RL-TRAINED CROP-AND-ZOOM IS OFTEN DECORATIVE: PIXEL BUDGETS, PROTOCOL COSTS, AND CAUSAL CREDIT

Anonymous authors
Paper under double-blind review

ABSTRACT

We ask when *crop-and-zoom* tools trained with reinforcement learning (RL) help and why they often fail. Many tool-equipped vision-language models (VLMs) turn out to behave *decoratively*: they localize accurately, yet the returned crop adds little causal information. The cause is a *pixel-budget gate*: a perfect crop's value is set by the gap between an image's resolution and the model's *working* pixel budget, and in every configuration we measure grows strictly monotonically as that budget tightens (from $+5.8$ to $+31.1$ points across datasets and scales). Real above-budget imagery (MME-RealWorld-Lite, DocVQA) confirms the degradation the gate rests on: whole-image accuracy falls under a tight budget only for models that can use the pixels. Standard benchmarks and RL training pools (median 0.19M pixels) sit below that budget, so crops that are useless there carry recoverable information at deployment budgets.

Where a crop does carry value, the standard additive tool-call serialization throws it away: a plain second image helps in all 16 Qwen dataset-by-budget comparisons ($+6.8$ to $+17.3$ points), but routing the identical pair through a tool-call exchange costs the untrained 3B base model 16–52 points across our budget sweeps, and larger untrained models 4–25 at 12.8M, by inducing tool-call loops; RL training only reduces that loss without learning to use the crop. Letting the returned crop *replace* the full image, or presenting it plainly, recovers most of the loss with no retraining.

We audit tool use with *counterfactual ablation credit*: a call counts only if replacing its returned crop with a shifted crop flips the answer. On V* it is empirically nested within probe-style credit ($p < 10^{-4}$); yet, under the standard group relative policy optimization (GRPO) recipe, even content-causal credit with above-budget data does not yield durable, selective tool use. Together, these results reduce to three conditions governing when crop-and-zoom can pay: images served above the model's working budget, the crop presented in place of the full image, and content-causal credit.

1 INTRODUCTION

Vision-language models (VLMs) are increasingly deployed on high-resolution imagery (remote-sensing scenes, document scans, and medical frames), where the evidence needed to answer a question can occupy a tiny fraction of the canvas. A standard remedy is to equip the model with a *crop-and-zoom tool*: the model names a region, the system returns an enlarged crop, and the model answers with both in context. Trained with reinforcement learning (RL), most commonly group relative policy optimization (GRPO) (Shao et al., 2024) (the reasoning-RL paradigm (Guo et al., 2025) extended to vision with verifiable rewards (Liu et al., 2025)), tool-equipped VLMs report strong gains on high-resolution benchmarks (Zheng et al., 2026).

Why a tool helps is harder to establish than whether it does: a tool-equipped model beating a tool-free one confounds whether a call was *used*, whether the crop was *usable* where it was served, and what the exchange itself *cost*. Two questions go unanswered. *Did this call matter?* Because the crop accompanies a full image that stays in context, an answer can survive the crop's removal. *Could a*

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.

crop matter here at all? Only if the full image was not already readable—a property of the model's visual budget that no benchmark reports. The first needs a per-call measurement, the second a theory of where value can exist. Neither has a reassuring answer: on V* (Wu & Xie, 2024), our trained 3B zoom-RL model (the DeepEyes recipe (Zheng et al., 2026) plus the anti-collapse exploration bonus (Wang et al., 2025a)) calls on every question and localizes accurately ($r = 0.83/0.89$), yet 97.6% of the calls it makes on questions it can already answer survive replacing the crop, and its causal share is $S_{\text{tool}} = 0.036$ (5 of 138). Tool use trained this way is largely *decorative*.

The crop does carry information, and its value is set by the *working* pixel budget. A VLM's visual encoder consumes a fixed token budget per image, so Qwen2.5-VL (Bai et al., 2025b) admits at most 12.8M pixels (16,384 visual tokens)—a *capability ceiling*, not the *operating point* (1–2M px is typical in deployment). Measuring a *perfect* crop's value across this budget with ground-truth boxes, crop accuracy is flat while full-image accuracy falls, so the crop's value rises **strictly monotonically** in every configuration we measure: $+5.8$ to $+31.1$ points (Table 1, Fig. 1). We call this the **pixel-budget gate**.

Standard benchmarks and training pools sit below the gate: essentially all benchmark images sit at or under the 12.8M ceiling, so an additive crop adds no pixels, and the RL *training* pool is further below (DeepEyes 47k(Zheng et al., 2026), median 0.19M px), leaving little for a credit to reward. Crop-and-zoom pays where real systems run, not where the field trains and evaluates.

Where value exists, the additive tool-call serialization discards it. Presenting a ground-truth crop as a plain second image helps in every comparison we run, while wrapping the identical pair in `tool_call`/`tool_response` costs the untrained 3B base model 16.3–51.8 points across the budget sweep (Table C.4). RL training moves the three image-reading conditions by at most 2 points, within the noise floor; its only measurable effect is shrinking the wrapper cost—teaching *when to stop* calling, not how to use the crop. The same crop given to the *untrained* model as a plain second image therefore beats the *trained* model routed through the tool—and letting the crop *replace* the full image recovers this loss wherever it is real.

To measure tool contribution honestly, we introduce counterfactual ablation credit. For each call, re-answer the question after replacing the returned crop with a same-sized *shifted* crop of the same image, and credit only calls whose answer flips. Because it substitutes the student's own crop at call level, it refines prior probe-style credit (Feng et al., 2026; Wang et al., 2026a): on the V* audit it never credits a call a probe would not, while a probe awards 17 pure re-asking artifacts (McNemar $p < 10^{-4}$, Section 6). It doubles as an inference-time test of whether a trained model's calls are genuine.

Three conditions for a working recipe. Given content-causal credit and above-budget data, one might expect durable tool use; under a standard recipe it does not emerge, and the failure is diagnostic rather than a tuning problem (Section 7). The findings isolate three conditions governing whether crop-and-zoom can pay in our setting (Fig. 1), and satisfying one or two does not guarantee durable, selective use: (i) images served above the pixel budget; (ii) the crop presented as a replacement, not an addition to a redundant full image; and (iii) content-causal credit.

Our contributions are:

1. **A protocol decomposition**: the wrapper, not the extra image, costs the untrained 3B base model 16–52 points across the budget sweep (4–25 on larger untrained models), and RL's only measurable effect is tolerating it.
2. **A measured pixel-budget gate**: a perfect crop's value rises strictly monotonically as the working budget tightens ($+5.8$ to $+31.1$ points across datasets, scales and families), turning *where a tool can pay* into a design choice readable before training.
3. **A causal credit measure**: an empirically nested refinement of probe-style credit on the audited V* calls ($p < 10^{-4}$), usable as a measurement and as a reward.
4. **The conditions for tool use**: even content-causal credit with above-budget data does not produce durable tool use, and the observed failure is *optimization-dependent* while changing the reward signal alone does not produce selective use: swapping the advantage estimator changes whether the tool survives (14–63 to 288–536 steps) but not whether its calls carry content (0.19–0.29).

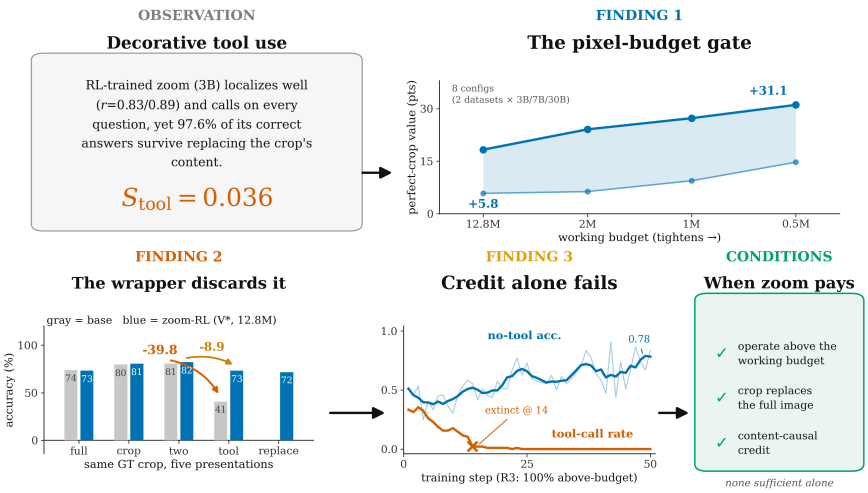


Figure 1: **A diagnostic chain for decorative crop-and-zoom.** RL-trained tool use is largely decorative ($S_{\text{tool}} = 0.036$). F1: the gate sets a perfect crop's value by the resolution–budget gap (+5.8 → +31.1 points across 8 configs). F2: the additive tool-call serialization discards that value; the wrapper costs the base model 39.8 points versus zoom-RL's 8.9 (V*, 12.8M). F3: content-causal credit alone still collapses the tool (R3). Bottom: the three conditions governing when the tool pays.

2 Related Work

Crop-and-zoom tools trained by reinforcement learning. Crop-and-zoom RL (Zheng et al., 2026; Wang et al., 2025a) trains a result reward on a crop tool, and, to counter cold-start collapse, recipes add an unconditional exploration bonus. Result-reward RL conflates two failures, *collapse* (never calling) and *decorativeness* (calling accurately but leaving the crop unused); the bonus converts the former into the latter, so an answer-only reward cannot distinguish them. Concurrent work (Kim et al., 2026) reports the same collapse, countering it with entropy regularization. Both lines treat tool use as a single scalar outcome; we measure *which* calls are content-causal.

Diagnosing the causal contribution of visual tools. MED (Ma et al., 2026) found tool use *coexists* with rather than is *mastered* by the base model; (Kumar et al., 2025) observes that a base model reads the downsized full image and oversamples hard questions instead. A concurrent audit (Wang et al., 2026c) reports a similar "calling without looking" mode at step level across six models. None treats the operating point as the variable that decides whether a crop can pay, or separates a crop's value from the cost of the exchange that delivers it.

Credit assignment for tool rewards. Probe-style credit (Feng et al., 2026) and teacher "benefit scores" (Wang et al., 2026a) reward calls whose answer improves without the tool, and can pay a call because re-asking helped rather than because the crop did. CFPO (Yu et al., 2026) uses an image-level counterfactual to regularize grounding rather than to credit calls; focuscrop (Zhao et al., 2026) makes the crop necessary but keeps an outcome-based reward and an aggregate diagnostic. Concurrent work (Peng et al., 2026) substitutes random same-size patches into the same call, as we do, but pairs it with a call-necessity term and a rent price. We test the substitution channel alone, and bound where any crop can pay at all (Appendix C).

Training-free zoom. HiDe (Liu et al., 2026) shows cropping helps below budget only when the crop *replaces* the full image and removes background, but at oracle precision and at a localization accuracy far below what a trained policy must supply; we ask what a learned, deciding model acquires under result-based RL.

High-resolution benchmarks (Wu & Xie, 2024; Wang et al., 2025b; Wei et al., 2026; Wang et al., 2026b; Yue et al., 2024; Fu et al., 2026; Zhang et al., 2025; Li et al., 2024; 2025) motivate fine-grained perception; we find essentially all of their images below Qwen2.5-VL's budget.

3 SETUP, PROTOCOL, AND METRICS

Models, budgets, and recipe. We study Qwen2.5-VL (Bai et al., 2025b) at two scales (3B, 7B), an untrained Qwen3-VL-30B(Bai et al., 2025a), and InternVL3-8B (Zhu et al., 2025) outside the Qwen family. Qwen2.5-VL's encoder cap is 12.8M pixels (16,384 tokens), a *capability ceiling* its sizes share; we treat the budget as a controllable *operating point*, capping `max_pixels` so each image is served at 0.25M–12.8M pixels. The "zoom-RL" model is a 3B trained with GRPO (Shao et al., 2024) on the DeepEyes 47k corpus with a result reward plus the standard unconditional exploration bonus (Wang et al., 2025a); every condition is also run on the untrained base model, with the untrained 7B and 30B confirming the scale claims.

All *training* results are on the 3B. Two arms ask how much of the outcome is specific to the optimizer: the identical recipe with GRPO's group-normalized advantage replaced by a batch-level moving baseline (REINFORCE++ (Hu et al., 2025)) or a leave-one-out baseline (RLOO (Ahmadian et al., 2024)), and one recipe repeated on Qwen2.5-VL-7B (Appendix A).

Protocol and data. Four conditions differ by exactly one variable, with system prompt and answer instruction fixed: `full` (image alone), `crop` ("crop-only"; the ground-truth (GT) crop alone), `two` (full and crop as two ordinary images), and `tool` (the same pair routed through `tool_call`/`tool_response`). A fifth, `replace`, runs the same exchange but lets the returned ground-truth crop be the only view. Ground-truth boxes, size-capped crops, deterministically shuffled options, and a frozen large language model (LLM) judge are used throughout, so comparisons are paired and the crop's value is a perfect-crop upper bound. Benchmarks must supply ground-truth boxes and be free of option-position artifacts; this selects V* (Wu & Xie, 2024) and ZoomBench (Wei et al., 2026) and excludes TreeBench (Wang et al., 2026b) (artifact) and HR-Bench (Wang et al., 2025b) (no boxes).

Standard benchmarks and training pools are below budget, so Section 4 additionally uses two above-budget corpora with no ground-truth crops — MME-RealWorld-Lite (Zhang et al., 2025) (photographs) and DocVQA (Mathew et al., 2021) (scans) — plus 100 hand-boxed DocVQA pages, for the whole-image degradation and for a real crop's value (Tables 2, C.3).

Judging reliability and localization. Three temperature-zero replays of one V* condition differ by at most $3/191$ questions (± 1.6 points); deltas below this floor are noise, paired comparisons use McNemar, and V* images are excluded from the RL pool by exact matching. *Localization* is the Pearson correlation between the requested crop's centre and the ground-truth box's centre, per axis, normalized by image size.

Measuring whether a call matters. Following MED (Ma et al., 2026), S_{tool} is the fraction of correct answers lost when the crop's content is ablated, measured per call: replace the returned crop with a same-sized *shifted* crop and re-answer greedily; a surviving correct answer is *decorative*, a flipped one *causal* (Section 6 builds the credit on this). During training these rates are reported only over steps with ≥ 20 testable calls.

4 FINDING 1: THE PIXEL-BUDGET GATE

Working budget, not capability ceiling. Within the encoder's fixed token budget we treat resolution as a controllable *working* point: an image above that budget is downsampled before the encoder sees it, and a crop that isolates a region recovers what the downsample destroyed.

The value of a perfect crop rises **strictly monotonically** as the working budget tightens in all eight crop-value configurations of Table 1 (untrained and RL-trained, 3B/7B/30B, V* and ZoomBench; monotone in 8 of 8, a count rather than a significance test since these share models and datasets); InternVL3-8B shows the same direction, only ZoomBench strictly ordered (Appendix B). Paired per-question tests put the entire effect on the full image: tightening the budget from 12.8M to 0.5M lowers full-image accuracy in every configuration ($p \leq 1.2 \times 10^{-3}$, McNemar; $< 10^{-11}$ on the largest models) while crop-only accuracy is unchanged ($p \geq 0.18$; on ZoomBench base full-image accuracy falls $47.1 \rightarrow 38.6$). We call this gap the **pixel-budget gate**.

	V*				ZoomBench			
working budget	base	zoom-RL	7B	30B*	base	zoom-RL	7B	30B*
0.5M	+14.7	+15.7	+30.9	+27.7	+21.1	+21.5	+24.3	+31.1
1M	+9.4	+10.5	+19.4	+19.4	+19.5	+20.1	+21.5	+27.3
2M	+6.3	+8.4	+13.6	+11.5	+16.5	+16.0	+16.8	+24.1
12.8M	+5.8	+7.3	+8.4	+8.9	+13.8	+13.8	+15.6	+18.3

Table 1: **Value of a perfect crop** (crop−full, points) across working budgets on V* and ZoomBench, from ground-truth-box compositions. *Qwen3-VL-30B base; `base` and `zoom-RL`: untrained and RL-trained Qwen2.5-VL-3B.

	MME-RealWorld-Lite				DocVQA			
working budget	base	zoom-RL	7B	30B*	base	zoom-RL	7B	30B*
0.5M	.372	.367	.390	.440	.849	.852	.889	.855
1M	.399	.404	.376	.486	.910	.907	.943	.931
2M	.358	.362	.454	.532	.925	.928	.946	.952
12.8M	.399	.404	.532	.555	.946	.940	.961	.946

Table 2: **Whole-image accuracy across working budgets on real above-budget imagery** (MME-RealWorld-Lite (Zhang et al., 2025), 218 natural photos; DocVQA (Mathew et al., 2021), 332 document scans). *Qwen3-VL-30B base; `base` and `zoom-RL`: untrained and RL-trained Qwen2.5-VL-3B.

Why RL-trained tool use cannot pay in the standard regime. Evaluation serves at the encoder ceiling while the training pool lies far below it, so on the redundant side of the gate a returned crop adds no pixels a credit could reward and tool use is decorative ($S_{\text{tool}} = 0.036$ on V*; Section 6).

Real imagery confirms the gate's driving term. Table 2 measures the whole-image degradation the gate rests on. On MME-RealWorld-Lite (natural photos; a quarter native $>$ 12.8M) accuracy rises with the budget only for models able to use the pixels: 7B gains $+14.2$ points from 0.5M to 12.8M ($p = 1.5 \times 10^{-5}$, McNemar), concentrated on images native above 4M ($+16.7$ vs. 1M on the 54 above-ceiling images, $p = 0.022$) and still rising to 12.8M (0.500 at 8M vs. 0.532; $p = 0.016$); 30B saturates at 4M (acc@$4/8/12.8\text{M} = 0.555$), a *working* budget below the shared 12.8M encoder ceiling; the 3B pair is flat (max ≈ 0.40; every $p \geq 0.12$). DocVQA (Mathew et al., 2021) triggers the gate for *every* model, since text detail is resolution-gated: cutting to 0.5M costs 7.2–9.6 points ($p \leq 2 \times 10^{-5}$, 3B included), while gains above 1M are ≤ 3.6 and native-above-12.8M scans stay legible (≈ 0). Real imagery thus reproduces the gate where it predicts and bounds the value a perfect crop could add on current real content. Residual low-budget dips in Table 2 are within paired-sample noise.

Real-image crop value. On the 100 hand-boxed native DocVQA pages (6–38M px) we compare crop-only with full-page accuracy across budgets (Table C.3). Crop-only accuracy is high and budget-insensitive (0.95 on 7B, 0.89 on zoom-RL), while full-page accuracy falls once the budget drops below the model's working point, so the ground-truth crop is worth $+8/+9$ points at 0.5M and $+14$–19 on the native 6–12.8M pages; by 4M the gap is gone. The gate's prediction therefore holds with real crops: a crop recovers what a tight budget destroys, and on documents — whose text stays readable near 1M — the magnitudes are smaller than on the synthetic fine-text compositions of Table 1.

5 Finding 2: The Additive Tool-Call Serialization Discards the Crop's Value

Adding a crop as a plain image helps. On the untrained and the RL-trained 3B models, presenting the ground-truth crop as a second ordinary image (`two`) improves accuracy over the image alone (`full`) in all 16 dataset-by-budget comparisons ($+6.8$ to $+17.3$ points), and the base 7B and 30B models show the same ordering at the 12.8M budget (Table 3), and so does InternVL3-8B(Zhu et al., 2025), outside the Qwen family, where all three signs of the decomposition appear (Appendix B).

Dataset	model	full	crop	two	tool	repl.	two−full	tool−two	repl.−tool
V*	base	0.738	0.796	0.806	0.408	—	+6.8	−39.8	—
V*	zoom-RL	0.733	0.806	0.822	0.733	0.717	+8.9	−8.9	−1.6
V*	7B base	0.822	0.906	0.922	0.686	0.880	+9.9	−23.6	+19.4
V*	30B base*	0.864	0.953	0.963	0.927	—	+10.0	−3.7	—
V*	InternVL3-8B†	0.696	0.906	0.901	0.649	0.927	+20.4	−25.1	+27.8
ZoomBench	base	0.471	0.609	0.564	0.401	—	+9.3	−16.3	—
ZoomBench	zoom-RL	0.468	0.606	0.557	0.486	0.571	+9.0	−7.2	+8.5
ZoomBench	7B base	0.489	0.646	0.632	0.514	0.636	+14.3	−11.8	+12.1
ZoomBench	30B base*	0.587	0.770	0.696	0.579	—	+10.8	−11.7	—
ZoomBench	InternVL3-8B†	0.473	0.730	0.696	0.544	0.704	+22.3	−15.1	+16.0

Table 3: **The additive tool-call serialization discards the crop's value; replacement recovers it** (V* and ZoomBench at 12.8M; accuracy, deltas in points). `two−full` > 0 and `tool−two` < 0 at every budget of the sweep; across the full four-budget sweep (0.5–12.8M) `two−full` ranges $+6.8$ to $+17.3$ points and `tool−two` from -16.3 to -51.8 points (full grid in Table C.4). `repl.` = `replace`. "—" = not run. *Qwen3-VL-30B base model; †InternVL3-8B (Appendix B).

Wrapping the identical pair in the tool-call exchange destroys this. Table 3 routes the same two images through the standard `tool_call` / `tool_response` exchange. The wrapper alone costs the untrained 3B base model 16.3–51.8 points relative to the plain `two` condition, large at every budget and peaking at an intermediate operating point (-51.8 at 2M on V* Table C.4; Table 3 shows the 12.8M column). The mechanism is behavioral, not perceptual: shown a tool-call exchange, the base model keeps issuing further tool calls instead of answering. On V* at the 2M point it averages 4.38 turns and hits the turn limit on 68% of questions, whereas the same model answers the same two plain images correctly 79.6% of the time. The defect shrinks with scale (V*: $-39.8 \rightarrow -23.6 \rightarrow -3.7$; ZoomBench: $-16.3 \rightarrow -11.8 \rightarrow -11.7$), so across the untrained models of Table 3 it spans 3.7–51.8 points, yet `two` still beats `tool` at every scale and dataset.

What RL training actually changes. Condition by condition, the three image-reading conditions move by at most 2 points on V* and 0.8 on ZoomBench, within the noise floor; RL produces no measurable reading improvement. Its only effect above noise is on the wrapper: the cost falls from -39.8 to -8.9 on V* and -16.3 to -7.2 on ZoomBench. zoom-RL thus learns when to *stop* issuing tool calls, not how to use the crop: the *identical pair* as two plain images scores 0.806 on V* (0.564 on ZoomBench), above the 0.733/0.486 the trained model reaches through the tool protocol, at zero training cost.

Present the crop plainly, or in place of the full image. The fix is on the presentation axis and needs no retraining: once a correct crop is in hand the full image adds little, and letting the returned crop *replace* it under the same protocol recovers the wrapper's loss wherever it is real and is never significantly worse. This is a ground-truth-crop upper bound that presumes the model localizes (zoom-RL does: $r = 0.83/0.89$); below the pixel budget no presentation change makes a call content-causal (Section 6).

6 Method: Counterfactual Ablation Credit

Why probe-style credit over-credits. Our credit replaces a call's returned crop with a same-sized *shifted* crop and re-answers greedily — a surviving correct answer is *decorative*, a flipped one *causal*; we use *causal* only in this operational, counterfactual sense, not as a claim about the model's internals (Section 3). Probe-style rewards (Feng et al., 2026; Wang et al., 2026a) ask a model to answer with and without the tool and credit calls whose answer improves. Because re-asking itself improves answers, a probe awards credit to calls whose crop was never used. Our intervention (Section 3) varies only the crop's *content* while keeping the questioning identical, so it isolates the crop's contribution: where probe over-credits (below budget, decorative), our credit is an empirical strict subset of the probe's (Table 4).

As an inference-time measurement. Applied to a trained model's own calls, the counterfactual ablation instantiates the causal share S_{tool} that MED (Ma et al., 2026) accounts for: on our trained model it yields $S_{\text{tool}} = 0.036$ on V* (5 of the 138 tool-using answers flip; MED reports 0.22–

sample set	probe credit	counterfactual credit	probe-only	ours-only	p
all-correct (n=138)	0.159 (22)	0.036 (5)	17	0	$< 10^{-4}$
decorative (n=126)	0.119 (15)	0.024 (3)	12	0	4.9×10^{-4}

Table 4: **On the decorative regime the counterfactual credit is empirically nested within probe-style credit** (V*, correct answers to a trained model's calls). The probe's 17 sole credits are re-asking artifacts the counterfactual credit refuses, and all of its 5 credits the probe also awards (McNemar $p < 10^{-4}$). The subset is empirical: above the budget the two diverge.

0.30)—and on the 126 questions it answers correctly *without* the tool, 97.6% of calls survive the same replacement. The estimate is robust to the ablation's form—gray, blurred, shifted, random-region and partly-overlapping replacements agree (0.029–0.044, pairwise Cohen $\kappa \geq 0.66$) and the share is flat in shift distance (Table C.2); the strongest variant is the most telling, since replacing the crop with a plain gray block moves accuracy from 0.723 to 0.728: the answer is largely insensitive to the crop's content. The same audit separates calls that genuinely depend on the crop: on an untrained 7B base whose returned crop is load-bearing, it returns $\approx$ 0.94–0.99 (Table C.2; representative calls in Fig. D.2).

As a reward signal. Following verifiable-reward practice for vision (Liu et al., 2025), we instantiate the credit in a GRPO reward $R = 0.8\,\text{acc} + 0.2\,\text{fmt} + 1.9\,\text{causal} + f \cdot [\text{tool called}] - p \cdot \text{decorative}$, where the causal and decorative terms come from the call-level counterfactual, and we sweep the exploration floor f and the decorative penalty p (six configurations, one seed each). The counterfactual signal stops paying for calls that did not matter: on the below-budget pool the pure reward (f=0) drives tool use to extinction, while an unconditional bonus saturates calls near 0.96, mostly decorative (Fig. C.1b); the floor f interpolates between these poles (Table C.1, Appendix C).

Whether the credit can *sustain* calls that do matter is answered in Section 7; under the standard recipe it cannot. On the protocol axis (Section 5), presenting the crop plainly or in place of the full image already removes the wrapper's loss, which credit alone cannot create.

7 Finding 3: Content-Causal Credit Alone Does Not Guarantee Selective Tool Use

Above-budget synthetic data is still not enough. The gate (Section 4) predicts tool value only above the budget, so we train on canvases genuinely above it (composited from below-budget sources, evidence region at native resolution) with the counterfactual credit of Section 6; Table 5 lists the settings and Figure 2 the trajectories. Tool use collapses in every GRPO setting. The most diagnostic is the 100% above-budget run, where calls hit zero by step 14 while *no-tool accuracy keeps rising past* 0.79: the model reads the degraded full image (the bypass MED's "coexist" finding anticipates), so calls are judged decorative, unrewarded, and the tool dies. Extinction lands on no fixed schedule across the six seeds—between step 14 and 63 (R3) and step 13 and 60 (R4)—and the seeds that keep calling longest saturate tool use before collapsing; the longest run stays at the floor through step 387, so slow re-emergence does not occur (Table C.5).

The advantage estimator changes tool persistence, not its measured causal share. Replacing GRPO's group-normalized advantage by a batch-level or leave-one-out baseline—one config field, everything else byte-identical—does not rescue the above-budget arm (R4+REINFORCE++ dies at step 23), but on the tool-necessary pool it changes the outcome outright: under REINFORCE++ and RLOO (three seeds each) the tool survives 288–536 steps at full call rate—five of the six runs to their end, one RLOO seed to step 443—where GRPO dies by step 68. The credit does not follow: 71–81% of the surviving calls are decorative over each run's audit-credible window, and no arm—RLOO included, though its best seed is the most accurate run on this pool (0.84)—moves the causal share above 0.29. Durability is not usefulness: a model can call persistently while remaining causally insensitive to what the tool returns.

Why it collapses. An offline oracle confirms the synthesized images contain recoverable information (crop 0.67 vs. degraded-full 0.33 at 1M), so the collapse is not for lack of signal: the model learns to read the degraded full image faster than it localizes, and any image readable enough to

run	f, mix, budget	start	peak@	end	ext.	causal	acc pk	acc end
R1	$f0.49$, 25%, 12.8M	.37	.43@13	.00	71	.14	.77	.66
R2	$f0.70$, 52%, 12.8M	.38	.47@2	.00	25	—	.76	.62
R3	$f0.70$, 100%, 12.8M	.23–.33	.35–.87	.00	14–63	.20	.79–.83	.69–.73
R4	$f0.70$, 100%, 1M	.33–.37	.55–.87	.00	13–60	.22–.28	.79–.86*	.72–.75
R4 ++	$f0.70$, 100%, 1M	.33	.47@3	.00	23	—	.80	.72
R5	necessity, 0.25M	.39	.69@44	.00	68	.20	.81	.74
R5 ++	necessity, 0.25M	.30–.39	1.00	1.00	**none**	.19–.28	.68–.76	.41–.67
R5 RLOO	necessity, 0.25M	.30–.39	1.00	1.00	**none** (2/3)	.22–.29	.80–.92	.71–.84
R5, masked	necessity+noise, 0.25M	.50	.88@29	.01	60	.18	.78	.67
R4 (7B)	$f0.70$, 100%, 1M	.10	.10@1	.00	16	—	.79	.66

Table 5: **Above-budget training collapses tool use; the advantage estimator changes whether the tool survives, not the credit it receives.** Extinction is the first step after which the call rate stays ≤ 0.05; causal rate is weighted over steps with ≥ 20 testable calls; acc peak is a 5-step moving average and acc end the last-20-step mean. Values below 1 omit the leading zero. "++" marks the REINFORCE++ advantage (Hu et al., 2025), one config field from GRPO. R3, R4, R5 ++ and R5 RLOO span three seeds (values give the range); the other rows are single runs. R5 is the tool-necessary pool (Appendix D); its masked variant replaces the evidence cell with noise. The last row repeats the R4 recipe on Qwen2.5-VL-7B (Appendix A), where the audit is never measurable. *R4's no-tool accuracy exceeds a perfect native crop (0.67).

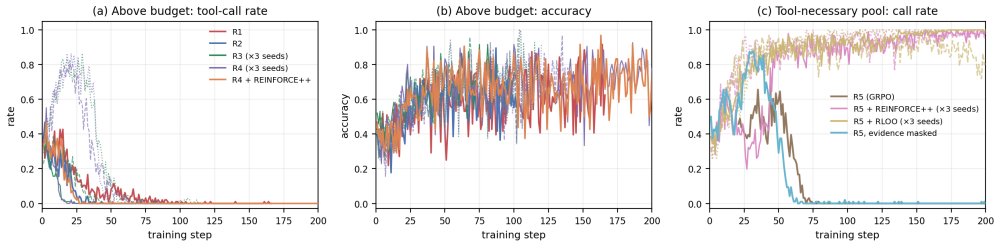


Figure 2: **Tool-call rate and accuracy over training.** (a) Tool-call rate and (b) accuracy for the synthetic above-budget runs R1–R4 (R3 and R4 each across three seeds; solid = seed 1, dashed/dotted = seeds 2 and 3) plus R4+REINFORCE++. Every seed collapses to zero on no fixed schedule (R3: steps 14–63; R4: 13–60) while no-tool accuracy stays at 0.79–0.86, and the longest run stays at the floor through step 387—an attractor, not a transient. (c) The tool-necessary pool: GRPO and its noise-masked variant both die (steps 68 and 60), whereas REINFORCE++ and RLOO (three seeds each) hold full call rate for 288–536 steps—one RLOO seed then collapses at 443; the other five run to their end—yet 71–81% of the calls they keep are decorative. R5 is omitted from (b): its pool's questions are coarse, so its accuracy is not comparable.

locate is readable enough to answer. Credit alone does not break it; the tool must be made *necessary*—and, as we show next, even that is not enough on its own.

Changing the working budget, or making the crop necessary, is not enough by itself. A 1M budget makes the degraded full image far weaker than a native crop (0.33 vs. 0.67), yet R4 still collapses by step 13 (Table 5): no-tool accuracy rises to 0.79–0.86, above a perfect native crop, consistent with memorization of, or shortcuts over, the small pool. Making the crop the only source of the clean image does not rescue it either (R5; Appendix D): with the evidence cell replaced by noise the policy still expands tool use from 0.50 to 0.88 and abandons it by step 60; on the blur-crush pool the audit certifies a load-bearing minority (20% of correct calls), the policy expands from 0.39 to 0.69, and abandons the tool by step 68. Result-based RL on this tool either fails to bootstrap tool use or converges to a decorative regime—which it then abandons, or (under a batch-level baseline) keeps without ever making the calls load-bearing.

8 Discussion and Limitations

Scope of the claims. Under the standard recipe (result or content-causal reward, additive tool protocol, below- or above-budget data), the model either abandons the tool or persists in a call-heavy, mostly decorative regime, with seed-level bistability at a marginal operating point (Section 7). We therefore report no training method that yields durable, selective tool use. The systems that do achieve it make the crop necessary at training time (Zhao et al., 2026), the direction our conditions (i)–(ii) prescribe; their outcome-based bonuses still pay decorative calls, and our credit audits whether a deployed model's calls are genuine.

Limitations. The gate and protocol decomposition are measured on three scales: untrained 3B/7B (Qwen2.5-VL) and 30B (Qwen3-VL) models, plus one RL-trained 3B, and Appendix B repeats the gate sweep outside the Qwen family. The RL-training experiments use a single crop-and-zoom tool and one model (Qwen2.5-VL-3B), with a 7B transfer in Appendix A. Swapping the advantage estimator under an otherwise byte-identical recipe (GRPO, REINFORCE++, RLOO) changes whether the tool survives, not whether its calls carry credit: the causal share stays within 0.19–0.29 throughout (Section 7). The above-budget settings R3 and R4, and the tool-necessary arm under REINFORCE++, were each run across three seeds (Fig. 2); every above-budget seed collapses to zero tool use, with seed-dependent timing. The training claims thus describe this recipe's trajectories, not an impossibility result for all RL recipes—an independent concurrent study reports the same collapse (Kim et al., 2026). The above-budget crop-value data are synthetic (a canvas composited from below-budget images), since standard benchmarks and training pools are all below budget and supply no ground-truth crops; Table 2 replicates the gate's driving term on real above-budget imagery (MME-RealWorld-Lite, DocVQA). The crop-value *magnitudes* remain synthetic. Benchmark images are excluded from the RL pool by exact matching (3), so every protocol and gate experiment evaluates models on data they did not train on. The above-budget training pools have no held-out split, however: the trajectory accuracies of Fig. 2 and Table 5 are means over the training batch, so their 0.79–0.86 no-tool accuracy is not a held-out estimate, and on a $4{,}000$-row pool memorization or shortcut learning remain open explanations alongside the reading explanation of Section 7. The counterfactual judgment reuses the accuracy judge, so judge error propagates into the causal/decorative labels. Option-position artifacts, which we found can invalidate cross-model comparisons, are controlled by the shuffled-choice and n-aware protocol of Section 3; TreeBench is excluded on this basis. We study one tool design (additive crop-and-zoom) and one serialization of it, and do not separate its schema cost from turn count or token budget (Appendix D); results may not transfer to replacement-style tools or imagery already above the encoder budget.

Implications. The gate turns the operating point into a measurable design choice rather than a hyperparameter: tool value exists only below the image resolution, and the working budget is model-specific, readable off real imagery (Table 2; Appendix A)—≈ 4M for 30B, above 12.8M for 7B, less on scans. With the protocol and credit conditions of Sections 5 and 6 this yields two checkable requirements: the gate statistics predict where value can exist for a given pool and budget, and S_{tool} measures whether a deployed model's calls are genuine. Conversely, a recipe that produced content-causal calls on below-budget benchmarks (S_{tool} far above zero) would falsify the gate's prediction that those benchmarks cannot reward tool use; across all of our runs we observed none.

9 Conclusion

We have shown why RL-trained crop-and-zoom is largely decorative, and where it is not. A perfect crop's value is gated by the working pixel budget ($+5.8 \rightarrow +31.1$ points as the budget tightens, across datasets, scales and families), and standard benchmarks and the training pool sit on the redundant side of that gate. Where value exists, the additive tool-call serialization discards it—the wrapper, not the extra image, costs the untrained 3B base model 16–52 points across our budget sweeps—while presenting the crop plainly or in place of the full image recovers most of that loss with no retraining. Credit alone is not sufficient either: under the standard recipe neither content-causal credit nor above-budget data yields durable, selective tool use. Our counterfactual ablation credit makes each of the three checkable: as a measurement it certifies whether a deployed model's calls are genuine, as a reward it pays only the calls whose content mattered.

ETHICS STATEMENT

Optional; does not count toward the page limit.

We use only publicly released benchmarks and corpora, each under its own license, and redistribute no benchmark image; the supplement ships generators and command lines instead. The one derived artifact we release is 100 ground-truth boxes we drew ourselves on DocVQA validation pages—we ship the boxes and the tool, not the pages. No human subjects were involved: the boxes were drawn by the authors, and the judge that scores answers is a model, not a person. Because a model judge can err in ways that correlate with fluency, its error propagates into the causal/decorative labels; we state this and report the measured noise floor (Section 8) rather than presenting those labels as ground truth. Nothing here is a safety or adversarial claim about any deployed system, and we have no conflicts of interest to declare.

REPRODUCIBILITY STATEMENT

Optional; does not count toward the page limit.

The training code is a patch against the public DeepEyes repository (`patches/zoomrl_cf.patch`, 13 files, verified to apply cleanly to the pristine upstream commit), and every run is one line of `train/train_deepeyes_3b_2gpu.sh`; `train/README.md` maps each run to its command. Synthetic pools are regenerated from `datagen/make_oversize_data.py` and the recipes in `datagen/recipes.md`. Benchmark images are not redistributed because their licenses forbid it, so we ship the preparation scripts, the commands and the raw per-item outputs instead. No number in the paper was transcribed by hand: each is read from an event file or a JSONL under `results/`, and `eval/` contains one command that recomputes the per-run table from those files. The two conventions that decide whether a rate is reported at all—extinction at a call rate that stays ≤ 0.05, and rates quoted only over steps with ≥ 20 testable calls—are implemented in code and stated in Section 3. `ENVIRONMENT.md` and `JUDGE.md` give the package versions, GPU model, judge endpoint and verbatim judge prompt; `README.md` §6 names what the supplement cannot rebuild.

AI USE STATEMENT

Required; this section does not count toward the page limit.

In this work, we used generative AI tools for drafting and revising parts of the paper text, writing and debugging the experimental-analysis and figure-generation code, and formatting the manuscript and references. We did not use generative AI to collect experimental data or to produce the quantitative results, tables, or figures, which are computed from our own training and evaluation runs and checked against the underlying logs and CSV/JSONL outputs. All AI-assisted output was reviewed, and the experimental design and the scientific claims in this work were verified by the authors. We take responsibility for the final content of this work, including any text, claims, code, or artifacts produced with the aid of generative AI.

REFERENCES

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.662. URL https://aclanthology.org/2024.acl-long.662/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

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, 2026. URL https://arxiv.org/abs/2606.30251.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=DgH9YCsqWm.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. Reinforce++: Stabilizing critic-free policy optimization with global advantage normalization, 2025. URL https://arxiv.org/abs/2501.03262.

Dong-Hee Kim, Reuben Tan, and Donghyun Kim. Diversity over frequency: Rethinking tool use in visual chain-of-thought agents. In *Workshop on Failure Modes of Agentic AI at ICML 2026*, 2026. URL https://openreview.net/forum?id=2fuiAuE4MR.

Sunil Kumar, Bowen Zhao, Leo Dirac, and Paulina Varshavskaya. Reinforcing vlms to use tools for detailed visual reasoning under resource constraints, 2025. URL https://arxiv.org/abs/2506.14821.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *International Conference on Learning Representations*, volume 2025, pp. 81182–81199, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/c9f95e9ec39fa5ad3d0a562b993b92aa-Paper-Conference.pdf.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26753–26763, 2024. doi: 10.1109/CVPR52733.2024.02527.

Xianjie Liu, Yiman Hu, Yixiong Zou, Liang Wu, Jian Xu, and Bo Zheng. Hide: Rethinking the zoom-in method in high resolution mllms via hierarchical decoupling, 2026. URL https://arxiv.org/abs/2510.00054.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044, 2025. doi: 10.1109/ICCV51701.2025.00197.

Yan Ma, Weiyu Zhang, Tianle Li, Du Linge, Xuyang Shen, and Pengfei Liu. What does vision tool-use reinforcement learning really learn? disentangling tool-induced and intrinsic effects for crop-and-zoom. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=iTWu7ziS5z.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 2199–2208, 2021. doi: 10.1109/WACV48630.2021.00225.

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.

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, 2024. URL https://arxiv.org/abs/2402.03300.

Chaoyang Wang, Kaituo Feng, Dongyang Chen, Zhongyu Wang, Zhixun Li, Sicheng Gao, Meng Meng, Xu Zhou, Manyuan Zhang, Yuzhang Shang, et al. AdaTooler-V: Adaptive tool-use for images and videos. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18058–18072, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.898. URL https://aclanthology.org/2026.findings-acl.898/.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani zheng, Sule Bai, Zijian Kang, Jiashi Feng, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *International Conference on Learning Representations*, volume 2026, pp. 148769–148794, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f09ab24aa37ae2664b3d923cdb0319b5-Paper-Conference.pdf.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning, 2025a. URL https://arxiv.org/abs/2505.15966.

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 Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025b. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i8.32852. URL https://doi.org/10.1609/aaai.v39i8.32852.

Zhiheng Wang, Bo Peng, Lai Wei, and Chaochao Lu. The illusion of visual tool-use: A causal audit of thinking with images, 2026c. URL https://arxiv.org/abs/2608.06270.

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. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=bVXx0K0j4c.

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, 2024. doi: 10.1109/CVPR52733.2024.01243.

ZhangYuan Yu, Wanran Sun, Guangjing Yang, Xiaohu Wu, and Qicheng Lao. CFPO: Counterfactual policy optimization for multimodal reasoning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=BfWFCVjsNe.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, June 2024.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. MME-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=k5VHHgsRbi.

Xuanpu Zhao, Zhentao Tan, Dianmo Sheng, Tianxiang Chen, Yao Liu, Yue Wu, Tao Gong, Qi Chu, and Nenghai Yu. Learning to focus and precise cropping:a reinforcement learning framework with information gaps and grounding loss for mllms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25938–25947, June 2026.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and XingYu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 126775–126798, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cdb347d7516d52a9280ea9d6708911f6-Paper-Conference.pdf.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

APPENDIX A THE WORKING BUDGET IS MODEL-DEPENDENT

The gate of Section 4 bounds a crop's value by the gap between the image's resolution and the *model's* working budget, so the budget at which a crop becomes worth taking is a property of the model rather than of the image. A direct contrast, holding data, reward, prompt and budget fixed: on the same 26M px pool at a 12.8M px budget, the 3B model calls the tool in 0.35 of rollouts at the first training step (R3, seed 3, before any update), whereas Qwen2.5-VL-7B calls it in 0.03 (3 of 96 rollouts), with trajectories averaging 4.3 observation tokens and 16.6k prompt tokens: the canvas is only $0.70\times$ downscaled at this budget, and the stronger model simply reads it. (That run logged only two steps; the contrast is a property of the pre-update policy, so it needs only step 1.) Lowering the same model to a 1M px budget—where both models see a 1,646-token prompt, identical to R4—narrows the gap but does not close it: the 3B starts at 0.35–0.40 (R4, three seeds, step 1) and the 7B at 0.10. Training does not close it either: over 113 steps the 7B call rate never rises above its step-1 value of 0.10, goes extinct at step 16, and never accumulates the 20 testable calls the audit needs (its peak is 8), while its accuracy still rises from 0.34 to 0.66 (Fig. A.1). We therefore report the 7B arm as a diagnostic rather than as evidence about credit—at a 10% call rate there is little for a credit signal to act on—and we state the training claims per (model, operating point) pair rather than per image size; Appendix D gives the calibration procedure.

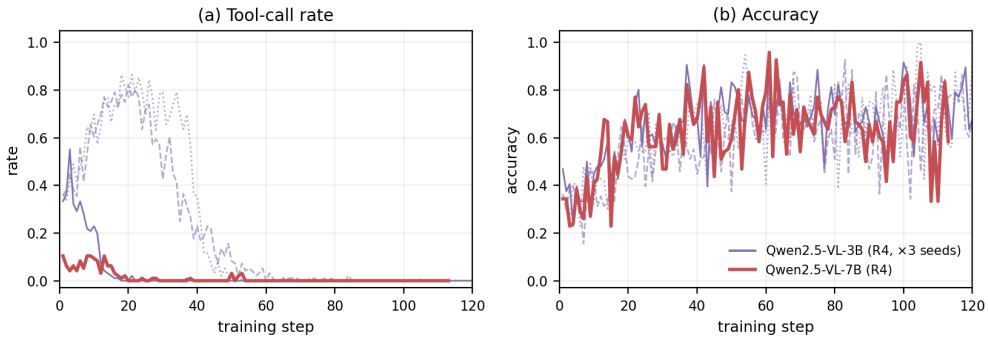


Figure A.1: **The 7B never bootstraps the tool at the budget where the 3B does.** Same above-budget pool, same 1M px working budget, same recipe; only the model differs. (a) The 3B's call rate starts at 0.33–0.36 and rises to 0.55–0.86 before collapsing; the 7B's is already at its peak (0.10) on step 1 and is gone by step 16. (b) Both reach comparable accuracy, so the 7B's failure to call costs it nothing on this pool—the stronger model reads the $0.70\times$-downscaled canvas directly. The 7B curve is a single seed; the 3B's are R4's three.

APPENDIX B A CROSS-FAMILY CHECK

The gate is measured on Qwen2.5-VL (3B, 7B) and Qwen3-VL (30B), so a reader may ask whether it is a property of that family rather than of the architecture–budget relationship it claims to describe. We therefore repeat the budget sweep of Table 1 on **InternVL3-8B**, a different family with a different image path (448-px dynamic tiling rather than a single resized view). The client serves every image at the same four working resolutions and compares a perfect ground-truth crop against the full image, exactly as in the Qwen sweep.

We also repeat the protocol decomposition of Table 3 on the same model, since Finding 2 is measured on Qwen alone as well.

Finally, the same model on the real above-budget imagery of Table 2. This is the weakest of the three cross-family checks and we report it as such: the direction holds on DocVQA and is monotone, but both effects are smaller than for the Qwen models and MME is not significant.

One difference between this arm and the Qwen arms should be stated rather than buried, because it bounds what the comparison shows. The Qwen processor turns a tighter working budget directly into fewer visual tokens (12.8M px = 16,384 tokens), so the Qwen sweep moves image resolution and the model's token budget together. The serving stack we used for InternVL3-8B holds the

run	0.25M	0.5M	1M	2M	12.8M
InternVL3-8B, ZoomBench	+28.0	+27.3	+25.3	+22.8	+22.8
InternVL3-8B, V*	+23.6	+22.0	+16.8	+18.9	+19.9
Qwen2.5-VL-3B `base`, ZoomBench	—	+21.1	+19.5	+16.5	+13.8
Qwen2.5-VL-3B `base`, V*	—	+14.7	+9.4	+6.3	+5.8

Table B.1: **The gate on InternVL3-8B.** Perfect-crop value (crop−full, points) for InternVL3-8B, with the Qwen2.5-VL-3B base rows of Table 1 repeated over the overlapping budgets for reference. On ZoomBench a crop is worth the most at the tightest budget and its value falls monotonically as the budget loosens—+28.0 to +22.8, a 5.2-point span—exactly as the gate predicts. On V* the two tightest budgets do give the largest values (+23.6, +22.0) but the interior is unordered (+16.8 at 1M against +18.9 and +19.9 above it); at $n = 191$ the interior differences are within one standard error, so we read the V* arm as directional rather than confirmatory. We stop at 0.25M px: below it the ground-truth crops themselves start to exceed the budget (6% of ZoomBench rows at 0.15M), which caps the crop arm and would break the comparison.

run	dataset	two−full	tool−two	replace−tool
InternVL3-8B	V*	+20.4	−25.1	+27.8
InternVL3-8B	ZoomBench	+22.3	−15.1	+16.0
Qwen2.5-VL-3B `base`	V*	+6.8	−39.8	—
Qwen2.5-VL-3B `base`	ZoomBench	+9.3	−16.3	—

Table B.2: **The protocol decomposition on InternVL3-8B** (points, at 12.8M px; the Qwen rows are from Table 3). All three signs replicate: adding the crop as an ordinary second image helps, routing the identical pair through a tool-call exchange hurts, and letting the crop *replace* the full image recovers most of that loss. The first gap is larger here than for Qwen—InternVL's effective visual budget is smaller (∼1.2k tokens against 16,384), so its images sit further above the budget and a perfect crop is worth more, which is what the gate predicts.

dataset	0.5M	1M	2M	12.8M	span
MME-RealWorld-Lite	0.486	0.477	0.482	0.509	+2.3 ($p = 0.26$)
DocVQA	0.889	0.892	0.910	0.922	+3.3

Table B.3: **Whole-image accuracy for InternVL3-8B on real above-budget imagery.** The Qwen rows of Table 2 span +2.7 (3B), +14.2 (7B) and +11.5 (30B) on MME and +9.7, +7.2 and +9.1 on DocVQA, so InternVL's spans are the smallest of any model measured—consistent with the fixed and small visual budget this arm runs under (Table B.1), which caps how much more it can extract from a larger served image. On MME it behaves like the 3B pair, which the main text already reports as flat; DocVQA is monotone.

visual token count fixed across input resolutions, so this arm moves the *served* resolution at a fixed model budget—a narrower manipulation, and the most likely reason its span is smaller. Within that narrower manipulation the direction is the gate's: the further the image sits above what the model actually consumes, the more a perfect crop is worth, and the effect is clean on ZoomBench and directional on V*.

We attempted a second out-of-family model, Molmo-7B, and report why it is absent rather than quietly omitting it. Its chat template rejects a system role outright, so the fixed system prompt every other condition shares could only have been delivered as a user turn; its serving stack admits one image per request, which removes the two-image and tool-wrapper conditions entirely; and its visual-token budget saturates above 0.5M px, so its usable sweep range does not overlap the one used here. An arm measured under a different prompt convention, without half the conditions, and over a disjoint budget range would not have been comparable, so we report InternVL3-8B alone.

APPENDIX C THE COUNTERFACTUAL CREDIT AS A REWARD SIGNAL

Section 6 instantiates the counterfactual credit in a GRPO reward $R = 0.8\,\text{acc} + 0.2\,\text{fmt} + 1.9\,\text{causal} + f \cdot [\text{tool called}] - p \cdot \text{decorative}$ and sweeps the exploration floor f and decorative penalty p.

How the per-call terms aggregate. Every executed call is ablated *independently*: we replay the trajectory up to that call's prefix, substitute a same-sized shifted crop for its returned observation, and regenerate the continuation *from that point* with the same greedy decoder, so later calls are regenerated rather than replayed from the original rollout. A call contributes causal=1 if its own answer flips from correct to incorrect and decorative=1 if a correct answer survives; a call that was already wrong under the real crop contributes to neither. The two terms are then averaged over the trajectory's executed calls, so 1.9 causal and p decorative are per-trajectory means in the same units as acc, not sums over calls; a rollout with no calls scores 0 on both. Ablation generation is chunked (8 per forward pass) and a call whose ablated continuation fails to parse is dropped from the mean rather than counted as decorative, which is why the per-run testable-call count in Table C.5 can be below the number of calls executed.

Sparsity of the causal signal. The causal term is sparse by construction: a call earns causal credit only when the real crop yields a correct answer *and* the shifted crop flips it, so a call that is already wrong under the real crop contributes to neither term. At step 1 of R4 the audit has 16 testable calls and a causal share of 0.188, i.e. three calls in the batch. This is a property of the reward we test, and it bounds what the negative result can support: we do not rule out that a denser counterfactual signal — a change in answer probability or margin rather than a hard correct-to-incorrect flip — would optimize differently, and we did not run that comparison.

Sparsity alone does not explain the outcome, however. On the tool-necessary pool the same reward reaches a causal share of 0.20 (Appendix D) and holds the tool at full call rate for 288–536 steps under REINFORCE++ and RLOO, so the signal is dense enough to keep calls in place; what does not follow is their content, which stays 71–81% decorative. A sparse reward that sustains calling but not causal calling is the failure we report.

Table C.1 and Figure C.1 give the full dose-response. Only an intermediate floor (f=0.10) produces a delayed decline; f=0 drives tool use to extinction and a large floor (f=0.15) keeps call rates near 0.9, with most of those calls decorative. Grey segments in Figure C.1 mark steps with fewer than 20 testable calls.

Config	s5	s10	s15	s20	s25	s30	s35
uncond. 0.7 (prior-art)	0.66	0.95	0.95	0.96	—	—	—
f=0, p=0	0.38	0.44	0.23	0.16	0.10	0.05	0.01
f=0.05, p=0.40	0.35	0.15	0.02	0.00	—	—	—
f=0.10, p=0.30	0.36	0.50	0.62	0.71	0.67	0.58	0.47
f=0.15, p=0.25	0.38	0.44	0.66	0.83	0.91	0.91	0.93
f=0.15, p=0	0.36	0.58	0.82	0.90	0.94	0.93	—

Table C.1: **Tool-call rate under counterfactual-ablation rewards (below-budget pool; per-sample, defined at every step).** The exploration floor f governs the outcome: f=0 drives calls to extinction, f=0.15 saturates them near 0.93, and only f=0.10 declines mid-course.

CONCURRENT COUNTERFACTUAL AUDITS

Two 2026 papers intervene on a returned observation and ask what the answer does without it.

(Wang et al., 2026c) replaces one observation and probes the model's immediate preference over the answer options at that step, across six models and five benchmarks, reporting a "calling without looking" mode. We regenerate the full answer under a same-size *shift* of the same image, and pair the audit with the gate that bounds where a crop can carry recoverable information.

(Peng et al., 2026) substitutes random same-size patches into the same call, as we do, and pairs the substitution with a call-necessity term, a rent on every executed call, and a dual-channel advantage. It reports $+6.3$ to $+9.4$ points over outcome-only GRPO while leaving 31–36% of calls spurious. Our experiments isolate the substitution channel on its own, and that residual is consistent with

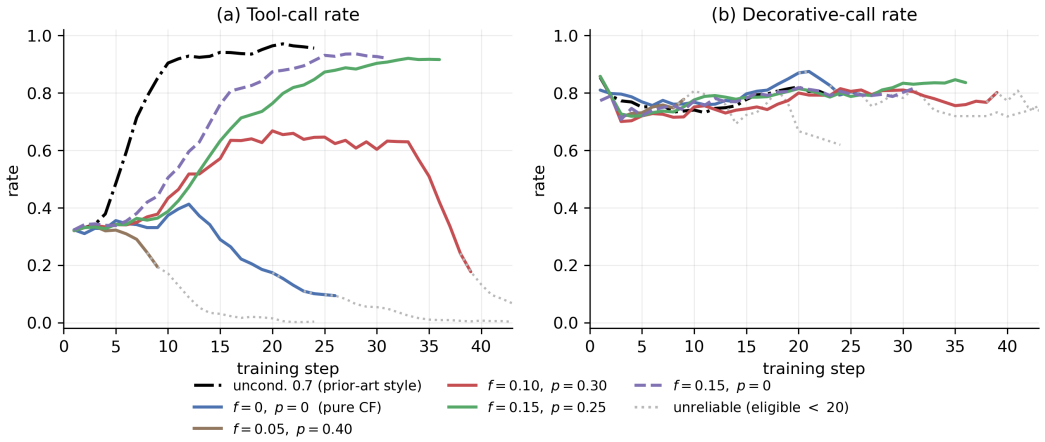


Figure C.1: **Reward ablation on the below-budget pool.** (a) Tool-call and (b) decorative-call rate over training for Table C.1 plus the prior-art baseline; gray segments have < 20 testable calls and are uninformative.

Replacement of the returned crop	3B zoom-RL	7B base
shift	0.037	0.99
gray	0.037	0.94
blur	0.044	0.95
random region	0.044	0.96
neighbor (partly overlapping)	0.029	0.95
distance 20% / 45% of width	0.051 / 0.029	0.96 / 0.97
probe-style (re-ask)	0.154	0.125

Table C.2: **Content-causal share of correct tool-using calls under different replacements of the returned crop** (V*; 136 correct calls for the RL-trained 3B, 112 for the untrained 7B base). On the 3B the five counterfactuals agree within 0.029–0.044 (pairwise Cohen $\kappa \geq 0.66$) and the share is flat in shift distance, so the audit does not hinge on the ablation's form. This robustness pass resamples the replacement crop and scores over the 136 calls the original trajectory got right; the primary audit of Section 6 scores over the 138 calls correct under replay, which is why it reports $S_{\text{tool}} = 0.036$ where the shift row here reads 0.037. Probe-style credit marks 0.154 of the same calls, four times higher. On the 7B base the same audit returns ≈ 0.94–0.99, because its calls genuinely depend on the returned crop; the measure therefore separates load-bearing from decorative calls.

what we find. Their benchmarks also sit below our gate, so a gain there is attributable to region reallocation or to policy calibration rather than to recovered resolution — a distinction our fixed-budget crop-versus-full comparison is designed to separate.

APPENDIX D THE TOOL-NECESSARY VARIANT: DOES CREDIT KEEP A LOAD-BEARING TOOL?

This arm is a *stress test*, not evidence about benchmark-level usefulness: it asks whether causal credit can survive when tool necessity is deliberately constructed, not whether the tool helps on real tasks.

To test whether content-causal credit suffices when the clean image reaches the model only through the tool, we build a pool in which the clean source is reachable *only* through the tool: the model-visible image is a crushed composite (the source pasted into a 34M px canvas, its evidence region low-passed, or replaced by noise in the masked variant), the carried tool image is the clean source, and a zoom returns it directly—the pool's annotations carry no box, so any zoom succeeds and the

condition	7B	zoom-RL
full page @ 0.5M	0.870	0.800
full page @ 1M	0.940	0.850
full page @ 2M	0.980	0.870
full page @ 4M	0.980	0.900
full page @ 8M	0.960	0.920
full page @ 12.8M	0.960	0.920
crop-only (GT region, native)	0.950	0.890

Table C.3: **Real-image crop value on DocVQA** (100 hand-boxed pages, native 6–38M px). Full-page accuracy falls at low budgets while the native GT crop stays accurate, so crop−full is positive at $\leq$ 1M ($+8$/$+9$ points at 0.5M) and vanishes by 4M; on the native 6–12.8M pages the 0.5M gap is $+14$–19 points.

model	dataset	budget	crop−full	two−full	oracle−two
base	V*	0.5M	+14.7	+12.0	−35.6
base	V*	1M	+9.4	+8.9	−45.6
base	V*	2M	+6.3	+7.8	−51.8
base	V*	12.8M	+5.8	+6.8	−39.8
base	ZoomBench	0.5M	+21.1	+17.3	−22.5
base	ZoomBench	1M	+19.5	+12.3	−20.0
base	ZoomBench	2M	+16.5	+10.5	−24.3
base	ZoomBench	12.8M	+13.8	+9.3	−16.3
zoom-RL	V*	0.5M	+15.7	+11.5	−4.2
zoom-RL	V*	1M	+10.5	+8.9	−2.1
zoom-RL	V*	2M	+8.4	+8.4	−3.1
zoom-RL	V*	12.8M	+7.3	+8.9	−8.9
zoom-RL	ZoomBench	0.5M	+21.5	+16.3	−12.7
zoom-RL	ZoomBench	1M	+20.1	+14.5	−9.8
zoom-RL	ZoomBench	2M	+16.0	+10.2	−8.0
zoom-RL	ZoomBench	12.8M	+13.8	+9.0	−7.2

Table C.4: **Full protocol sweep** (V* and ZoomBench; untrained 3B base and RL-trained zoom-RL; points). Deltas are crop−full = crop-only minus full-only accuracy, two−full = plain two-image minus full-only, and oracle−two = the tool-call wrapper cost over the plain two-image condition. Two−full is positive in all 16 comparisons (ranging $+6.8$ to $+17.3$) and oracle−two negative on the base model (ranging -16.3 to -51.8), the source of the ranges in Table 3 and the abstract. TreeBench is excluded (option-position artifact).

run	seed	tool-call peak	extinction step
R3	s1	0.35	14
R3	s2	0.39	39
R3	s3	0.87	63
R4	s1	0.55	13
R4	s2	0.82	60
R4	s3	0.87	49

Table C.5: **Seed-level collapse of the two above-budget runs.** Tool-call peak and extinction step (the first step after which the call rate stays ≤ 0.05) for each of six seeds; every run converges to extinction, but the timing varies by more than $4\times$.

variant is aimed at readability-necessity rather than localization-necessity. The counterfactual arm still crops a same-size region from the *crushed* canvas, so a genuinely load-bearing call is one whose clean copy changes the answer.

The working budget must be calibrated, not assumed. The model sees the canvas scaled to the working budget, so the evidence region occupies about budget/cols^2 pixels, with cols $= \lceil\sqrt{\text{target/source}}\rceil \in [7, 22]$ across rows here. Measured on strided samples of the pool, the re-

gion's median size in the model view is ∼6.9k px (∼83 px on a side) at 1M, ∼3.5k (∼58) at 0.5M, ∼1.7k (∼41) at 0.25M and ∼0.7k (∼26) at 0.1M. We release the sampler used for this (`make_samples.py`), which renders that region at 1:1 for any candidate budget. An uncalibrated 1M budget is not a valid manipulation check: the call rate starts at 0.23—comparable to R3–R4—and the audit never accumulates 20 testable calls per step, so nothing can be concluded about credit at all. The pool's questions are also coarse—left/right relations account for 93% of a 150-row strided sample and survive a 10× downscale—so we report the *causal share*, which needs no assumption about what a human could read, rather than accuracy alone. The budget at which a crop starts to pay also moves with the model: Appendix A contrasts 3B and 7B on identical data and budget.

At a calibrated budget the tool grows and then dies. At 0.25M px the crop starts to matter: 20% of correct calls are content-causal, measured over the 32 steps that carry ≥ 20 testable calls, and the call rate *rises* from 0.39 to 0.69 by step 44. It then decays to zero by step 68, and the audit shows why: after step 32 the causal share stays below 0.4 throughout, i.e. the policy drifts into a *decorative* regime *before* it abandons the tool. Content-causal credit does not keep even a crop that the audit itself certifies as load-bearing for a fifth of the calls: the calls it can pay for are too few to survive the decorative majority (Table D.1). Masking the evidence cell does not isolate necessity either: without the tool the model still answers well above the 0.47 chance level of this pool (0.67), and the tool still goes extinct (step 60). Nor does the outcome depend on the credit being group-normalized: with REINFORCE++ the tool is *kept* for the whole run, yet 72–81% of the calls it keeps are decorative and accuracy is lower than in the arms that dropped the tool. Durability and selectivity come apart, and our strongest attempt to make the crop necessary did not make it so (Fig. D.1).

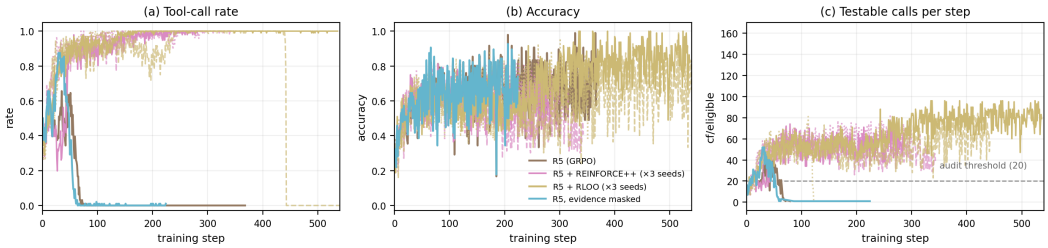


Figure D.1: **Surviving the audit is not passing it.** The tool-necessary pool, all arms on the calibrated 0.25M px budget. (a) GRPO and its evidence-masked variant collapse (steps 68 and 60); swapping only the advantage estimator keeps the tool at full call rate for 288–536 steps (one RLOO seed collapses at 443, the rest hold to their end). (b) Accuracy separates the arms in the other direction: the three REINFORCE++ seeds end at 0.41–0.67, below the arms that dropped the tool, while RLOO's best seed ends highest (0.84) despite calling on three of four questions. (c) The audit's own budget: once an arm stops calling, testable calls fall to zero, so the decorative/causal shares of Table D.1 are quoted only over the window that clears the threshold.

What the wrapper cost does and does not isolate

Across the four conditions of Table 3 the image pair is held fixed and only its presentation changes, which isolates the cost of the exchange from the value of the crop. It does not further decompose that cost. The `two` and `tool` conditions differ in more than the wrapper: they differ in serialization (a `tool_call`/`tool_response` schema with an action field versus two ordinary images), in how many turns the model may take, in the token budget available to the continuation, and in how familiar the format is to a model that was never trained on it. The base models in Table 3 are untrained, so format unfamiliarity is a live alternative to the mechanism we identify (the model continuing to emit calls instead of answering).

Separating them needs controls we did not run: a fixed-turn wrapper that forbids further calls, a plain textual multi-turn condition with no action schema, matched maximum generation budgets, and a wrapper-format adaptation baseline that fine-tunes on the schema alone. What the reported numbers establish is that *this* serialization of the additive interface costs the untrained 3B base model 16–52 points on the same image pair across our budget sweeps, and that the loop behaviour is the

budget	arm	call$_1$	peak	extinct	decorative	causal	testable
1M	GRPO	0.23	0.32	≤ 15	1.00 (step 1)	—	0
0.25M	GRPO	0.39	0.69	68	0.80	0.20	32
0.25M	+REINFORCE++	0.30–0.39	1.00	**none**	0.72–0.81	0.19–0.28	262–331
0.25M	+RLOO	0.30–0.39	1.00	**none** (2 of 3)	0.71–0.78	0.22–0.29	105–526
0.25M, masked	GRPO	0.50	0.88	60	0.82	0.18	42

Table D.1: **The tool-necessary arm.** *call*$_1$ is the call rate at step 1; *extinct* is the first step after which the call rate stays ≤ 0.05; *testable* counts the steps with ≥ 20 testable calls, and decorative and causal shares are eligible-weighted over those steps, the threshold below which the audit is noise (Appendix C). At 1M px the crop carries nothing (every correct call is decorative) and the audit never becomes measurable. At the calibrated 0.25M px it certifies a load-bearing minority and the policy *expands* tool use—then loses it anyway, whether or not the evidence cell is masked. Substituting a batch-level or leave-one-out baseline keeps the tool instead, for 288–536 steps; the causal share still never rises above 0.29, so durability and selectivity come apart. One RLOO seed is also the most accurate arm on this pool (0.84 end accuracy, above every tool-free arm) while still calling on three of four questions without the crop changing the answer; the other two end at 0.71 and 0.74.

mechanism visible in the trajectories (5). Whether a schema-adapted or turn-limited variant of the same interface would cost less is open, and would bound the wrapper cost from below rather than remove it: the RL-trained 3B, which was trained on this schema, still pays 8.9 points on V* where the untrained base pays 39.8.

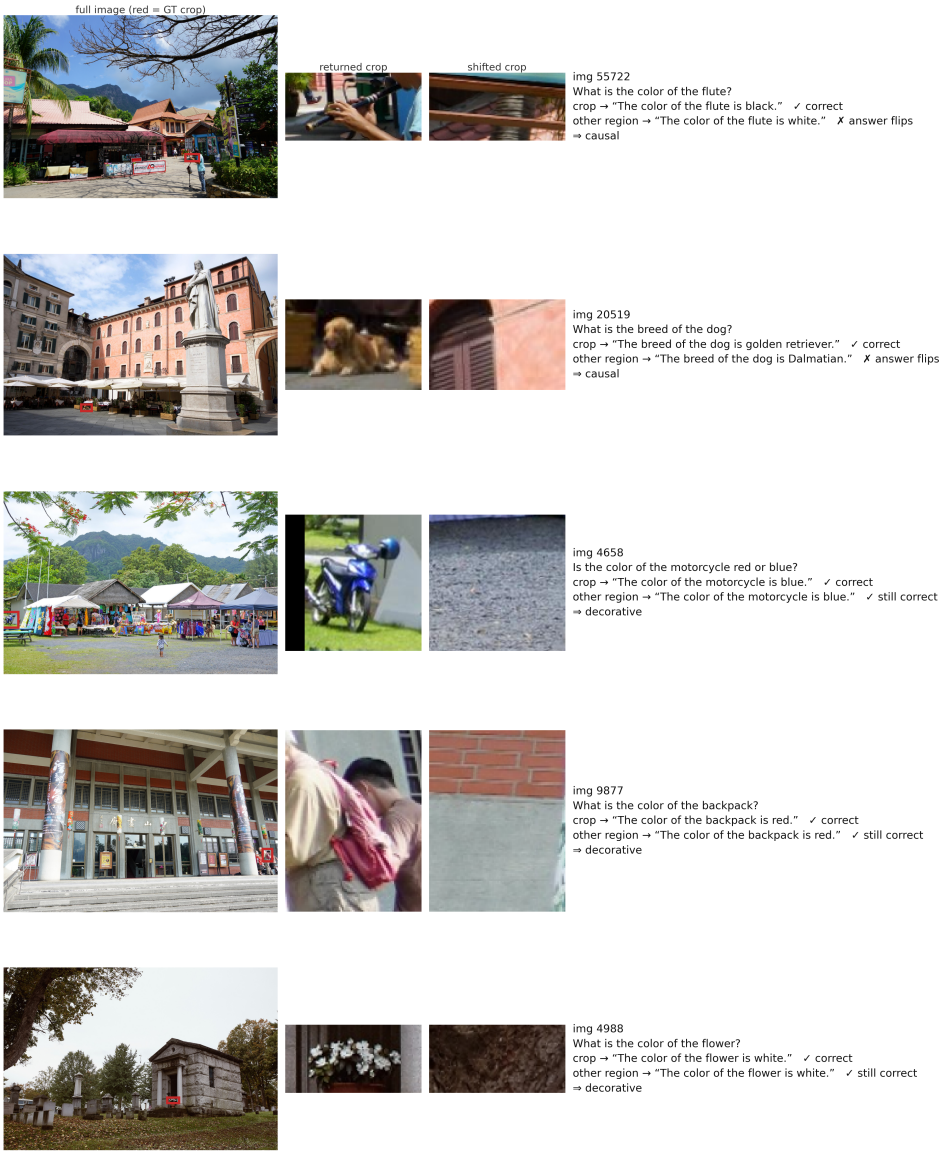


Figure D.2: **Decorative vs. causal calls under the counterfactual ablation** (V*, ckpt40, $S_{\text{tool}} = 0.036$). Each row shows the full image with the ground-truth crop region (red box), the returned crop, and a same-size *shifted* crop used in its place. In the first and second rows the shifted crop flips a correct answer (causal calls); in the third-fifth rows the answer is unchanged (decorative calls).