TOWARD COUNTERFACTUAL LEARNING FOR SELECTIVE RE-EXAMINATION IN AGENTIC VISUAL SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual search over high-resolution images requires a vision–language model (VLM) not only to determine *where* to inspect, but also *whether* another observation is warranted before committing to an answer. Existing prompting, demonstrations, and outcome-based reinforcement learning can encourage re-examination, yet do not directly supervise whether it is beneficial at a particular search state. We introduce RESEE, a benchmark of re-examination decisions comprising $3{,}227$ visual-search trajectories and $8{,}827$ explicit commit-or-re-examine decisions, with counterfactual adjudication that distinguishes states where additional inspection is warranted from those where commitment is preferable. Building on this formulation, we propose CARE (**C**ounterfactual **A**dvantage for **R**e-**E**xamination), a decision-level reinforcement learning method that compares paired continuations under commit and re-examine verdicts from the same search state. Their cost-aware return difference estimates the counterfactual value of an additional observation and is assigned only to the corresponding verdict tokens, providing localized credit for whether search should continue. Experiments expose limited re-examination selectivity in existing VLMs and show that CARE significantly improves accuracy and decision selectivity over GRPO across two backbones, with higher accuracy and fewer visual requests on six unseen OOD benchmarks.

1 INTRODUCTION

Visual search is a core capability for vision–language models (VLMs) operating on real images, where the evidence required to answer a question may occupy only a small fraction of a high-resolution frame (Zhang et al., 2025). Recent work therefore formulates visual understanding as a multi-step interaction in which the model selects a region, observes a crop, updates its belief, and decides where to inspect next (Zheng et al., 2026; Lai et al., 2026). This interaction introduces a trade-off. Committing too early can lock in errors before sufficient evidence has been gathered, whereas unnecessary inspection consumes a finite interaction budget. Agentic VLMs now support crop, zoom, and backtrack actions within a single episode (Wang et al., 2026a; Sarch et al., 2026), making this trade-off central to long-horizon visual search. Beyond learning *where* to look, a visual search policy must therefore also determine *whether an additional observation is warranted*.

Existing visual search agents are largely optimized through trajectory-level outcomes (Zheng et al., 2026; Su et al., 2026; Lai et al., 2026). Reinforcement learning with answer-based rewards can encourage a policy to locate informative regions, interpret their content, and ultimately produce a correct, grounded answer. Such supervision, however, does not identify the marginal value of an individual observation. A trajectory in which re-examination corrects an erroneous answer and one in which it merely confirms an already-correct answer may receive the same final reward. Prompted verification (Madaan et al., 2023; Huang et al., 2024), explicit rewards for additional checking (Su et al., 2026), and demonstrations that include re-examination can all encourage the behavior itself, but none directly supervises whether re-examination is warranted at a particular search state (Figure 1a–c). Consequently, the ability to re-examine does not imply the ability to decide *whether* re-examination is warranted at the current search state (Figure 1, left).

This limitation points to a more precise learning target: the state-dependent value of re-examination. At each search state, the desired signal should quantify the benefit of acquiring an additional observation relative to committing with the evidence already available. Trajectory-level answer rewards can-

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.

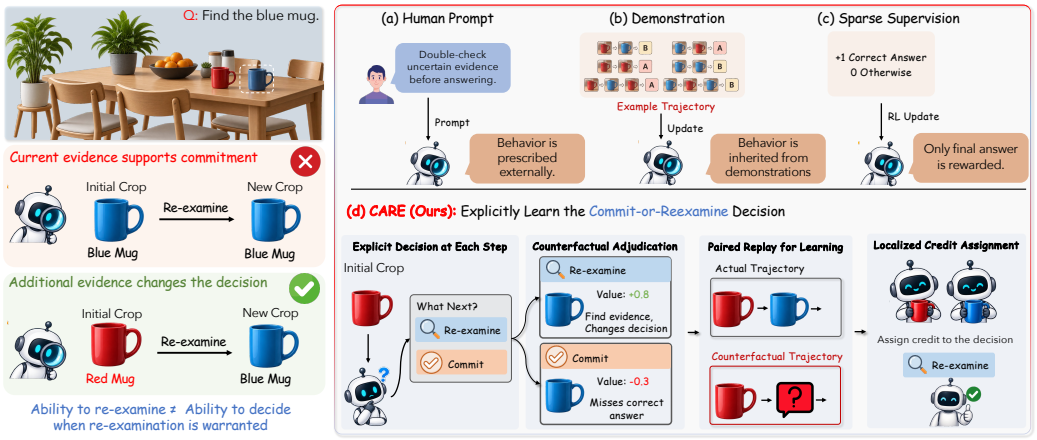


Figure 1: Re-examination as an explicit decision. The left panel contrasts warranted and unwarranted re-examination. The right panel shows how CARE evaluates the commit or re-examine verdict through paired continuations and credits that verdict directly.

not provide this supervision (Kazemnejad et al., 2024; Feng et al., 2026a), because re-examinations with different counterfactual utility may nevertheless receive the same final reward. Nor can the policy's own confidence or uncertainty directly supply this value, as such signals describe its current belief rather than the counterfactual effect of observing additional evidence. Learning this decision therefore requires both an explicit representation of the commit-or-re-examine choice and supervision that attributes the consequence of an additional observation directly to that choice.

To address this gap, we introduce RESEE, a benchmark of explicit commit-or-reexamine decisions, and CARE, a learning method for assigning decision-level credit to re-examination (Figure 1d). RESEE contains trajectories in which a conservatively searching VLM records a commit-or-reexamine verdict at each decision state, exposing both decisions to continue searching and decisions to commit. Rather than treating the model's verdict as supervision, we counterfactually adjudicate each state by comparing the consequences of re-examination and immediate commitment, yielding cases where additional inspection is beneficial and where commitment is preferable. Building on these decisions, CARE constructs paired continuations under the two verdicts from the same search state and uses their cost-adjusted return difference to estimate the value of re-examination. This advantage is assigned only to the verdict tokens that determine whether search continues, rather than to the subsequent trajectory, providing localized credit for commit-or-reexamine decision. Contributions:

- We formulate re-examination in long-horizon visual search as a decision-level credit-assignment problem, with the learning target defined as the counterfactual value of an additional observation relative to immediate commitment.
- We introduce RESEE, a benchmark of re-examination decisions comprising 3,227 visual-search trajectories and 8,827 explicit commit-or-re-examine decisions, including 1,862 counterfactually adjudicated cases. Its positive and negative cases separate policy self-assessment from the measured utility of re-examination.
- We propose CARE, a decision-level RL method that constructs paired continuations under opposite verdicts and assigns their cost-aware return difference only to the corresponding verdict tokens, providing localized credit for the commit-or-re-examine decision.
- Experiments on RESEE expose limited decision selectivity in existing VLMs. Across two backbones, CARE improves macro accuracy and balanced decision accuracy over GRPO with fewer visual requests. Further analyses show accuracy gains on six unseen OOD benchmarks and across a range of model capacities.

2 RELATED WORK

Long-Horizon Visual Search. Long-horizon visual search extends visual question answering beyond single-view inference by acquiring evidence through sequential interaction. Early approaches guide frozen VLMs through targeted or tree-structured exploration (Wu & Xie, 2024; Shen et al.,

2025), while grounded chain-of-thought supervision associates intermediate reasoning with relevant image regions (Shao et al., 2024a). More recent agentic VLMs learn the search process itself, using multi-turn crop or zoom actions (Zheng et al., 2026; Su et al., 2026; Lai et al., 2026), grounding reasoning in image coordinates or traceable evidence regions (Sarch et al., 2026; Wang et al., 2026a), or delegating localization to dedicated search agents (Li et al., 2025). Many such systems optimize trajectory-level correctness with verifiable outcome rewards and group-relative reinforcement learning objectives. Recent analysis further suggests that crop-and-zoom training improves intrinsic reasoning more readily than it enables additional observations to correct intrinsic failures (Ma et al., 2026). Interaction efficiency is typically handled separately through fixed budgets or mechanisms that regulate computation, including length penalties and adaptive allocation based on the expected return of further reasoning (Chen et al., 2024; Lyu et al., 2025; Zhou et al., 2026). These approaches regulate search effort at the trajectory or budget level, but do not directly estimate the state-dependent value of an additional visual observation.

Self-Correction and Credit Assignment. Self-correction encourages models to verify, critique, and revise their outputs through prompting or verbal feedback (Madaan et al., 2023; Shinn et al., 2023). Without external feedback, however, intrinsic self-correction does not consistently improve accuracy and can even degrade it (Huang et al., 2024). Reinforcement learning can train corrective behavior directly (Kumar et al., 2025), while verification and backtracking may also emerge from outcome-based training (Guo et al., 2025; Gandhi et al., 2025), although their occurrence does not reveal whether an individual correction was warranted. Credit-assignment methods address this limitation with denser supervision, including process-level scores from human or rollout-derived signals (Lightman et al., 2024; Wang et al., 2024), Monte Carlo estimates from intermediate states (Kazemnejad et al., 2024), and advantages defined over turns or comparable states (Wei et al., 2025; Feng et al., 2026a). More recent work estimates the effect of tool use or individual steps through before–after comparisons and controlled counterfactual interventions (Feng et al., 2026b; Meng, 2026), while token-level methods concentrate optimization on selected positions (Wang et al., 2026b; Jin et al., 2026). In contrast to assigning credit over generic steps, turns, or tokens, our setting exposes the commit-or-re-examine choice as an explicit decision at a known search state, allowing its alternatives to be evaluated directly and credit to be localized to the decision itself.

3 THE RESEE BENCHMARK

Task Formulation. RESEE evaluates the commit-or-re-examine decision that recurs throughout long-horizon visual search. At a decision state h, the model has accumulated a search history and formed a candidate answer, but may either commit or acquire another observation. We denote this decision by $v \in \{\text{RECOVER}, \text{SKIP}\}$. The benchmark asks whether re-examination has positive state-dependent utility relative to immediate commitment:

$$\tau(h) = R_{\text{rec}}(h) - R_{\text{skip}}(h) - \lambda c(h), \tag{1}$$

where $R_{\text{rec}}(h)$ and $R_{\text{skip}}(h)$ are the rewards reached by re-examining and by committing at h, and $\lambda c(h)$ is the priced cost of the additional observation. Re-examination at h is warranted when $\tau(h) > 0$, and immediate commitment is preferable otherwise. Unlike evaluation by final answers, this directly measures whether more evidence is warranted at the decision state.

Benchmark Construction. RESEE contains 3,227 visual-search trajectories drawn from six high-resolution benchmarks: V* (Wu & Xie, 2024), HR-Bench (Wang et al., 2025b), MME-RealWorld (Zhang et al., 2025), TreeBench (Wang et al., 2026a), O3-Bench (Li et al., 2025), and VisualProbe (Lai et al., 2026). The resulting questions span attribute recognition, spatial reasoning, text reading, relational reasoning, and counting, with the target evidence occupying a median of only 0.8% of a 3.39-MP image. Across these trajectories, we record $8{,}827$ explicit decisions, including $1{,}131$ re-examinations and $7{,}696$ commitments. Each state preserves the accumulated evidence, candidate answer, remaining uncertainty, and the model's proposed verdict, and each re-examination further specifies its target region and search operation. The resulting decisions cover six recurring re-examination patterns, summarized in Figure 2 and Appendix A.2.

Counterfactual Adjudication. We counterfactually adjudicate $1{,}862$ decision states by replaying the same state under both verdicts and comparing their downstream returns. Labels therefore follow

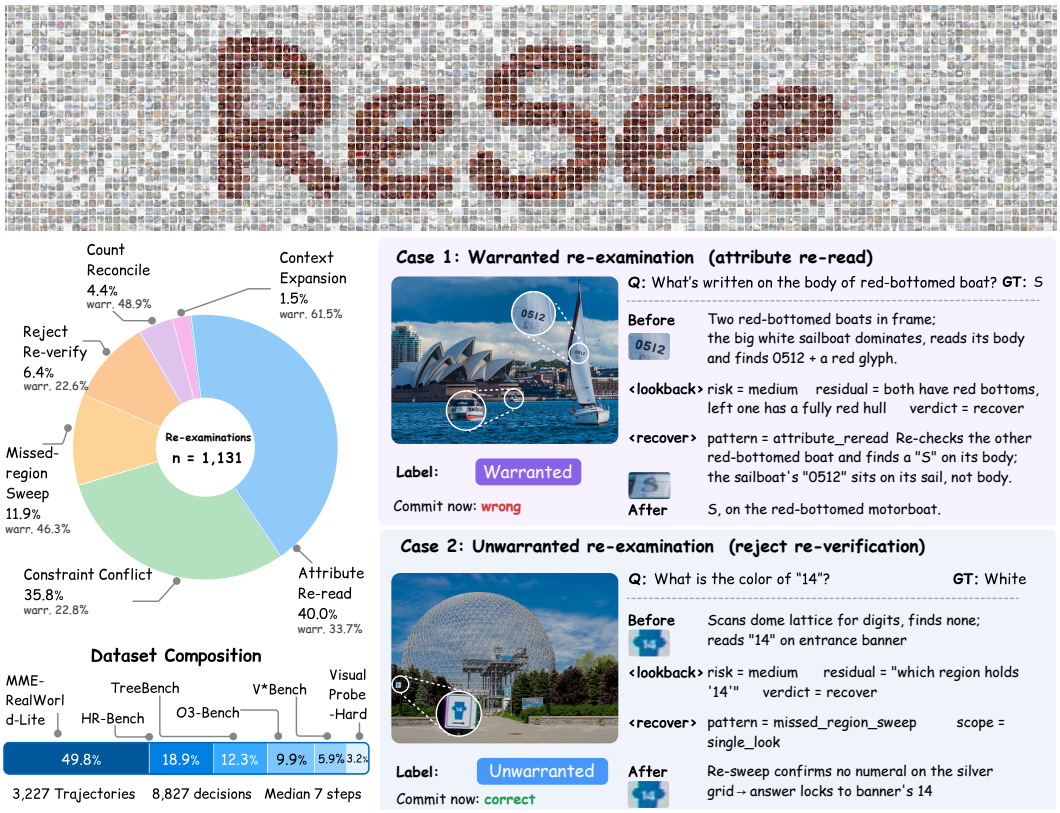


Figure 2: Overview of ReSee. The left panel shows the sources and six re-examination patterns, with their warranted rates. The right panel contrasts a warranted and an unwarranted re-examination and shows the fields recorded at each decision. Some demo samples are available at *this link*.

from the consequence of the decision rather than the model's own verdict or a subjective judgment of whether it is warranted. Among 883 adjudicated re-examinations, 279 (31.6%) carry positive utility, whereas 918 of 979 adjudicated commitments (93.8%) are safe to retain. Both classes are required to measure selectivity, since a policy that always re-examines attains perfect recall on the positive class alone. ReSee therefore evaluates not re-examination itself, but the ability to distinguish states in which further evidence is valuable from those in which commitment is preferable.

4 Methodology

Problem Formulation. Consider a visual-search rollout $\zeta = (u_1, a_1, \ldots, u_T, a_T)$ generated by a policy π, where u_t and a_t denote the observation and action at turn t. At each decision state h_j, the model has accumulated a search history and formed a candidate answer, and chooses a verdict $v_j \in \{\text{RECOVER}, \text{SKIP}\}$, corresponding to acquiring another observation or committing with the current evidence. Let $R(\zeta)$ denote the trajectory reward and $Q^\pi(h_j, v)$ the expected return after taking verdict v at h_j and following π thereafter. We define the state-dependent utility of re-examination as

$$\tau^\pi(h_j) = Q^\pi(h_j, \text{RECOVER}) - Q^\pi(h_j, \text{SKIP}) - \lambda c(h_j), \tag{2}$$

where $c(h_j)$ is the cost of the additional observation and $\lambda \geq 0$ sets its price. This policy-dependent quantity is the expected-return counterpart of the paired-outcome benchmark utility $\tau(h)$ defined in Section 3. Re-examination is warranted when $\tau^\pi(h_j) > 0$ and immediate commitment is preferable otherwise. The objective is to preserve task-level performance while selecting re-examination only at states where its expected benefit exceeds its interaction cost. Since a rollout realizes only one verdict at each state, this utility is not directly available from trajectory-level supervision.

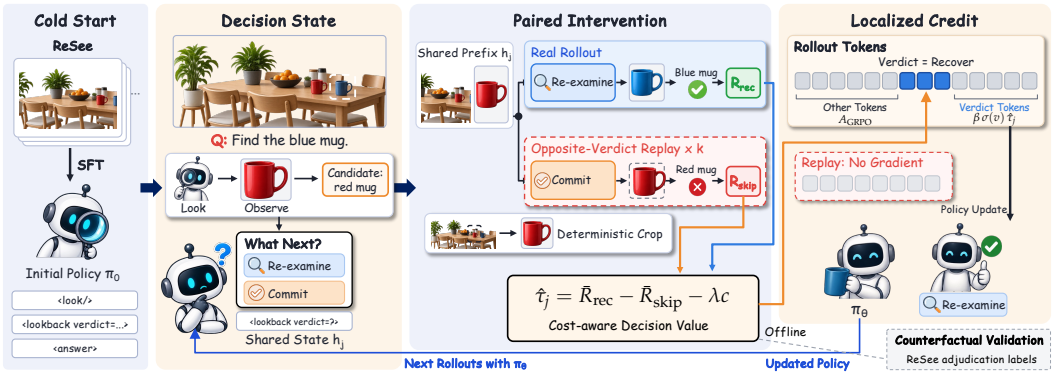


Figure 3: CARE. Each on-policy verdict is paired with a replay under the opposite verdict from the same decision state. Their cost-adjusted return difference credits only the verdict tokens.

Overview. Figure 3 summarizes CARE. The curated RESEE cold start establishes the visual-search protocol and an explicit commit-or-re-examine verdict. At up to M decision states in each on-policy rollout, the observed continuation is paired with an opposite-verdict replay from the same state. Their cost-adjusted return difference estimates $\tau^{\pi}(h_j)$ and credits only the verdict tokens. Other tokens keep GRPO's group-relative advantage (Shao et al., 2024b) (Algorithm 1).

Paired Counterfactual Estimation. At a decision state h_j, one verdict is taken by the on-policy rollout, while the alternative is imposed after resetting to the same environment state. This reset is exact because each visual observation is a deterministic crop of a fixed image: identical image–box pairs reproduce identical observations. The two branches therefore share the complete pre-decision context, including all preceding tokens and observations, while their post-verdict continuations are sampled independently from the current policy under identical decoding settings. Let $\bar{R}_{\text{rec}}$ and $\bar{R}_{\text{skip}}$ denote the empirical returns of the two branches at h_j, where the replayed branch is averaged over k independent replays. We estimate the cost-aware value of re-examination as

$$\hat{\tau}_j = \text{clip}\big(\bar{R}_{\text{rec}} - \bar{R}_{\text{skip}} - \lambda c(h_j), -\kappa, \kappa\big)\,, \tag{3}$$

where $\kappa > 0$ bounds the estimate. Before clipping, this estimator is unbiased conditional on h_j when both continuations follow their corresponding conditional policies, and increasing k reduces the sampling variance of the replayed branch. Importantly, the observation cost enters only this decision contrast rather than the trajectory reward, so it changes the threshold for re-examination without penalizing unrelated search or reasoning tokens.

Localized Credit Assignment. The commit-or-re-examine verdict occupies an explicit span in the model output, so its token positions can be identified directly from the interaction protocol rather than inferred from entropy, attention, or state similarity. For the i-th rollout ζ_i of a sampling group, let $R_i = R(\zeta_i)$ be its reward, $v_{i,j}$ and $\hat{\tau}_{i,j}$ the verdict and estimate at its j-th decision state, and $\mathcal{V}_{i,j}$ the positions of that verdict's tokens. Define the sign $s(\text{RECOVER}) = +1$ and $s(\text{SKIP}) = -1$. We replace the trajectory-level advantage on these positions with the signed counterfactual value,

$$\tilde{A}_{i,t} = \begin{cases} \beta\, s(v_{i,j})\, \hat{\tau}_{i,j}, & t \in \mathcal{V}_{i,j}, \\ \dfrac{R_i - \mu_{g(i)}}{\sigma_{g(i)} + \delta}, & \text{otherwise}, \end{cases} \tag{4}$$

where $\beta > 0$ scales the counterfactual credit, $\mu_{g(i)}$ and $\sigma_{g(i)}$ are the mean and standard deviation of the rewards in the sampling group $g(i)$, and δ is a small constant. Thus, a positive $\hat{\tau}_{i,j}$ reinforces RECOVER and suppresses SKIP, whereas a negative value induces the opposite update, and the counterfactual credit is isolated to the decision that controls whether search continues.

Policy Optimization Objective. We optimize the resulting token-level advantages with the clipped policy-gradient objective used by GRPO. To preserve support for both verdicts during training, an entropy bonus is applied only to the verdict positions $\mathcal{V} = \bigcup_{i,j} \mathcal{V}_{i,j}$:

$$\mathcal{L}(\theta) = -\mathbb{E}_t\Big[\min\Big(r_t(\theta)\tilde{A}_t,\ \text{clip}(r_t(\theta), 1-\varepsilon_{\text{lo}}, 1+\varepsilon_{\text{hi}})\,\tilde{A}_t\Big)\Big] + \gamma\,\mathbb{E}_t[\text{KL}_t] - \eta\,\mathbb{E}_{t\in\mathcal{V}}[H_t], \tag{5}$$

where y_t is the t-th output token, $r_t(\theta) = \pi_\theta(y_t \mid y_{<t})/\pi_{\text{old}}(y_t \mid y_{<t})$, and $1 - \varepsilon_{\text{lo}}$ and $1 + \varepsilon_{\text{hi}}$ are the lower and upper clipping bounds. KL_t is the token-level divergence from the cold-start policy weighted by γ, and H_t is the policy entropy weighted by η and applied only to verdict tokens. Replays receive no gradient, and adjudication labels are used only to evaluate $\hat{\tau}$ offline.

Proposition 1 (Verdict-Contrast Alignment) *Fix a search state h under a binary verdict choice, with $p = \pi(\text{RECOVER} \mid h) \in (0, 1)$ and $z = \log[p/(1-p)]$. Hold the conditional continuation policies fixed. Assume exact reset, independent finite-variance returns, and opposite-verdict replays from the corresponding conditional policy. For the unclipped paired contrast D_v and $k \geq 1$ replays, the expected verdict score gradient is*

$$\mathbb{E}[\beta s(v) D_v \, \partial_z \log \pi(v \mid h) \mid h] = 2\beta p(1-p)\tau^\pi(h). \tag{6}$$

The contrast has conditional variance $\sigma_v^2 + \sigma_{\neg v}^2/k$, where σ_v^2 is the return variance under v. Clipping preserves the expected gradient's sign for $|\tau^\pi(h)| > B_\kappa(h)$. Appendix B.3 characterizes this distortion and derives a finite-replay bound on the probability of a clipped verdict-score sign error.

5 Experiments

Benchmarks and Metrics. We evaluate on our 980-question split drawn from six source benchmarks (Wu & Xie, 2024; Wang et al., 2026a; Lai et al., 2026; Wang et al., 2025b; Zhang et al., 2025; Li et al., 2025): 587 annotated records and a 393-record zero-shot set unseen in training. Accuracy is Avg@32 under a common scorer, reported as the macro average over the six sources so that each contributes equally. These results use our split and scoring protocol. The separate 468-question uniform subset is the comparison set for published benchmark numbers. Against adjudication labels, we report recall, precision and the over-examination rate $P(\text{re-examine} \mid \text{unwarranted})$ for the re-examination decision. Appendix C gives the grading rules and full metric definitions.

Baselines. Under the ReSee tool protocol, we benchmark eight instruction-tuned VLMs and five released visual-search agents: Pixel Reasoner (Su et al., 2026), DeepEyes (Zheng et al., 2026), ViGoRL (Sarch et al., 2026), TreeVGR (Wang et al., 2026a), and Mini-o3 (Lai et al., 2026). We also include their common Qwen2.5-VL-7B backbone (Bai et al., 2025b). We further evaluate state-agnostic controls that instruct the policy to re-examine, commit, or sample a random verdict, scoring the actions actually taken. For learning-based comparisons, Zero-shot uses the backbone without task-specific prompting or training, Prompted adds a re-examination instruction without parameter updates, and SFT stops after the common ReSee cold start. All RL methods follow the same protocol: GRPO (Shao et al., 2024b) uses trajectory-level advantages, GiGPO (Feng et al., 2026a) adds step-level credit from repeated states, and Entropy (Wang et al., 2026b) restricts updates to high-entropy tokens. GRPO-Pruned removes counterfactually unwarranted cold-start demonstrations to isolate data filtering from decision-level credit assignment. More details are given in Appendix C.3.

Implementation Details. The two backbones are Qwen3-VL-8B-Instruct (Bai et al., 2025a) and InternVL3-8B (Zhu et al., 2025), with Qwen3-VL-8B used for framework analyses. Cold starts use the six variants of Appendix B and are full-parameter SFT with the vision tower and projector frozen, ZeRO-3 with optimizer offload (Rajbhandari et al., 2020; Ren et al., 2021), learning rate 1×10^{-5}, three epochs, and context length 32,768. RL uses GRPO (Shao et al., 2024b) with learning rate 1×10^{-6}, group size 8, batch size 128, sampling temperature 1.0, one epoch over 11,500 prompts, clipping $(\varepsilon_{\text{lo}}, \varepsilon_{\text{hi}}) = (0.2, 0.3)$, a token-level KL weight $\gamma = 10^{-3}$, and at most six generation rounds per trajectory. Both stages share a pixel budget of $[4\times10^4, 2\times10^6]$ per image. Care uses $k = 1$ counterfactual replay per decision, at most $M = 3$ decisions per trajectory, $\beta = 1$, clip $\kappa = 1$, an observation cost weight $\lambda = 2\times10^{-4}$, and an entropy weight $\eta = 10^{-2}$ on the verdict span.

5.1 Main Results

Benchmarking on ReSee. Table 1 shows that answer accuracy alone does not predict whether a model re-examines at the right decision states. Several instruction-tuned models remain near the 50% balanced-accuracy chance level under the CoT prompt, despite strong answer accuracy on the benchmark sources. Prompts that favor re-examination increase recall but also increase over-examination, while some released search agents attain higher precision by making few requests and

leave warranted states unexamined. The resulting gap between answer quality and decision quality supports evaluating the verdict against its downstream consequence.

Table 1: Benchmarking on ReSee. Balanced accuracy is shaded relative to the 50% chance level.

	Answer accuracy ↑						Re-examination decision				Cost
Model	V*	HR-Bench	MME-RW	TreeBench	O3-Bench	**Macro**	Recall	Prec.	Over-ex. ↓	Bal. acc.	Looks ↓
Instruction-tuned vision–language models											
Qwen3.8-27B	**92.6**	**88.4**	48.8	**61.6**	**59.1**	**70.1**	4.4	14.3	5.9	49.3	3.4
DeepSeek-V4-Flash	79.5	80.0	53.9	54.0	36.6	60.8	4.4	11.1	7.7	48.4	25.4
Qwen3-VL-235B-A22B	81.6	77.6	43.6	49.5	35.9	57.6	31.4	18.0	31.9	49.7	5.0
Muse-Glimmer-30B	55.3	69.1	47.8	44.5	40.9	51.5	0.0	0.0	1.5	49.3	8.2
Qwen3-VL-32B	77.4	74.0	44.7	47.0	26.2	53.9	**31.6**	33.3	18.2	**56.7**	0.8
MiMo-V2.5	69.5	78.4	43.8	37.9	21.9	50.3	6.2	33.3	1.2	52.5	0.3
GLM-4.6V	37.4	7.4	20.4	10.3	5.6	16.2	30.6	26.8	19.2	55.7	3.9
Grok-4.6	7.9	9.7	4.4	2.8	1.6	5.3	4.0	25.0	2.8	50.6	5.1
Released visual search agents											
TreeVGR-7B (Wang et al., 2026a)	70.2	57.9	**57.9**	42.7	33.1	52.4	18.8	18.0	18.6	50.1	**0.01**
Mini-o3-7B (Lai et al., 2026)	73.8	57.1	54.6	43.6	32.4	52.3	12.5	24.6	8.7	51.9	4.22
DeepEyes-7B (Zheng et al., 2026)	63.7	54.2	51.1	35.7	27.6	46.5	2.9	**43.8**	**0.5**	51.2	0.27
Pixel-Reasoner-7B (Su et al., 2026)	64.9	53.4	46.2	39.0	25.4	45.8	6.2	22.0	4.6	50.8	0.40
Qwen2.5-VL-7B (Bai et al., 2025b)	61.3	53.4	43.6	34.5	24.3	43.4	1.9	22.5	1.3	50.3	0.22
ViGoRL-7B (Sarch et al., 2026)	61.9	50.6	44.2	33.8	22.4	42.6	7.7	13.0	10.9	48.4	4.24

Learning on ReSee. Table 2 shows that Care achieves the highest macro accuracy and balanced accuracy on both backbones while using fewer looks than GRPO. On Qwen3-VL, it gains 2.5 macro points and 11.1 balanced-accuracy points, with the same direction of improvement on InternVL3. The answer-accuracy gains are larger on O3-Bench and VisualProbe. These gains coincide with higher recall of warranted requests and lower over-examination. Care also improves on GRPO-Pruned, indicating that removing unwarranted demonstrations alone does not account for the full result. Together, the two backbones support verdict-level credit as a way to improve both answer quality and selective visual re-examination.

Table 2: Main results on our six-source split. Macro subscripts report run-to-run standard deviations.

	Answer accuracy ↑							Re-examination decision				Cost
Method	V*	HR-Bench	MME-RW	TreeBench	O3-Bench	VisualProbe	**Macro**	Recall	Prec.	Over-ex. ↓	Bal. acc.	Looks ↓
Qwen3-VL-8B												
Always-prompt	53.8	57.6	46.5	32.4	23.6	19.7	$38.9_{\pm 1.8}$	**66.5**	17.3	70.0	48.2	2.16
Never-prompt	56.0	58.8	45.7	31.6	21.1	22.9	$39.3_{\pm 2.0}$	0.7	24.2	0.4	50.1	2.19
Random verdict	53.7	57.7	47.0	32.1	22.2	**23.6**	$39.4_{\pm 1.4}$	17.5	17.3	18.4	49.5	2.20
Zero-shot	**68.5**	56.9	23.4	**41.1**	29.6	0.8	$36.7_{\pm 1.1}$	0.0	–	**0.0**	50.0	**0.00**
Prompted	46.9	50.6	34.1	33.4	11.5	17.2	$32.3_{\pm 1.5}$	11.1	22.9	8.2	51.5	3.53
SFT	55.7	59.3	48.5	33.2	23.4	15.5	$39.3_{\pm 1.8}$	30.4	22.1	23.5	53.5	2.14
GRPO	65.8	67.1	58.2	36.8	31.5	16.8	$46.0_{\pm 1.6}$	23.1	27.3	13.6	54.8	2.09
GiGPO	65.9	67.4	58.3	37.0	31.8	17.4	$46.3_{\pm 1.6}$	23.1	26.7	14.0	54.5	2.15
Entropy	65.5	67.0	58.1	37.1	32.0	16.9	$46.1_{\pm 1.7}$	21.2	28.2	11.9	54.6	2.06
GRPO-Pruned	65.1	67.2	58.7	37.4	32.9	18.1	$46.6_{\pm 1.5}$	25.0	37.1	9.3	57.8	1.99
Care (Ours)	66.8	**68.1**	**59.9**	39.0	**35.1**	21.9	$\mathbf{48.5}_{\pm 1.3}$	42.3	**46.8**	10.6	**65.9**	1.97
InternVL3-8B												
Always-prompt	51.5	57.7	50.4	37.6	12.7	21.2	$38.5_{\pm 1.5}$	**54.7**	21.0	45.4	54.7	2.54
Never-prompt	53.2	58.4	49.7	36.8	14.9	22.6	$39.3_{\pm 1.7}$	1.9	20.0	1.7	50.1	2.50
Random verdict	52.4	58.0	50.2	37.2	14.2	22.0	$39.0_{\pm 1.5}$	19.2	17.9	19.5	49.9	2.51
Zero-shot	54.8	54.8	32.1	35.3	**35.6**	23.8	$39.4_{\pm 0.2}$	0.0	–	**0.0**	50.0	**0.00**
Prompted	57.1	58.3	34.9	31.7	29.4	15.5	$37.8_{\pm 0.2}$	5.3	**46.6**	1.5	51.9	1.06
SFT	53.0	58.7	51.0	38.1	15.4	21.7	$39.7_{\pm 1.6}$	32.7	20.7	27.5	52.6	2.49
GRPO	58.9	62.5	55.3	41.8	23.8	25.2	$44.6_{\pm 1.5}$	26.9	26.9	16.1	55.4	2.34
GiGPO	59.2	62.9	55.2	42.4	24.5	25.9	$45.0_{\pm 1.5}$	28.8	28.8	15.7	56.6	2.32
Entropy	58.8	62.8	55.6	42.2	24.1	25.3	$44.8_{\pm 1.6}$	25.0	28.9	13.6	55.7	2.27
GRPO-Pruned	59.0	62.7	55.8	42.0	25.0	25.7	$45.0_{\pm 1.4}$	28.8	34.9	11.9	58.5	2.20
Care (Ours)	**59.8**	**63.6**	**57.0**	**44.1**	27.6	**29.3**	$\mathbf{46.9}_{\pm 1.2}$	44.2	43.4	12.7	**65.8**	2.15

5.2 Framework Analysis

Ablation of Decision-Level Credit. Table 3 shows that the paired verdict contrast and its localization are central to Care. Removing the verdict flip lowers macro accuracy by 4.3 points, the largest ablation effect. Its recall also falls most sharply, consistent with paired comparison helping identify states that warrant another look. Without observation pricing, over-examination nearly doubles, while trajectory-level pricing restores selectivity only partly and diffuse credit reduces recall and precision. This pattern supports localizing credit and cost to the verdict under evaluation.

Hyperparameter Sensitivity. Figure 4 shows broad performance peaks for verdict credit and entropy, with a flatter response to clipping. Reducing β to 0.25 lowers macro accuracy by 2.7 points,

Table 3: Ablation of CARE. Macro subscripts give run-to-run standard deviations. Red subscripts show relative degradation from the full method, with increases for Over-ex. and Looks.

Variant	Macro ↑	Recall ↑	Prec. ↑	Over-ex. ↓	Bal. acc. ↑	Looks ↓
CARE	**48.5**$_{\pm 1.3}$	**42.3**	**46.8**	**10.6**	**65.9**	**1.97**
w/o localization	45.8$_{\pm 1.7}$	30.8$_{-27.2\%}$	34.0$_{-27.4\%}$	13.1$_{+23.6\%}$	58.8$_{-10.8\%}$	2.08$_{+5.6\%}$
w/o verdict flip	44.2$_{\pm 1.9}$	21.2$_{-49.9\%}$	25.6$_{-45.3\%}$	13.6$_{+28.3\%}$	53.8$_{-18.4\%}$	2.12$_{+7.6\%}$
w/o cost	45.4$_{\pm 1.6}$	29.3$_{-30.7\%}$	24.4$_{-47.9\%}$	19.9$_{+87.7\%}$	54.7$_{-17.0\%}$	2.14$_{+8.6\%}$
w/ trajectory cost	46.5$_{\pm 1.6}$	34.6$_{-18.2\%}$	38.3$_{-18.2\%}$	12.3$_{+16.0\%}$	61.2$_{-7.1\%}$	2.07$_{+5.1\%}$
w/o verdict entropy	47.2$_{\pm 1.4}$	34.6$_{-18.2\%}$	40.0$_{-14.5\%}$	11.4$_{+7.5\%}$	61.6$_{-6.5\%}$	2.05$_{+4.1\%}$

while balanced accuracy falls and visual requests rise. Increasing β beyond its reference value also reduces answer and decision quality, giving the sweep a clear interior optimum. Entropy follows a similar pattern, whereas clipping causes smaller variation across its tested range. Additional replays yield diminishing accuracy gains: macro accuracy rises from 48.5 at $k = 1$ to 49.0 at $k = 4$ and remains there at $k = 8$. The curves indicate that verdict-credit calibration matters more than fine adjustments to clipping, while a small replay budget captures most of the observed gain.

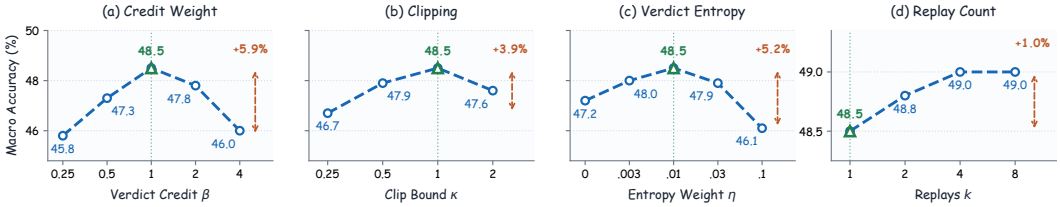


Figure 4: Qwen3-VL-8B macro accuracy across four parameter sweeps. Green triangles mark the main setting, and arrows report the percentage gain between the extremes.

Counterfactual Utility. Counterfactual utility tests whether paired-rollout return differences rank decisions by the benefit of re-examination. Figure 5 shows that the single-replay estimator reaches 0.69 AUC, compared with 0.55 for self-assessed answer-flip risk and 0.54 for unresolved-constraint count. The paired estimate therefore ranks warranted decisions more reliably than either annotator-recorded signal. Additional replays raise AUC to 0.73 at $k = 8$, with smaller gains beyond the single-replay setting. This pattern supports estimating the consequence of the verdict rather than relying on stated uncertainty alone.

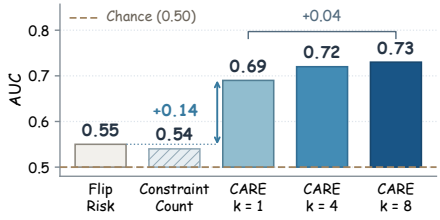


Figure 5: Utility-ranking AUC.

Decision Quality at Equal Visual Cost. Figure 6 shows that CARE retains higher accuracy and selectivity than cost-tuned GRPO at the same mean number of visual requests. At 1.97 looks per trajectory, it leads by 2.7 macro points and 10.2 balanced-accuracy points. Across the nonzero prices, increasing observation cost reduces requests for both policies, while CARE retains higher balanced accuracy. At the strongest price, both policies lose answer accuracy as looks fall, consistent with useful evidence being suppressed. The gap at comparable cost indicates that request budgeting alone does not explain the decision-quality gains, supporting the role of assigning observation cost and paired return contrast directly to each verdict.

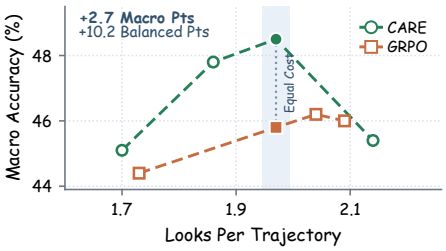


Figure 6: Macro accuracy and visual cost for CARE and GRPO.

Out-of-Distribution Generalization. To assess transfer beyond the RESEE training split, we evaluate six test sets in high-resolution perception and extended visual search under their official scorers (Appendix C.2). Table 4 shows that CARE scores higher than GRPO with fewer visual requests on all six sets. The largest gains are on VistaHop and VisualNeedle, at 4.0 and 3.6 points, respectively. Gains on remote-sensing and multi-image tasks extend this pattern across visual-search settings.

Table 4: OOD results for CARE with Qwen3-VL-8B-Instruct. Obs. reports mean visual tool calls, and Gain is the relative score improvement over GRPO.

		Zero-shot		SFT		GRPO		CARE		
Benchmark	**Metric**	Score↑	Obs.↓	Score↑	Obs.↓	Score↑	Obs.↓	Score↑	Obs.↓	Gain (%)
High-Resolution Perception										
HR-Bench 8K (Wang et al., 2025b)	Acc.	66.5	0.00	68.0	2.35	69.5	2.24	**71.0**	2.11	+2.2
MagnifierBench (Li et al., 2023)	Acc.	52.3	0.00	56.2	2.56	58.7	2.41	**61.5**	2.25	+4.8
XLRS-Bench-Lite (Wang et al., 2025a)	Acc.	42.4	0.00	44.1	3.02	46.0	2.90	**48.3**	2.72	+5.0
Extended Visual Search										
VistaHop (He et al., 2026)	Pass@1	8.3	0.00	11.7	5.64	14.0	5.52	**18.0**	5.18	+28.6
VisualNeedle (Chen et al., 2026)	Acc.	17.0	0.00	36.3	3.10	40.7	2.94	**44.3**	2.70	+8.8
Visual Haystacks-SN-50 (Wu et al., 2025)	Acc.	50.7	0.00	52.6	4.84	54.0	4.58	**55.8**	4.26	+3.3

Scaling Behaviours. We evaluate four scaling dimensions: model capacity, training data, update budget, and complete-trace coverage. Figure 7 shows that CARE retains its advantage from 8B through 235B-A22B and across the three training sweeps. Its macro-accuracy lead reaches 2.9 points on the largest variant, while extra prompts and complete traces yield steeper gains than for GRPO. These trends suggest that decision-level credit remains useful at larger scales.

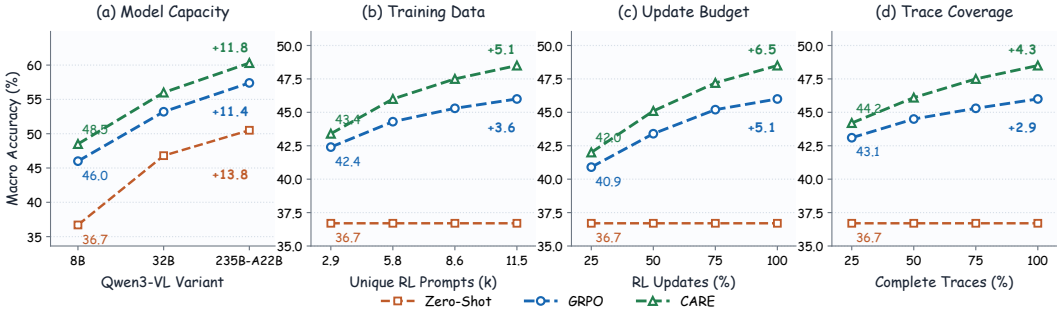


Figure 7: Six-source macro Avg@32 under the evaluation protocol of Table 2. Model scaling includes 235B-A22B sparse experts, and other sweeps use 8B.

Case Studies. Figure 8 contrasts useful re-examination with timely commitment. On the route map, CARE checks the missing connection before answering, whereas GRPO commits after the first transfer. In the beach scene, both objects are located and their spatial relation is settled. CARE stops, while GRPO unnecessarily repeats the umbrella inspection. Across map connectivity and object relations, the contrast shows a shared decision criterion: re-examine when required evidence is missing and commit once it is resolved.

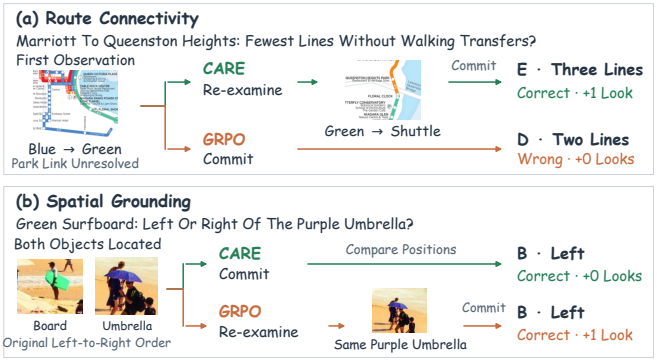


Figure 8: CARE and GRPO trajectories. Extra looks start at the shared state. Full cases appear in Appendix E.1.

6 CONCLUSION

This paper proposes a framework for evaluating and learning selective re-examination in visual search. RESEE provides explicit commit-or-re-examine decisions and counterfactual adjudication of their utility, supporting both benchmarking and policy learning. CARE compares paired continuations from the same search state and assigns their cost-aware return difference to the verdict tokens. Extensive experiments on real-world visual-search tasks across two VLM backbones demonstrate that CARE outperforms GRPO in accuracy and decision selectivity while reducing visual requests, with accuracy gains extending to six out-of-distribution benchmarks.

AI USE STATEMENT

In this work, we used LLM tools to polish wording and grammar, generate the icons in Figure 1, and revise code. We also used a vision-language model to generate trajectories for the RESEE benchmark. We carefully reviewed all AI-assisted content and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

Section 5 specifies the evaluation split, metrics, baselines, and training configuration. Appendix A documents dataset construction and counterfactual adjudication. Appendix B gives the training algorithm and cold-start variants, while Appendix C defines answer grading and decision metrics. Our code and a subset of RESEE demo samples are available at `https://anonymous.4open.science/r/CARE-D42D`.

REFERENCES

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

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

Jingru Chen, Yiming Liu, Mingtao Chen, Sijie Chen, Richeng Xuan, Liang Yang, Zhichao Hu, and Fanyang Lu. Visualneedle: Benchmarking active visual search in information-dense scenes. *arXiv preprint arXiv:2605.26380*, 2026.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *Advances in Neural Information Processing Systems*, 38:46375–46408, 2026a.

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

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li,

Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 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.

Hang He, Chuhuai Yue, Chengqi Dong, Chengcheng Wan, Ting Su, Haiying Sun, Jiajun Chai, Xiaohan Wang, and Guojun Yin. Vistahop: Benchmarking multi-hop visual reasoning for visual deepsearch. *arXiv preprint arXiv:2606.03273*, 2026.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International conference on learning representations*, volume 2024, pp. 32808–32824, 2024.

Senjie Jin, Peixin Wang, Boyang Liu, Xiaoran Fan, Shuo Li, Zhiheng Xi, Jiazheng Zhang, Yuhao Zhou, Tao Gui, Qi Zhang, et al. Entropy is not enough: Unlocking effective reinforcement learning for visual reasoning via vision-anchored token selection. *arXiv preprint arXiv:2606.03937*, 2026.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in rl training of llms. *arXiv preprint arXiv:2410.01679*, 2024.

Klaus Krippendorff. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*, 30(3):411–433, 2004.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, volume 2025, pp. 54523–54549, 2025.

Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. In *International Conference on Learning Representations*, volume 2026, pp. 76722–76746, 2026.

Bo Li, Peiyuan Zhang, Jingkang Yang, Yuanhan Zhang, Fanyi Pu, and Ziwei Liu. Otterhd: A high-resolution multi-modality model. *arXiv preprint arXiv:2311.04219*, 2023.

Kaican Li, Lewei Yao, Jiannan Wu, Tiezheng Yu, Jierun Chen, Haoli Bai, Lu Hou, Lanqing Hong, Wei Zhang, and Nevin L Zhang. Insight-o3: Empowering multimodal foundation models with generalized visual search. In *The Fourteenth International Conference on Learning Representations*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison 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*, volume 2024, pp. 39578–39601, 2024.

Shangke Lyu, Linjuan Wu, Yuchen Yan, Xingyu Wu, Hao Li, Yongliang Shen, Peisheng Jiang, Weiming Lu, Jun Xiao, and Yueting Zhuang. Hierarchical budget policy optimization for adaptive reasoning. *arXiv preprint arXiv:2507.15844*, 2025.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Renwei Meng. Policy-conditioned counterfactual credit for verifiable reinforcement learning of long-horizon language agents. *arXiv preprint arXiv:2606.05263*, 2026.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pp. 1–16. IEEE, 2020.

Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. {Zero-offload}: Democratizing {billion-scale} model training. In *2021 USENIX Annual Technical Conference (USENIX ATC 21)*, pp. 551–564, 2021.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *Advances in Neural Information Processing Systems*, 38:150977–151013, 2026.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024a.

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

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6613–6629, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

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

Fengxiang Wang, Hongzhen Wang, Zonghao Guo, Di Wang, Yulin Wang, Mingshuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, et al. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336. IEEE, 2025a.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *International Conference on Learning Representations*, volume 2026, pp. 148769–148794, 2026a.

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 the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, et al. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning. *Advances in Neural Information Processing Systems*, 38:115452–115486, 2026b.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7907–7915, 2025b.

Quan Wei, Siliang Zeng, Chenliang Li, William Brown, Oana Frunza, Wei Deng, Anderson Schneider, Yuriy Nevmyvaka, Yang Katie Zhao, Alfredo Garcia, et al. Reinforcing multi-turn reasoning in llm agents via turn-level reward design. *arXiv preprint arXiv:2505.11821*, 2025.

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

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, David Chan, et al. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *International Conference on Learning Representations*, volume 2025, pp. 64592–64615, 2025.

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 *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025.

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

Zhaomeng Zhou, Lan Zhang, Junyang Wang, Mu Yuan, and Junda Lin. Nice fold or hero call: Learning budget-efficient thinking for adaptive reasoning. *arXiv preprint arXiv:2605.11625*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

Roadmap. Appendix A details dataset construction and human evaluation, Appendix B gives the training procedure and alignment proofs, and Appendix C specifies the experimental protocol. Appendices D, E, and F provide released-model benchmarking, case studies, and additional results, respectively.

A Details of ReSee

A.1 Annotation Protocol

Trajectory Generation. Qwen3-VL-Plus (Bai et al., 2025a) writes each trajectory from the image, question and reference answer, which anchors the search to the target evidence. A trajectory contains at most twelve steps. A step parses the question, scans the image, inspects a region, compares candidates, re-examines an earlier region, confirms the answer, or flags uncertainty. Every step that reads the image records its inspected box.

Conservative Verdicts. At each decision state, the annotator records a verdict, unresolved constraints, estimated answer-flip risk and estimated cost. The protocol calls for re-examination when a constraint remains unresolved, with at most three re-examinations per trajectory. All 1,131 re-examinations record an unresolved constraint, while 7,505 of the 7,696 commitments record none. Counterfactual adjudication compares continuation outcomes to assign the labels. Table 5 lists the recorded fields and six re-examination patterns. Each re-examination also records the revisited step, region and effort.

Counterfactual Adjudication. Decision labels compare re-examination and commitment from the same state using the utility $\tau(h)$ of Section 3. The demonstrated re-examination continuation reaches the reference answer in 99.3% of trajectories. The commitment outcome truncates the trajectory at the decision state and requires an answer from the evidence gathered so far. With answer correctness as the reward and priced cost $\lambda c(h) \in (0, 1)$, a successful re-examination has $\tau(h) > 0$ exactly when commitment produces an incorrect answer. The same comparison labels re-examinations as warranted or unwarranted and commitments as premature or safe.

Table 5: The top block lists fields recorded at a decision state, and the bottom block describes the six re-examination patterns.

Decision record		
Field	**Values**	**Meaning**
Verdict	recover, skip	whether the annotator proposes to re-examine before committing
Candidate step	step index	the step that produced the answer currently held
Residual constraints	free text, possibly empty	constraints of the question not yet verified on the image
Flip risk	low, medium, high	self-assessed risk that the answer would change (7,538 / 637 / 652)
Cost estimate	four integers and flags	extra steps, attention units, need to zoom again, distance to the revisited step
Rationale	free text	justification of the verdict
Label	warranted, unwarranted, safe, premature	outcome of the adjudication, present on 1,862 decisions
Target step and region	step index, box	where a re-examination returns to and what it looks at
Scope	single look, re-inspect, re-sweep	effort the re-examination may spend (648 / 437 / 46)

Re-examination patterns			
Pattern	**The re-examination . . .**	**Count**	**Warranted (%)**
Attribute re-read	reads again an attribute, such as text or color, that was read under doubt	452	33.7
Constraint conflict	checks a candidate against a constraint that the evidence seems to contradict	405	22.8
Missed-region sweep	searches a region that the earlier search left unvisited	135	46.3
Reject re-verify	returns to a candidate that was rejected earlier	72	22.6
Count reconcile	recounts after partial counts disagree	50	48.9
Context expansion	widens the view around a candidate to recover its context	17	61.5

A.2 Dataset Statistics

Table 6 reports the composition of ReSee by source. Images have a median area of 3.39 MP and reach 86.25 MP, and the median target box occupies 0.8% of the frame. Each trajectory becomes a multi-turn episode in which every look and every re-examination ends the assistant turn. The requested region is cut at full resolution with a 10% margin, bounded at 1024 pixels on its longer side and returned as the next observation, giving 6,333 observations over 2,197 episodes. Two filters

separate training from the evaluation split: the 459 trajectories whose reasoning quotes the reference answer leave supervised training, and a perceptual comparison against the evaluation images removes the 397 training records built on 207 near-duplicate images. On the 883 adjudicated re-examinations, self-assessed flip risk has an AUC of 0.545, unresolved-constraint count has an AUC of 0.533, and each cost estimate has an AUC no greater than 0.51.

Table 6: Composition of RESEE by source benchmark.

			Decisions			Adjudicated			
Source	Traj.	Images	All	Re-ex.	Commit	Warr.	Unwarr.	Safe	Premature
MME-RealWorld	1,606	1,422	4,380	588	3,792	144	300	207	6
HR-Bench	611	193	1,509	72	1,437	9	54	156	1
TreeBench	398	395	1,205	172	1,033	41	84	176	1
O3-Bench	320	197	982	199	783	55	112	139	9
V*	190	190	437	21	416	1	15	145	0
VisualProbe	102	102	314	79	235	29	39	95	44
Total	3,227	2,499	8,827	1,131	7,696	279	604	918	61

A.3 HUMAN EVALUATION

Three expert raters from our research group, who received no compensation, assess the five dimensions in Table 7 on source-stratified trajectories, verdict-stratified decisions, and pattern-stratified re-examinations, including the three rarest patterns. The adjudication sample balances warranted and unwarranted decisions. Raters see the images, transcripts, and returned crops without adjudication labels, reference answers, or other ratings. Agreement is measured by Fleiss' κ for binary dimensions (Fleiss, 1971) and Krippendorff's α for ordinal dimensions (Krippendorff, 2004). Table 8 reports high question and decision-state validity and 83.3% agreement with adjudication labels, with inter-rater $\kappa = 0.73$. Lower agreement on thought–action faithfulness suggests that trajectory wording provides a less stable supervision signal than outcome-based decision labels.

A.4 EXAMPLES

Figure 9 shows six trajectories of RESEE, each with a single re-examination. Boxes mark the inspected regions, the re-examined one in orange, and the observations appear in the order requested. The annotator proposes re-examination in all six cases and rates its risk as medium or high, yet only the first three change the outcome.

B DETAILS OF CARE

B.1 TRAINING PROCEDURE

Algorithm 1 summarizes one training iteration of CARE in the notation of Section 4.

Locating the Verdict Span. The verdict is an attribute of the `<lookback>` tag, so its text is found by matching that attribute in the decoded response. The matched characters are then mapped to token positions on the sequence over which the loss is computed, that is, after the placeholders of the returned observations have been expanded. Each decision yields exactly one span, which contains neither the surrounding rationale nor the answer.

B.2 COLD-START SUPERVISION

All six variants use the same 2,197 episodes and change re-examination supervision. *Full* retains all 450 demonstrated re-examinations across 437 episodes. *Pruned* removes those adjudicated as unwarranted and changes the deciding verdict to SKIP, leaving 95; both variants retain the same 5,883 ordinary looks. *Masked* retains the full text but excludes unwarranted re-examinations from

Table 7: Rubric used by the RESEE raters.

QUESTION VALIDITY — trajectory, binary

"Given the question and the reference region outlined on the full image, is the answer settled by that region alone?"

POSITIVE The region holds the evidence that settles the question, and an attentive reader could answer from the image without further context.

FAILURE The region is empty or off-target, several regions answer equally well, or the question admits more than one answer.

DECISION-STATE VALIDITY — decision, binary

"The search has returned some evidence. Has a candidate answer formed here, so that committing and searching further are both real options?"

POSITIVE A candidate answer is stated or implied, and stopping at this point would produce an answer rather than a refusal.

FAILURE No candidate has formed, the step is the first look at the image, or an answer was already committed earlier.

THOUGHT–ACTION FAITHFULNESS — step, three-point ordinal

"Does the reasoning at this step describe what is visible in the returned crop, and does it justify the action that follows?"

POSITIVE *Faithful* names objects, text or spatial relations present in the crop and ties them to the next action. *Partial* does only one of the two.

FAILURE *Unfaithful* describes content absent from the crop, falls back on generic wording such as *"examining the region"*, or contradicts the crop.

PATTERN FIDELITY — re-examination, five-point Likert

"How clearly does this re-examination follow its recorded pattern, one of attribute re-read, constraint conflict, missed-region sweep, reject re-verify, count reconcile or context expansion?"

POSITIVE *5* the recorded pattern is unambiguous, *4* it is clear apart from one minor deviation, *3* it is partially recognizable.

FAILURE *2* the pattern is weak or inconsistent, *1* the re-examination does what a different pattern describes.

ADJUDICATION AGREEMENT — decision, binary

"Here are the search state and its two continuations, one that re-examines and one that commits. Which of them reaches the reference answer?"

POSITIVE The choice made without seeing the label matches the adjudicated label of that decision.

FAILURE Both continuations reach the same outcome, or the choice contradicts the adjudicated label.

Table 8: Human evaluation of RESEE. Favorable rate pools the $3n$ judgments, counting scores of 4 or 5 for pattern fidelity and agreement with the counterfactual label for adjudication. The final column reports agreement among raters.

Dimension	Unit	n	Favorable rate	Agreement
Question validity	Trajectory	120	93.3%	$\kappa = 0.73$
Decision-state validity	Decision	120	89.2%	$\kappa = 0.65$
Thought–action faithfulness	Step	240	76.4%	$\alpha = 0.61$
Pattern fidelity	Re-examination	120	73.6%	$\alpha = 0.79$
Adjudication agreement	Decision	120	83.3%	$\kappa = 0.73$

the loss. *Stripped* removes the protocol tags while retaining the reasoning. *Single* uses the one-turn form. *External* is a size-matched external cold start re-expressed in our tag vocabulary; its trajectories are $3.4\times$ shallower and contain no re-examination decisions.

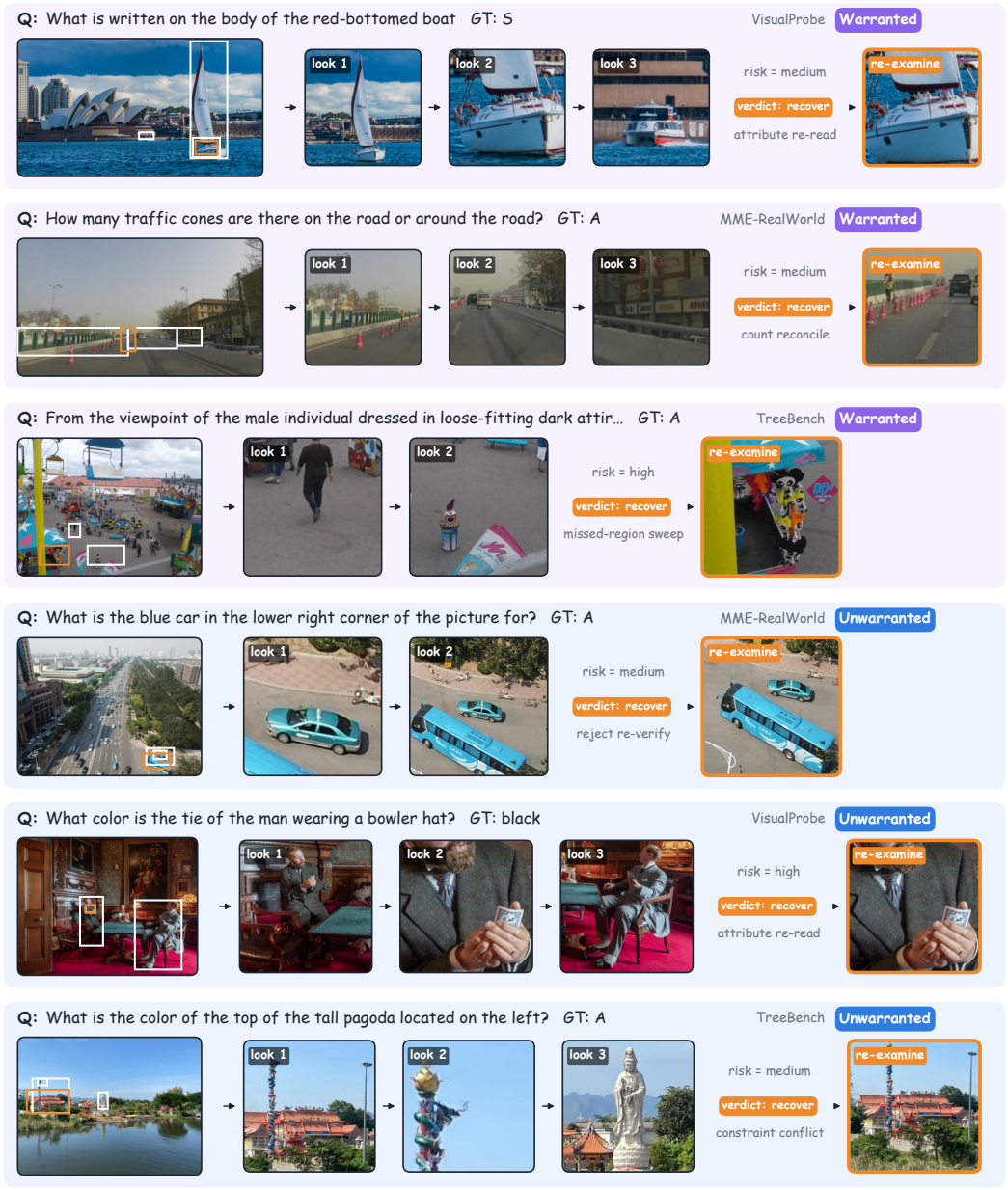


Figure 9: Six RESEE trajectories. Re-examination is warranted in the first three cases and unwarranted in the last three.

B.3 DECISION-LEVEL ALIGNMENT

Proof of Proposition 1. Fix h and write $q_v = Q^\pi(h, v)$ for the mean continuation reward and σ_v^2 for its finite variance under verdict v. The continuation policies are held fixed while differentiating the current verdict probability. Write $c = c(h)$ and $\tau = q_{\text{rec}} - q_{\text{skip}} - \lambda c$. Let R_v be the on-policy return conditional on the sampled verdict v, and let $R_{\neg v}^{(1)}, \ldots, R_{\neg v}^{(k)}$ be independent returns from resets to the same state with the opposite verdict imposed. We assume that a forced replay has the same conditional return law as an on-policy continuation given that verdict. This requires the same model prefix and decoding settings after the forced verdict, with fresh sampling randomness. The

raw paired contrast is

$$D_v = \begin{cases} R_{\text{rec}} - k^{-1}\sum_{j=1}^{k} R_{\text{skip}}^{(j)} - \lambda c, & v = \text{RECOVER}, \\ k^{-1}\sum_{j=1}^{k} R_{\text{rec}}^{(j)} - R_{\text{skip}} - \lambda c, & v = \text{SKIP}. \end{cases}$$

Exact reset and the conditional replay law place both branches at h under the corresponding verdict distributions, so $\mathbb{E}[D_v \mid h, v] = \tau$ for either verdict. Conditional independence makes the covariance between the observed and replayed returns zero and makes the variance of the replay average $\sigma_{\neg v}^2/k$. Hence $\text{Var}(D_v \mid h, v) = \sigma_v^2 + \sigma_{\neg v}^2/k$.

For the binary verdict abstraction, $\partial_z \log \pi(\text{RECOVER} \mid h) = 1 - p$ and $\partial_z \log \pi(\text{SKIP} \mid h) = -p$. The CARE verdict credit is $\beta s(v) D_v$, with $s(\text{RECOVER}) = 1$ and $s(\text{SKIP}) = -1$. Taking expectation over the sampled verdict and then over its two continuations gives

$$\begin{aligned} \mathbb{E}[\beta s(v) D_v \, \partial_z \log \pi(v \mid h) \mid h] &= \beta p(1-p)\mathbb{E}[D_{\text{rec}} \mid h, \text{RECOVER}] \\ &\quad + \beta(1-p)p\mathbb{E}[D_{\text{skip}} \mid h, \text{SKIP}] \\ &= 2\beta p(1-p)\tau. \end{aligned}$$

The cost-adjusted decision value, with continuation policies fixed, is $J_h(z) = p(z)(q_{\text{rec}} - \lambda c) + (1 - p(z))q_{\text{skip}}$. Since $\partial_z p = p(1-p)$, its derivative is $\partial_z J_h = p(1-p)\tau$. The raw CARE verdict term therefore has the same sign as this local derivative for $p \in (0, 1)$ and $\beta > 0$.

For the implemented clipping, let $C_v = \text{clip}(D_v, -\kappa, \kappa)$ and define

$$B_\kappa(h) = \frac{1}{2} \sum_{v \in \{\text{RECOVER}, \text{SKIP}\}} \mathbb{E}[(|D_v| - \kappa)_+ \mid h, v]\,.$$

For every real d, $|\text{clip}(d, -\kappa, \kappa) - d| = (|d| - \kappa)_+$. Write $e_v = \mathbb{E}[C_v - D_v \mid h, v]$. Replacing D_v by C_v in the score identity gives

$$\mathbb{E}[\beta s(v) C_v \, \partial_z \log \pi(v \mid h) \mid h] = 2\beta p(1-p)\left(\tau + \frac{e_{\text{rec}} + e_{\text{skip}}}{2}\right).$$

The triangle inequality and $|\mathbb{E}X| \le \mathbb{E}|X|$ give

$$\left|\frac{e_{\text{rec}} + e_{\text{skip}}}{2}\right| \le \frac{1}{2}\sum_v \mathbb{E}[|C_v - D_v| \mid h, v] = B_\kappa(h).$$

The finite second moments assumed in the proposition ensure that these expectations exist. For $|\tau| > B_\kappa(h)$, the perturbation cannot change the sign of τ, proving the clipping claim. A complete verdict span has this binary score if its two allowed sequences exhaust the normalized choice under a fixed prefix. □

Corollary 1 (Finite-Replay Direction Reliability) *Under the assumptions of Proposition 1, let $\tau = \tau^\pi(h) \neq 0$, $\beta, \kappa > 0$, and $V_v(k) = \sigma_v^2 + \sigma_{\neg v}^2/k$. For the clipped verdict score $G_v = \beta s(v)\text{clip}(D_v, -\kappa, \kappa)\,\partial_z \log \pi(v \mid h)$,*

$$\Pr(\tau G_v \le 0 \mid h, v) \le \frac{V_v(k)}{V_v(k) + \tau^2}. \tag{7}$$

The upper bound is nonincreasing in the replay count and converges to $\sigma_v^2/(\sigma_v^2 + \tau^2)$.

Proof. Conditional on v, the multiplier $s(v)\partial_z \log \pi(v \mid h)$ equals $1 - p$ for RECOVER and p for SKIP. Both are strictly positive. Symmetric clipping with $\kappa > 0$ preserves the sign of each realized contrast, so $\tau G_v \le 0$ is equivalent to $\tau D_v \le 0$. Set $t = |\tau|$ and $X = -\text{sign}(\tau)(D_v - \tau)$. Proposition 1 gives $\mathbb{E}[X \mid h, v] = 0$ and $\mathbb{E}[X^2 \mid h, v] = V_v(k)$, and the adverse event is $X \ge t$.

For every $a \ge 0$, the event $X \ge t$ implies $(X + a)^2 \ge (t + a)^2$. Applying Markov's inequality to the nonnegative random variable $(X + a)^2$ yields

$$\Pr(X \ge t \mid h, v) \le \frac{\mathbb{E}[(X + a)^2 \mid h, v]}{(t + a)^2} = \frac{V_v(k) + a^2}{(t + a)^2}.$$

For $V_v(k) > 0$, choose $a = V_v(k)/t$ and simplify the ratio to $V_v(k)/(V_v(k) + t^2)$. For $V_v(k) = 0$, $X = 0$ almost surely and $t > 0$, so the event has probability zero. This proves Eq. (7). The function $x/(x + \tau^2)$ is increasing for $x \ge 0$, and $V_v(k)$ is nonincreasing and converges to σ_v^2 as k grows. Substitution gives the stated limit. □

Corollary 2 (Value Signal under Same-Verdict Replay) *Keep the state, binary score and continuation assumptions of Proposition 1. Replace opposite-verdict replay by k independent continuations under the observed verdict and define $D_v^{\text{same}} = s(v)(R_v - k^{-1}\sum_j R_v^{(j)}) - \lambda c(h)$. Then*

$$\mathbb{E}[\beta s(v) D_v^{\text{same}} \partial_z \log \pi(v \mid h) \mid h] = -2\beta p(1-p)\lambda c(h). \tag{8}$$

For $k = 1$, $\kappa > 0$ and $\lambda c(h) \geq 0$, the expected clipped score is nonpositive and equals zero at zero cost.

Proof. All $k+1$ returns have conditional mean q_v, so

$$\mathbb{E}[D_v^{\text{same}} \mid h, v] = s(v)\left(q_v - \frac{1}{k}\sum_{j=1}^{k} q_v\right) - \lambda c(h) = -\lambda c(h).$$

Substitution into the two terms of the binary score expectation gives Eq. (8). The difference $q_{\text{rec}} - q_{\text{skip}}$ has canceled because each return is compared with another continuation under the same action.

For $k = 1$, write $Z_v = R_v - R_v^{(1)}$ and $a = \lambda c(h)$. Exchanging the independent, identically distributed returns changes Z_v to $-Z_v$ without changing its law. Consequently, the odd function $\text{clip}(\cdot, -\kappa, \kappa)$ has zero expectation under Z_v. Define $F_v(b) = \mathbb{E}[\text{clip}(Z_v + b, -\kappa, \kappa) \mid h, v]$. This expectation is finite since clipping is bounded. Pointwise monotonicity gives $F_v(-a) \leq F_v(0) = 0 \leq F_v(a)$ for $a \geq 0$. Oddness also gives

$$s(v)\text{clip}(s(v)Z_v - a, -\kappa, \kappa) = \text{clip}(Z_v - a\, s(v), -\kappa, \kappa).$$

The expected clipped score is therefore $\beta p(1-p)[F_{\text{rec}}(-a) - F_{\text{skip}}(a)] \leq 0$. At $a = 0$, both terms vanish. □

Opposite-verdict pairing retains the difference in decision values that same-verdict replay cancels.

C Experimental Setup

C.1 Evaluation Protocol

Answer Grading. Answers are graded by exact normalized match, followed by a bracketed-letter match when the reference is a single letter. Free-text references also allow word-bounded containment. An optional LLM stage sees the question and response, without the reference, and may rescue

Algorithm 1 One training iteration of CARE

Input: policy π_θ initialized by the cold start on RESEE, prompts $\mathcal{D}$, group size G, replays k, maximum decisions M, price λ, bound κ, weights β, γ, η
Output: updated policy π_θ

for each prompt $x \in \mathcal{D}$ **do**
  // Rollout
  sample a group of rollouts $\{\zeta_i\}_{i=1}^{G} \sim \pi_{\text{old}}(\cdot \mid x)$ and score $R_i = R(\zeta_i)$
  // Paired intervention
  for each rollout ζ_i and each of its first M decision states $h_{i,j}$ **do**
    read the written verdict $v_{i,j}$ and its token positions $\mathcal{V}_{i,j}$ ▷ located from the protocol
    reset the environment to $h_{i,j}$ and impose the opposite verdict $\neg v_{i,j}$ ▷ exact reset
    sample k replays from π_{old} and average their rewards into $\bar{R}_{i,j}^{\text{cf}}$ ▷ no gradient
    set $(\bar{R}_{\text{rec}}, \bar{R}_{\text{skip}}) \leftarrow (R_i, \bar{R}_{i,j}^{\text{cf}})$ if $v_{i,j} =$ RECOVER, and $(\bar{R}_{i,j}^{\text{cf}}, R_i)$ otherwise
    $\hat{\tau}_{i,j} \leftarrow \text{clip}\big(\bar{R}_{\text{rec}} - \bar{R}_{\text{skip}} - \lambda c(h_{i,j}),\ -\kappa,\ \kappa\big)$ ▷ Eq. (3)
  end for
  // Localized credit
  for each token position t of each rollout ζ_i **do**
    $\tilde{A}_{i,t} \leftarrow \beta\, s(v_{i,j})\, \hat{\tau}_{i,j}$ if $t \in \mathcal{V}_{i,j}$, and $(R_i - \mu_{g(i)})/(\sigma_{g(i)} + \delta)$ otherwise ▷ Eq. (4)
  end for
end for
  // Policy update
update θ by descending $\mathcal{L}(\theta)$ on the rollouts $\{\zeta_i\}$ only ▷ Eq. (5)

answers marked incorrect by the rules. Both scores are reported. The two untrained baselines may omit the <answer> tag, while trained policies must emit it. ΔAcc is reported when both sides contain at least twenty records. Table 2 and the framework analyses use our 980-question split, six-source macro average, and a common six-round generation budget. Table 1 uses the arithmetic mean of five displayed sources. VisualProbe is excluded from that comparison because its free-text references do not match the option template.

Macro Aggregation. Let $a_{b,r}$ be the accuracy on source b in evaluation run r. The six-source macro accuracy, its mean over 32 runs, and its sample standard deviation are

$$A_r = \frac{1}{6}\sum_{b=1}^{6} a_{b,r}, \qquad \bar{A} = \frac{1}{32}\sum_{r=1}^{32} A_r,$$
$$s_A = \left[\frac{1}{31}\sum_{r=1}^{32}(A_r - \bar{A})^2\right]^{1/2}.$$

Table 2 reports $\bar{A}$ and s_A. Aggregating sources within each run preserves their covariance in the reported spread.

Decision Metrics. We score the 288 adjudicated evaluation decisions against their labels. A re-examination counts as taken when the policy executes a <recover> action at that state, since <lookback> is emitted at every decision, including commitments. Writing D for these decisions, $W \subseteq D$ for the warranted ones and $\hat{W} \subseteq D$ for those at which the policy re-examines,

$$\text{recall} = \frac{|\hat{W} \cap W|}{|W|}, \qquad \text{precision} = \frac{|\hat{W} \cap W|}{|\hat{W}|}, \qquad \text{over-examination} = \frac{|\hat{W} \setminus W|}{|D \setminus W|},$$

and balanced accuracy is the mean of recall and $1 - \text{over-examination}$. Always re-examining gives 100% recall, 18.1% precision and 100% over-examination. Never re-examining gives zero recall and over-examination, with undefined precision. Both policies have 50% balanced accuracy. We pool decision metrics across sources, which contain between one and sixteen positive cases each. Grounding IoU is reported separately and its reward uses independently annotated boxes.

C.2 Unseen Benchmarks

High-Resolution Perception. HR-Bench 8K (Wang et al., 2025b) contains 200 questions on fine-grained perception in 8K images, split equally between single-instance properties and cross-instance relations. MagnifierBench (Li et al., 2023) contains 283 questions over 166 images and tests small-object identification, counting and attributes that often require a closer view. XLRS-Bench-Lite (Wang et al., 2025a) contributes 2,810 multiple-choice VQA items from ultra-high-resolution remote-sensing imagery, extending the evaluation to land-use, object and spatial reasoning.

Extended Visual Search. We use the initial 350-task release of VistaHop (He et al., 2026), which tests repeated inspection and multi-hop reasoning over visual anchors. VisualNeedle (Chen et al., 2026) tests crop-based search for small clues in 300 questions covering text, attributes, relations and occlusion. Visual Haystacks (Wu et al., 2025) tests retrieval across unrelated images. We evaluate its 1,000 binary questions in the 50-image single-needle setting, with one anchor-bearing image among distractors.

Scoring and Observations. VistaHop uses VistaArena Pass@1 and the other benchmarks report accuracy. VisualNeedle follows its tool-enabled protocol. We report request counts by benchmark from the harness traces. All six evaluation sets are disjoint from ReSee training.

C.3 Compared Methods

Trained Policies. Within each backbone, the trained comparisons in Table 2 share the RL data and reward. All use the *Full* cold start except GRPO-Pruned, which uses the *Pruned* variant. The methods differ as follows.

- **Always-prompt, Never-prompt, Random.** The cold-started backbone instructed to favor RECOVER, favor SKIP, or select RECOVER with a fixed random probability. We score the actions each policy executes.
- **Zero-shot, Prompted.** The untrained backbone under the common contract and under the selective prompt of Appendix D.1.
- **SFT.** The cold-start policy without reinforcement learning.
- **GRPO** (Shao et al., 2024b). Group-relative policy optimization, which normalizes the trajectory reward within a group of rollouts and assigns the resulting advantage to every token.
- **GiGPO** (Feng et al., 2026a). A two-level advantage that adds to the episode-level group advantage a step-level one, obtained by grouping actions taken from the same environment state across rollouts.
- **Entropy** (Wang et al., 2026b). GRPO with policy-gradient updates restricted to the 20% of tokens with the highest entropy.
- **GRPO-Pruned.** GRPO from the *Pruned* cold start, which removes the demonstrated re-examinations adjudicated as unwarranted and therefore uses the labels of RESEE.

Ablation Protocol. Table 3 uses Qwen3-VL-8B. The verdict-flip ablation replays the observed verdict while retaining the branch ordering and cost term of Eq. (3), with the same $k = 1$ replay and clipping as the main setting. The localization ablation assigns the signed contrast throughout the trajectory, while the trajectory-cost variant moves the observation penalty from the paired contrast into the trajectory reward.

Released Agents. We evaluate the agents in Table 1 from their public checkpoints and include their shared backbone.

- **Pixel Reasoner** (Su et al., 2026). Instruction tuning on synthesized traces with zoom-in and select-frame operations, followed by RL with a curiosity-driven reward.
- **DeepEyes** (Zheng et al., 2026). End-to-end RL without a cold start, in which the model crops regions with its own grounding ability.
- **ViGoRL** (Sarch et al., 2026). RL that anchors each reasoning step to image coordinates, with a multi-turn variant that zooms into the predicted coordinates.
- **TreeVGR** (Wang et al., 2026a). RL that supervises localization and reasoning jointly through a box-level reward.
- **Mini-o3** (Lai et al., 2026). A diverse cold start followed by RL with over-turn masking, which scales visual search to tens of interaction turns.
- **Qwen2.5-VL-7B** (Bai et al., 2025b). The instruction-tuned backbone of the five agents above, without any visual-search training.

D EXISTING VLMS ON RESEE

D.1 MODELS, PROMPTS AND FULL RESULTS

Query Setup. We query the eight models of Section 5.1 through the OpenRouter API in their instruction-tuned public releases, with at most $1{,}024$ output tokens per turn, the provider's default temperature and twelve rounds. Each turn returns the requested crop, and a trajectory ends at the first answer. We exclude trajectories abandoned by the API after repeated time-outs from scoring. Table 9 reports every model under every prompt.

Prompts. All prompts except the selective one extend a common contract that defines `<look>`, `<recover>` and `<answer>` with boxes normalized to $[0, 1]$ and requires reasoning inside `<think>`. The default prompt asks the model to reason about the evidence already seen, the requirements of each remaining option and the constraint the next action should settle. The always and never prompts require at least one `<recover>` or forbid it. The random prompt tells the model, per question and with probability 0.22 matched to the annotated re-examination rate, whether it must

re-examine or decide from the crops already seen. The selective prompt uses a plain protocol description and asks the model to re-examine when the evidence does not settle the question, noting that looking back costs a turn.

Common Contract, Shared by All Prompts Except the Selective One

You are performing visual search on an image to answer a multiple-choice question. You examine the image by requesting crops; the environment returns the requested crop each round. You may also re-examine a region you have already seen.

You may reason before each action, but put all reasoning inside `<think>` tags. Never mention any tag name in your reasoning text.

Format – one action per turn:

- `<look bbox="[x0, y0, x1, y1]"/>` Request a crop. Coordinates are normalized [0,1]: 0 = left/top, 1 = right/bottom, with x0 < x1 and y0 < y1.
- `<recover pattern="..." bbox="[x0, y0, x1, y1]"/>` Go back and re-examine a region. pattern is a short reason for the recheck.
- `<answer>X</answer>` Give the final answer, where X is the chosen option letter.

Continue issuing look/recover actions until you can answer, then output `<answer>`.

Default Prompt: Explicit Reasoning, Appended to the Contract

Before each action, work through the evidence explicitly inside `<think>`: what you have already seen, what each remaining option would require, and which unresolved constraint your next action is meant to settle. Then emit the action.

Selective Prompt: Protocol Description That Replaces the Contract

You answer questions about a high-resolution image by searching it.

Think inside `<think>` and `</think>`. To inspect a region more closely, emit `<look bbox="[x0, y0, x1, y1]"/>` and stop; a crop of that region is returned to you as the next image. Coordinates are normalized to the image, between 0 and 1.

Before committing, judge whether the evidence you have already gathered settles the question. Go back and re-examine a region with `<recover pattern="..." bbox="[x0, y0, x1, y1]"/>` when it does not. Looking back costs a turn, so do it only when it would change your answer.

Put the final answer inside `<answer>` and `</answer>`.

Control Prompts: Instructions Appended to the Contract

Always Re-examine

After your first look, always go back and re-examine a region you have already seen before you answer. Take at least one `<recover>` action on every question, even when you believe you already have enough evidence.

Never Re-examine

Never go back to a region you have already examined. Do not emit `<recover>`. Decide from the crops you have already been shown and answer as soon as you can.

Random Verdict, Drawn Once per Question

A coin has been flipped for this question. You will be told below whether you are in a re-examine question or a decide-now question; follow that instruction.

With probability 0.22. This is a re-examine question: take a `<recover>` action on a region you have already seen before you answer.

With probability 0.78. This is a decide-now question: do not `<recover>`; answer from the crops you have already been shown.

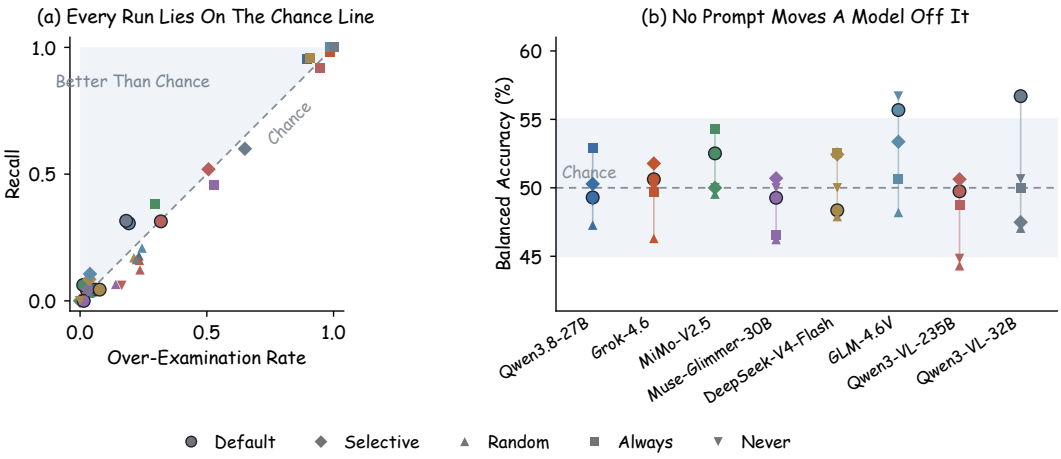


Figure 10: Eight VLMs under five prompts on 288 adjudicated decisions. Panel (a) compares recall with over-examination, with default prompts outlined. Panel (b) shows balanced accuracy against the 50% chance level.

D.2 Decision Showcase

Figure 11 shows six decisions from Section 5.1 with actions quoted from the trajectories. A re-examination requires an executed `<recover>` action. Each model and prompt is a single sampled run, and the adjudication label refers to the decision in the reference trajectory.

In case (a), Grok-4.6 initially commits to `1916` after inspecting one of two signs marked "SINCE". Re-examination of the other sign yields `1934`. In cases (b) and (d), Qwen3.8-27B revisits a region after the answer is settled, adding an unnecessary action. In cases (c), (e) and (f), additional observations overturn correct answers by shifting attention to a different region. These cases illustrate both missed warranted re-examinations and unnecessary ones.

E Case Studies of the Trained Policy

E.1 Full Trajectories for the Main-Text Cases

Figures 12 and 13 expand the two comparisons in Figure 8. Each includes the original scene, the evidence available at the shared decision state, and both policies' continuations through the final answer. Additional observation counts start at that shared state.

E.2 Additional Trajectories

Figures 14 to 17 follow four trajectories of the trained policy on the evaluation split from the first observation to the answer, two successful and two unsuccessful. Each overlay marks the requested regions in order, in blue for an observation and orange for a re-examination.

F Additional Results

Table 10 shows that pruning unwarranted demonstrations raises GRPO macro accuracy by 0.6 points, while CARE from the full cold start exceeds pruned GRPO by 1.9 points. Masking unwarranted actions while retaining their surrounding trajectories preserves most of CARE's gain, consistent with learning from search context while correcting which re-examinations to imitate.

Table 9: All model and prompt combinations on RESEE. Of the n scorable adjudicated decisions, "Pos." are warranted. "Failed" counts incomplete provider trajectories.

	Answer	Re-examination decision					Behaviour	Coverage		
Prompt	Acc. ↑	Recall ↑	Prec. ↑	Over-ex. ↓	F1 ↑	Bal. acc. ↑	Re-ex.	n	Pos.	Failed
Qwen3.8-27B										
Default	73.1	4.4	14.3	5.9	6.8	49.3	0.06	250	45	54
Selective	**75.3**	4.7	**20.0**	4.1	7.5	50.3	0.03	239	43	76
Random	70.8	17.8	14.8	23.2	16.2	47.3	0.27	243	45	66
Always	72.6	**95.6**	18.4	89.7	**30.8**	**52.9**	1.03	258	45	43
Never	71.3	0.0	0.0	**2.0**	–	49.0	0.02	243	43	65
Grok-4.6										
Default	70.2	4.0	25.0	2.8	6.9	50.6	0.02	268	50	28
Selective	69.0	5.8	**37.5**	2.2	10.0	**51.8**	0.03	278	52	14
Random	69.1	16.0	13.6	23.4	14.7	46.3	0.23	268	50	24
Always	**71.5**	**98.1**	18.3	98.7	**30.9**	49.7	1.01	282	52	12
Never	71.4	0.0	–	**0.0**	–	50.0	0.00	267	49	28
MiMo-V2.5										
Default	69.3	6.2	**33.3**	1.2	10.5	52.5	0.01	180	16	191
Selective	**69.9**	0.0	–	**0.0**	–	50.0	0.00	206	28	157
Random	66.3	3.0	12.5	4.0	4.9	49.5	0.05	210	33	126
Always	67.6	**38.2**	21.0	29.7	**27.1**	**54.3**	0.24	199	34	158
Never	69.0	0.0	–	**0.0**	–	50.0	0.00	184	28	163
Muse-Glimmer-30B										
Default	66.0	0.0	0.0	1.5	–	49.3	0.02	251	45	54
Selective	**69.9**	4.5	**25.0**	3.2	7.7	**50.7**	0.03	234	44	99
Random	66.7	6.5	9.4	14.1	7.7	46.2	0.36	252	46	63
Always	63.6	**45.8**	15.9	52.7	**23.7**	46.6	1.12	268	48	36
Never	66.4	0.0	–	**0.0**	–	50.0	0.00	246	44	71
DeepSeek-V4-Flash										
Default	65.7	4.4	11.1	7.7	6.3	48.4	0.22	252	45	63
Selective	**65.9**	8.5	**33.3**	3.6	13.6	52.4	0.11	267	47	28
Random	61.7	17.0	15.1	21.2	16.0	47.9	0.40	259	47	49
Always	61.1	**95.8**	19.0	90.7	**31.7**	**52.5**	2.56	264	48	42
Never	60.7	0.0	–	**0.0**	–	50.0	0.00	249	45	51
GLM-4.6V										
Default	53.4	30.6	26.8	19.2	28.6	55.7	0.21	262	49	37
Selective	54.3	10.6	**38.5**	**3.9**	16.7	53.4	0.05	252	47	54
Random	47.4	20.8	15.6	24.4	17.9	48.2	0.33	269	48	30
Always	**55.5**	**100.0**	18.7	98.7	**31.5**	50.7	1.65	276	51	20
Never	49.3	31.2	28.3	17.8	29.7	**56.7**	0.21	261	48	37
Qwen3-VL-235B-A22B										
Default	50.2	31.4	18.0	31.9	22.9	49.7	0.50	280	51	13
Selective	**54.3**	51.9	**19.3**	50.7	28.1	**50.6**	1.04	275	52	17
Random	52.7	12.2	10.3	23.6	11.2	44.3	0.39	269	49	29
Always	**54.3**	**92.0**	18.1	94.5	**30.3**	48.7	1.99	270	50	25
Never	53.0	6.1	7.7	**16.4**	6.8	44.8	0.27	268	49	30
Qwen3-VL-32B										
Default	44.7	31.6	**33.3**	18.2	**32.4**	**56.7**	0.11	85	19	444
Selective	**55.5**	60.0	17.1	65.0	26.7	47.5	1.83	273	50	21
Random	45.7	16.3	13.8	22.2	15.0	47.1	1.50	274	49	21
Always	46.7	**100.0**	18.5	100.0	31.2	50.0	1.72	281	52	12
Never	45.6	4.0	25.0	**2.7**	6.9	50.7	0.04	273	50	21

Table 10: Cold-start variants evaluated before and after reinforcement learning.

		Cold start only			+ GRPO			+ CARE		
Variant	Re-ex. kept	Acc. ↑	Prec. ↑	Over. ↓	Acc. ↑	Prec. ↑	Over. ↓	Acc. ↑	Prec. ↑	Over. ↓
Label-free										
Full	450	**39.3**	22.1	23.5	46.0	27.3	13.6	**48.5**	**46.8**	10.6
Stripped	0	37.0	18.2	25.4	43.9	24.5	19.1	45.1	34.7	15.3
Single-turn	450	37.9	20.2	26.0	44.5	25.2	16.5	46.0	37.8	13.1
External	0	34.7	17.9	12.0	42.8	23.7	**9.0**	44.0	32.0	**8.1**
Label-assisted										
Pruned	95	38.8	**35.0**	**10.2**	**46.6**	**37.1**	9.3	47.4	45.1	8.9
Masked	450	39.1	29.0	16.7	45.9	31.8	11.5	48.1	45.0	9.8

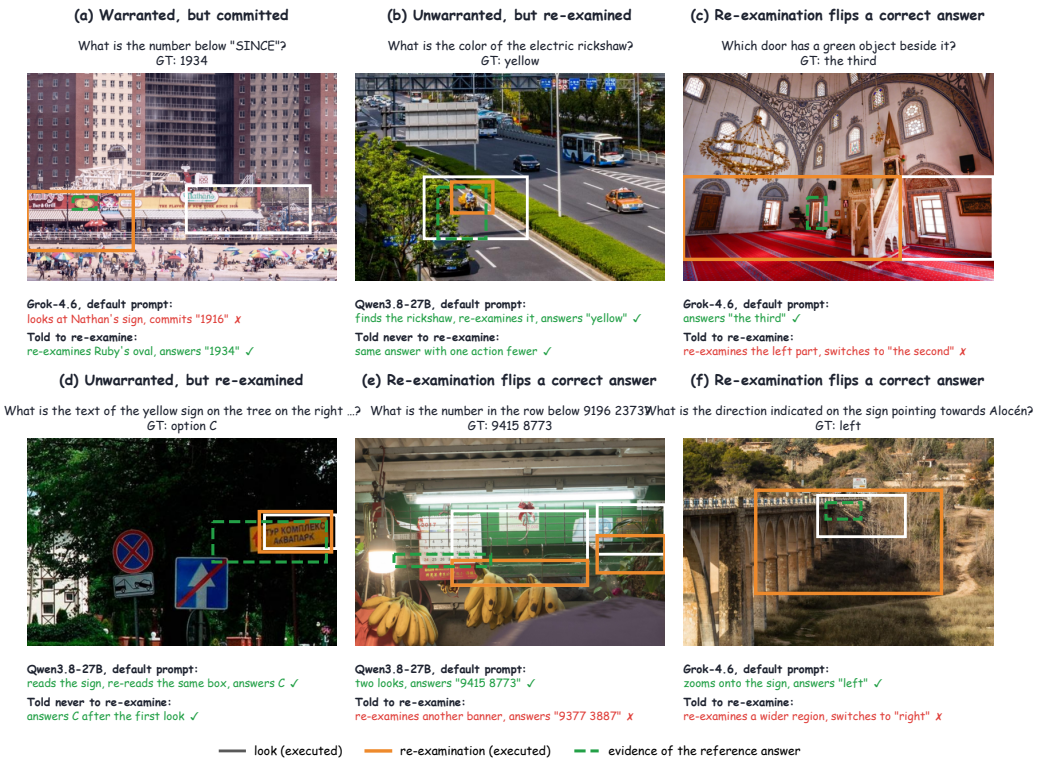
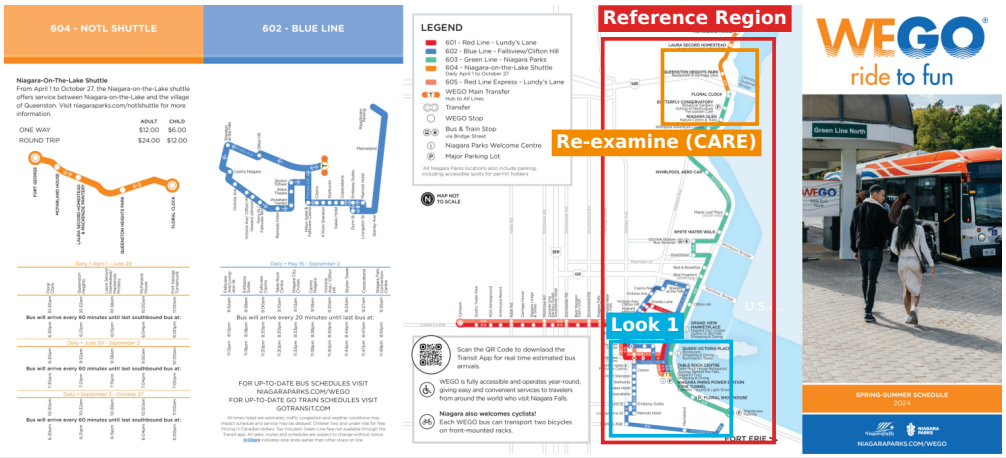


Figure 11: Six decisions from existing VLMs. White and orange boxes mark executed looks and re-examinations, while dashed green boxes mark the reference evidence.

Table 11: Per-source answer accuracy for the ablations in Table 3. Macro is the unweighted mean of the six sources.

Variant	V*	HR-Bench	MME-RW	TreeBench	O3-Bench	VisualProbe	**Macro**
CARE	**66.8**	**68.1**	**59.9**	**39.0**	**35.1**	**21.9**	**48.5**
w/o localization	65.1	66.8	57.5	36.7	31.7	17.0	45.8
w/o verdict flip	64.5	65.7	55.8	35.0	29.0	15.2	44.2
w/o cost	65.6	66.5	57.0	36.0	31.0	16.3	45.4
w/ trajectory cost	66.1	67.3	58.3	37.3	32.6	17.4	46.5
w/o verdict entropy	66.3	67.5	59.0	38.1	33.5	18.8	47.2

Route Planning: Resolving the Final Transfer



■ **QUESTION.** If I am staying at the Marriott Hotel and plan to go to Queenston Heights Park, can I reach it by transferring between buses without walking in between, and how many lines must I take at minimum? (A) No, 2. (B) No, 3. (C) Yes, 1. (D) Yes, 2. (E) Yes, 3. (F) None of the above.

■ **INITIAL SCENE.** The map separates the hotel district in the south from Queenston Heights Park in the north. A valid answer requires identifying the hotel's line, the destination's line and a connected sequence of transfer stops.

■ **SHARED OBSERVATION.** The southern crop places Marriott Hotel on the Blue Line (602). The Blue Line reaches Table Rock Centre, where it connects to the Green Line (603). This observation establishes the first transfer but leaves the destination connection unresolved.

■ **CARE VERDICT.** Re-examine the northern map before answering. The outstanding relation is whether the Green Line reaches the park or requires another transfer.

■ **CARE OBSERVATION.** The Green Line reaches Floral Clock. A transfer there connects to the Niagara-on-the-Lake Shuttle (604), which serves Queenston Heights Park. The two transfer locations connect Blue, Green and Shuttle without a walking segment.

■ **CARE ANSWER.** Commit to E, three lines. The additional observation supplies the missing transfer and completes the route.

■ **GRPO VERDICT AND ANSWER.** Commit from the shared southern observation and answer D, two lines. This continuation treats the Green Line as serving the destination and omits the shuttle connection.

■ **COMPARISON.** The reference answer is E. From the shared state, CARE adds one observation and answers correctly, while GRPO adds none and answers incorrectly.

Figure 12: Full route-planning comparison underlying Figure 8(a). Red marks the reference region, cyan the shared observation and orange the additional CARE observation.

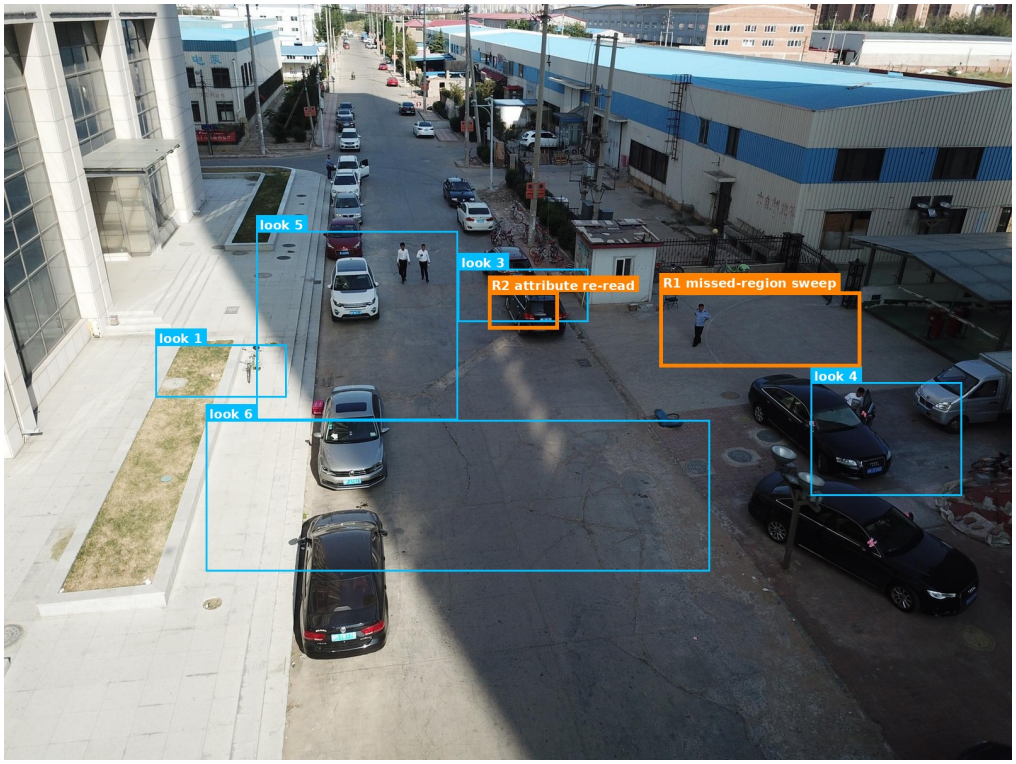
Spatial Grounding: Commitment After Both Objects Are Located



- **QUESTION.** Is the green surfboard on the left or right side of the purple umbrella? (A) Right. (B) Left.
- **INITIAL SCENE.** The beach contains many umbrellas, people and boards. The requested relation requires grounding both the green surfboard and the purple umbrella, rather than comparing arbitrary nearby objects.
- **SHARED OBSERVATION 1.** Inspect the person carrying a green surfboard near the left side of the shoreline. The crop verifies the board's identity and color. Its position is retained in the original image.
- **SHARED OBSERVATION 2.** Inspect the purple umbrella carried by the group farther right along the shoreline. The crop distinguishes this target from the many stationary pale umbrellas. Both object identities and their original horizontal order are now resolved.
- **CARE VERDICT AND ANSWER.** Commit to B, left. The green surfboard is to the left of the purple umbrella in the original scene. No additional observation is requested.
- **GRPO VERDICT AND OBSERVATION.** Re-examine the purple umbrella. The new crop returns to the same object and confirms the same color and location, adding no new relation needed for the answer.
- **GRPO ANSWER.** Commit to B, left, after the repeated inspection.
- **COMPARISON.** Both answers match the reference B. From the shared state, CARE adds no observation and GRPO adds one. The difference concerns stopping after the evidence is complete.

Figure 13: Full spatial-grounding comparison underlying Figure 8(b). Red marks the reference region, cyan the shared observations and orange GRPO's repeated inspection.

Success: counting bicycles under occlusion, nine turns



QUESTION. How many bicycles are in the image? (A) 9 (B) 22 (C) 18 (D) 16 (E) none.

PLAN. Bicycles are small next to the vehicles and buildings and may be occluded or clustered along the fences, so the left pavement, the right gate area and the cluster by the guardhouse are inspected in turn.

LOOKS 1–3. A green bicycle stands beside the pavement. Behind the metal fence at the gate, four bicycles are parked in a row and two more lean against the posts, six in that zone. By the guardhouse another five are visible, including a red and a black one. The running tally is twelve.

LOOKS 4–5. The right pavement and the left row of parked cars contain none.

RE-EXAMINE 1. *Missed-region sweep.* The gate zone was counted at six while the total is still short of every option, and the earlier crop stopped at the fence line. Re-scanning it finds a black-framed bicycle behind the silver car and another against the far wall, so that zone holds eight.

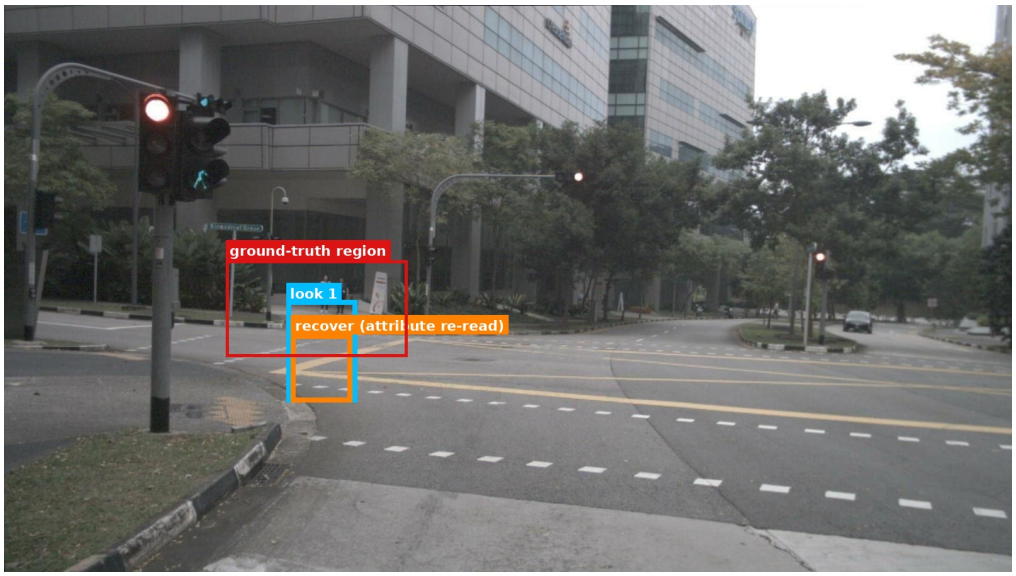
LOOK 6. The guardhouse zone is confirmed at five.

RE-EXAMINE 2. *Attribute re-read.* Fourteen are counted and the nearest options are sixteen and eighteen, so the shadowed area is read again. A bicycle is partly hidden under the rear wheel of the brown SUV and the front wheel of another shows behind a white car, which brings the count to sixteen.

ANSWER: D, 16. Both re-examinations are warranted. The tally before them would have answered fourteen, which is not among the options.

Figure 14: A counting question answered after two re-examinations.

Success: separating two motion attributes, three turns



■ **QUESTION.** The image is the front view of the ego car. Predict its behavior. (A) going straight, fast (B) steering right, very fast (C) slightly steering right, very fast (D) slightly steering left, normal speed (E) none.

■ **PLAN.** The lane position, the curvature of the markings and the motion blur carry the answer. The lateral offset already points to a rightward steer, and the gentle curve suggests a slight one, which separates option C from option B.

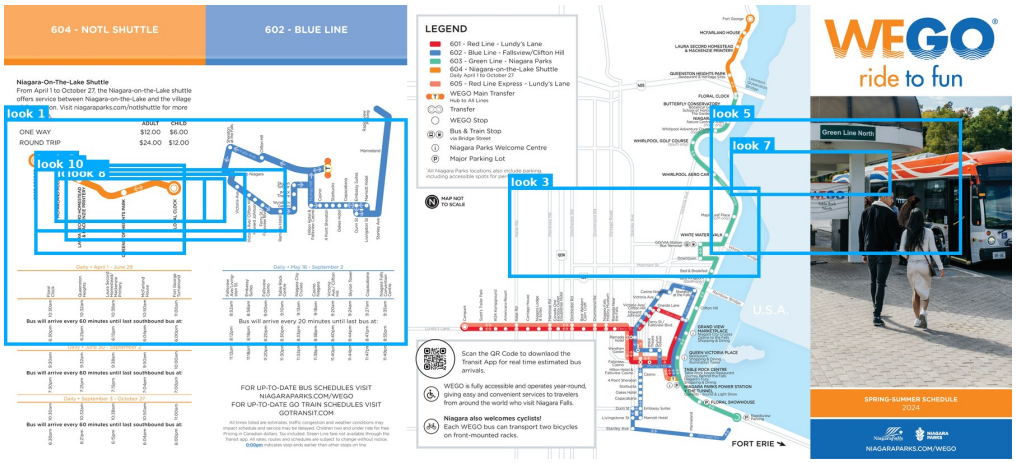
■ **DECISION.** *Verdict recover, risk medium.* One constraint is left open, whether the offset counts as a slight or a strong right steer. The two remaining options agree on the speed and differ only in that magnitude, so the region is read again before committing.

■ **RE-EXAMINE 1.** *Attribute re-read.* At higher resolution the left marking sits about half a meter from the bumper and the right one about four tenths, and the centerline is offset by roughly eight percent of the lane width. The asymmetry is small, which is mild corrective steering rather than an aggressive maneuver.

■ **ANSWER: C.** One observation localizes the lane, and the re-examination measures the quantity that separates the two surviving options.

Figure 15: A motion-attribute question settled by re-reading the lane markings.

Failure: the anchor entity is never grounded



■ **QUESTION.** Which major parking lot is nearest to the Niagara Falls History Museum? The answer is a labeled parking icon beside a labeled terminus.

■ **LOOKS 1–2.** The first crop covers the casino stops and reports that no museum label appears there, although a parking icon sits just west of one of them. The second finds a second parking icon by the bridge, and again no museum label.

■ **LOOKS 3–10.** Seven further crops sweep the central corridor, the terminal and the eastern river edge. Every observation repeats the same finding, that the museum is not labeled in the region examined.

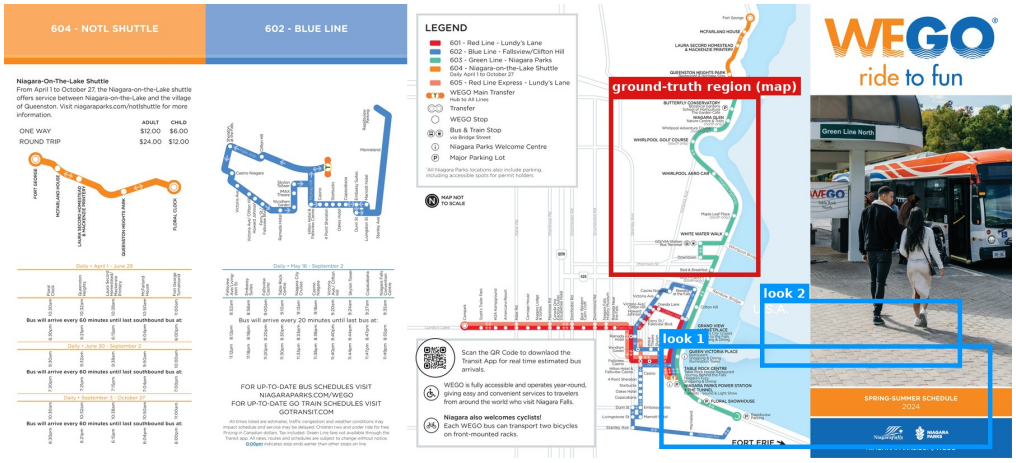
■ **DECISION.** *Verdict recover, on each of these turns.* The gate reports that evidence is still missing, and every re-examination it triggers returns to the same corridor.

■ **COMMITMENT.** The model then reasons from outside the image, stating that the museum stands at Table Rock House and that its primary parking is Rapidsview, and treats that as confirmation.

■ **ANSWER: C, INCORRECT.** The reference answer is A. Ten observations are spent confirming that a label is absent, and the question is finally resolved from prior knowledge of the city rather than from the map.

Figure 16: A failure in which the re-examinations never leave the region that lacks the evidence.

Failure: a confident commitment after the wrong region



■ **QUESTION.** A traveler starts at the Marriott Hotel and ends at Queenston Heights Park without walking between transfers. How should they travel? The reference answer is E.

■ **LOOK 1.** The crop reports the hotel labels and enumerates the stops of one line, then reasons about where two lines might meet.

■ **LOOK 2.** A transfer hub is reported where the two lines intersect, with the destination park on the second of them.

■ **DECISION.** *Verdict skip, risk low.* No constraint is left open. The transfer is described as visually confirmed, neither line is said to serve the hotel directly, and no other path is thought to exist.

■ **ANSWER: D, INCORRECT.** Both crops land on the photographic inset in the lower right of the map rather than on the route diagram, so the confirmed transfer hub and the park label are read from a region that contains neither. The gate reports low risk precisely because the mislocalization was never noticed.

Figure 17: A failure in which the model commits with low risk after reading the wrong region.

Table 12: Sensitivity of Qwen3-VL-8B to the four principal CARE parameters. Each block varies one parameter with the others at their main-experiment values.

Parameter	Value	Macro acc. ↑	Bal. acc. ↑	Looks ↓
Credit weight β	0.25	45.8	55.8	2.11
	0.5	47.3	61.0	2.04
	1	**48.5**	**65.9**	**1.97**
	2	47.8	62.4	2.03
	4	46.0	56.8	2.13
Clip κ	0.25	46.7	59.6	2.04
	0.5	47.9	63.2	1.99
	1	**48.5**	**65.9**	**1.97**
	2	47.6	61.4	2.08
Entropy η	0	47.2	61.6	2.05
	0.003	48.0	64.0	2.00
	0.01	**48.5**	**65.9**	**1.97**
	0.03	47.9	63.5	2.04
	0.1	46.1	56.7	2.16
Replays k	1	48.5	65.9	1.97
	2	48.8	66.3	1.96
	4	**49.0**	66.6	**1.95**
	8	**49.0**	**66.7**	**1.95**

Table 13: Utility ranking on 288 unseen decisions, including 52 warranted re-examinations. Higher AUC indicates better discrimination between warranted and unwarranted decisions.

Decision signal	**AUC** ↑
Self-assessed answer-flip risk	0.55
Unresolved-constraint count	0.54
CARE, $k = 1$	0.69
CARE, $k = 4$	<u>0.72</u>
CARE, $k = 8$	**0.73**

Table 14: Observation-price sweeps for Qwen3-VL-8B. Every row is separately trained. CARE prices the verdict contrast, while GRPO prices the trajectory reward.

Method	λ	Macro acc. ↑	Bal. acc. ↑	Looks ↓
CARE	0	45.4	54.7	2.14
	2×10^{-4}	**48.5**	**65.9**	1.97
	10^{-3}	<u>47.8</u>	<u>63.0</u>	1.86
	5×10^{-3}	45.1	59.6	**1.70**
GRPO	0	46.0	54.8	2.09
	2×10^{-4}	46.2	55.6	2.04
	10^{-3}	45.8	55.7	1.97
	5×10^{-3}	44.4	54.5	<u>1.73</u>