BEYOND PIXELS: WHAT BLIND REPLACEMENT MEASURES IN VISUAL AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual agents gather evidence through crops and other intermediate observations. A common process score compares a judge's belief under a real image and a blind reference, then interprets the contrast as visual credit. We show that this interpretation mixes answer parameterization with the structure of a crop turn. On V*-Bench, annotated crops improve answer accuracy over matched distractors by 8.9–10.5 points, yet the tagged-sequence score places them below black replacements. Across 32 serialization, path, and judge settings, structural increments exceed pixel increments in 27 cases; their median magnitudes are 4.72 and 0.31 nats. We separate two choices hidden by the original score: the belief target and the credited part of the turn. Under a shared GRPO setup, option-posterior turn credit reaches 52.5% and 48.0% exact match on V*-Bench and HR-Bench-4K. Matching total reward mass retains nearly all of this performance, whereas uniform or shuffled placement does not. Pixel-level credit yields the strongest target retrieval, and its advantage remains after controlling for crop area. Blind replacement is therefore an operator-specific comparison, not a self-defining measure of visual credit.

1 INTRODUCTION

Vision-language agents often act before answering. They zoom, crop, and inspect returned images (Wu & Xie, 2024; Liu et al., 2024b; Wang et al., 2025). A final-answer reward supervises the complete trajectory; process rewards try to identify which intermediate turns helped. If that credit responds to the syntax of a crop request rather than the evidence it retrieves, training can reinforce the appearance of visual search instead of visual search itself.

This construction is simple, but its interpretation depends on the components of a crop turn. Figure 1 illustrates the central ambiguity. A crop turn simultaneously adds action text, creates a returned-image message, and supplies pixels; its whole-turn belief gain therefore mixes all three. Replacing the crop with a black, noisy, blurred, or mean-color image targets the visual component. The contrast uses outcome labels in place of step-level annotations. Its form resembles leave-one-out attribution in text (DeYoung et al., 2020; Elazar et al., 2021) and occlusion in vision (Zeiler & Fergus, 2014; Ribeiro et al., 2016; Fong & Vedaldi, 2017).

The measurement therefore depends on how visual absence is represented. Deletion changes the conversation structure. Replacement preserves the image slot and supplies reference content. Image dimensions can also change the token budget of dynamic-resolution encoders (Liu et al., 2024a; Bai et al., 2025; Chen et al., 2024). Each operation defines a distinct comparison. Our starting question is therefore simple: *when a blind-replacement score changes, what exactly received credit?*

We study crop-tool VQA agents on V*-Bench and HR-Bench-4K. The main agents and judges are Qwen2.5-VL-3B/7B-Instruct. Turn-0 replays add SmolVLM2 as a second model family. Gold crops outperform matched same-image distractors by 10.5 and 8.9 percentage points for Qwen and SmolVLM2, but their raw real-versus-black scores fall below the black reference. Across the tested serialization, path, and judge settings, structural increments exceed pixel increments in $27/32$ cases. Their median magnitudes are 4.72 and 0.31 nats.

The interventions expose two design decisions. The *belief target* selects the answer event: a tagged sequence, the answer token, or the option posterior. The *intervention scope* selects the credited

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.

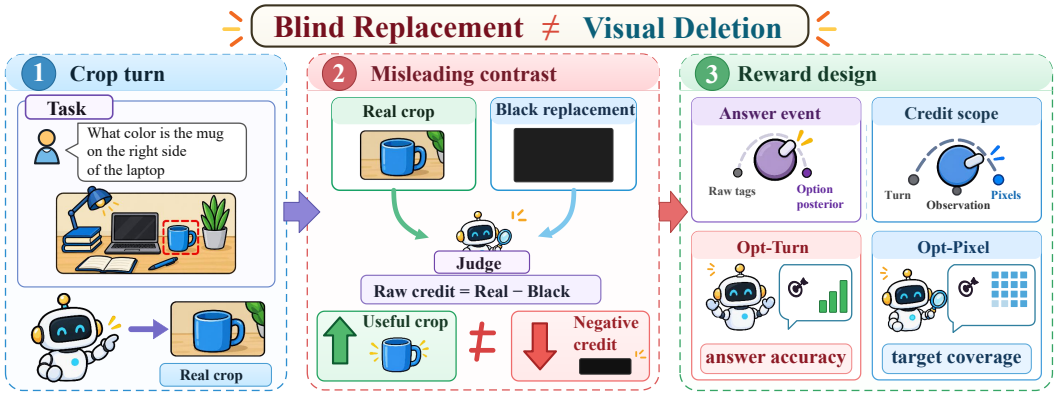


Figure 1: **A crop turn combines structural and visual evidence.** A crop turn changes action text, message-slot structure, and pixels at once, so its whole-turn belief gain combines all three. Blind replacement defines an operator-specific visual comparison whose sign varies with the reference. On Qwen2.5-VL-3B/V*-Bench first crop turns, the real-versus-black increment is 2.9% of the gain along the delete→black→real path.

object: the complete turn, the returned observation, or its pixels. This view yields two predictions. Content-focused beliefs should align more closely with answer quality. With the belief target fixed, narrower intervention scopes should favor target-containing crops. The GRPO study exhibits both patterns: Opt-Turn gives the best final-answer accuracy, whereas Opt-Pixel gives the highest target coverage.

Our contributions are threefold. First, we separate crop utility from blind-reference scoring: annotated crops improve answers over target-free distractors even when their raw scores fall below black replacements. Second, we identify two choices hidden by a single process score, the belief target and the credited part of the turn. Third, we connect these choices to optimization. Reward-matched training isolates the value of turn alignment, while crop geometry shows that pixel-level credit retrieves targets more precisely without enlarging the crop.

2 Related Work

Process rewards and step-level credit. Process reward models score solution prefixes with learned or model-based verifiers (Cobbe et al., 2021; Lightman et al., 2024; Uesato et al., 2022), with scalable labels and step-error evaluation developed in Wang et al. (2024); Setlur et al. (2025); Zheng et al. (2025). Interactive-agent methods refine outcome-only RL (DeepSeek-AI, 2025; Jin et al., 2025) through belief deviation, message information, or separated perception and reasoning signals (Zou et al., 2026b;a; Yoon et al., 2026a). Visual objectives also reweight rollouts, tokens, or teacher views by fine-grained visual dependence (Yoon et al., 2026b; Liu et al., 2026; Aniri et al., 2026; Li et al., 2026). Optimizing any learned score raises the familiar risk of reward gaming (Skalse et al., 2022; Gao et al., 2023); our concern is whether the score names the component it actually rewards.

Blind-state objectives target different estimands. VIG compares with-image and without-image next-token predictions (Luo et al., 2026). VCSD uses a same-size content-erased image and contrasts full next-token distributions (Liang et al., 2026). VIGIL masks text–vision attention to create an internal blind state (Xiao et al., 2026). Our experiments hold the recorded trajectory and gold-answer target fixed while varying the message-level comparison and answer belief. These methods instantiate particular blind operators or belief targets. Our focus is the interaction between the two: changing either axis changes the estimand, and the resulting cells produce different learned behaviors under a shared training setup.

Visual search and zoom agents. Tool-using agents interleave reasoning with external actions (Yao et al., 2023; Schick et al., 2023; Yang et al., 2023). Zoom/crop agents make perception an

explicit sequence (Wu & Xie, 2024; Liu et al., 2024b; Hu et al., 2024; Zheng et al., 2026), while high-resolution benchmarks measure the gap they close (Wang et al., 2025). Scoring whether a crop helped is where blind-state estimation enters; our results show that this score can respond to discourse structure as well as visual evidence.

Information gain as an intrinsic signal. Expected information gain is commonly used as an exploration bonus under learned world models with well-defined state variables (Houthooft et al., 2016; Pathak et al., 2017; Bellemare et al., 2016). Here, the "state" is the judge's belief over an answer string. Message-level interventions update this belief while also changing the conversation structure.

Faithfulness and rationale evaluation. Faithfulness work asks whether an explanation reflects the computation it describes (Jain & Wallace, 2019; Serrano & Smith, 2019; Wiegreffe & Pinter, 2019; DeYoung et al., 2020; Jacovi & Goldberg, 2020). Visual attribution often uses occlusion (Zeiler & Fergus, 2014; Ribeiro et al., 2016). Agent histories add a complication: removing an observation also changes message structure, patch budgets, action tags, and image slots. Feature-removal methods are known to be sensitive to this distribution shift (Hooker et al., 2019). Our answer analysis also connects to MCQ selection bias (Zheng et al., 2024) and VLM reliance tests (Li et al., 2023).

3 Measurement Setting and Evaluation Design

3.1 Agent, judge, and belief

The agent is Qwen2.5-VL-3B/7B-Instruct (Bai et al., 2025) with a crop tool and an eight-turn limit. At each turn, it either requests a crop or answers. A crop request has the form <box>$x_1\ y_1\ x_2\ y_2$</box>. The environment then returns the cropped image. An answer begins with <answer> and contains an option letter. Appendix A gives the exact system prompt and message templates. Turn-0 replays also use SmolVLM2-2.2B-Instruct (Marafioti et al., 2025) as a second VLM family. V*-Bench contains 191 multiple-choice questions: 107 have four options and 84 have two (Wu & Xie, 2024). We sample four rollouts per question at temperature 1. This yields 764 trajectories and 715/730 crop turns at 3B/7B. We repeat the main analysis on 200 questions from HR-Bench-4K (Wang et al., 2025) (§4.3). The diagnostic trajectories have per-rollout accuracy of 26.8% at 3B and 38.7% at 7B (chance: 36.0%). The end-to-end training study starts from a separate 3B checkpoint with 46.1% EM on V*-Bench and 42.0% on HR-Bench-4K (Section 4.2). The judge (same model as the agent) is teacher-forced: for trajectory prefix H_t,

$$B_t \;=\; \log p_M\big(\langle\text{answer}\rangle\, y^*\, \langle/\text{answer}\rangle \;\big|\; H_t\big), \tag{1}$$

where y^* is the gold option letter and M the judge. Under Qwen's tokenizer, the tagged string has six tokens. The tag's closing bracket is merged with the letter. Exactly one token distinguishes the answer options; the other five encode the surrounding answer tag. The deployed B mixes these components. The known letter position gives two content-based beliefs from the same forward pass. B^{ans} is the log-probability of the merged letter token alone. B^{opt} is the option posterior $\log \frac{p(>y^*)}{\sum_{X\in\mathcal{O}} p(>X)}$ over the bracket-merged option tokens. All three beliefs condition on the gold option and serve as training-time or offline scores. Primary results use the deployed B. Section 4.3 compares both content-based definitions and checks letter-position priors.

3.2 Whole-turn gain and blind-replacement contrast

The *whole-turn belief gain* is $\Delta_t^{\text{turn}} = B_t - B_{t-1}$. Here, B_{t-1} conditions on the history before the action and returned observation at turn t. For blind mode $m \in \{\text{black}, \text{noise}, \text{blur}, \text{mean}\}$, we re-render the returned image at the same pixel dimensions. This preserves `image_grid_thw` and the patch count under Qwen2.5-VL. We define the operator-indexed *blind-replacement contrast*

$$C_t^m = B_t - B_t^m. \tag{2}$$

The target construct is the effect of visual content with all non-visual features fixed. C_t^m is the observed difference induced by operator m. We call the remaining whole-turn gain $D_t^m = \Delta_t^{\text{turn}} -$

C_t^m the *remainder*. The two quantities have different scopes. Δ_t^{turn} crosses the action/observation boundary and credits both the search action and its returned observation. In contrast, C_t^m compares two versions of the returned image message. The *delete* operator makes this boundary explicit. B_t^{del} retains the assistant's action text but removes the returned user image message. Along this path, $B_t - B_t^{\text{del}}$ measures the *observation channel* (§4.1).

3.3 Four operators, three structural dials

Figure 2 summarizes the comparison space. It organizes the interventions by the structural features they preserve: action text, image-slot existence, vision-token budget, and pixels. Blind replacement is one point in this space.

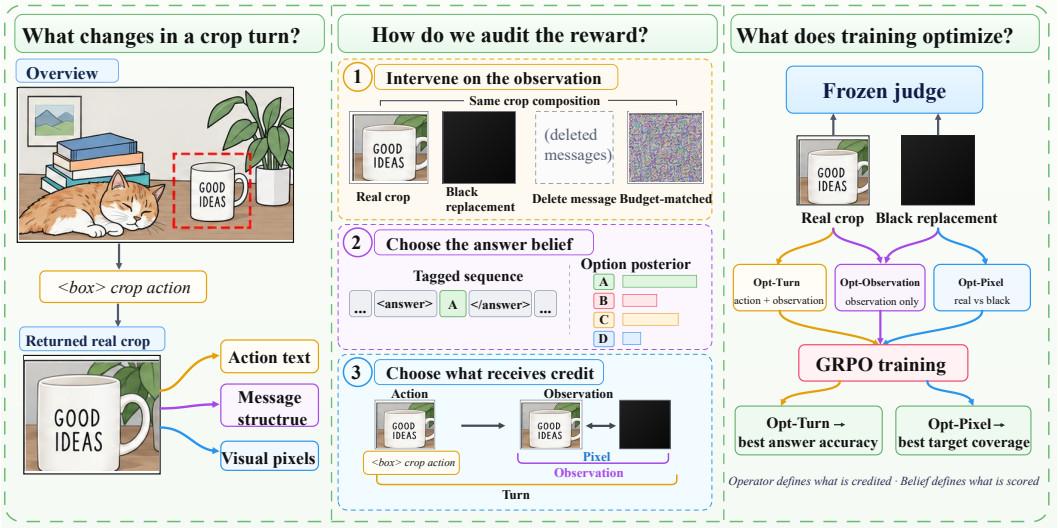


Figure 2: **What changes under each blind-image operation.** For each recorded trajectory, a fixed judge M scores the same gold answer after the returned observation is deleted or replaced. The matrix marks changes to action text, image-slot existence, vision-token budget, and pixels.

We compare it with three alternatives:

- **Replace** (above): preserves message structure and patch budget; injects content.
- **Delete**: removes the user message carrying the returned crop while keeping the assistant's box text, the message-space analogue of leave-one-out; changes message structure and budget, and fixes the action/observation boundary (§4.1).
- **Budget-match canvas**: pastes each crop onto a canvas so that all compared images share one patch budget; decouples content from budget.
- **Text ladder**: adds the gold-region message in pieces (template sentence, permuted box, true box, and true image), separating tag existence, coordinate content, and image content.

Evaluation setup. We apply all image operations to the agent's own trajectories. We also replay each question's annotated gold region at turn 0, following the benchmarks' annotated-region setup (Wu & Xie, 2024; Wang et al., 2025). Each replay combines a fixed template, the gold box, and the gold crop. We score the resulting observation. All 191 gold-region turn-0 messages share the template; only the box coordinates vary, a property the ladder exploits. For 190 V*-Bench questions, we compare the gold crop with three alternatives: a target-free crop from the same image, a crop from another question, and a black image. The conditions share action text, image-slot structure, dimensions, and patch budget. We measure forced-answer accuracy, B^{opt} gain, and the real-versus-black contrast. We render each gold crop with eight action formats: empty action, true/random/zero <box> forms, plain coordinates, a progress sentence, a natural crop request, and JSON. Each format is paired with deletion, a black image, and the real observation. A 2×3 comparison uses empty and

true-box actions with the same observation conditions. We evaluate both orders of adding the action and observation.

Aggregation. We average eligible crop turns within each question and give each question equal weight. Appendix B gives the implementation details. We repeat the analysis with two content-based belief definitions (§4.3).

4 RESULTS

4.1 BLIND REPLACEMENT MIXES STRUCTURAL AND VISUAL EFFECTS

Figure 3 shows the central mismatch. Most of the tagged-sequence change lies outside real-versus-black image content, and high whole-turn scores remain common when the final answer is wrong. We next test whether the crops themselves remain useful. Appendix Table 7a gives the exact values for all four blind operators and the corresponding remainder shares.

Table 1 separates crop utility from the chosen blind reference. Gold crops raise forced-answer accuracy over same-image distractors by 10.53 percentage points for Qwen and 8.95 points for SmolVLM2. The same gold crops have mean real-versus-black contrasts of -0.492 and -1.462 nats. Thus the utility control and the blind comparison answer different questions: the first tests whether the crop helps, while the second tests where the score places that crop relative to black. The raw score ranks gold above a matched distractor on 59.21% of Qwen pairs and 39.21% of SmolVLM2 pairs. With B^{opt}, these rates rise to 65.79% and 57.37%.

The original 191-question gold-region replay yields the same raw sign reversal: the black-replacement contrast is -0.483 nats. A 2×2 comparison varies content and resolution. Content is either gold or model-selected; resolution is either native (111/204 px) or rescaled (252/112 px). The four contrasts are -0.483 for native/gold, -0.831 for rescaled/gold, $+0.193$ for native/model,

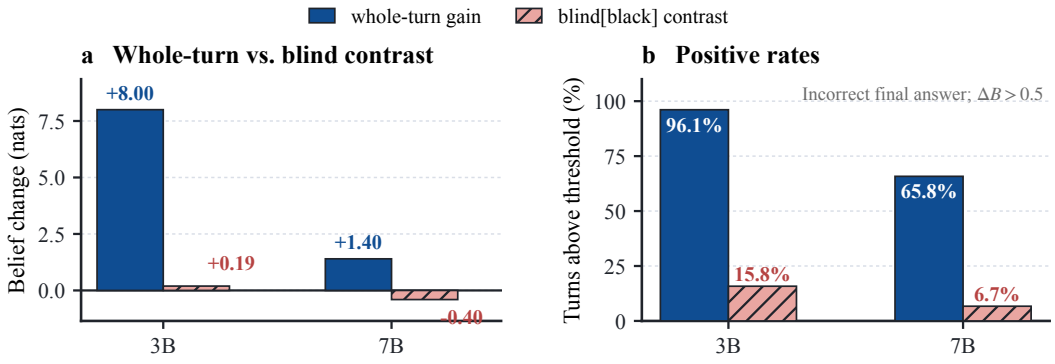


Figure 3: **Whole-turn gain and blind-replacement contrast.** Bars show mean per-crop-turn changes in the judge's gold-answer belief on V*-Bench. The whole-turn gain is positive at both Qwen2.5-VL scales. The black-replacement contrast is near zero at 3B and changes sign at 7B. Panel (b) reports the percentage of crop turns in trajectories with incorrect final answers whose score exceeds 0.5 nats.

Table 1: Gold crops versus matched distractors on 190 V*-Bench questions. Accuracy differences are in percentage points (pp); belief changes are in nats. Ranking compares the gold crop with same-image and cross-question distractors (ties count as 0.5).

judge	gold−distractor accuracy (pp)	raw real−black	B^{opt} gain	B^{opt} ranking (%)
Qwen2.5-VL-3B	+10.53	−0.492	+0.141	65.79
SmolVLM2-2.2B	+8.95	−1.462	+0.060	57.37

Table 2: Blind-replacement contrasts in nats under patch-budget equalization. Rows vary crop resolution and content; columns specify the blind operator. All canvas conditions use 81 patches.

protocol	black	noise	blur	mean
native gold region	-0.483	$+0.452$	$+0.408$	$+0.060$
rescaled gold region	-0.831	$+0.164$	$+0.139$	-0.379
canvas (equalized)	$\mathbf{-1.192}$	-0.098	-0.323	-0.664
7B canvas	-0.714	-0.814	-0.664	-0.734

and $+0.191$ for rescaled/model. The two changes act in different directions. Replacing a gold crop with a model-selected crop at matched resolution shifts the contrast by $+0.674$, whereas rescaling the gold crop shifts it by -0.348. Gold-region evaluation therefore combines content selection with a resolution change.

Equalizing patch budgets preserves the sign ordering and shifts the black contrast from -0.483 to -1.192 nats (Table 2). Across 164 upscaled crops, blind[black] falls 0.469 nats per log-unit of the blind image's patch budget. This is roughly 0.33 nats per budget doubling. The slopes are -0.521 for mean, -0.147 for noise, and -0.255 for blur (Appendix E). Under the canvas protocol, the four blind-replacement contrasts range from -1.192 to -0.098 nats. The matched budget places real and blind images on the same vision-token scale. Unlike text deletion, visual replacement introduces a resolution choice. In the tested range, that choice systematically changes the contrast.

Figure 4a localizes most of the gold-region replay gain to the action tag; correcting the coordinates and adding the gold image contribute little. Appendix Table 7b gives the full sequence of states and increments.

The ladder fixes one interaction template. We next compare delete, black, and real observations under eight action serializations (Appendix Table 6). For Qwen, the raw real-versus-black contrast changes sign. It ranges from -0.765 for a progress sentence to $+0.392$ for a natural crop request. SmolVLM2's eight contrasts range from -2.435 to -1.401 nats. The score depends on both action format and model family.

Coordinate correctness has little effect within the matched <box> format. True and random boxes yield raw contrasts of -0.463 versus -0.456 for Qwen. For SmolVLM2, the values are -1.490 and -1.487. Their free-generation accuracies differ by at most 2.1 percentage points. Under B^{opt}, Qwen's eight real-versus-black contrasts shrink to $+0.036$–$+0.060$; SmolVLM2 ranges from -0.010 to $+0.101$. Variation beyond tag presence comes from serialization and belief definition.

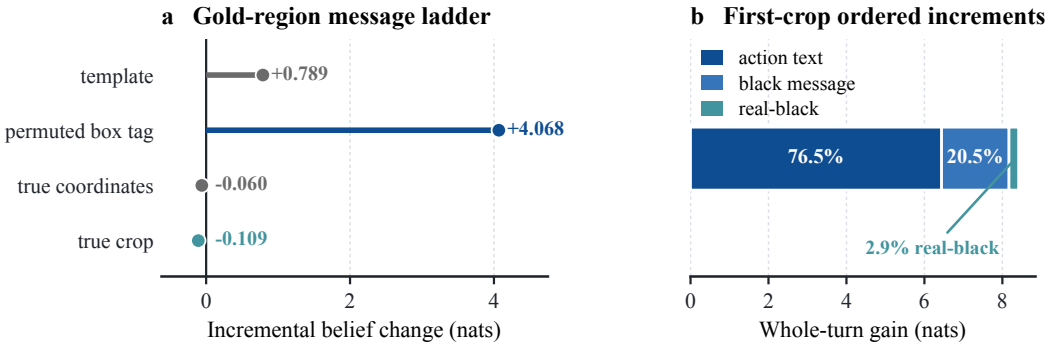


Figure 4: **Action and message structure account for most of the raw score. (a)** On 191 gold-region turn-0 questions, adding a permuted <box> action tag increases gold-answer belief by 4.068 nats. Correct coordinates and the true crop add approximately zero. **(b)** On 635 first crop turns from 3B trajectories, the specified delete→black→real path assigns 76.5%, 20.5%, and 2.9% of the 8.416-nat gain to action text, the black-image message, and the real image.

Figure 4b separates action text, image-message creation, and real-versus-black content along one ordered path. The first two account for most of the gain on that path, while the pixel increment is small. Appendix Table 7c gives the exact decomposition. We reverse the order to expose non-additive interactions rather than treat either path as a unique allocation.

For Qwen, the raw action increment is $+5.978$ when the action is added first and $+0.597$ when it is added last. The corresponding pixel increments are -0.463 and -0.691 nats. For SmolVLM2, the action increments are $+8.458$ and $+4.244$. Its pixel increments are -1.490 and -2.435. The decomposition is path-dependent; Appendix C gives the full comparison. Across eight serializations, two addition orders, and two Qwen judges, structural magnitude exceeds pixel magnitude in $27/32$ settings (84.4%). The median structural and pixel magnitudes are 4.72 and 0.31 nats, a median difference of 4.36 nats. The whole-turn gain is also nearly unchanged between eventually correct and incorrect trajectories ($+8.76$ vs. $+8.41$ nats). These results make the broad pattern clear without assigning a unique percentage to every path. Deletion and replacement define distinct baselines. On gold-region replays, black replacement raises gold belief by 0.371 nats relative to deletion. The image message therefore contributes independently of its pixels.

4.2 Belief target and intervention scope shape learned behavior

Having varied the intervention scope, we now vary the other axis: the belief target. We use a gold-conditioned best-of-four diagnostic to test training-time and offline score alignment (Wang et al., 2023). The deployed whole-turn and blind-replacement scores stay within 0.026 exact match of random. With B^{opt}, the whole-turn score improves over random by $+0.085$ at 3B and $+0.080$ at 7B. Blind variants gain $+0.045$–$+0.090$, depending on the operator and model. Final-state gold-answer belief gains $+0.302$ and $+0.136$ (Table 9, App. D). The endpoint score ranks completed trajectories. Process scores rank trajectories by accumulated intermediate support. Option normalization aligns both rankings more closely with answer content.

The diagnostics predict that changing either axis should change what the policy learns. We examine this connection by training Qwen2.5-VL-3B-Instruct with GRPO in the DeepEyes/VeRL multi-turn environment (Zheng et al., 2026). All reward variants use the same data and training setup. The total reward is $r_{\mathrm{ans}} + 0.1 r_{\mathrm{fmt}} + \lambda r_{\mathrm{proc}}$, with $\lambda = 0.3$; outcome-only training sets $r_{\mathrm{proc}} = 0$. Raw-Turn uses tagged-sequence belief gain. Ans-Turn scores only the gold answer token. Opt-Turn normalizes that token over the answer options. We attach each turn-level signal to the assistant tokens generated at the corresponding crop turn. Terminal-Opt instead places one option-posterior change on the pre-answer turn. VIG-style adapts the two-pass entropy reward of Luo et al. (2026). It deletes the current returned crop while retaining the overview and earlier history, then averages the entropy difference over the following assistant tokens. Opt-Observation excludes the crop action and scores the returned observation. Opt-Pixel compares the real observation with a same-size black image. We standardize process rewards within each prompt group and clip them to $[-3, 3]$.

Table 3: End-to-end reward training tests the two design axes. Raw-Turn→Ans-Turn→Opt-Turn changes the belief target; Opt-Turn→Opt-Observation→Opt-Pixel changes the intervention scope. EM is per-rollout exact match, averaged over four temperature-1 rollouts per question. Accuracy and target coverage are percentages; the real–black gap is in points.

Reward	V* EM (%)	HR EM (%)	Target coverage	Real–black gap	Crops
Initial checkpoint	46.1	42.0	38.0	11.0	1.61
Outcome-only	49.0	44.5	40.5	12.5	1.65
VIG-style	50.7	46.0	43.3	14.9	1.70
Terminal-Opt	51.3	46.8	41.5	13.3	1.65
Raw-Turn	48.3	43.8	39.8	11.8	2.02
Ans-Turn	50.4	45.9	42.4	13.9	1.78
Opt-Observation	51.9	47.5	45.5	16.6	1.69
Opt-Turn	**52.5**	**48.0**	44.6	15.6	1.70
Opt-Pixel	51.7	47.2	**46.2**	**17.2**	1.67

Table 4: Turn alignment under matched process-reward mass. EM is reported on V*-Bench and HR-Bench-4K; mass is relative to Terminal-Opt.

Reward	V* EM	HR EM	Relative mass
Terminal-Opt	51.3	46.8	1.00
Uniform-Turn	51.4	46.9	1.01
Shuffled-Turn	50.9	46.3	0.99
Mass-Matched Opt-Turn	52.3	47.8	1.02
Opt-Turn	**52.5**	**48.0**	1.18

Table 5: Crop geometry and target retrieval. Area, IoU, and precision are normalized to the original image. Coverage is adjusted for crop area.

Reward	Area	Target IoU	Precision	Coverage (%)
Outcome-only	0.186	0.223	0.194	40.4
Opt-Turn	0.183	0.241	0.216	44.3
Opt-Observation	0.181	0.258	0.235	45.1
Opt-Pixel	0.184	**0.276**	**0.251**	**45.7**

Changing the belief target improves answer alignment. Raw-Turn, Ans-Turn, and Opt-Turn hold the turn-level credit boundary fixed. Ans-Turn reaches 50.4% EM on V*-Bench and 45.9% on HR-Bench-4K (Table 3), 2.1 points above Raw-Turn on both benchmarks. Opt-Turn reaches 52.5% and 48.0%, adding 2.1 points over Ans-Turn on both benchmarks. Removing shared tag tokens focuses the score on the answer, while normalization focuses it on the decision among valid options.

Turn alignment matters after the belief target is fixed. Terminal-Opt reaches 51.3% and 46.8%. Uniform placement gives 51.4% and 46.9%, while shuffling the same signal gives 50.9% and 46.3%. Mass-Matched Opt-Turn reaches 52.3% and 47.8%, within 0.2 points of Opt-Turn on both benchmarks (Table 4). Preserving turn alignment retains the gain after matching reward mass. Uniform or shuffled placement does not.

Narrowing the intervention scope favors target-containing crops. Opt-Turn, Opt-Observation, and Opt-Pixel hold the option-posterior target fixed. Opt-Turn gives the highest final-answer accuracy. Opt-Observation reaches 51.9% and 47.5% EM while raising target coverage to 45.5%. Opt-Pixel gives the highest target coverage (46.2%) and real–black gap (17.2 points), with 51.7% and 47.2% EM. This pattern matches the second prediction. Turn credit can reward the action and its downstream answer support. Observation- and pixel-level credit instead place more weight on retrieving crops that contain the annotated target. The intervention scope changes the observed balance between answer accuracy and target coverage rather than merely rescaling one reward.

Opt-Pixel and Opt-Turn select nearly identical crop areas (0.184 versus 0.183). Opt-Pixel improves target IoU from 0.241 to 0.276, crop precision from 0.216 to 0.251, and area-adjusted coverage from 44.3% to 45.7% (Table 5). Its target-retrieval advantage therefore does not come from larger crops.

4.3 Alternative beliefs and benchmark transfer

The deployed B mixes answer content with tag mass (Eq. 1). We compare it with the two content-based definitions from Section 3.1. Under B^{ans}, the whole-turn gain is $+0.02$ nats and blind[black] is $+0.09$. Under B^{opt}, the values are $+0.078$ and $+0.012$.

The reparameterization removes most of the raw $+8.00$-nat scale. At 3B, it compresses the four operator contrasts to $[-0.004, +0.012]$. The 7B range is $[-0.070, -0.023]$ (Appendix Table 8). The belief definition also changes channel attribution. On the gold-region ladder, B^{opt} assigns $+0.057$ to the action tag and $+0.030$ to the true image. The deployed belief assigns $+4.068$ nats to the tag step. Appendix C gives the full decomposition.

We repeat the main analysis on 200 native-4K HR-Bench questions (Wang et al., 2025). Four rollouts per question produce 798 trajectories. The 3B agent reaches 42.0% EM. Whole-turn gain is

$+6.70$ nats, whereas blind[black] is $+0.18$. Gold crops improve forced-answer accuracy over same-image distractors by 9.0 points. With B^{opt}, the gold-crop gain is $+0.10$ nats and the whole-turn selector improves best-of-four exact match by $+0.065$ over random. Appendix Table 10 gives the remaining results.

5 Discussion

What blind-state contrasts measure. The judge responds to action tags, returned-image messages, patch budgets, pixels, and answer parameterization. A blind state therefore defines an operator-relative contrast, not an operator-independent quantity called visual credit. Structural increments exceed pixel increments in 84.4% of the tested settings, with median magnitudes of 4.72 and 0.31 nats. Option normalization reduces the four trajectory-level operator contrasts to a 0.016-nat range at 3B. The same structural elements occur in other tool-using VLMs (Chen et al., 2024; Wu & Xie, 2024).

Scope. Our experiments study gold-conditioned credit in option-based crop-tool agents. The option posterior is directly available in this setting; open-ended agents require a different answer event. The broader design principle does not depend on multiple choice: a reward must still state both what belief event it scores and which intervention defines the credited component. We view the two-axis framework as the transferable contribution, not any single blind image or answer encoding.

Reward design. Reward design requires an explicit answer event and credit boundary. The answer event may be a tagged sequence, an answer token, or a posterior over valid options. The boundary may cover the complete action–observation turn, the returned observation, or pixels relative to a named reference. Option normalization resolves the first choice. Turn-, observation-, and pixel-level rewards resolve the second. Reward-matched training shows that temporal alignment matters, while the crop-geometry results connect a narrower boundary to more precise target retrieval.

6 Conclusion

Blind replacement is not a self-defining measure of visual credit. Tagged-sequence likelihood responds strongly to action text and message structure even when the intended target is useful pixels. Reward design therefore needs two explicit choices: the belief target and the credited part of the turn. Option-posterior turn credit gives the highest final-answer accuracy. Pixel credit gives the strongest target retrieval without increasing crop area. Matching reward mass preserves the benefit of turn alignment. These outcomes reflect different reward definitions, not competing estimates of one universal visual-credit scalar.

Ethics Statement

This work studies how process rewards assign credit in crop-tool vision-language agents. A misspecified reward can favor the form of a crop action or message structure over the visual evidence it retrieves, making apparent visual search less faithful to the information used by the model. Our analysis is intended to expose this failure mode; when such scores are used to train or compare agents, the intervention operator and belief target should be stated explicitly. The experiments use open-weight models and public V*-Bench and HR-Bench-4K data, with no human subjects or personal data.

AI Use Statement

Generative AI tools were used to improve readability, reorganize sections, assist with literature organization, refine scientific figures, and provide feedback on experimental presentation and result interpretation. The authors reviewed all AI-assisted text and suggestions, independently ran the experiments, and verified the citations, figures, and numerical claims against the original sources and experimental artifacts. The authors take responsibility for the final content of this work.

REFERENCES

Aniri, Jinhe Bi, Peng Liao, Zengjie Jin, Volker Tresp, Fei Shen, Yunpu Ma, and Tat-Seng Chua. OPD-V: Visual on-policy self-distillation with modality balance. *arXiv preprint arXiv:2608.05131*, 2026.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Jiapeng Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites. In *Proceedings of CVPR*, 2024.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of ACL*, 2020.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of ICCV*, 2017.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 10835–10866, 2023.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of ACL*, 2020.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of NAACL-HLT*, 2019.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Conference on Language Modeling (COLM)*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of EMNLP*, 2023.

Yijiang Li, Yijun Liang, Yunjie Tian, Bingyang Wang, Ke Zhang, Zhenfei Yin, Di Fu, Philip Torr, and Nuno Vasconcelos. Self-supervised visual on-policy distillation. *arXiv preprint arXiv:2608.14144*, 2026.

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

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024a.

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

Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-spot: Interactive reasoning improves large vision-language models. In *Proceedings of ECCV*, 2024b.

Wen Luo, Xiaohan Yi, Xiaotao Huang, and Liqun Huang. VIG: Visual information gain as a reward signal for multimodal chain-of-thought compression. *arXiv preprint arXiv:2608.21883*, 2026.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of ICML*, 2017.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of KDD*, 2016.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of ACL*, 2019.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. Rewarding progress: Scaling automated process verifiers for LLM reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471, 2022.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of ACL*, 2024.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 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. In *International Conference on Learning Representations (ICLR)*, 2023.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of EMNLP-IJCNLP*, 2019.

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

Xi Xiao, Chen Liu, Chih-Ting Liao, Yunbei Zhang, Qizhen Lan, Yuxiang Wei, Lin Zhao, Janet Wang, Jianyang Gu, Muchao Ye, Tianyang Wang, and Hao Xu. Staying VIGILant: Mitigating visual laziness via counterfactual visual alignment in MLLMs. *arXiv preprint arXiv:2606.26387*, 2026.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-REACT: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Hee Suk Yoon, Eunseop Yoon, Ji Woo Hong, SooHwan Eom, Gwanhyeong Koo, Mark Hasegawa-Johnson, Qi Dai, Chong Luo, and Chang D. Yoo. PDCR: Perception-decomposed confidence reward for vision-language reasoning. *arXiv preprint arXiv:2605.13467*, 2026a.

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. In *International Conference on Machine Learning (ICML)*, 2026b.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*, 2014.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *International Conference on Learning Representations (ICLR)*, 2024.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, 2025. doi: 10.18653/v1/2025.acl-long.50.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2026.

Deyu Zou, Yongqiang Chen, Fan Feng, Mufei Li, Pan Li, Yu Gong, and James Cheng. On information self-locking in reinforcement learning for active reasoning of LLM agents. In *International Conference on Machine Learning (ICML)*, 2026a.

Deyu Zou, Yongqiang Chen, Jianxiang Wang, Haochen Yang, Mufei Li, Qing Da, James Cheng, Pan Li, and Yu Gong. T^3: Reducing belief deviation in reinforcement learning for active reasoning. In *International Conference on Learning Representations (ICLR)*, 2026b.

A PROMPTS AND TEMPLATES

The system prompt gives the agent eight turns. It may zoom by emitting <box>$x_1\ y_1\ x_2\ y_2$</box>, or answer by emitting <answer> followed by the option letter. The overview image is rendered at display resolution (short side 288 px); crop coordinates are mapped back to native resolution before cropping. The oracle template is "I will zoom into the relevant region to inspect the details." followed by the gold box; the sequence is defined in §3.3.

B IMPLEMENTATION DETAILS

Belief is the teacher-forced log-probability of <answer>y^*</answer> given the prefix (Eq. 1), computed in a single forward pass per condition. We use four blind variants. ***Black*** is all-zero RGB. ***Noise*** uses i.i.d. Gaussian pixels with μ=0.5 and σ=0.25, clipped to $[0, 1]$. ***Blur*** is Gaussian with radius $\max(8, \min(w, h)/20)$. ***Mean*** uses the per-image mean color. Each variant retains the replaced image's dimensions, `image_grid_thw`, and patch count. Belief levels are computed in bf16. Delete removes the returned user image message while retaining the assistant box action, matching the observation-channel boundary used in Table 7c. The canvas protocol pastes each image onto a common canvas yielding 81 patches in every condition. BoN uses per-question trajectory sums of each signal, with the random baseline given by per-question expected exact match; the oracle ceiling is the best-of-4 oracle.

End-to-end training. The policy and frozen judge are both Qwen2.5-VL-3B-Instruct. Training follows the DeepEyes/VeRL multi-turn path with GRPO, eight rollouts per question, and an eight-turn cap. We retain public DeepEyes training examples with verifiable answers, crop access, and explicit answer options, then remove overlap with V*-Bench and HR-Bench-4K. All reward variants use the same optimization setup.

For a crop turn t, Raw-Turn, Ans-Turn, and Opt-Turn use $s_t = B^q(H_t) - B^q(H_{t-1})$, with $q \in \{\text{tag}, \text{ans}, \text{opt}\}$. After group standardization and clipping, s_t is assigned to the policy tokens generated at turn t. Terminal-Opt applies the same procedure once to $B^{\text{opt}}(H_{\text{preans}}) - B^{\text{opt}}(H_0)$ and assigns it to the pre-answer turn. The history H_{preans} excludes generated answer tokens. The two variants therefore differ in where the option-posterior signal enters the multi-turn objective.

The VIG-style control keeps the recorded assistant tokens fixed and makes two scoring passes for each crop turn. The first pass includes the returned crop. The second deletes only that crop while retaining the overview, question, and earlier history. For the assistant tokens T_t that follow the observation, the scalar reward is $\text{clip}(|T_t|^{-1} \sum_{j \in T_t} [H(p_j^{\text{del}}) - H(p_j^{\text{real}})], -1, 1)$. This expression matches the token-entropy aggregation in VIG (Luo et al., 2026). We then handle the scalar like the other turn-level process rewards.

Process rewards are standardized over the eight rollouts in each prompt group and clipped to $[-3, 3]$ before they enter the GRPO advantage. We set $\lambda = 0.3$ on a held-out split of the training corpus. V*-Bench and HR-Bench-4K remain evaluation-only. End-to-end evaluation uses four temperature-1 rollouts per question and averages per-rollout exact match with equal question weight. Target coverage requires the first crop to cover at least 50% of the annotated target area. The real–black gap is the exact-match difference between matched real-crop and black-crop evaluations.

C GOLD-CROP AND ACTION-FORMAT DETAILS

For the gold-crop comparison, every condition asks the model to stop cropping and emit one option letter inside the answer tags. Decoding is greedy with a 16-token limit. The same-image distractor is a target-free window with the gold crop's dimensions. Valid target-free windows exist for 190 questions and have zero overlap with the annotated target. The remaining question lies outside this comparison. Gold improves accuracy over the same-image distractor by 10.53 percentage points for Qwen and 8.95 for SmolVLM2. The gains over the cross-question distractor are 13.16 and 10.00 points.

Table 6: Action serialization on 191 gold-region replays. Entries are mean real-versus-black contrasts in nats; BD denotes black-minus-delete.

(a) Qwen2.5-VL-3B

serialization	raw BD	raw real−black	B^{opt} BD	B^{opt} real−black
empty	5.935	−0.691	0.020	0.048
true box	0.326	−0.463	−0.023	0.056
random box	0.244	−0.456	−0.013	0.058
plain coords	−0.095	0.104	0.004	0.047
zero box	0.160	−0.462	−0.009	0.059
progress	0.693	−0.765	−0.019	0.036
natural crop	0.338	0.392	0.008	0.060
crop JSON	2.851	0.241	−0.006	0.057

(b) SmolVLM2-2.2B

serialization	raw BD	raw real−black	B^{opt} BD	B^{opt} real−black
empty	5.876	−2.435	−0.083	0.047
true box	0.718	−1.490	−0.048	0.101
random box	0.728	−1.487	−0.060	0.094
plain coords	0.298	−1.807	−0.024	0.008
zero box	1.203	−1.503	−0.036	0.092
progress	0.790	−2.434	0.004	−0.010
natural crop	1.284	−1.791	−0.030	0.016
crop JSON	−0.132	−1.401	−0.074	0.063

Table 6 compares the action serializations. Within each row, delete, black, and real share the same action text. Black and real also share dimensions and patch budgets. Cross-row variation therefore isolates serialization effects from the matched pixel comparison.

The 2×3 comparison uses the empty and true-box actions. For each belief definition, we add the action both before and after the image message. We likewise add the pixels both before and after the action. The paired results measure order dependence while allowing interactions among action, slot, and pixel effects.

The raw score ranks gold above a matched distractor on 59.21% of Qwen pairs and 39.21% of SmolVLM2 pairs. The original gold-region black contrast is -0.483 nats. In the content×resolution comparison, rescaled/gold is -0.831 and native/model is $+0.193$. The content and resolution shifts are $+0.674$ and -0.348. Across serializations, the Qwen endpoints are -0.765 and $+0.392$. The SmolVLM2 endpoints are -2.435 and -1.401. Qwen's action increment is $+5.978$ when added first and $+0.597$ when added last.

Under B^{ans}, the action-tag step is $+0.723$ and the full replay gain is $+0.36$. Under B^{opt}, the action-tag and true-image steps are $+0.057$ and $+0.030$. The B^{ans} observation increment is -0.523. On 3B first crop turns, the B^{opt} observation, message, and pixel increments are $+0.003$, -0.009, and $+0.013$.

D FULL PER-MODE TABLES

Table 7 gives the values summarized by Figures 3 and 4. Table 8 reports the option-posterior results across image operations. Table 9 gives the trajectory-selection results.

Table 7: Full values summarized by the main-text figures. **(a)** Mean per-crop-turn belief changes on V*-Bench; remainder share is $1 - C^{\text{black}}/\Delta^{\text{turn}}$. **(b)** Gold-region crop-turn components. **(c)** First-crop whole-turn decomposition on 3B V*-Bench trajectories (n=635). All changes are in nats.

(a) Blind-operator comparison

	3B (715 turns)	7B (730 turns)
whole-turn $B_t - B_{t-1}$	$+8.001$	$+1.399$
blind[black] C_t^{black}	$+0.193$	$\mathbf{-0.404}$
blind[noise] C_t^{noise}	-0.095	-0.458
blind[blur] C_t^{blur}	-0.075	-0.063
blind[mean] C_t^{mean}	-0.003	-0.433
remainder share	0.976	$\mathbf{1.289}$

(b) Gold-region message ladder

state	added content	change (nats)
B_{q}	question+overview	-10.561 (level)
B_{tpl}	+template only	$+0.789$
B_{scr}	+**permuted** box	$\mathbf{+4.068}$
B_{box}	+true box	-0.060
B_{img}	+true image	-0.109

(c) First-crop ordered increments

increment	gain (nats)	share
whole-turn	$+8.416$	100%
action-text	$+6.441$	76.5%
observation	$+1.975$	23.5%
black-message	$+1.729$	20.5%
real-versus-black	$+0.246$	**2.9%**

Table 8: Mean option-posterior belief changes on V*-Bench in nats.

	3B (715 turns)	7B (730 turns)
whole-turn $\Delta_{\text{opt}}^{\text{turn}}$	$+0.078$	$+0.119$
blind[black]	$+0.012$	-0.025
blind[noise]	$+0.002$	-0.070
blind[blur]	-0.004	-0.035
blind[mean]	$+0.011$	-0.023

Table 9: Best-of-4 exact-match gain over random (§4.2); B^{opt} normalizes over answer options.

belief	selector	3B	7B
deployed B	whole-turn	-0.022	$+0.021$
	blind[black]	$+0.009$	$+0.016$
	blind[noise]	$+0.004$	$+0.026$
	blind[blur]	$+0.009$	-0.010
	blind[mean]	-0.007	$+0.021$
B^{opt}	whole-turn	$\mathbf{+0.085}$	$+0.080$
	blind[black]	$+0.060$	$+0.070$
	blind[noise]	$+0.045$	$\mathbf{+0.090}$
	blind[blur]	$+0.050$	$+0.060$
	blind[mean]	$+0.045$	$+0.055$
final-state belief B_{final}		$\mathbf{+0.302}$	$\mathbf{+0.136}$
oracle@4 ceiling		$+0.418$	$+0.314$

E HR-BENCH DETAILS

Table 10: HR-Bench-4K corrected-score utility.

comparison or selector	gain	selected EM / ranking
gold vs. same-image distractor	$+.090$	—
raw real$-$black	$-.400$	—
B^{opt} gold-crop gain	$+.100$	63.0%
B^{opt} whole-turn BoN	$+.065$	.485
B^{opt} blind[black] BoN	$+.050$	.470
B^{opt} blind[noise] BoN	$+.075$	.495
final-state belief	$+.180$	.600
oracle@4	$+.330$	.750