $q_R(\cdot \mid s, o)$ is its distribution over the answer options at the first answer token. The held-out gain G averages $q_R(y^\star \mid s, o) - q_R(y^\star \mid s, \varnothing)$ over the four (repaired) sampled crops of a state and then over states, for a frozen Qwen3.5-27B reader at temperature 1 without truncation. Reader–actor swaps (Table 9) score a reader on another model's recorded crops by p_{correct}, the mass of $y^\star$ under q_R after the answer decoding (temperature 0.7, top-k 20, top-p 0.8, renormalized over the surviving options; zero if $y^\star$ is truncated away), which equals the expected accuracy of the sampled answers.

Sampling statistics. Samples = greedy (Table 16) is the share of sampled actions whose tokens equal the greedy action's, and near-duplicate pairs (Fig. 5a) are pairs of valid siblings whose pixel boxes have an IoU of at least 0.9, among all 28 sibling pairs of a state.

Statistics. All contrasts are paired: both models are evaluated on the same states with the same sampling seeds. A 95% interval spans the 2.5th to 97.5th percentiles of the difference over 10,000

bootstrap resamples of source images (seed 42), so that questions about the same image are resampled together (Efron, 1992); the intervals reflect evaluation uncertainty only, and we do not correct for multiple comparisons.

D DIAGNOSIS DETAILS

States and crops. The diagnosis uses 1,000 of the 6,240 training states, drawn uniformly (seed 42); for each state, Base, OPD (after its 65 updates), and the frozen teacher each sample $K = 8$ crops with the action decoding of Table 6, the same prompt, and shared sampling seeds. The evidence boxes of these states are never used for training; for OPD, the states are in-sample, since one epoch covers every training state.

Crop geometry. Table 7 gives the numbers behind Figs. 2a and 3, with the sibling IoU; its last block is the second training seed of OPD and both Exec-ED strengths. The teacher's greedy crop is as good as Base's, although its samples overlap the evidence less (34.4 against 40.9), and Exec-ED's sampled crops overlap the evidence as much as those of the label-based controls (IoU 39.3–39.8 against 38.0–39.0).

Table 7: Crop geometry on the 1,000 diagnosis states: IoU with the evidence box of the sampled crops (over valid crops) and of the greedy crop, and sibling IoU, the mean pixel-box IoU between two valid sampled crops of a state. Last block: seed 43 (Fig. 3).

	IoU (%)		Sibling
Model	sampled crops	greedy crop	IoU
Teacher (as actor)	34.4	47.3	0.40
Base	40.9	47.4	0.53
OPD	37.1	50.6	0.43
TA-OPD	33.0	49.0	0.35
VA-OPD	36.6	49.6	0.43
Exec-Acc	38.0	49.9	0.44
OPD+RLVR	39.0	51.5	0.45
Exec-ED (λ=0.1)	39.3	50.6	0.47
Exec-ED (λ=0.2)	39.8	49.9	0.48
Seed 43			
OPD	35.5	49.9	0.40
Exec-ED (λ=0.1)	38.5	50.5	0.44
Exec-ED (λ=0.2)	39.9	52.0	0.47

Readouts and OPD's action KL. OPD's action KL on a crop is the student–teacher reverse KL summed over its 27 action tokens. Each readout uses one teacher pass on the executed crop and one on the null view per state: the entropy drop of Eq. (4), the maximum option probability $\max_y q_T(y \mid s, o)$, the belief change $D_{\mathrm{KL}}\big(q_T(\cdot \mid s, o) \,\|\, q_T(\cdot \mid s, \varnothing)\big)$, the teacher's probability $q_T(y^\star \mid s, o)$ of the correct answer, and the student's accuracy over eight answers. Each signal's Spearman correlation with IoU is computed over the valid crops of a state, if at least three are valid and neither sequence is constant, and averaged over the states on which all six correlations are defined (389 for Base's crops, 564 for OPD's), with intervals from 10,000 bootstrap resamples of states (seed 42). Among the label-free readouts, the entropy drop ranks crops best and the belief change worst (Table 8).

Table 8: Within-state Spearman correlation between a crop's IoU and OPD's action KL on the crop or a post-execution readout, with 95% intervals (Base's crops: 389 states; OPD's crops: 564 states).

Signal of a sampled crop	Label	Base's crops	OPD's crops
OPD's action KL on the crop	no	-0.062 [-0.108, -0.017]	-0.004 [-0.044, 0.034]
Entropy drop (ED)	no	+0.179 [0.134, 0.223]	+0.203 [0.163, 0.241]
Maximum probability	no	+0.146 [0.099, 0.194]	+0.166 [0.128, 0.205]
Belief change	no	+0.062 [0.019, 0.107]	+0.112 [0.072, 0.153]
Teacher $p_T(y^\star \mid s, o)$	yes	+0.156 [0.110, 0.201]	+0.196 [0.156, 0.235]
Student accuracy (8 answers)	yes	+0.126 [0.080, 0.172]	+0.154 [0.115, 0.193]

Reader–actor swaps. In Table 9, only the reader changes along a row and only the actor down a column. Beyond the comparisons of Sections 3.2 and 6.1, Exec-ED's crops do not gain over OPD's under Base's weaker reader (-0.23).

Table 9: Reader–actor swaps on ZoomBench-MCQ: probability that the reader answers correctly from the actor's sampled crops (%, malformed crops repaired; Exec-ED at λ=0.1).

	Reader		
Actor	Base	OPD	Exec-ED
Base	46.56	56.03	56.20
OPD	47.45	56.18	56.20
Exec-ED	47.22	56.93	57.24

E ADDITIONAL RESULTS

Paired differences and complete results. Table 10 gives the paired differences behind Tables 3 and 4. Table 11 lists every label-free model trained with the 9B teacher on ZoomBench-MCQ, Table 12 the models of Table 3 trained with the 4B teacher, and Table 13 the models of Table 3 on all 170 V* questions, pooled and by type. VA-OPD + Exec-ED, which adds the Exec-ED credit to VA-OPD, reaches the highest IoU of any model with the 9B teacher but improves neither accuracy nor G over Exec-ED and costs greedy accuracy, so a higher IoU need not mean a more useful crop.

Table 10: Exec-ED against OPD and VA-OPD, the comparisons behind Tables 3 and 4: paired differences (points) with 95% bootstrap intervals, malformed actions repaired. V*: the 96 four-option questions.

Difference	Accuracy	Precision (V*: hit)	G
ZoomBench-MCQ			
Teacher Qwen3.5-9B			
Exec-ED (λ=0.1) − OPD	+1.69 [0.35, 3.02]	+2.34 [1.35, 3.33]	+0.85 [0.15, 1.57]
Exec-ED (λ=0.2) − OPD	+1.40 [0.04, 2.80]	+3.74 [2.83, 4.68]	+1.07 [0.31, 1.84]
Exec-ED (λ=0.1) − VA-OPD	+0.39 [−1.04, 1.82]	+3.36 [2.33, 4.38]	+0.66 [−0.12, 1.43]
Exec-ED (λ=0.2) − VA-OPD	+0.10 [−1.38, 1.57]	+4.76 [3.81, 5.75]	+0.87 [0.03, 1.70]
Teacher Qwen3.5-4B			
Exec-ED (λ=0.1) − OPD	+0.70 [−0.51, 1.89]	+1.47 [0.70, 2.21]	+0.47 [−0.14, 1.07]
Exec-ED (λ=0.2) − OPD	−0.70 [−1.95, 0.52]	+3.90 [2.99, 4.83]	+1.06 [0.34, 1.78]
Exec-ED (λ=0.1) − VA-OPD	+0.64 [−0.79, 2.08]	+1.15 [0.46, 1.84]	+0.31 [−0.30, 0.91]
Exec-ED (λ=0.2) − VA-OPD	−0.76 [−2.27, 0.76]	+3.58 [2.72, 4.50]	+0.90 [0.20, 1.58]
Teacher Qwen3.5-9B, seed 43			
Exec-ED (λ=0.1) − OPD	+0.72 [−0.62, 2.03]	+4.58 [3.70, 5.47]	+1.20 [0.48, 1.93]
Exec-ED (λ=0.2) − OPD	+0.39 [−1.20, 1.97]	+6.62 [5.59, 7.68]	+1.14 [0.34, 1.94]
Exec-ED (λ=0.1) − VA-OPD	−0.17 [−1.58, 1.28]	+2.83 [1.98, 3.73]	+0.73 [−0.04, 1.48]
Exec-ED (λ=0.2) − VA-OPD	−0.50 [−2.05, 1.04]	+4.87 [3.88, 5.87]	+0.66 [−0.17, 1.48]
V* Bench			
Teacher Qwen3.5-9B			
Exec-ED (λ=0.1) − OPD	+2.60 [−0.65, 6.25]	+4.17 [1.04, 7.55]	+0.77 [−1.42, 2.77]
Exec-ED (λ=0.2) − OPD	+2.60 [−0.26, 5.47]	+5.73 [2.08, 9.64]	+2.31 [0.30, 4.26]
Exec-ED (λ=0.1) − VA-OPD	+1.30 [−1.95, 4.95]	+3.65 [0.78, 6.77]	+1.13 [−0.71, 2.84]
Exec-ED (λ=0.2) − VA-OPD	+1.30 [−1.17, 3.78]	+5.21 [2.08, 8.33]	+2.66 [0.98, 4.33]
Teacher Qwen3.5-4B			
Exec-ED (λ=0.1) − OPD	0.00 [−2.60, 2.73]	+2.60 [0.26, 5.21]	+0.59 [−1.04, 2.25]
Exec-ED (λ=0.2) − OPD	+0.65 [−1.30, 2.60]	+1.04 [−1.56, 3.65]	+1.54 [−0.05, 3.24]
Exec-ED (λ=0.1) − VA-OPD	+1.17 [−1.70, 4.17]	+2.60 [0.52, 4.69]	+0.18 [−1.07, 1.42]
Exec-ED (λ=0.2) − VA-OPD	+1.82 [−1.56, 5.21]	+1.04 [−1.30, 3.39]	+1.14 [−0.10, 2.42]
Teacher Qwen3.5-9B, seed 43			
Exec-ED (λ=0.1) − OPD	+1.30 [−1.30, 4.04]	+5.47 [2.08, 9.38]	+1.06 [−0.59, 2.69]
Exec-ED (λ=0.2) − OPD	+3.39 [0.26, 7.03]	+6.77 [2.34, 11.46]	+1.98 [0.26, 3.70]
Exec-ED (λ=0.1) − VA-OPD	−2.08 [−5.21, 0.78]	+1.56 [−1.30, 4.69]	+0.17 [−1.73, 2.14]
Exec-ED (λ=0.2) − VA-OPD	0.00 [−2.99, 3.26]	+2.86 [0.00, 5.99]	+1.09 [−0.39, 2.63]

Greedy decoding and majority voting. Tables 14 and 15 rescore the evaluation records of Table 3; no new inference is run. Of Exec-ED's gains over OPD at λ=0.1 (Section 5.2), the interval excludes zero for greedy accuracy on V* with the 9B teacher and for the greedy vote in three of the four blocks. Among the greedy and vote comparisons, one interval excludes zero against Exec-ED: at λ=0.2, its greedy accuracy is 2.2 points below VA-OPD's on ZoomBench with the 9B teacher. Oracle@8 measures the coverage of the samples rather than the typical crop and rewards spreading them; Exec-ED does not raise it consistently, and lowers it by 2.6 points at λ=0.2 with the 4B teacher on ZoomBench.

Sampled and greedy actions. Table 16 separates sampled and greedy actions for the runs of Section 6; the runs that switch objective after update 32 (Appendix F) match the objective of their second half on the sampled action and barely differ on the greedy action.

Table 11: All label-free models on ZoomBench-MCQ with the Qwen3.5-9B teacher (%); the label-based controls are in Table 18. Malf.: malformed sampled actions; IoU and precision of the sampled crops against the evidence box; G: held-out gain.

	Accuracy					
Method	sampled	greedy	Malf.	IoU	Precision	G
Base	44.99	48.39	4.83	33.36	59.18	6.39
OPD	53.69	58.62	4.66	31.09	55.04	6.64
TA-OPD	52.54	57.47	4.29	28.41	51.61	5.58
VA-OPD	54.99	59.74	3.88	32.28	54.02	6.84
Exec-ED (λ=0.1)	55.38	60.19	4.25	33.15	57.38	7.49
Exec-ED (λ=0.2)	55.09	57.59	2.68	34.09	58.78	7.70
Exec-ED (λ=0.4)	56.46	59.82	3.30	34.69	61.11	8.42
Exec-ED$_S$	54.93	59.90	3.51	33.30	57.90	7.49
VA-OPD + Exec-ED	54.02	57.76	4.00	35.19	56.87	7.32
OPD w/o action KL	48.08	55.57	5.78	18.17	48.88	1.98
Exec-ED w/o action KL	58.56	59.24	0.91	34.28	66.01	11.02
Exec-ED→OPD	53.40	58.09	4.66	31.00	54.30	6.97
OPD→Exec-ED	54.97	58.91	4.04	33.01	57.00	7.60

Table 12: The models of Table 3 trained with the Qwen3.5-4B teacher, on ZoomBench-MCQ (%); columns as in Table 11.

	Accuracy					
Method	sampled	greedy	Malf.	IoU	Precision	G
Base	44.99	48.39	4.83	33.36	59.18	6.39
OPD	53.63	55.32	3.34	35.66	56.18	6.94
TA-OPD	49.79	54.33	3.38	32.90	54.57	5.87
VA-OPD	53.69	56.39	3.18	35.27	56.50	7.10
Exec-ED (λ=0.1)	54.33	56.02	2.39	35.79	57.65	7.41
Exec-ED (λ=0.2)	52.93	54.91	1.65	36.00	60.07	8.00

Table 13: The models of Table 3 on V* Bench (%). Top: mean over the 170 questions (measures as in Appendix C). Bottom: by question type; chance accuracy is 25% with four options and 50% with two.

	Accuracy						
Method	sampled	greedy	Hit	Area | hit	Recall	Malf.	G
Teacher Qwen3.5-9B							
Base	56.91	56.91	41.32	1.54	43.49	3.24	2.16
OPD	71.84	72.50	44.26	1.00	48.14	1.91	2.47
TA-OPD	70.15	74.56	41.03	1.01	44.87	2.21	1.50
VA-OPD	75.00	76.62	45.44	1.07	49.73	1.47	2.87
Exec-ED (λ=0.1)	72.65	75.15	47.35	0.77	51.97	1.76	2.90
Exec-ED (λ=0.2)	73.82	77.50	48.82	1.00	53.14	1.18	4.31
Teacher Qwen3.5-4B							
Base	56.91	56.91	41.32	1.54	43.49	3.24	2.16
OPD	72.79	74.41	47.94	1.90	52.65	1.18	3.09
TA-OPD	71.54	74.85	43.09	1.73	47.62	1.32	2.26
VA-OPD	71.99	73.68	47.50	2.04	52.31	1.47	3.02
Exec-ED (λ=0.1)	72.87	75.15	50.88	3.24	55.26	1.18	3.78
Exec-ED (λ=0.2)	73.46	73.97	50.00	1.75	53.86	1.03	4.60

	Attribute, 4 options (96)			Attribute, 2 options (6)			Relative position (68)		
Method	Acc.	Hit	G	Acc.	Hit	G	Acc.	Hit	G
Teacher Qwen3.5-9B									
Base	59.64	57.55	5.39	35.42	87.50	7.06	54.96	14.34	−2.85
OPD	72.14	61.20	5.19	100.00	91.67	6.97	68.93	16.18	−1.77
TA-OPD	70.05	56.77	4.64	100.00	87.50	6.22	67.65	14.71	−3.33
VA-OPD	73.44	61.72	4.84	100.00	87.50	6.68	75.00	18.75	−0.23
Exec-ED (λ=0.1)	74.74	65.36	5.96	100.00	91.67	7.67	67.28	18.01	−1.84
Exec-ED (λ=0.2)	74.74	66.93	7.49	100.00	91.67	7.65	70.22	19.49	−0.48
Teacher Qwen3.5-4B									
Base	59.64	57.55	5.39	35.42	87.50	7.06	54.96	14.34	−2.85
OPD	75.13	67.19	6.20	100.00	79.17	5.46	67.10	18.01	−1.51
TA-OPD	72.79	61.98	5.45	95.83	75.00	4.20	67.65	13.60	−2.42
VA-OPD	73.96	67.19	6.60	100.00	79.17	5.29	66.73	16.91	−2.24
Exec-ED (λ=0.1)	75.13	69.79	6.78	100.00	91.67	8.07	67.28	20.59	−0.84
Exec-ED (λ=0.2)	75.78	68.23	7.74	100.00	91.67	8.18	67.83	20.59	−0.16

Table 14: Greedy decoding and majority voting (%), rescored from the records of Table 3. Accuracy and majority vote of the eight answers from the sampled actions (as in Table 3) and of the four answers from the greedy action; Oracle@8: at least one of the eight sampled answers is correct. Malformed actions cast no vote, a question without votes scores zero, and a tie among m options scores $1/m$ if it includes the correct one.

	Sampled actions		Greedy action		
Method	Accuracy	Vote	Accuracy	Vote	Oracle@8
ZoomBench-MCQ, teacher Qwen3.5-9B					
Base	44.99	47.61	48.39	49.09	63.20
OPD	53.69	57.81	58.62	59.41	79.04
TA-OPD	52.54	56.52	57.47	58.66	82.51
VA-OPD	54.99	58.17	59.74	60.07	77.56
Exec-ED (λ=0.1)	55.38	59.65	60.19	61.72	77.72
Exec-ED (λ=0.2)	55.09	58.00	57.59	58.25	80.20
ZoomBench-MCQ, teacher Qwen3.5-4B					
Base	44.99	47.61	48.39	49.09	63.20
OPD	53.63	56.24	55.32	56.02	79.04
TA-OPD	49.79	52.98	54.33	55.53	78.22
VA-OPD	53.69	55.53	56.39	57.43	76.73
Exec-ED (λ=0.1)	54.33	57.59	56.02	56.85	77.23
Exec-ED (λ=0.2)	52.93	55.24	54.91	55.61	76.40
V Bench, teacher Qwen3.5-9B*					
Base	59.64	59.38	61.20	60.94	73.96
OPD	72.14	77.08	74.48	74.48	88.54
TA-OPD	70.05	74.48	78.12	79.17	91.67
VA-OPD	73.44	77.60	75.52	75.52	85.42
Exec-ED (λ=0.1)	74.74	79.17	78.39	78.12	91.67
Exec-ED (λ=0.2)	74.74	77.69	79.17	80.21	88.54
V Bench, teacher Qwen3.5-4B*					
Base	59.64	59.38	61.20	60.94	73.96
OPD	75.13	79.17	77.08	76.56	90.62
TA-OPD	72.79	78.65	77.86	78.12	87.50
VA-OPD	73.96	77.60	77.86	78.65	90.62
Exec-ED (λ=0.1)	75.13	79.95	78.91	80.73	88.54
Exec-ED (λ=0.2)	75.78	80.21	79.17	80.21	88.54

Table 15: Exec-ED against OPD and VA-OPD under greedy decoding and majority voting (Table 14): paired differences (points) with 95% bootstrap intervals below.

Difference	Greedy accuracy	Greedy vote	Sampled vote	Oracle@8
ZoomBench-MCQ, teacher Qwen3.5-9B				
Exec-ED (λ=0.1) − OPD	+1.57 [−0.12, 3.29]	+2.31 [0.24, 4.43]	+1.84 [−0.39, 4.08]	−1.32 [−3.76, 1.15]
Exec-ED (λ=0.2) − OPD	−1.03 [−2.90, 0.74]	−1.16 [−3.46, 1.06]	+0.19 [−2.03, 2.36]	+1.16 [−1.33, 3.65]
Exec-ED (λ=0.1) − VA-OPD	+0.45 [−1.43, 2.36]	+1.65 [−0.58, 3.88]	+1.49 [−0.90, 3.93]	+0.17 [−2.13, 2.47]
Exec-ED (λ=0.2) − VA-OPD	−2.15 [−4.24, −0.12]	−1.82 [−4.28, 0.59]	−0.17 [−2.42, 2.05]	+2.64 [0.33, 5.08]
ZoomBench-MCQ, teacher Qwen3.5-4B				
Exec-ED (λ=0.1) − OPD	+0.70 [−1.02, 2.38]	+0.83 [−1.31, 2.89]	+1.35 [−0.61, 3.27]	−1.82 [−4.28, 0.66]
Exec-ED (λ=0.2) − OPD	−0.41 [−2.24, 1.40]	−0.41 [−2.71, 1.82]	−1.00 [−2.96, 0.93]	−2.64 [−5.13, −0.17]
Exec-ED (λ=0.1) − VA-OPD	−0.37 [−2.45, 1.69]	−0.58 [−3.05, 1.89]	+2.06 [−0.08, 4.20]	+0.50 [−1.99, 2.98]
Exec-ED (λ=0.2) − VA-OPD	−1.49 [−3.56, 0.58]	−1.82 [−4.38, 0.75]	−0.29 [−2.66, 2.02]	−0.33 [−2.97, 2.30]
V Bench, teacher Qwen3.5-9B*				
Exec-ED (λ=0.1) − OPD	+3.91 [0.78, 7.29]	+3.65 [0.52, 7.81]	+2.08 [−3.65, 7.81]	+3.12 [−1.04, 7.29]
Exec-ED (λ=0.2) − OPD	+4.69 [0.00, 9.90]	+5.73 [0.52, 11.98]	+0.61 [−4.08, 5.21]	0.00 [−4.17, 4.17]
Exec-ED (λ=0.1) − VA-OPD	+2.86 [0.78, 5.21]	+2.60 [0.00, 5.21]	+1.56 [−4.17, 7.29]	+6.25 [2.08, 11.46]
Exec-ED (λ=0.2) − VA-OPD	+3.65 [−0.27, 8.33]	+4.69 [0.00, 9.90]	+0.09 [−3.91, 3.91]	+3.12 [0.00, 7.29]
V Bench, teacher Qwen3.5-4B*				
Exec-ED (λ=0.1) − OPD	+1.82 [−0.52, 4.43]	+4.17 [0.52, 8.33]	+0.78 [−3.12, 4.69]	−2.08 [−5.21, 0.00]
Exec-ED (λ=0.2) − OPD	+2.08 [−0.26, 4.43]	+3.65 [0.52, 7.29]	+1.04 [−1.56, 3.65]	−2.08 [−8.33, 3.12]
Exec-ED (λ=0.1) − VA-OPD	+1.04 [−2.08, 4.43]	+2.08 [−1.56, 6.25]	+2.34 [−2.34, 7.29]	−2.08 [−6.25, 2.08]
Exec-ED (λ=0.2) − VA-OPD	+1.30 [−1.56, 4.43]	+1.56 [−2.08, 5.73]	+2.60 [−2.60, 7.81]	−2.08 [−7.29, 3.12]

Table 16: Sampled and greedy actions on ZoomBench-MCQ (Qwen3.5-9B teacher, %). Samples = greedy: share of sampled actions identical to the greedy action. Bold and underline: best and second best per column.

	Sampled actions			Greedy action			Samples
Method	Acc.	IoU	G	Acc.	IoU	G	= greedy
Base	44.99	<u>33.36</u>	6.39	48.39	39.21	8.48	0.41
OPD	53.69	31.09	6.64	58.62	43.07	9.93	0.04
Exec-ED	<u>55.38</u>	33.15	7.49	**60.19**	**44.10**	<u>10.57</u>	0.08
Exec-ED w/o action KL	**58.56**	**34.28**	**11.02**	<u>59.24</u>	34.59	**10.91**	23.72
Exec-ED→OPD	53.40	31.00	6.97	58.09	43.96	10.45	0.04
OPD→Exec-ED	54.97	33.01	<u>7.60</u>	58.91	<u>44.07</u>	10.31	0.04

Sampling temperature. Table 17 lowers the action temperature of OPD and Exec-ED (λ=0.1) from 1.0 to 0.5; Exec-ED's difference from OPD excludes zero at every temperature for IoU and at 1.0, 0.8, and 0.7 for G (Section 6.3).

Table 17: Action temperature on ZoomBench-MCQ (Qwen3.5-9B teacher, %). Top: OPD and Exec-ED (λ=0.1) with actions sampled at temperature T, where T=1.0 is the protocol of Table 3; the greedy action does not depend on T. Width: mean IoU between a sampled and the greedy crop (higher is narrower). Bottom: Exec-ED − OPD at the same T (points), with 95% paired bootstrap intervals.

Model	T	Width	Accuracy	Precision	IoU	G
OPD	1.0	0.506	53.69	55.04	31.09	6.64
	0.9	0.546	54.95	57.41	33.10	7.51
	0.8	0.589	56.02	59.18	34.92	7.49
	0.7	0.626	56.02	60.69	36.57	7.93
	0.5	0.714	57.30	63.80	39.89	9.05
Exec-ED (λ=0.1)	1.0	0.537	55.38	57.38	33.15	7.49
	0.9	0.571	55.84	59.27	34.84	8.11
	0.8	0.611	56.66	61.25	36.77	8.44
	0.7	0.652	57.08	62.71	38.29	9.12
	0.5	0.722	58.07	65.01	40.70	9.65

T	Accuracy	IoU	Recall	G
1.0	+1.69 [0.35, 3.02]	+2.06 [1.39, 2.74]	+2.59 [1.75, 3.47]	+0.85 [0.15, 1.57]
0.9	+0.89 [−0.45, 2.26]	+1.74 [1.05, 2.47]	+2.04 [1.19, 2.98]	+0.60 [−0.12, 1.35]
0.8	+0.64 [−0.74, 2.03]	+1.85 [1.14, 2.60]	+1.84 [1.00, 2.77]	+0.95 [0.20, 1.73]
0.7	+1.05 [−0.33, 2.43]	+1.72 [0.98, 2.50]	+2.22 [1.32, 3.18]	+1.19 [0.42, 2.00]
0.5	+0.76 [−0.68, 2.23]	+0.81 [0.01, 1.66]	+0.89 [0.00, 1.82]	+0.60 [−0.18, 1.40]

Label-based controls. This paper addresses only the label-free setting: Exec-Acc and OPD+RLVR, the diagnostic controls of Section 3, and Exec-p_T, which credits the teacher's probability of the correct answer, use the label and are not competing methods; Table 18 lists them for reference.

Table 18: Label-based controls on ZoomBench-MCQ with the Qwen3.5-9B teacher (%), next to Exec-ED (λ=0.1), for reference. Malformed: malformed sampled actions.

Method	Change from Exec-ED	Label	Accuracy	Precision	G	Malformed
Exec-ED	–	–	55.38	57.38	7.49	4.25
Exec-p_T	$u = p_T(y^\star \mid s, o)$	✓	55.86	59.45	8.05	1.49
Exec-Acc	$u = \mathbf{1}[\hat{y} = y^\star]$	✓	56.39	56.98	8.07	1.94
OPD+RLVR	Exec-Acc on $\mathcal{A} \cup \mathcal{Y}$	✓	59.49	58.90	7.62	1.82

Treatments of malformed actions. Table 19 repeats the seed-42 comparisons of Table 10 under the three treatments of Appendix C. Scoring malformed actions as zero, the stricter test, keeps every difference whose interval excludes zero, adds two, both in favor of the model with fewer malformed actions, and changes no sign. Excluding them changes no sign either, but leaves one interval just including zero (G of Exec-ED at λ=0.2 against VA-OPD on ZoomBench with the 9B teacher, $+0.84$ $[-0.01, 1.67]$) and, as it also reweights the states, lowers the share of OPD's precision loss that Exec-ED recovers at λ=0.1 and 0.2, from 57% and 90% to 50% and 61% with the 9B teacher and from 49% and 130% to 24% and 77% with the 4B teacher. Under exclusion, one interval barely excludes zero (Exec-ED at λ=0.1 against VA-OPD in precision with the 4B teacher; lower end $+0.004$); we do not rely on it.

Table 19: Paired differences under three treatments of malformed actions (points; *: the 95% bootstrap interval excludes zero). Repaired: our protocol; Zero: malformed actions score zero; Excluded: valid actions only (Appendix C). V*: hit rate on the 96 four-option questions.

	Precision (V*: hit)			G		
Difference	Repaired	Zero	Excluded	Repaired	Zero	Excluded
ZoomBench-MCQ						
Teacher Qwen3.5-9B						
Exec-ED (λ=0.1) − OPD	+2.34*	+2.34*	+2.20*	+0.85*	+1.14*	+0.88*
Exec-ED (λ=0.2) − OPD	+3.74*	+3.74*	+2.67*	+1.07*	+2.02*	+0.95*
Exec-ED (λ=0.1) − VA-OPD	+3.36*	+3.39*	+3.76*	+0.66	+0.46	+0.77
Exec-ED (λ=0.2) − VA-OPD	+4.76*	+4.79*	+4.23*	+0.87*	+1.35*	+0.84
Teacher Qwen3.5-4B						
Exec-ED (λ=0.1) − OPD	+1.47*	+1.51*	+0.98*	+0.47	+0.92*	+0.42
Exec-ED (λ=0.2) − OPD	+3.90*	+4.02*	+3.10*	+1.06*	+1.73*	+0.88*
Exec-ED (λ=0.1) − VA-OPD	+1.15*	+1.13*	+0.69*	+0.31	+0.61	+0.31
Exec-ED (λ=0.2) − VA-OPD	+3.58*	+3.64*	+2.80*	+0.90*	+1.42*	+0.76*
V* Bench						
Teacher Qwen3.5-9B						
Exec-ED (λ=0.1) − OPD	+4.17*	+4.17*	+4.41*	+0.77	+0.54	+0.81
Exec-ED (λ=0.2) − OPD	+5.73*	+5.47*	+5.56*	+2.31*	+2.30*	+2.29*
Exec-ED (λ=0.1) − VA-OPD	+3.65*	+3.65*	+3.71*	+1.13	+1.02	+1.03
Exec-ED (λ=0.2) − VA-OPD	+5.21*	+4.95*	+4.86*	+2.66*	+2.78*	+2.51*
Teacher Qwen3.5-4B						
Exec-ED (λ=0.1) − OPD	+2.60*	+2.60*	+2.82*	+0.59	+0.52	+0.67
Exec-ED (λ=0.2) − OPD	+1.04	+0.78	+0.79	+1.54	+1.52	+1.54
Exec-ED (λ=0.1) − VA-OPD	+2.60*	+2.86*	+2.55*	+0.18	+0.65	+0.14
Exec-ED (λ=0.2) − VA-OPD	+1.04	+1.04	+0.52	+1.14	+1.65*	+1.01

F TRAINING DYNAMICS

Figure 5 shows the training curves behind Sections 6.2 and 6.3. In the runs that switch objective after update 32 (Exec-ED→OPD and OPD→Exec-ED), replacing Exec-ED by OPD drags the batch IoU below that of the Exec-ED run within a few updates and keeps it about 2 points lower, whereas replacing OPD by Exec-ED closes a gap of the same size within two updates; the batch IoU is measured against the training set's evidence boxes, which no objective uses.

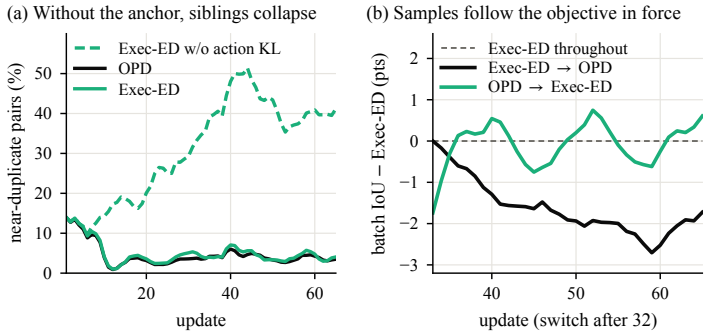

Figure 5: Training dynamics with the Qwen3.5-9B teacher (moving averages over 3 and 7 updates). (a) Near-duplicate sibling pairs: pixel-box IoU ≥ 0.9 between two valid siblings, among all 28 pairs. (b) Batch IoU of runs switching objective after update 32, minus Exec-ED's on the same batches.

G QUALITATIVE EXAMPLES

Figure 6 shows the crops that OPD, VA-OPD, OPD+RLVR, and Exec-ED (λ=0.1), all trained with the 9B teacher, sample for three ZoomBench-MCQ states with the same seeds. The first two states come from the 20 of 606 on which at least three of Exec-ED's four samples hit while OPD's greedy crop hits and at least two of OPD's samples miss, and the third from the 14 on which Exec-ED's samples hit at most once while OPD's hit at least twice, or at most twice while all of OPD's hit; among the candidates, we chose states whose crops are legible at this size. On the first state, every greedy crop covers the teddy bear, but two sampled crops each of OPD, VA-OPD, and OPD+RLVR land beside it, whereas all four of Exec-ED's cover it; OPD and OPD+RLVR still answer correctly three times out of four from the crops that miss, so accuracy alone does not reveal acquisition. On the second, the samples of OPD and VA-OPD scatter over the crowd and hit once each, whereas those of OPD+RLVR and Exec-ED concentrate on the dog and the person behind it. On the third, Exec-ED fails: two of its samples slide below the badge, and it answers correctly four times out of eight, against seven for OPD, whose four larger crops all include the badge.



Figure 6: Crops of four students (Qwen3.5-9B teacher, same sampling seeds) on three ZoomBench-MCQ states, each row showing one window around the evidence. Dashed amber: evidence box; thin: the four sampled crops, white if they hit (cover at least half of the evidence box), red if they miss; thick: the greedy crop. Below each panel: hits and correct answers (two per sampled crop, four for the greedy crop). Faces and a name tag are blurred in the third row.