Hider–Seeker Self-Play: Geometry-Verifiable Process Rewards for Long-Horizon Visual Search

Yu Mao
Department of Computer Science
The University of Hong Kong
cici.yumao@gmail.com

Shengchao Chen*
Australian Artificial Intelligence Institute
University of Technology Sydney
shengchao.chen.uts@gmail.com

Abstract

Long-horizon visual search requires a vision-language policy to execute multi-step spatial actions over high-resolution images, with performance depending on the entire trajectory rather than any single decision. Existing supervision breaks at this horizon. Outcome-only RLVR collapses many geometric decisions into a single delayed signal and discards the per-step geometry the environment already computes, while LLM-judged step rewards recover density at the cost of verifiability, scale linearly with trajectory length, and conflate distinct failure modes. We propose HSSP, a self-play framework in which a Hider adaptively weights challenging skeletons from a fixed trajectory pool and a Seeker learns to solve them through grounded spatial actions. Instead of relying on external judges, HSSP supervises every step with a geometry-verifiable process reward built from target-directed progress and exploration coverage, with trajectory length and spatial redundancy enforced as separate Constrained Markov Decision Process (CMDP) constraints. We pair HSSP with V-TRACE, a trajectory-level companion dataset that augments skeleton seeds from established benchmarks with target-region annotations and explicit search-regime labels, directly usable by other trajectory-level training frameworks. The reward design admits closed-form guarantees on backtracking, find-versus-explore separation, and policy invariance. Extensive results on eleven challenging multimodal benchmarks, supported by controlled ablations and qualitative case studies, show that HSSP consistently outperforms advanced baselines and transfers zero-shot to OOD tasks. Code at *Link*.

1 Introduction

Long-horizon visual search casts high-resolution image understanding as a multi-step spatial trajectory [1, 2], in which errors in early actions propagate through the rollout and redundant inspection wastes the limited interaction budget. Recent vision-language model (VLM) agents [2–4] make such interaction loops practical through crop, zoom, and backtrack actions, achieving strong gains over single-pass inference on challenging benchmarks. Once these actions are available, however, the primary bottleneck shifts to supervision [5, 6], since training must assign credit to the geometry of the search process rather than to the final answer or heuristic judgments of intermediate states.

Existing multi-turn visual-search training relies on two forms of supervision. Outcome-only reinforcement learning with verifiable rewards assigns credit solely to the final answer [2, 3, 7], discarding the rich per-step geometry already computed during rollouts, such as intersection-over-union, coverage, and spatial redundancy. In contrast, process-level rewards based on a language-model judge [8, 9] provide dense per-step feedback but suffer from three limitations, namely that they are not grounded in verifiable quantities, that their cost scales with trajectory length, and that a single scalar score

*Project leader and corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

entangles distinct failure modes. In particular, redundant exploration and early mislocalization require opposite corrections, yet are merged into a single feedback signal.

Self-play curricula have delivered substantial gains in related regimes. In text-domain reasoning, agents are trained through self-play on zero-sum games [10], proposer–solver loops over code reasoning [11], and proposer–solver search [12], while in vision-language settings, single-turn models benefit from social-deduction games [13], questioner–reasoner pairs [14], and reinforcement learning on arcade-style games [15]. These successes rely on self-play to generate challenging instances and to shape training distributions adaptively. However, all such frameworks operate over discrete, non-spatial states and do not expose a grounded multi-step spatial action space. As a result, they lack access to geometric signals required for verifiable per-step credit assignment, and cannot adapt the training distribution to localization difficulty. Extending self-play to long-horizon visual search therefore requires more than an adaptive curriculum. It must couple self-play with a geometry-aware process reward that (i) provides dense, verifiable supervision without an external judge, (ii) remains policy-invariant under a non-stationary opponent, and (iii) disentangles trajectory length and repeated inspection as independent cost axes rather than one scalar.

To address this gap, we propose **Hider–Seeker Self-Play (HSSP)**, a training framework that combines adversarial self-play over grounded spatial trajectories with a *geometry-verifiable, judge-free* process reward. Each step's reward is a deterministic function of the crop history and the recorded target region. A Hider policy adaptively weights challenging skeletons from a fixed trajectory pool, while a Seeker learns to solve them through multi-step spatial actions. HSSP replaces external judges with dense supervision derived from environment-computed geometry, including target-directed progress, region coverage, and penalties for redundant search. The resulting reward preserves optimal answer policies while enabling stable training over long trajectories. In addition, the objective explicitly separates trajectory length and spatial redundancy as independent control axes, allowing exploration, verification, and backtracking to be shaped by distinct signals. At inference time, only the Seeker is deployed. Our contributions are fourfold.

- We formulate long-horizon visual search as a two-player adversarial policy-learning problem and propose HSSP, a training framework in which a Hider samples skeletons from a fixed trajectory pool and a Seeker executes the multi-step search through grounded spatial actions.
- We propose a geometry-verifiable process reward that fuses target-approach progress with exploration coverage and admits closed-form guarantees on backtracking robustness, find-versus-explore separation, and policy invariance under a non-stationary opponent.
- We build V-TRACE, a trajectory-level supervision dataset that augments skeleton seeds from established benchmarks with target-region annotations and explicit search-regime labels, verified by a full-coverage human evaluation.
- Extensive results on 11 real-world multimodal benchmarks show that HSSP outperforms strong baselines across seven method families, transfers zero-shot to OOD tasks, and concentrates its gains on the backtrack-heavy long-horizon scenarios that the theory predicts.

2 Related Work

Multi-Turn Visual Search. High-resolution visual question answering has been increasingly reformulated as an iterative localization process rather than a single forward pass. Early works [16, 17] introduced crop-and-zoom primitives that enable VLMs to focus on successive sub-regions, and subsequent systems extend this paradigm along three axes, namely longer interaction horizons via trajectory-level curricula [3], structured verifier–searcher decompositions on information-dense images [2], and interleaved reasoning across turns [18]. Building on this shared interaction loop, a parallel line of work explores alternative training signals, including tool-use rewards granted only on correct answers [5], curiosity bonuses that encourage pixel-space operations [6], and language-mediated focus refinement over adaptive crops [19]. However, the per-step spatial structure produced by the interaction loop is typically collapsed into a terminal-correctness scalar before optimization, and deriving a dense, policy-invariant training signal directly from this structure remains unresolved.

Process-Level Supervision. Process-level supervision addresses long-horizon credit assignment by assigning per-step rewards to intermediate observations, rather than relying solely on terminal correctness. Prior works instantiate this paradigm with a learned process reward model in language

reasoning [20, 21] or with a prompted LLM judge that scores intermediate states, as in VAGEN [8], which couples it with bi-level generalized advantage estimation [22]. Such judges supply dense signals but introduce three compounding limitations as trajectory length grows, namely (i) rewards reflect the judge's preferences rather than any ground-truth quantity the task itself supplies, (ii) judge inference cost scales linearly with the number of steps, and (iii) a single scalar judgment cannot independently regulate failure modes that call for opposite corrections. Deriving a process-level reward from signals that the environment determines, without an external judge, remains open.

Self-Play for Visual Search. Self-play has driven substantial gains in related domains, including zero-sum games [10], proposer–solver loops over code reasoning [11], and proposer–solver search agents [12]. In vision-language settings, it appears as social-deduction games [13], co-evolving questioner–reasoner pairs [14], and reinforcement learning on arcade-style games [15]. These frameworks use self-play to generate challenging instances and shape training distributions adaptively. However, they operate over non-spatial discrete states and lack access to grounded geometric signals, preventing verifiable per-step credit assignment and curricula tied to localization difficulty. As a result, existing self-play paradigms cannot directly support long-horizon spatial reasoning. **Table 1** situates HSSP within this landscape along three conceptual axes (i.e., multi-step spatial actions, self-play curricula, and step-level feedback), and HSSP is the only method in the table that combines all three.

Table 1: Positioning of HSSP along three conceptual axes.

Method	Multi-Step Spatial Act.	Self-Play Curriculum	Step-Level Feedback
Single-turn RLVR			
R1-OneVision [23]	✗	✗	*none*
MM-Eureka [24]	✗	✗	*none*
VLAA-Thinker [25]	✗	✗	*none*
OpenVLThinker [26]	✗	✗	*none*
Multi-turn visual search			
V*/SEAL [16]	✓	✗	*none*
Chain-of-Spot [17]	✓	✗	*none*
Mini-o3 [3]	✓	✗	*none*
Insight-o3 [2]	✓	✗	*none*
Pixel-Reasoner [6]	✓	✗	*heuristic (curiosity)*
Chain-of-Focus [19]	✓	✗	*heuristic (format)*
DeepEyes [5]	✓	✗	*heuristic (tool use)*
VAGEN [8]	✓	✗	*learned (LLM-judge)*
Self-play curricula			
Vision-Zero [13]	✗	✓	*none*
VisPlay [14]	✗	✓	*none*
SSP(text) [12]	✗	✓	*learned (answerability)*
HSSP (Ours)	✓	✓	*verifiable (geometry)*

3 V-TRACE: A Trajectory-Annotated Companion Dataset

Training a policy for long-horizon visual search requires supervision at the granularity of entire spatial trajectories rather than at the granularity of individual answers. Existing visual-search benchmarks supply images, questions, and ground-truth answers, but the multi-step rollouts that a crop-and-backtrack agent would take are not part of the annotation. We build V-TRACE, a companion dataset of trajectory-level supervision, constructed so that every trajectory carries target-region annotations that let the environment compute per-step geometry during training, and every trajectory is labeled with an explicit search regime that exposes a specific failure mode of long-horizon rollouts.

Skeletons and Sources. The unit of V-TRACE is a *skeleton* $\tau = \langle I, q, a^*, \mathbf{b}^*, \mathcal{D} \rangle$, consisting of an image I, a question q, the ground-truth answer a^*, a set $\mathbf{b}^* \subseteq [0,1]^4$ of target bounding boxes (empty for absent targets), and an optional set $\mathcal{D}$ of distractor regions. V-TRACE builds its skeleton seeds from the questions of O3-Bench [2], V*Bench [16], TreeBench [27], VisualProbe [3], HR-Bench [28], and MME-RealWorld-Lite [29]. Together the six sources span the visual-search regimes that our method targets, covering small-target localization on natural images, object-dense reasoning over spatial hierarchies, ultra-high-resolution perception, and real-world multi-scenario questioning. Three sources provide native bounding-box supervision, and the remaining three are grounded by an advanced VLM with human verification. After expansion into Type-A/B/C trajectories and multi-agent quality filtering, V-TRACE contains 10,773 accepted trajectories with per-source acceptance rates of 91.7% to 100%, and a full-coverage human evaluation rates every accepted trajectory (Appendix C.4). Details about the construction and benchmarks are in Appendix C.

Trajectory Types. Each skeleton is expanded into one or more *trajectory types* that target distinct long-horizon failure modes. Type-A (backtrack) inserts a deliberate wrong-then-correct crop sequence separated by a `<BACKTRACK>`, exposing the policy to the cost of recovering from early mislocalization. Type-B (distractor) pairs the target with a plausible wrong region of comparable salience and inserts

a VERIFY step that compares the two regions before the Seeker backtracks to the target, exposing the policy to confirmation bias. Type-C (absent) asks a counterfactual question whose answer is that the target is not present, requiring the Seeker to establish sufficient coverage before declaring absence. The type label is used only for data construction and Hider bucketing and is never supplied to the Seeker, which must infer the search regime from the rollout. Expansion, multi-agent enrichment, and quality filtering are reported in Appendix C.

4 Methodology: Hider–Seeker Self-Play

Problem Formulation. A long-horizon visual-search task is a finite-horizon MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$ in which the state encodes the input image, the conversation history, and the multiset of patches already inspected. The action space is the token-level vocabulary $\mathcal{A} = \{\langle \texttt{PLAN} \rangle, \langle \texttt{CROP}\ x_1\ y_1\ x_2\ y_2 \rangle, \langle \texttt{BACKTRACK}\ k \rangle, \langle \texttt{VERIFY} \rangle, \langle \texttt{STOP}\ y \rangle\}$ augmented with the agent's natural-language thoughts, and the unshaped primary reward is the answer-correctness indicator $\mathbb{1}[\hat{y} = a^*]$. Each episode is indexed by a skeleton $\tau = \langle I, q, a^*, \mathbf{b}^*, \mathcal{D} \rangle \in \mathcal{T}$, with $\mathcal{T}$ the V-TRACE skeleton pool over which the Hider induces a sampling distribution ρ_{π_H}. The skeleton instantiates the per-episode MDP $\mathcal{M}(\tau)$ and fixes the image, question, and target region(s). Unless stated otherwise, expectations below are taken over $\tau \sim \rho_{\pi_H}$ and rollouts $h \sim \pi_S(\cdot \mid \tau)$, with T denoting the random episode length and L a fixed comparison horizon used in theorem statements. The central difficulty is that supervision arrives only at the terminal answer, while informative geometric quantities are already computed by the environment at every intermediate step.

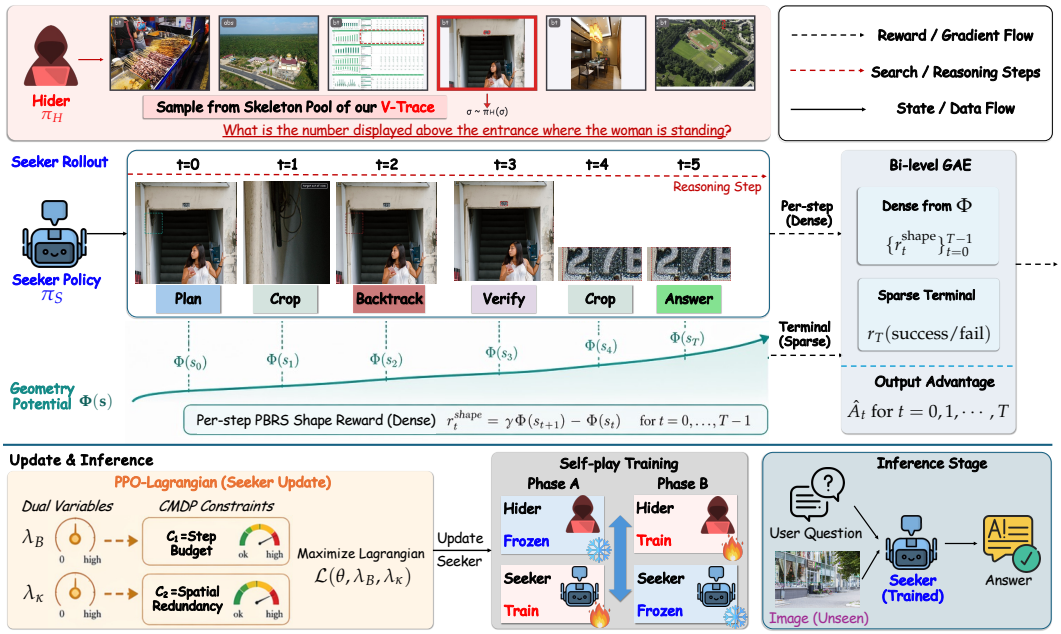


Figure 1: **Overview of HSSP.** A Hider π_H samples a skeleton τ from the V-TRACE pool, and the Seeker π_S rolls out over $\{\texttt{PLAN}, \texttt{CROP}, \texttt{BACKTRACK}, \texttt{VERIFY}, \texttt{STOP}\}$. The state-only potential Φ produces a dense canonical-PBRS shaping reward that Bi-Level GAE fuses with the sparse terminal correctness reward into per-step advantages. The Seeker is updated by PPO-Lagrangian under two CMDP constraints on step budget B and spatial redundancy κ, with dual variables λ_B and λ_κ. The two players alternate between Phase A (Seeker trains, Hider frozen) and Phase B (Hider trains, Seeker frozen), and only π_S is used at inference.

Overview. HSSP is a self-play framework for long-horizon visual search problems that couples a Hider[2], which adaptively weights challenging skeletons from a fixed trajectory pool, with a Seeker, which solves them via multi-step spatial actions (**Fig. 1**). Given a skeleton $\tau \in \mathcal{T}$ sampled by a Hider policy π_H, the Seeker policy π_S executes a search trajectory on $\mathcal{M}(\tau)$. Since the primary reward

[2]Each V-TRACE skeleton specifies a hiding configuration (target placement, distractors, or absence), and the Hider selects among these to challenge the current Seeker. The adversary thus controls the training distribution over configurations rather than generating new images or annotations online.

arrives only at termination, it gives failed localization and absent exploration identical feedback. To address this, HSSP introduces two components. First, a geometry-aware objective augments the Seeker with a state-only potential, yielding dense credit via canonical PBRS under two orthogonal CMDP constraints. Second, two-timescale updates alternate between Seeker and Hider, inducing a zero-sum game over the training distribution.

4.1 Game, Geometric Potential, and Constraints

Terminal correctness separates success from failure but cannot distinguish productive from redundant inspection. We augment the Seeker's objective with a geometry-derived shaping reward, while keeping the Hider on the zero-sum complement so that the adversarial game remains unshaped. The primary reward is defined as a zero-sum signal on answer correctness. It can be formulated as

$$r^{\text{primary}}(\tau, h) = \mathbb{1}\big[\hat{y}(h) = a^*(\tau)\big], \qquad r_H^{\text{primary}}(\tau, h) = -r^{\text{primary}}(\tau, h), \tag{1}$$

where h denotes the Seeker's rollout on $\mathcal{M}(\tau)$. The shaping reward is then canonical Potential-Based Reward Shaping [30] on a state-only max-pool potential. It can be formulated as

$$\begin{aligned} r_t^{\text{shape}} &= \gamma\, \Phi(s_{t+1}) - \Phi(s_t), \\ \Phi(s) &= \max\Big(\underbrace{\alpha \cdot \text{maxIoU}(s) + (1-\alpha) \cdot \text{cov}_{\text{tgt}}(s)}_{\text{Target-Approach Potential} \le 1}, \quad \underbrace{c \cdot \text{cov}_{\text{img}}(s)}_{\text{Exploration Potential} \le c} \Big). \end{aligned} \tag{2}$$

Here $\text{maxIoU}(s)$ denotes the running maximum intersection-over-union between cumulative inspected crops and any target box (cumulative across the rollout, retained under `BACKTRACK`, see Appendix A.1), $\text{cov}_{\text{tgt}}(s)$ the covered fraction of the target, and $\text{cov}_{\text{img}}(s)$ the covered fraction of the image under the normalization $|I| = 1$. Consequently, for Type-C absent skeletons ($\mathbf{b}^* = \varnothing$), $\text{maxIoU} = \text{cov}_{\text{tgt}} = 0$, so Φ reduces to the exploration branch $c\,\text{cov}_{\text{img}}$. Each quantity lies in $[0, 1]$ and is monotone non-decreasing along a rollout, so $\Phi(s) \in [0, 1]$ and every non-terminal shaped increment satisfies $r_t^{\text{shape}} \in [-(1-\gamma),\, \gamma]$. Under the PBRS convention $\Phi(s_T) = 0$, the last increment is $-\Phi(s_{T-1})$, and $\sum_{t=0}^{T-1} \gamma^t r_t^{\text{shape}} = -\Phi(s_0) = 0$ since no crop is inspected at s_0. Shaping therefore leaves the objective unchanged while redistributing credit across steps (Theorem 1).

The construction yields invariance under non-stationary opponents (Theorem 1), separation between target-directed progress and exhaustive exploration, and independent regularization of trajectory length and repeated inspection. The canonical form $\gamma\Phi(s_{t+1}) - \Phi(s_t)$ preserves policy invariance for every $\gamma \in (0, 1]$ (Theorem 1). Within the potential, a weighted sum would conflate target approach and exploration coverage, while the pointwise maximum keeps the two branches separate and avoids task-type conditioning. The saturation level $c < 1$ caps the exploration branch, so a prefix that explores without localizing the target cannot match the prefix shaping mass of a successful localization (Theorem 3). Finally, trajectory length and repeated inspection constitute distinct long-horizon costs, and merging them into a single scalar leads to imbalanced regularization. We therefore impose orthogonal CMDP constraints,

$$\max_{\pi_S} \ \mathbb{E}\big[r^{\text{primary}}\big] + \mathbb{E}\Big[\textstyle\sum_t \gamma^t\, r_t^{\text{shape}}\Big] \quad \text{s.t.} \quad \mathbb{E}[T] \le B, \ \ \mathbb{E}\Big[\textstyle\sum_t \text{Redun}_t\Big] \le \kappa, \tag{3}$$

where $\text{Redun}_t = |\,\text{patch}_t \cap \bigcup_{k<t} \text{patch}_k\,| / |\text{patch}_t|$ measures the overlap of a valid `CROP` with prior regions, and $\text{Redun}_t = 0$ if no valid crop is issued. The budget bounds trajectory length, while the redundancy term penalizes revisits without forbidding them.

4.2 Optimization

We approach Eq. 3 by alternating updates on separate timescales. The Seeker performs fast on-policy improvement under dense shaping, and the Hider tracks the Seeker's weaknesses on a slower clock that keeps the curriculum adversarial. The Seeker's reference objective is the Lagrangian of Eq. 3,

$$\begin{aligned} \max_{\pi_S} \min_{\lambda_B, \lambda_\kappa \ge 0} \ \mathcal{L}_S(\pi_S, \lambda_B, \lambda_\kappa) = \mathbb{E}_{\pi_S}\Big[r^{\text{primary}} + \textstyle\sum_{t=0}^{T-1} \gamma^t r_t^{\text{shape}}\Big] \\ - \lambda_B\big(\mathbb{E}_{\pi_S}[T] - B\big) - \lambda_\kappa\big(\mathbb{E}_{\pi_S}[\textstyle\sum_t \text{Redun}_t] - \kappa\big), \end{aligned} \tag{4}$$

and the Hider minimizes $\mathbb{E}[r^{\text{primary}}]$ against the current Seeker, with an entropy bonus.

Seeker Update (Phase A). We implement the Seeker as an actor-critic pair, with policy parameters θ_S and value parameters ϕ_S, and write $\pi_S = \pi_S^{\theta_S}$ if the parameterization matters. Phase A applies the standard PPO clipped surrogate to the penalized per-turn reward of Algorithm 1, which charges λ_B for each turn beyond B. Since $-\lambda_B(T-B)_+ = -\lambda_B(T-B) - \lambda_B(B-T)_+$, this cost is the budget term of Eq. 4 without its bonus for stopping early, so a search receives no budget credit for ending before B. The duals $\lambda_B, \lambda_\kappa$ perform dual ascent on the respective constraint residuals,

$$\lambda_B \leftarrow \Pi_{[0,\lambda_{\max}]}\big(\lambda_B + \eta_\lambda(\mathbb{E}[T] - B)\big), \qquad \lambda_\kappa \leftarrow \Pi_{[0,\lambda_{\max}]}\big(\lambda_\kappa + \eta_\lambda\big(\mathbb{E}[\textstyle\sum_t \mathrm{Redun}_t] - \kappa\big)\big), \quad (5)$$

where $\Pi_{[0,\lambda_{\max}]}$ projects onto the compact dual interval required by Theorem 4, and per-turn advantages are computed via Bi-Level GAE [8, 22], which applies the GAE recursion separately at the turn level (across action steps) and the token level (within each step) over the dense shape signal of Eq. 2.

Hider Update (Phase B). The Hider is a curriculum-weighting policy over the fixed V-TRACE skeleton pool. Its action is a categorical distribution over 54 buckets, indexed by $(\mathrm{source}, \mathrm{type}, \mathrm{difficulty})$ with 6 sources, 3 task types, and 3 difficulty bins. Its observation combines the windowed per-bucket Seeker success rate with the mean pre-terminal potential, the CMDP constraint residuals, and per-bucket visitation counts, and a per-bucket lower bound with entropy regularization prevents collapse to a narrow subset (details in Appendix B.1). It is updated with RL against $-r^{\mathrm{primary}}$ using role-advantage estimation [10] to offset the role-conditional information asymmetry between the two players. Phase switching is hysteretic with minimum dwell $K_{\min}$. Phase A switches to B once the Seeker's held-out accuracy on skeletons drawn from the current Hider distribution reaches τ_{hi}, and Phase B returns to A once the Hider drives it down to τ_{lo}.

Algorithm 1 (Appendix B) gives the full training loop. At inference, only the Seeker π_S is deployed, since the Hider only shapes the training distribution over skeletons. The resulting alternating primal-dual scheme induces an adversarial curriculum, provides dense geometry-verifiable credit throughout rollouts, and treats trajectory length and spatial redundancy as disentangled control axes.

4.3 Theoretical Guarantees

Theorems 1–3 are guarantees on the reward design. Theorem 1 adapts the PBRS argument of Ng et al. [30] to the Hider–Seeker game, and Theorems 2 and 3 are stated for the cumulative max-pool potential of Eq. 2. Theorem 4 adapts the two-timescale analysis of Borkar [31] and is stated for an idealized recursion with a fixed Hider and unbiased estimators. Full proofs are in Appendix A.

Theorem 1 (Policy invariance under a non-stationary opponent)**.** *Fix any Hider policy π_H and any state-only potential $\Phi : \mathcal{S} \to \mathbb{R}$ extended to terminal states by $\Phi(s_T) = 0$. Adding the canonical PBRS shape reward of Eq. 2 to the Seeker's per-step reward (and to the Seeker's reward only) leaves both the Seeker best-response set and the two-player Nash equilibrium set unchanged.*

Theorem 2 (Backtracking robustness)**.** *For any non-terminal transition $s_t \to s_{t+1}$ ($t \le T-2$) in a target-present skeleton, the max-pool potential satisfies $\Phi(s_{t+1}) \ge \Phi(s_t)$, hence $r_t^{\mathrm{shape}} \ge -(1-\gamma)\Phi(s_t) \ge -(1-\gamma)$. This extends to absent skeletons, for which $\mathrm{maxIoU} \equiv \mathrm{cov}_{\mathrm{tgt}} \equiv 0$. At the terminal step, $\Phi(s_T) = 0$ gives $r_{T-1}^{\mathrm{shape}} = -\Phi(s_{T-1})$.*

Theorem 3 (Saturation-asymmetry gap)**.** *Fix $L \in \mathbb{N}$ and a target-present skeleton. Let h_{find} and h_{miss} be length-L prefixes that end before termination, with $\Phi(s_L^{\mathrm{find}}) = 1$ and $\Phi(s_L^{\mathrm{miss}}) \le c$. Denote the discounted shaping sum by $\Sigma F(h) = \sum_{t=0}^{L-1} \gamma^t r_t^{\mathrm{shape}}$. Then*

$$\Sigma F(h_{\mathrm{find}}) - \Sigma F(h_{\mathrm{miss}}) \;\ge\; \gamma^L(1-c). \quad (6)$$

Theorem 4 (CMDP convergence under a fixed Hider, idealized SA recursion)**.** *Fix a Hider policy π_H and hold its skeleton distribution fixed, and consider the two-timescale recursion of Eq. 24 with unbiased estimators (Assumption 5). Under Borkar step-size conditions (slow dual, fast primal), Slater feasibility of the Seeker CMDP (Eq. 3), and Assumptions 1–8 (Appendix A.5), the iterates $(\theta_S^{(n)}, \lambda^{(n)})$ converge almost surely to a saddle $(\theta_S^{**}, \lambda^{**})$ satisfying complementary slackness, with equality on every coordinate i with $\lambda_i^{**} > 0$. The slower-timescale Hider is treated as quasi-static.*

5 Experiments

Baselines. We compare HSSP with advanced baselines across seven families, namely (i) base VLMs Qwen2.5-VL-7B, InternVL3-8B, and InternVL-14B, (ii) single-turn RLVR post-training

with R1-OneVision [23], MM-Eureka [24], VLAA-Thinker [25], and OpenVLThinker [26], (iii) game-based VLM self-play via Vision-Zero [13] and VisPlay [14], (iv) multi-turn visual search with Mini-o3 [3], Insight-o3 [2], DeepEyes [5], Pixel-Reasoner [6], and Chain-of-Focus [19], (v) the process-reward multi-turn baseline VAGEN [8], (vi) text-domain self-play adapted to visual search via SSP-Visual [12], an adaptation in which <CROP> replaces the text search-engine call, and (vii) proprietary references GPT-4o and Gemini-2.5-Flash. Multi-turn methods (iii)–(vi) are retrained on V-TRACE under matched backbone, cold-start, and compute (Appendix D.2) and carry the direct comparison. The single-pass methods (i)–(ii) bound base-model capability without iterative search, the proprietary models (vii) are an external reference excluded from best and runner-up marking, Appendix D.8 compares supervision regimes, and **Table 24** isolates single components.

Implementation Details. Following [13], we use Qwen2.5-VL-7B-Instruct [32] as the primary backbone, with InternVL3-8B [33] and InternVL-14B [34] as cross-architecture controls. Each backbone first undergoes SFT on 10% of V-TRACE trajectories to acquire the action vocabulary. HSSP is implemented in verl [35] with bi-level GAE [8], using FSDP for actor–critic training and vLLM [36] for rollout generation on 32×A100-80GB. We evaluate under two settings, namely *in-domain* on held-out splits of the six V-TRACE source benchmarks and *out-of-distribution* on BLINK [37], CV-Bench [38], MMMU [39], RefCOCOg [40], and MathVista [41], none of which contributes data to V-TRACE. Multi-turn methods use a shared Seeker prompt, while base VLMs and single-turn RLVR methods are evaluated in a single pass. Decoding is greedy with at most 20 actions per trajectory. Main results are three-seed means with seed standard deviations at most 1.1 points. Hyperparameters are listed in **Table 15**, and full details are in Appendix D.

5.1 Main Results

Table 2: In-domain accuracy on the six V-TRACE source benchmarks. Entries are three-seed means. Best and runner-up among open-weight methods are **bold** and underlined, with proprietary models excluded. Green Δ denotes gains of HSSP over base backbones. The rightmost column shows mean improvement over each method's base model (Qwen2.5-VL-7B unless noted, N/A if unavailable). Positive (negative) values are green (red), with the same highlighting for the top two.

Method	O3-Bench	V*	TreeBench	VisualProbe-Hard	HR-Bench 4K	MME-RW-Lite	**Mean**	**Δ Mean**
GPT-4o [42]	27.5	68.1	46.7	27.4	64.6	51.3	47.6	N/A
Gemini-2.5-Flash	49.0	72.3	49.6	18.9	74.9	55.1	53.3	N/A
Qwen2.5-VL-7B [32]	27.0	74.8	36.7	24.3	68.6	47.2	46.4	–
InternVL3-8B [33]	25.5	72.3	36.0	21.7	64.9	45.3	44.3	–
InternVL-14B [34]	28.4	73.8	38.3	24.5	69.4	48.4	47.1	–
R1-OneVision-7B [23] (ICCV'25)	27.3	74.2	36.5	24.5	68.5	47.0	46.3	−0.1
MM-Eureka-8B (InternVL2.5) [24] (arXiv'25)	25.5	71.9	35.5	22.6	64.0	45.5	44.2	N/A
VLAA-Thinker [25] (TMLR'25)	26.5	71.9	36.0	22.6	66.3	46.8	45.0	−1.4
OpenVLThinker [26] (NeurIPS'25)	27.9	74.3	37.5	24.5	68.8	47.8	46.8	+0.4
Chain-of-Focus [19] (arXiv'25)	26.0	82.0	37.5	28.0	72.5	46.2	48.7	+2.3
Pixel-Reasoner [6] (NeurIPS'25)	26.5	80.1	38.0	27.0	71.5	46.1	48.2	+1.8
DeepEyes [5] (ICLR'26)	27.5	83.1	38.0	29.6	74.0	45.5	49.6	+3.2
Mini-o3 [3] (ICLR'26)	28.6	**85.0**	38.5	**32.0**	76.0	44.1	50.7	+4.3
Insight-o3 [2] (ICLR'26)	28.8	83.4	39.5	31.1	**77.0**	43.2	50.5	+4.1
Vision-Zero [13] (ICLR'26)	28.1	75.2	37.9	24.5	68.6	47.1	46.9	+0.5
VisPlay [14] (arXiv'25)	27.6	75.2	37.7	24.5	68.5	46.9	46.7	+0.3
SSP-Visual (adapted from [12]) (ICLR'26)	27.5	75.6	36.9	24.8	68.4	46.4	46.6	+0.2
VAGEN [8] (NeurIPS'25)	30.0	77.7	38.8	25.8	70.7	46.4	48.2	+1.8
HSSP-Qwen-7B (Ours)	37.4	83.6	42.7	31.3	76.4	49.6	53.5	+7.1
Δ *vs. base (Qwen2.5-VL-7B)*	+10.4	+8.8	+6.0	+7.0	+7.8	+2.4	+7.1	
HSSP-InternVL3-8B (Ours)	35.6	80.3	41.2	29.6	72.0	47.8	51.1	+6.8
Δ *vs. base (InternVL3-8B)*	+10.1	+8.0	+5.2	+7.9	+7.1	+2.5	+6.8	
HSSP-InternVL-14B (Ours)	**39.9**	81.7	**45.0**	31.8	76.8	**52.0**	**54.5**	**+7.4**
Δ *vs. base (InternVL-14B)*	+11.5	+7.9	+6.7	+7.3	+7.4	+3.6	+7.4	

Table 2 reports in-domain accuracy on the six V-TRACE source benchmarks. HSSP attains the best six-source mean with the best results on O3-Bench, TreeBench, and MME-RW-Lite. Under a paired bootstrap ($p < 0.05$), HSSP-Qwen-7B exceeds the strongest retrained multi-turn baseline on both means (+2.8 in-domain, +3.4 OOD) and on 8 of the 11 benchmarks, with no significant difference on V*, VisualProbe-Hard, and HR-Bench 4K (**Table 17**). HSSP outperforms VAGEN on every benchmark because our geometry-verifiable max-pool potential credit-assigns per-step IoU and coverage gains to the corrective action an LLM judgment struggles to identify. The largest in-domain gains over VAGEN, on O3-Bench (+7.4) and V* (+5.9), trace to our cumulative-multiset potential, which retains progress through BACKTRACK and rewards recovery from confident-but-wrong crops, mirroring the +4.7 Type-A advantage in **Table 20**. Mini-o3 and Insight-o3 lead on small-target /

high-resolution sources (V*, VisualProbe-Hard, HR-Bench 4K) via a trajectory-length curriculum (Mini-o3) and verifier-guided search (Insight-o3), with HSSP runner-up on all three. The pattern holds under both InternVL backbones, indicating gains track our reward design, not the backbone.

5.2 Out-of-Distribution Generalization

Zero-shot accuracy on five OOD benchmarks is reported in **Table 3**. HSSP-Qwen-7B improves the OOD mean from 48.6 to 55.1, surpassing Mini-o3 by +3.4. The +10.3 gain on BLINK traces to our PBRS shaping, which trains a spatial-search behavior transferable to BLINK's perceptual primitives. CV-Bench (+7.6) and RefCOCOg (+6.2) follow from our cumulative-multiset coverage, aligned with 2D/3D spatial reasoning and box prediction. MathVista (+3.5) gains least, as math reasoning depends less on spatial search, with OpenVLThinker's math-RLVR specialization runner-up. The OOD gain (+6.5) nearly matches the in-domain gain (+7.1), indicating that our reward shapes transferable search behavior rather than benchmark memorization.

Table 3: Zero-shot OOD accuracy on five held-out benchmarks. The rightmost column reports mean improvement over each method's base model (Qwen2.5-VL-7B unless noted, `N/A` if unavailable). Positive (negative) values are shown in green (red), and best and runner-up are in **bold** and underline, respectively.

Method	BLINK	CV-Bench	MMMU	RefCOCOg	MathVista	**Mean**	**Δ Mean**
Qwen2.5-VL-7B [32]	39.7	52.9	43.8	61.3	45.2	48.6	–
InternVL3-8B [33]	37.3	50.6	42.4	58.8	44.0	46.6	–
InternVL-14B [34]	41.3	54.0	45.4	63.9	47.1	50.3	–
R1-OneVision-7B [23] (ICCV'25)	38.0	51.0	44.8	60.0	49.5	48.7	+0.1
MM-Eureka-8B (InternVL2.5) [24] (arXiv'25)	36.0	49.2	42.8	58.0	45.0	46.2	N/A
VLAA-Thinker [25] (TMLR'25)	35.5	48.5	43.0	59.0	43.5	45.9	−2.7
OpenVLThinker [26] (NeurIPS'25)	40.5	53.0	44.1	61.5	50.5	49.9	+1.3
DeepEyes [5] (ICLR'26)	45.0	57.0	44.5	64.5	45.5	51.3	+2.7
Mini-o3 [3] (ICLR'26)	46.0	57.5	45.0	64.0	46.0	51.7	+3.1
Vision-Zero [13] (ICLR'26)	33.8	50.0	43.6	57.5	44.0	45.8	−2.8
VAGEN [8] (NeurIPS'25)	43.8	56.0	44.0	66.2	47.0	51.4	+2.8
HSSP-Qwen-7B (Ours)	50.0	60.5	48.8	67.5	48.7	55.1	**+6.5**
Δ *vs. base*	+10.3	+7.6	+5.0	+6.2	+3.5	+6.5	
HSSP-InternVL3-8B (Ours)	47.5	58.0	46.8	65.5	47.1	53.0	+6.4
Δ *vs. base*	+10.2	+7.4	+4.4	+6.7	+3.1	+6.4	
HSSP-InternVL-14B (Ours)	**51.0**	**61.5**	**49.5**	**69.2**	**50.8**	**56.4**	+6.1
Δ *vs. base*	+9.7	+7.5	+4.1	+5.3	+3.7	+6.1	

5.3 Framework Analysis

Ablation. **Table 4** shows every component in HSSP is necessary, with the largest drop from the potential composition. Exploration-only ($c \cdot \mathrm{cov}_{\mathrm{img}}$) loses 3.5 since the target-approach branch carries the localization signal exploration struggles to supply. Removing PBRS shaping and replacing the max-pool with a linear-scalar merge lose comparably, so the max-pool combination, and not only the presence of shaping, carries the gain. Ablating both CMDP constraints loses 1.4 since our step-budget and redundancy duals regulate length and revisits orthogonally. The uniform-Hider drop confirms our adaptive Hider focuses training on hard skeletons. $c=1$ yields −1.4, the saturation regime in which the bound of Theorem 3 vanishes.

Table 4: HSSP-Qwen-7B ablation. Full per-source results in Appendix D.9.

Configuration	**Mean**	**Δ**
Full HSSP	**53.5**	–
Self-Play (Hider)		
uniform Hider	52.0	−1.5
frozen Hider	52.7	−0.8
Reward		
w/o PBRS shaping	51.1	−2.4
linear-scalar merge	51.4	−2.1
CMDP		
w/o duals ($\lambda_B{=}\lambda_\kappa{=}0$)	52.1	−1.4
Potential Φ		
$\Phi{=}$ maxIoU only	51.8	−1.7
$\Phi{=}c{\cdot}\mathrm{cov}_{\mathrm{img}}$ only	50.0	−3.5
sum-pool variant	53.0	−0.5
$c{=}1$ (sat. symmetry)	52.1	−1.4

Alternatives to Each Component. **Table 5** sets each design choice of HSSP against its strongest alternatives. For the reward signal, a judge-free CLIP-progress potential reaches 49.8 in-domain and 50.5 OOD, so the gain comes from geometric verifiability rather than from dense supervision as such. It also correlates less with answer correctness (Pearson 0.43 against 0.62) and nearly doubles the high-reward wrong-answer rate (14.7% against 7.8%). For the curriculum, four heuristic samplers and a matched learned sampler reach at most 52.9, so the gain stems from the adversarial objective rather than from reweighting alone. For the constraints, the best

Table 5: Each HSSP-Qwen-7B design choice against its strongest alternatives, as in-domain and OOD mean accuracy. Full results in Appendices D.8, D.10, and D.11.

	Reward signal			*Curriculum*			*Constraints*	
	Outcome only	LLM judge (VAGEN)	CLIP progress	Uniform	Best heuristic	Learned sampler	Best fixed penalty	**HSSP**
In-domain	46.4	48.2	49.8	52.0	52.6	52.9	52.8	**53.5**
OOD	50.7	51.4	50.5	53.9	54.2	54.6	54.0	**55.1**

of 16 tuned fixed penalties reaches 52.8, and its coefficients transferred to InternVL3-8B reach 49.5 against 51.1 for the duals, whose targets B and κ transfer across backbones unchanged.

Training Dynamics. **Fig. 2** shows HSSP and VAGEN training dynamics under an identical training budget. Accuracy and the pre-terminal potential $\Phi(s_{T-1})$ rise jointly under HSSP and plateau through frozen-Seeker Phase B, while the potential under VAGEN stays below the exploration saturation level c. The duals $\lambda_B, \lambda_\kappa$ peak in Phase A and then settle as the constraints are met. The repeated-crop rate falls below VAGEN's under the joint effect of the geometric potential and the κ-redundancy constraint, which regulate a revisit cost that LLM-judge supervision misses.

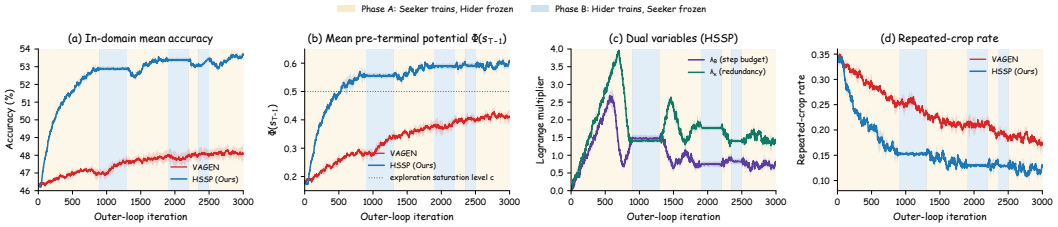


Figure 2: Training dynamics of HSSP and VAGEN over three seeds. *(a)* Mean accuracy. *(b)* Pre-terminal potential $\Phi(s_{T-1})$. *(c)* Duals $\lambda_B, \lambda_\kappa$. *(d)* Repeated-crop rate. Shading marks the Phase A/B schedule.

Hyperparameter Sensitivity. **Fig. 3** shows hyperparameter sensitivity. Accuracy stays within 0.8 points of the default for $\alpha \in [0.25, 0.75]$, $c \in [0.35, 0.65]$, every tested γ, $B \geq 8$, $\kappa \in [0.28, 0.40]$, and $K_{\min} \in [25, 100]$, and the larger drops occur toward the ends of the tested ranges, at most 3.3 points at $\alpha = 0$. The exploration saturation level c exhibits the predicted asymmetry, with the saturation gap Δ scaling linearly in $1-c$ and accuracy decreasing toward $c = 1$, consistent with Theorem 3. The step budget B saturates beyond 10, κ peaks near its default, and $K_{\min}$ has a flat optimum near 50.

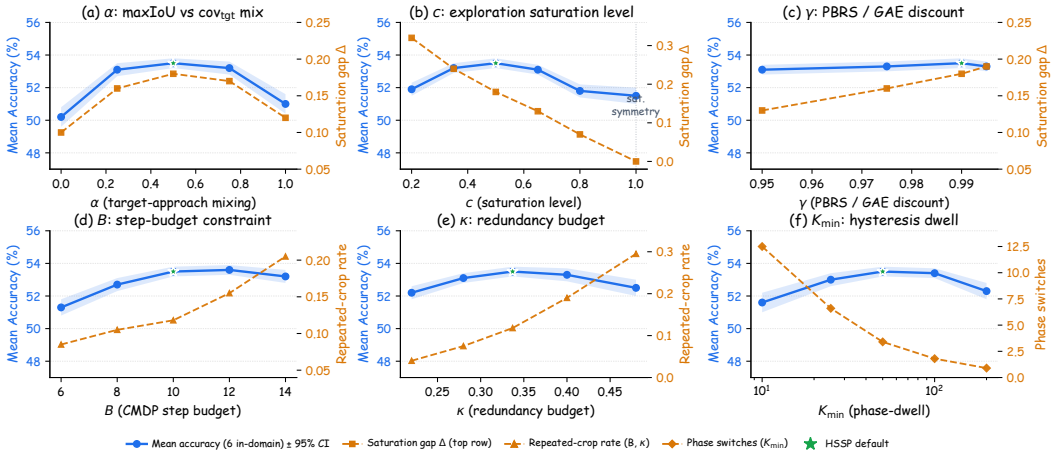


Figure 3: Hyperparameter sensitivity. Top: reward parameters α, c, γ with saturation gap Δ on the right axis. Bottom: control parameters $B, \kappa, K_{\min}$ with repeated-crop rate (or phase-switch count for $K_{\min}$).

Resolution Robustness. As shown in **Table 6**, the HSSP-Qwen-7B policy remains robust down to $0.75\times$ resolution and degrades at $0.5\times$. The largest drops occur on HR-Bench 4K, whose small

targets become unresolvable under downsampling. Upsampling to $1.5\times$ yields a negligible 0.1-point gain, indicating that the policy does not benefit from synthetic detail beyond the original capture.

Table 6: HSSP-Qwen-7B accuracy across evaluation-time image rescaling factors.

Scale	O3-B	V*	TreeB	VP-Hard	HR-4K	MME-RW	Mean
1.50× (upsample)	37.4	83.6	42.8	31.4	76.5	49.7	53.6
1.00× (default)	**37.4**	**83.6**	**42.7**	**31.3**	**76.4**	**49.6**	**53.5**
0.75×	35.6	82.7	41.7	29.9	73.7	48.4	52.0
0.50×	31.5	79.1	38.5	26.1	65.1	45.4	47.6
0.25×	25.7	69.8	32.8	19.5	50.6	40.2	39.8

Robustness to Box Source and Target Category. The boxes behind Φ enter training only, and the deployed Seeker uses no boxes. With unverified Qwen3.5 grounding on all six sources or Grounding-DINO detections in their place, HSSP still exceeds Mini-o3, the strongest retrained multi-turn baseline (**Table 7**, left). The geometric terms never reference target identity. With animals, vehicles, and text or signage removed from training, HSSP still gains $+6.4$ on these categories, 86% of the size-matched random-removal gain (right), so the learned search policy is category-agnostic.

Table 7: HSSP-Qwen-7B with training-time boxes from alternative sources (left, IoU against human boxes) and on target categories held out from training (right). Full results in Appendices C.5 and D.12.

Box source	IoU	In-dom.	OOD
Annotated (default)	0.92	53.5	55.1
VLM grounding	0.78	51.7	53.6
Open-vocab. detector	0.69	50.9	52.8
Mini-o3 (baseline)	–	50.7	51.7

Training data	Acc.	Δ vs. base
Base model	41.5	–
Full training pool	49.3	+7.8
w/o random subset	48.9	+7.4
w/o target categories	47.9	+6.4

Empirical Validation of the Saturation Gap. **Fig. 4** compares target-approach (Type-A) and exhaustive-explore (Type-C) rollouts on V-TRACE. The potential of the former climbs toward the target-branch upper bound, while the latter levels off at the exploration saturation level $c=0.5$, and the class-wise telescoped mass differs by $\Delta = 0.18$. Theorem 3 compares two prefixes on the same target-present skeleton, so the figure illustrates the asymmetry it formalizes rather than testing its constant. Since $\Phi(s_0) = 0$, the prefix shaping mass equals $\gamma^L \Phi(s_L)$ (Eq. 20), and the closed form $\gamma^L(1-c)$ is the margin reached once a find prefix attains $\Phi = 1$, which the target-approach class approaches without reaching. The separation stays strictly positive, and exhaustive coverage does not substitute for localization.

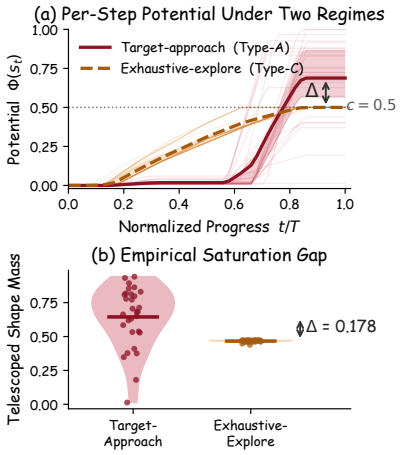


Figure 4: Empirical saturation gap: (a) $\Phi(s_t)$ over t/T (60 rollouts, mean/IQR), (b) class-wise telescoped mass, gap Δ.

The appendix further reports representative trajectories (**Fig. 6**), per-type decomposition (**Table 20**), failure modes (**Fig. 9**), reward-integrity analysis (Appendix D.7), compute-cost breakdown (**Table 16**), and case studies (**Appendix E**). It also covers Hider training dynamics (**Fig. 5**), test-time constraint adherence (**Table 18**), and the Seeker prompt (**Appendix D.4**).

6 Conclusion

We propose HSSP, a Hider–Seeker self-play framework with *judge-free, geometry-verifiable* process rewards for long-horizon visual search. By combining a state-only max-pool potential with PBRS and separate CMDP constraints on budget and redundancy, HSSP provides dense policy-invariant credit that favors finding the target over exhaustive exploration. Across eleven benchmarks, HSSP attains the best in-domain and zero-shot OOD means and outperforms advanced visual-search baselines on both means under the same training data and compute allocation. V-TRACE makes trajectory-level geometric supervision available to other training frameworks.

Acknowledgments and Disclosure of Funding

This work received no external funding. The authors declare no competing interests.

References

[1] Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. DyFo: A training-free dynamic focus visual search for enhancing LMMs in fine-grained visual understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9098–9108. IEEE, 2025.

[2] 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.

[3] 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, pages 76722–76746, 2026.

[4] Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang. ScreenAgent: A vision language model-driven computer control agent. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, page 6433–6441. International Joint Conferences on Artificial Intelligence Organization, August 2024. doi: 10.24963/ijcai.2024/711. URL `http://dx.doi.org/10.24963/ijcai.2024/711`.

[5] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

[6] 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.

[7] Fang Wu, Weihao Xuan, Danny Xing, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why RLVR may or may not escape its origin, 2026. URL `https://arxiv.org/abs/2507.14843`.

[8] Kangrui Wang, Pingyue Zhang, Zihan Wang, Yaning Gao, Linjie Li, Qineng Wang, Hanyang Chen, Yiping Lu, Zhengyuan Yang, Lijuan Wang, et al. VAGEN: Reinforcing world model reasoning for multi-turn VLM agents. *Advances in Neural Information Processing Systems*, 38: 172871–172933, 2026.

[9] Md Tahmid Rahman Laskar, Mohammed Saidul Islam, Ridwan Mahbub, Ahmed Masry, Mizanur Rahman, Amran Bhuiyan, Mir Tafseer Nayeem, Shafiq Joty, Enamul Hoque, and Jimmy Xiangji Huang. Judging the judges: Can large vision-language models fairly evaluate chart comprehension and reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1203–1216, 2025.

[10] Bo Liu, Simon Yu, Zichen Liu, Leon Guertler, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, et al. SPIRAL: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pages 35407–35434, 2026.

[11] Andrew Zhao, Yiran Wu, Tong Wu, Quentin Xu, Yang Yue, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute Zero: Reinforced self-play reasoning with zero data. *Advances in Neural Information Processing Systems*, 38:105816–105879, 2026.

[12] Hongliang Lu, Yuhang Wen, Pengyu Cheng, Ruijin Ding, Jiaqi Guo, Haotian Xu, Chutian Wang, Haonan Chen, et al. Search self-play: Pushing the frontier of agent capability without supervision. In *International Conference on Learning Representations*, volume 2026, pages 65132–65165, 2026.

[13] Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, and Wentian Zhao. Vision-Zero: Scalable VLM self-improvement via strategic gamified self-play, 2026. URL https://arxiv.org/abs/2509.25541.

[14] Yicheng He, Chengsong Huang, Zongxia Li, Jiaxin Huang, and Yonghui Yang. VisPlay: Self-evolving vision-language models from images, 2025. URL https://arxiv.org/abs/2511.15661.

[15] Yunfei Xie, Yinsong Ma, Shiyi Lan, Alan Yuille, Junfei Xiao, and Chen Wei. Play to generalize: Learning to reason through game play. In *International Conference on Learning Representations*, volume 2026, pages 17213–17243, 2026.

[16] 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)*, pages 13084–13094. IEEE, 2024.

[17] Zuyan Liu, Yuhao Dong, Yongming Rao, Jie Zhou, and Jiwen Lu. Chain-of-Spot: Interactive reasoning improves large vision-language models, 2024. URL https://arxiv.org/abs/2403.12966.

[18] Ye Wang, Qianglong Chen, Zejun Li, Siyuan Wang, Shijie Guo, Zhirui Zhang, and Zhongyu Wei. Simple o3: Towards interleaved vision-language reasoning, 2025. URL https://arxiv.org/abs/2508.12109.

[19] Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, et al. Adaptive Chain-of-Focus reasoning via dynamic visual search and zooming for efficient VLMs. *arXiv preprint arXiv:2505.15436*, 2025.

[20] 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, pages 39578–39601, 2024.

[21] 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)*, pages 9426–9439, 2024.

[22] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2018. URL https://arxiv.org/abs/1506.02438.

[23] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-Onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2376–2385. IEEE, 2025.

[24] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning, 2025. URL https://arxiv.org/abs/2503.07365.

[25] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. SFT or RL? An early investigation into training R1-like reasoning large vision-language models, 2025. URL https://arxiv.org/abs/2504.11468.

[26] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. OpenVL-Thinker: Complex vision-language reasoning via iterative SFT-RL cycles. *Advances in Neural Information Processing Systems*, 38:123817–123846, 2026.

[27] 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, pages 148769–148794, 2026.

[28] 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, pages 7907–7915, 2025.

[29] 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, pages 89655–89701, 2025.

[30] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287. Citeseer, 1999.

[31] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 2. Springer, 2008.

[32] 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, 2025. URL `https://arxiv.org/abs/2502.13923`.

[33] 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`.

[34] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024.

[35] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025.

[36] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

[37] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer, 2024.

[38] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

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

[40] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016.

[41] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pages 23439–23554, 2024.

[42] OpenAI. Hello GPT-4o. `https://openai.com/index/hello-gpt-4o/`, 2024. Announcement.

[43] Michał Borowski and Błażej Miasojedow. Convergence of projected stochastic approximation algorithm. *arXiv preprint arXiv:2501.08256*, 2025.

[44] Harold J Kushner and G George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer, 2003.

[45] Christoph Zauner. Implementation and benchmarking of perceptual image hash functions. 2010. URL `https://api.semanticscholar.org/CorpusID:17075066`.

[46] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024.

[47] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003. IEEE, 2023.

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021.

Appendix Contents

A Theorems and Proofs **16**

A.1 Notation . . . 16

A.2 Policy Invariance under a Non-Stationary Opponent . . . 16

A.3 Backtracking Robustness . . . 17

A.4 Saturation-Asymmetry Gap . . . 18

A.5 Seeker-Side CMDP Convergence under Fixed Hider . . . 19

B HSSP Algorithm **22**

B.1 Hider Policy Specification . . . 22

B.2 Hider Training Dynamics . . . 23

C V-TRACE Construction Details **24**

C.1 Source Datasets . . . 24

C.2 Skeleton Synthesis . . . 25

C.3 Multi-Agent Enrichment and Validation . . . 25

C.4 Human Evaluation and Annotation . . . 29

C.5 Reliability of Training-Time Bounding Boxes . . . 30

D Supplementary Experimental Details **34**

D.1 Evaluation Datasets . . . 34

D.2 Baselines . . . 34

D.3 Training and Hyperparameter Settings . . . 36

D.4 Inference-Time Seeker Prompt . . . 38

D.5 Per-Trajectory-Type Decomposition . . . 40

D.6 Failure-Mode Taxonomy . . . 41

D.7 Reward-Integrity Analysis . . . 42

D.8 Judge-Free Non-Geometric Progress Control . . . 42

D.9 Full Six-Source Ablation Matrix . . . 43

D.10 Alternative Hider Samplers . . . 44

D.11 Fixed Penalties in Place of the Duals . . . 44

D.12 Transfer to Unseen Target Categories . . . 45

E Case Studies **45**

E.1 HSSP Rollouts on V-TRACE Sources . . . 45

E.2 HSSP Rollouts on Out-of-Distribution Benchmarks . . . 55

E.3 HSSP versus Single-Pass RLVR and VAGEN . . . 57

F Limitations and Broader Impacts **60**

A Theorems and Proofs

Detailed proofs of Theorems 1–4 follow. Each proof is preceded by an unnumbered restatement that makes every scope condition and regularity assumption explicit. Theorems 1 and 4 extend classical results of Ng et al. [30] and Borkar [31] to the Hider–Seeker setting, and the preconditions of those results are verified at the point of each invocation.

A.1 Notation

A *skeleton* $\tau = \langle I, q, a^*, \mathbf{b}^*, \mathcal{D} \rangle$ with target-box set $\mathbf{b}^* = \{b_1, \ldots, b_m\} \subseteq [0,1]^4$ parameterizes the per-instance MDP $\mathcal{M}(\tau) = \langle \mathcal{S}_\tau, \mathcal{A}, P_\tau, r_\tau, \gamma \rangle$. Trajectories in $\mathcal{M}(\tau)$ are denoted $h = (s_0, a_0, s_1, \ldots, a_{T-1}, s_T)$, the Seeker's terminal answer by $\hat{y}(h)$, and the primary reward by $r^{\text{primary}}(\tau, h) = \mathbb{1}[\hat{y}(h) = a^*]$. Episode length T is finite and pathwise bounded by the rollout turn limit $T \le T_{\max} = 20$ (**Table 15**), which gives the noise bound in Theorem 4.

Each state s_t exposes three aggregate geometric functionals. Write the cumulative multiset of inspected patches up to but not including step t as $C_t = \{\text{patch}_k : a_k = \texttt{CROP},\ 0 \le k < t\}$, with $C_0 = \varnothing$, so the action-induced update reads $C_{t+1} = C_t \cup \{\text{patch}_t\}$ if $a_t = \texttt{CROP}$ and $C_{t+1} = C_t$ otherwise. <BACKTRACK> retires patches from the active context stack but never removes them from the cumulative record. The running maximum IoU of inspected crops with any target is

$$\text{maxIoU}(s_t) \;=\; \max_{p \in C_t} \max_{b \in \mathbf{b}^*} \text{IoU}(p, b), \tag{7}$$

and the cumulative target and image coverages are

$$\text{cov}_{\text{tgt}}(s_t) \;=\; \frac{\left|\bigcup_{p \in C_t} p \,\cap\, \bigcup_{b \in \mathbf{b}^*} b\right|}{\left|\bigcup_{b \in \mathbf{b}^*} b\right|}, \qquad \text{cov}_{\text{img}}(s_t) \;=\; \frac{\left|\bigcup_{p \in C_t} p\right|}{|I|}. \tag{8}$$

We work throughout in normalized image coordinates with $|I| = 1$, and every target box and every issued crop is an axis-aligned rectangle of positive area inside the image (a CROP with invalid geometry is a no-op and leaves C_t unchanged). A maximum over an empty set is taken as 0, so $C_0 = \varnothing$ gives $\text{maxIoU}(s_0) = \text{cov}_{\text{tgt}}(s_0) = \text{cov}_{\text{img}}(s_0) = 0$ and $\Phi(s_0) = 0$ on every skeleton. *Empty-target convention.* On Type-C skeletons with $\mathbf{b}^* = \varnothing$, both $\text{maxIoU}(\cdot)$ and $\text{cov}_{\text{tgt}}(\cdot)$ are defined to be identically 0 (the zero-by-zero coverage ratio is replaced by 0). The potential then reduces to $\Phi(s) = c\,\text{cov}_{\text{img}}(s)$ on absent skeletons. Theorem 3 concerns target-present skeletons ($\mathbf{b}^* \neq \varnothing$). Since all three geometric quantities lie in $[0,1]$, the max-pool potential of Eq. 2,

$$\Phi(s) \;=\; \max\big(\alpha\ \text{maxIoU}(s) + (1-\alpha)\ \text{cov}_{\text{tgt}}(s),\ c\ \text{cov}_{\text{img}}(s)\big) \;\in\; [0,1], \tag{9}$$

is well-defined with $\alpha \in [0,1]$ and $c \in (0,1)$ (the open upper endpoint makes the saturation gap of Theorem 3 strictly positive, while $c = 1$ is the saturation-symmetric boundary discussed as an ablation). The canonical PBRS shaping reward of Eq. 2 is $r_t^{\text{shape}} = \gamma\Phi(s_{t+1}) - \Phi(s_t)$. We adopt the standard finite-horizon PBRS terminal convention $\Phi(s_T) = 0$, equivalently treating the post-terminal absorbing state as zero-potential, under which the discounted telescoped shape return of a length-T trajectory becomes

$$\Sigma F(h) \;=\; \sum_{t=0}^{T-1} \gamma^t r_t^{\text{shape}} \;=\; \gamma^T \Phi(s_T) - \Phi(s_0) \;=\; -\Phi(s_0). \tag{10}$$

The terminal convention $\Phi(s_T) = 0$ is the finite-horizon analogue of the absorbing-state convention used in the original PBRS proof, and it makes the shaped Seeker objective a policy-independent shift of the unshaped objective in expectation. The convention overrides the geometric value only at s_T. Earlier shaping rewards use the geometric potential at both endpoints, and the last is $r_{T-1}^{\text{shape}} = -\Phi(s_{T-1}) \in [-1, 0]$. Because $\Phi(s_0) = 0$, Eq. 10 gives $\Sigma F(h) = 0$, so $J_{\pi_H}^{\text{shaped}} = J_{\pi_H}^{\text{primary}} \in [0,1]$.

We write $J_{\pi_H}^{\text{primary}}(\pi_S) = \mathbb{E}_{\tau \sim \rho_{\pi_H}} \mathbb{E}_{h \sim \pi_S | \tau}[r^{\text{primary}}(\tau, h)]$ and $J_{\pi_H}^{\text{shaped}}(\pi_S) = J_{\pi_H}^{\text{primary}}(\pi_S) + \mathbb{E}_{\tau,h}[\Sigma F(h)]$ for the unshaped and shaped Seeker objectives, and the Hider maximizes $-J_{\pi_H}^{\text{primary}}$.

A.2 Policy Invariance under a Non-Stationary Opponent

Theorem (1 restated with explicit scope)**.** *Fix any Hider policy π_H with induced skeleton distribution ρ_{π_H}, and any state-only potential $\Phi : \mathcal{S} \to \mathbb{R}$ extended to terminal states by the convention*

$\Phi(s_T) = 0$. *Assume that (i) the shaping reward ΣF enters the Seeker's objective but not the Hider's, and (ii) the Seeker best-response set* $\arg\max_{\pi_S} J^{\text{primary}}_{\pi_H}(\pi_S)$ *and the Hider best-response set* $\arg\min_{\pi_H} J^{\text{primary}}_{\pi_H}(\pi_S)$ *over the admissible policy class are non-empty (a standard compactness and continuity condition on the admissible policy class, which makes the arg-extrema well defined). Then*

$$\arg\max_{\pi_S} J^{\text{shaped}}_{\pi_H}(\pi_S) \;=\; \arg\max_{\pi_S} J^{\text{primary}}_{\pi_H}(\pi_S), \tag{11}$$

and the Nash equilibrium sets of the shaped and unshaped two-player games coincide.

Proof of Theorem 1. Fix a skeleton $\tau \in \text{supp}(\rho_{\pi_H})$ and the per-instance MDP $\mathcal{M}(\tau)$ with initial state s_0^τ. The shaped Seeker return on a single rollout h is $r^{\text{primary}}(\tau, h) + \Sigma F(h)$. By the canonical-PBRS telescoping identity, valid for any state-only potential Φ on a finite-horizon trajectory,

$$\Sigma F(h) \;=\; \sum_{t=0}^{T-1} \gamma^t \big[\gamma\Phi(s_{t+1}) - \Phi(s_t)\big] \;=\; \gamma^T \Phi(s_T) - \Phi(s_0), \tag{12}$$

which under the terminal convention $\Phi(s_T) = 0$ collapses to $\Sigma F(h) = -\Phi(s_0^\tau)$. The right-hand side depends on the skeleton-determined initial state s_0^τ but not on the Seeker policy, so taking expectations under any rollout distribution yields

$$\mathbb{E}_{h\sim\pi_S|\tau}\big[r^{\text{primary}}(\tau, h) + \Sigma F(h)\big] \;=\; \mathbb{E}_{h\sim\pi_S|\tau}\big[r^{\text{primary}}(\tau, h)\big] \;-\; \Phi(s_0^\tau), \tag{13}$$

matching the absorbing-state shift of Theorem 1 of [30] under the terminal convention.

Integrating Eq. 13 against ρ_{π_H} gives

$$J^{\text{shaped}}_{\pi_H}(\pi_S) \;=\; J^{\text{primary}}_{\pi_H}(\pi_S) \;-\; C(\rho_{\pi_H}), \qquad C(\rho_{\pi_H}) := \mathbb{E}_{\tau\sim\rho_{\pi_H}}\big[\Phi(s_0^\tau)\big], \tag{14}$$

where $C(\rho_{\pi_H})$ depends on the Hider distribution and on Φ but not on π_S. Adding a π_S-independent constant preserves the set of maximizers of a real-valued function over π_S, hence

$$\arg\max_{\pi_S} J^{\text{shaped}}_{\pi_H}(\pi_S) \;=\; \arg\max_{\pi_S} J^{\text{primary}}_{\pi_H}(\pi_S), \tag{15}$$

which proves the first statement.

For the Nash equilibrium statement, let (π_H^*, π_S^*) be a Nash equilibrium of the unshaped game, so that $\pi_S^* \in \arg\max_{\pi_S} J^{\text{primary}}_{\pi_H^*}(\pi_S)$ and $\pi_H^* \in \arg\min_{\pi_H} J^{\text{primary}}_{\pi_H}(\pi_S^*)$. Substituting $\pi_H = \pi_H^*$ into Eq. 15 shows that π_S^* also best-responds to π_H^* under the shaped reward. The zero-sum Hider objective $-J^{\text{primary}}_{\pi_H}(\pi_S)$ contains no shape term and is identical between shaped and unshaped games, so the argmin set is unchanged and π_H^* remains a best response to π_S^*. The symmetric argument shows any Nash equilibrium of the shaped game is also a Nash equilibrium of the unshaped one, so the two equilibrium sets coincide. □

Scope conditions of Theorem 1. Condition (i) holds by construction, since $r_H = -r^{\text{primary}}$ (Appendix B.1), and it is sufficient for the theorem. A Hider rewarded on ΣF would instead face the ρ_{π_H}-dependent offset $C(\rho_{\pi_H})$ of Eq. 14, which can change its best response if the initial potential depends on the skeleton. The potential of Eq. 2 instead has $\Phi(s_0) = 0$ and no offset. The entropy bonus of the implemented Hider depends only on its own sampling distribution, so it enters the shaped and unshaped games identically and the equilibrium statement carries over to the entropy-regularized game.

A.3 Backtracking Robustness

Theorem (2 restated with explicit scope)**.** *Fix a target-present skeleton ($\mathbf{b}^* \neq \varnothing$) and let Φ be the max-pool potential of Eq. 2, with* $\text{maxIoU}, \text{cov}_{\text{tgt}}, \text{cov}_{\text{img}}$ *evaluated on the cumulative patch multiset C_t of Eq. 7–8 and parameters $\alpha \in [0, 1]$, $c \in (0, 1)$. Then $\Phi(s_{t+1}) \geq \Phi(s_t)$ for every non-terminal transition ($0 \leq t \leq T-2$), and the PBRS shaping reward of these steps obeys*

$$r_t^{\text{shape}} \;=\; \gamma\Phi(s_{t+1}) - \Phi(s_t) \;\geq\; -(1-\gamma)\Phi(s_t) \;\geq\; -(1-\gamma). \tag{16}$$

The conclusion extends to absent skeletons ($\mathbf{b}^* = \varnothing$) by the same argument applied to the exploration branch alone under the empty-target convention of Appendix A.1. At the terminal step the convention $\Phi(s_T) = 0$ gives $r_{T-1}^{\text{shape}} = -\Phi(s_{T-1}) \in [-1, 0]$, so $r_t^{\text{shape}} \in [-1, \gamma]$ along every trajectory.

Proof of Theorem 2. The action semantics of Section 4 induce three kinds of transition $s_t \to s_{t+1}$. If $a_t = \texttt{CROP}$ with inspected patch patch_t, we have $C_{t+1} = C_t \cup \{\mathrm{patch}_t\}$. If $a_t = \texttt{BACKTRACK}$, patches leave the active context stack but stay in the cumulative record, so $C_{t+1} = C_t$. The other actions (`PLAN`, `VERIFY`, `STOP`) issue no patch, so $C_{t+1} = C_t$. In all three cases $C_t \subseteq C_{t+1}$ as multisets.

Each of the three geometric quantities in Eq. 7–8 is monotone non-decreasing as a set functional on C. For maxIoU, adding elements enlarges the index set over which the maximum is taken, so $\max_{p\in C, b\in \mathbf{b}^*} \mathrm{IoU}(p,b)$ is non-decreasing in C. For $\mathrm{cov}_{\mathrm{tgt}}$, the inspected region $R(C) := \bigcup_{p\in C} p$ satisfies $R(C) \subseteq R(C')$ for all $C \subseteq C'$, and intersection with the fixed target union $\bigcup_b b$ preserves inclusion. Because Lebesgue area is monotone under set inclusion and the normalizing area $|\bigcup_b b|$ is a positive skeleton-fixed constant, $\mathrm{cov}_{\mathrm{tgt}}$ inherits the same monotonicity. The argument for $\mathrm{cov}_{\mathrm{img}}$ is identical up to the removal of the target intersection.

The inner target-approach potential $\Phi_1(s) := \alpha\,\mathrm{maxIoU}(s) + (1-\alpha)\,\mathrm{cov}_{\mathrm{tgt}}(s)$ is a non-negative linear combination of non-decreasing quantities with non-negative coefficients $\alpha, 1-\alpha \in [0,1]$, and is therefore non-decreasing. The exploration potential $\Phi_2(s) := c\,\mathrm{cov}_{\mathrm{img}}(s)$ is a non-negative scalar multiple of a non-decreasing quantity. Pointwise maximum preserves monotonicity, since for any f_1, f_2 with $f_i(s_t) \le f_i(s_{t+1})$ we have $f_i(s_t) \le f_i(s_{t+1}) \le \max(f_1,f_2)(s_{t+1})$ for both i, so $\max(f_1,f_2)(s_t) \le \max(f_1,f_2)(s_{t+1})$. Applied to Φ_1 and Φ_2 this yields $\Phi(s_t) \le \Phi(s_{t+1})$.

The shape-reward bound is then immediate. Using $\Phi(s_{t+1}) \ge \Phi(s_t)$ and $\gamma \in (0,1]$,

$$r_t^{\mathrm{shape}} \;=\; \gamma\Phi(s_{t+1}) - \Phi(s_t) \;\ge\; \gamma\Phi(s_t) - \Phi(s_t) \;=\; -(1-\gamma)\Phi(s_t). \tag{17}$$

Because $\mathrm{maxIoU}, \mathrm{cov}_{\mathrm{tgt}}, \mathrm{cov}_{\mathrm{img}} \in [0,1]$ we have $\Phi_1 \in [0,1]$ (convex combination of $[0,1]$ values) and $\Phi_2 \in [0,c] \subseteq [0,1]$, so $\Phi \in [0,1]$ and the state-independent bound $r_t^{\mathrm{shape}} \ge -(1-\gamma)$ follows for every non-terminal step, together with $r_t^{\mathrm{shape}} \le \gamma\Phi(s_{t+1}) \le \gamma$. The terminal step is not covered by this argument, because the convention $\Phi(s_T) = 0$ replaces the geometric value there, and it contributes $r_{T-1}^{\mathrm{shape}} = -\Phi(s_{T-1}) \in [-1,0]$. □

Cumulative crops as the domain of Φ. If Φ were defined over the *active* crop stack, a `<BACKTRACK>` could retire patches that contributed to maxIoU, $\mathrm{cov}_{\mathrm{tgt}}$, or $\mathrm{cov}_{\mathrm{img}}$, lowering Φ and violating Theorem 2. The cumulative record keeps Φ monotone and leaves the Seeker's observation unchanged.

A.4 Saturation-Asymmetry Gap

Telescoping in Theorems 1 and 3. The two telescoping identities used in this appendix differ in their stopping convention. Theorem 1 applies the finite-horizon absorbing convention $\Phi(s_T) = 0$, under which the full-episode identity Eq. 10 collapses to $\Sigma F(h) = -\Phi(s_0)$, a π_S-independent shift. Theorem 3 fixes a *prefix* horizon $L < T$ before any termination and uses no terminal convention, so the prefix identity $\Sigma F(h) = \gamma^L\Phi(s_L) - \Phi(s_0)$ retains a policy-dependent endpoint term $\gamma^L\Phi(s_L)$. The two identities are consistent because appending the absorbing step to the prefix identity recovers the full-episode identity. The closed-form margin $\gamma^L(1-c)$ of Theorem 3 is a prefix-regime quantity compatible with the policy-invariant shift of Theorem 1.

Theorem (3 restated)**.** *Fix a comparison horizon $L \in \mathbb{N}$ and a target-present skeleton ($\mathbf{b}^* \neq \varnothing$). Let h_{find} and h_{miss} be two trajectory prefixes of length L that end before termination ($L \le T-1$ on both trajectories) and share initial potential $\Phi(s_0^{\mathrm{find}}) = \Phi(s_0^{\mathrm{miss}}) = 0$, with endpoint potentials $\Phi(s_L^{\mathrm{find}}) = 1$ and $\Phi(s_L^{\mathrm{miss}}) \le c$. Then*

$$\Sigma F(h_{\mathrm{find}}) - \Sigma F(h_{\mathrm{miss}}) \;\ge\; \gamma^L(1-c). \tag{18}$$

The endpoint-potential bound $\Phi(s_L^{\mathrm{miss}}) \le c$ is implied by $\mathrm{maxIoU}(s_t^{\mathrm{miss}}) = \mathrm{cov}_{\mathrm{tgt}}(s_t^{\mathrm{miss}}) = 0$ for every $t \le L$, since the target-approach branch vanishes and the exploration branch is at most c.

Proof of Theorem 3. For any state-only potential Φ, expanding the canonical PBRS shaping reward inside the discounted shape return and distributing the outer factor,

$$\Sigma F(h) \;=\; \sum_{t=0}^{L-1} \gamma^t\big[\gamma\Phi(s_{t+1}) - \Phi(s_t)\big] \;=\; \sum_{t=0}^{L-1} \gamma^{t+1}\Phi(s_{t+1}) - \sum_{t=0}^{L-1} \gamma^t\Phi(s_t). \tag{19}$$

Reindexing the first sum with $u = t + 1$ and isolating the endpoint terms,

$$\Sigma F(h) \;=\; \sum_{u=1}^{L} \gamma^u \Phi(s_u) - \sum_{t=0}^{L-1} \gamma^t \Phi(s_t) \;=\; \gamma^L \Phi(s_L) - \Phi(s_0), \tag{20}$$

since the two inner sums $\sum_{u=1}^{L-1} \gamma^u \Phi(s_u)$ are identical and cancel.

Applying Eq. 20 under the theorem's endpoint hypotheses gives $\Sigma F(h_{\text{find}}) = \gamma^L \cdot 1 - 0 = \gamma^L$ and $\Sigma F(h_{\text{miss}}) \le \gamma^L c - 0 = \gamma^L c$. Subtracting the two,

$$\Sigma F(h_{\text{find}}) - \Sigma F(h_{\text{miss}}) \;\ge\; \gamma^L - \gamma^L c \;=\; \gamma^L (1 - c), \tag{21}$$

which is Eq. 18. More generally, Eq. 20 gives the identity $\Sigma F(h_{\text{find}}) - \Sigma F(h_{\text{miss}}) = \gamma^L \big(\Phi(s_L^{\text{find}}) - \Phi(s_L^{\text{miss}})\big)$ for any two such prefixes, and the endpoint hypotheses turn it into the bound. □

Corollary 1 (Numerical instantiation)**.** *With the default $\alpha = c = 0.5$, $\gamma = 0.99$, and horizon $L = 8$, Eq. 18 gives $\Sigma F(h_{\text{find}}) - \Sigma F(h_{\text{miss}}) \ge 0.99^8 \times 0.5 \approx 0.4614$.*

> **Tightness and boundary behavior.** Given the endpoint hypotheses, equality in Eq. 18 holds if and only if $\Phi(s_L^{\text{miss}}) = c$. At $\alpha = c = 0.5$, a crop covering exactly half of a single target gives $\Phi = 0.5$ and hence equality, while a prefix that never touches the target has $\Phi(s_L^{\text{miss}}) \le c\big(1 - |\bigcup_{b \in \mathbf{b}^*} b|\big) < c$, so the bound is strict for it. An exhaustive scan on a target-present skeleton has $\text{cov}_{\text{img}} = \text{cov}_{\text{tgt}} = 1$ and $u := \text{maxIoU}(s_L) > 0$, hence $\Phi(s_L) = \max\big(1 - \alpha + \alpha u, c\big)$. It lies outside the miss hypothesis if and only if $1 - \alpha + \alpha u > c$, in particular if $c \le 1 - \alpha$ as in our setting, and its margin to a find prefix is $\gamma^L \min\{\alpha(1-u),\, 1-c\}$. As $c \to 1$ the bound vanishes, which is the saturation symmetry the max-pool potential avoids, and as $c \to 0$ it approaches γ^L but absent skeletons lose all shaping signal. Both endpoints lower accuracy in the ablation (**Table 24**, -1.9 at $c = 0$ and -1.4 at $c = 1$).

A.5 Seeker-Side CMDP Convergence under Fixed Hider

Theorem 4 analyzes the idealized Seeker-side CMDP recursion of Eq. 24 under a fixed Hider, the regime that Phase A of Algorithm 1 realizes by freezing the Hider parameters. The Hider is held quasi-static on the slowest self-play timescale while the Seeker's primal-dual iterates evolve on the two faster timescales. The recursion uses unbiased estimators (Assumption 5 below). The scope remark after the proof relates it to the implemented PPO-Lagrangian update. We freeze a Hider policy π_H and analyze the Seeker Lagrangian

$$\mathcal{L}_S(\pi_S, \lambda_B, \lambda_\kappa \mid \pi_H) \;=\; J_{\pi_H}^{\text{shaped}}(\pi_S) - \lambda_B \big(\mathbb{E}_{\pi_H, \pi_S}[T] - B\big) - \lambda_\kappa \big(\mathbb{E}_{\pi_H, \pi_S}[\textstyle\sum_t \text{Redun}_t] - \kappa\big), \tag{22}$$

which is the inner game of the joint primal-dual scheme. The regularity and feasibility conditions are stated in Assumptions 1–8, and the theorem statement refers to them by label.

Assumption 1 (Policy parameterization)**.** The Seeker policy $\pi_S^{\theta_S}$ is parameterized by $\theta_S \in \Theta_S \subseteq \mathbb{R}^{d_S}$ with Θ_S compact, and the mapping $(\theta_S, \lambda) \mapsto \nabla_{\theta_S} \mathcal{L}_S$ is Lipschitz-continuous on $\Theta_S \times \Lambda$.

Assumption 2 (Iterate stability)**.** Almost surely, $\theta_S^{(n)} \in \Theta_S$ for every n. Together with the compactness of Θ_S and the dual projection onto Λ, it gives $\sup_n \big(\|\theta_S^{(n)}\| + \|\lambda^{(n)}\|\big) < \infty$ almost surely.

Assumption 3 (Slater feasibility for the fixed Hider)**.** For the fixed Hider policy π_H of Theorem 4, there exist Seeker parameters $\theta_S^\dagger \in \Theta_S$ inducing a policy $\pi_S^{\theta_S^\dagger}$ with

$$\mathbb{E}_{\pi_H, \pi_S^{\theta_S^\dagger}}[T] \;<\; B, \qquad \mathbb{E}_{\pi_H, \pi_S^{\theta_S^\dagger}}\big[\textstyle\sum_t \text{Redun}_t\big] \;<\; \kappa. \tag{23}$$

Assumption 4 (Saddle existence and slow-ODE stability)**.** For fixed π_H, the Seeker Lagrangian $\mathcal{L}_S(\theta_S, \lambda \mid \pi_H)$ admits a saddle point $(\theta_S^{**}, \lambda^{**}) \in \Theta_S \times \Lambda$, and the slow projected ODE Eq. 27 has λ^{**} as its unique globally asymptotically stable equilibrium with the fast variable at $\theta_S^*(\lambda)$.

Assumption 5 (Stochastic-estimator regularity)**.** The primal estimator $g_n^\theta \in \mathbb{R}^{d_S}$ and dual estimator $g_n^\lambda \in \mathbb{R}^2$ used in Eq. 24 are adapted to the filtration $\mathcal{F}_n = \sigma\big(\theta_S^{(0)}, \lambda^{(0)}, \{g_k^\theta, g_k^\lambda : 1 \le k \le n\}\big)$ and conditionally unbiased,

$$\mathbb{E}\big[g_n^\theta \,\big|\, \mathcal{F}_{n-1}\big] = h_\theta(\theta_S^{(n-1)}, \lambda^{(n-1)}), \qquad \mathbb{E}\big[g_n^\lambda \,\big|\, \mathcal{F}_{n-1}\big] = h_\lambda(\theta_S^{(n-1)}),$$

and have uniformly bounded conditional second moments $\sup_n \mathbb{E}[\|g_n^\theta\|^2 + \|g_n^\lambda\|^2 \mid \mathcal{F}_{n-1}] \le C_M < \infty$ almost surely on $\Theta_S \times \Lambda$. The residuals $M_n^\theta = g_n^\theta - h_\theta$, $M_n^\lambda = g_n^\lambda - h_\lambda$ are then martingale-difference sequences with respect to $\{\mathcal{F}_n\}$ with the same uniform second-moment bound.

Assumption 6 (Two-timescale step sizes)**.** The primal and dual step-size sequences $\{\eta_n\}$ and $\{\eta_{\lambda,n}\}$ satisfy the Robbins–Monro conditions $\sum_n \eta_n = \sum_n \eta_{\lambda,n} = \infty$, $\sum_n \eta_n^2 < \infty$, $\sum_n \eta_{\lambda,n}^2 < \infty$, together with the time-scale separation $\eta_{\lambda,n} = o(\eta_n)$.

Assumption 7 (Dual clipping)**.** The dual variables are projected onto the compact cube $\Lambda = [0, \lambda_{\max}]^2$ with $\lambda_{\max} > \max_{i\in\{B,\kappa\}} 1/\big(c_i - \mathbb{E}_{\pi_H, \pi_S^{\theta_S^\dagger}}[g_i]\big)$, with $\theta_S^\dagger$ the Slater point of Assumption 3, $g_B = T$, $c_B = B$, $g_\kappa = \sum_t \mathrm{Redun}_t$, and $c_\kappa = \kappa$. Boundary saddles with $\lambda_i^{**} = 0$ for inactive constraints are permitted.

Assumption 8 (Fast-scale well-posedness)**.** For every $\lambda \in \Lambda$, the fast-scale ODE of Eq. 26 below admits a unique globally asymptotically stable fixed point $\theta_S^*(\lambda)$, which is the unique maximizer of $\mathcal{L}_S(\,\cdot\,, \lambda \mid \pi_H)$ over Θ_S for the frozen Hider π_H. The selector $\lambda \mapsto \theta_S^*(\lambda)$ is Lipschitz on Λ, and the slow drift $h_\lambda(\theta_S) = (\mathbb{E}[T] - B,\ \mathbb{E}[\sum_t \mathrm{Redun}_t] - \kappa)$ is Lipschitz on Θ_S, so that the composed map $\lambda \mapsto h_\lambda(\theta_S^*(\lambda))$ is Lipschitz and the slow projected ODE Eq. 27 is well-posed on Λ.

Theorem (4 restated, Seeker-side under fixed Hider)**.** *Fix a Hider policy π_H and hold its skeleton distribution ρ_{π_H} fixed. Consider the idealized two-timescale stochastic-approximation recursion of Eq. 24, in which the primal update uses an unbiased estimator of $\nabla_{\theta_S}\mathcal{L}_S(\theta_S, \lambda \mid \pi_H)$ and the dual update uses unbiased estimators of the two constraint residuals (Assumption 5). Under Assumptions 1–8, the iterates $(\theta_S^{(n)}, \lambda^{(n)})$ converge almost surely to the saddle point $(\theta_S^{**}, \lambda^{**})$ of $\mathcal{L}_S(\,\cdot\,, \cdot \mid \pi_H)$ guaranteed by Assumption 4, and the primal limit satisfies the CMDP constraints with complementary slackness in expectation, with equality on every coordinate i with $\lambda_i^{**} > 0$.*

Proof of Theorem 4. The stochastic iterations on the Seeker side decompose into a deterministic drift plus a martingale-difference noise term,

$$\begin{aligned} \theta_S^{(n+1)} &= \theta_S^{(n)} + \eta_n\big[h_\theta(\theta_S^{(n)}, \lambda^{(n)}) + M_{n+1}^\theta\big], \\ \lambda^{(n+1)} &= \Pi_\Lambda\Big[\lambda^{(n)} + \eta_{\lambda,n}\big[h_\lambda(\theta_S^{(n)}) + M_{n+1}^\lambda\big]\Big], \end{aligned} \tag{24}$$

where the drift terms are the conditional expectations

$$\begin{aligned} h_\theta(\theta_S, \lambda) &:= \nabla_{\theta_S}\mathcal{L}_S(\theta_S, \lambda \mid \pi_H), \\ h_\lambda(\theta_S) &:= \big(\mathbb{E}_{\pi_H, \pi_S^{\theta_S}}[T] - B,\ \mathbb{E}_{\pi_H, \pi_S^{\theta_S}}[\textstyle\sum_t \mathrm{Redun}_t] - \kappa\big), \end{aligned} \tag{25}$$

and M_{n+1}^θ, M_{n+1}^λ are the (rollout-driven) zero-mean residuals adapted to the estimator filtration $\mathcal{F}_n$ of Assumption 5, with respect to which every iterate $(\theta_S^{(n)}, \lambda^{(n)})$ is measurable. The slow drift h_λ is the signed constraint-residual vector, consistent with the dual-ascent update Eq. 5 of the method section, so that h_λ pushes λ up if a constraint is violated and toward zero (within the projection cone) if it is slack. Trajectory rewards are pathwise bounded ($r^{\text{primary}} \in \{0, 1\}$, $r_t^{\text{shape}} \in [-1, \gamma]$ by Theorem 2 and the terminal convention, $T \le T_{\max}$ by the rollout turn limit, and $\mathrm{Redun}_t \in [0, 1]$), which directly gives the dual-residual second-moment bound. The primal residual M_n^θ is martingale-difference and uniformly second-moment-bounded by Assumption 5.

Because $\eta_{\lambda,n} = o(\eta_n)$, on the fast timescale $\tau_n^{\text{fast}} = \sum_{k\le n} \eta_k$ the dual $\lambda^{(n)}$ is asymptotically constant. Conditional on $\lambda^{(n)} \approx \lambda$, the θ_S-iterate is a Robbins–Monro scheme for the fast ODE

$$\dot\theta_S \;=\; h_\theta(\theta_S, \lambda) \;=\; \nabla_{\theta_S}\mathcal{L}_S(\theta_S, \lambda \mid \pi_H), \tag{26}$$

whose RHS is Lipschitz on $\Theta_S \times \Lambda$ by Assumption 1 and admits a unique globally asymptotically stable fixed point $\theta_S^*(\lambda)$ by Assumption 8. The tracking lemma of Borkar [31, Ch. 6, Lemma 1], applied under Assumptions 1, 2, 5, 6, and 8, gives $\|\theta_S^{(n)} - \theta_S^*(\lambda^{(n)})\| \to 0$ almost surely.

On the slow timescale $\tau_n^{\text{slow}} = \sum_{k \le n} \eta_{\lambda,k}$ the primal iterate is already at its fast-scale equilibrium, so the dual tracks the slow ODE

$$\dot{\lambda} \;=\; \Pi_{T\Lambda}\big[-\nabla_\lambda \mathcal{L}_S(\theta_S^*(\lambda), \lambda \mid \pi_H)\big] \;=\; \Pi_{T\Lambda}\big[\big(\mathbb{E}_{\theta_S^*(\lambda)}[T] - B,\, \mathbb{E}_{\theta_S^*(\lambda)}[\textstyle\sum_t \mathrm{Redun}_t] - \kappa\big)\big], \quad (27)$$

where $\Pi_{T\Lambda}$ denotes projection onto the tangent cone of Λ (inactive for interior λ). Because the dual recursion is projected onto the compact convex set Λ, the slow recursion can be written as $\lambda^{(n+1)} = \Pi_\Lambda\big[\lambda^{(n)} + \eta_{\lambda,n}\big(H(\lambda^{(n)}) + M_{n+1}^\lambda + r_n\big)\big]$ with $H(\lambda) := h_\lambda(\theta_S^*(\lambda))$ and bias $r_n := h_\lambda(\theta_S^{(n)}) - H(\lambda^{(n)})$. The fast-scale tracking above and the Lipschitz continuity of h_λ give $r_n \to 0$ almost surely, and $\sum_n \eta_{\lambda,n} M_{n+1}^\lambda$ converges almost surely because it is a martingale bounded in L^2 under Assumptions 5 and 6. Since Λ is a hyperrectangle and H is bounded and continuous on it, the projected-SA results for box constraints of Borowski and Miasojedow [43, Thm. 3.1 and Prop. 3.2], which complete the argument of Kushner and Yin [44, Ch. 5, Thm. 2.1], identify every limit of the slow iterates with a solution of the projected ODE Eq. 27. By Assumption 4 that ODE has λ^{**} as its unique globally asymptotically stable point, so $\lambda^{(n)} \to \lambda^{**}$ almost surely.

By Assumption 4 the Seeker Lagrangian $\mathcal{L}_S$ admits a saddle point $(\theta_S^{**}, \lambda^{**})$, and since θ_S^{**} maximizes $\mathcal{L}_S(\cdot, \lambda^{**} \mid \pi_H)$, the uniqueness clause of Assumption 8 gives $\theta_S^{**} = \theta_S^*(\lambda^{**})$, so the coupled limit is the saddle. The saddle inequalities

$$\mathcal{L}_S(\theta_S', \lambda^{**} \mid \pi_H) \;\le\; \mathcal{L}_S(\theta_S^{**}, \lambda^{**} \mid \pi_H) \;\le\; \mathcal{L}_S(\theta_S^{**}, \lambda' \mid \pi_H) \quad \forall \theta_S' \in \Theta_S,\ \lambda' \in \Lambda \quad (28)$$

yield feasibility and complementary slackness as follows. Write $\rho(\theta_S) = \big(\mathbb{E}[g_B] - c_B,\, \mathbb{E}[g_\kappa] - c_\kappa\big)$ for the residual vector, so that $\mathcal{L}_S(\theta_S, \lambda \mid \pi_H) = J_{\pi_H}^{\text{shaped}}(\theta_S) - \lambda \cdot \rho(\theta_S)$. The right inequality gives $\lambda^{**} \cdot \rho(\theta_S^{**}) \ge \lambda' \cdot \rho(\theta_S^{**})$ for every $\lambda' \in \Lambda$, and $\lambda' = 0$ gives $\lambda^{**} \cdot \rho(\theta_S^{**}) \ge 0$. The left inequality at the Slater point $\theta_S^\dagger$ then gives

$$\textstyle\sum_i \lambda_i^{**}\big(c_i - \mathbb{E}_{\theta_S^\dagger}[g_i]\big) \;\le\; J_{\pi_H}^{\text{shaped}}(\theta_S^{**}) - J_{\pi_H}^{\text{shaped}}(\theta_S^\dagger) - \lambda^{**} \cdot \rho(\theta_S^{**}) \;\le\; 1,$$

because $J_{\pi_H}^{\text{shaped}} = J_{\pi_H}^{\text{primary}} \in [0,1]$ (Appendix A.1). Every summand is non-negative, so $\lambda_i^{**} \le 1/(c_i - \mathbb{E}_{\theta_S^\dagger}[g_i]) < \lambda_{\max}$ by Assumption 7, and the upper face of Λ is inactive. If $\rho_i(\theta_S^{**}) > 0$, raising λ_i' above λ_i^{**} would increase $\lambda' \cdot \rho(\theta_S^{**})$, so $\rho_i(\theta_S^{**}) \le 0$, which is feasibility. If $\lambda_i^{**} > 0$ and $\rho_i(\theta_S^{**}) < 0$, lowering λ_i' to 0 would increase $\lambda' \cdot \rho(\theta_S^{**})$, so the coordinatewise KKT relation holds,

$$\lambda_i^{**}\,\big(\mathbb{E}_{\pi_H, \pi_S^{\theta_S^{**}}}[g_i(\pi_S)] - c_i\big) \;=\; 0, \qquad i \in \{B, \kappa\}, \quad (29)$$

with $g_B(\pi_S) = T$, $c_B = B$, $g_\kappa(\pi_S) = \sum_t \mathrm{Redun}_t$, $c_\kappa = \kappa$. Equality of constraint i in expectation therefore holds for every i with $\lambda_i^{**} > 0$, while a coordinate with $\lambda_i^{**} = 0$ may stay strictly slack.

Finally, the preconditions of the two-timescale argument are verified directly. The fast drift $\nabla_{\theta_S} \mathcal{L}_S$ is Lipschitz on the compact product $\Theta_S \times \Lambda$ by Assumption 1, the iterates stay in the domain of these bounds by Assumption 2 and the dual projection, and the selector $\theta_S^*(\cdot)$ and the slow drift $h_\lambda(\theta_S^*(\lambda))$ are Lipschitz on Λ by Assumption 8. The primal and dual residuals M_n^θ, M_n^λ are martingale-difference with uniformly bounded conditional second moments by Assumption 5 and the pathwise bounds on $r^{\text{primary}}, T, \mathrm{Redun}_t$. Robbins–Monro and timescale separation hold by Assumption 6. The fast-scale ODE has a unique globally asymptotically stable point by Assumption 8, and the slow-scale projected ODE has a unique globally asymptotically stable point at λ^{**} by Assumption 4. All preconditions are therefore satisfied, and the coupled iterates converge almost surely to $(\theta_S^{**}, \lambda^{**})$. The saddle-point structure and the KKT relation Eq. 29 give the claimed complementary slackness in expectation at the limit, with equality on every coordinate i with $\lambda_i^{**} > 0$. □

Empirical Timescale Separation. Theorem 4 treats the Hider as fixed, and Phase A of Algorithm 1 realizes this regime closely. The Hider's parameters are frozen in Phase A, so its sampling distribution moves only through the moving-window statistics in its state (Appendix B.1). The mean total-variation drift between consecutive Seeker updates is 0.033 within Phase A (95th percentile 0.086), against 0.21 per update in Phase B and 0.47 over all of training, so the Hider is near-static on the Seeker timescale and evolves on the slower self-play timescale. Training is also stable across dwell times (**Table 8**). Halving $K_{\min}$ doubles the switch rate while accuracy stays within 0.5 points, and even at $K_{\min} = 10$ the dual oscillation stays bounded and no seed diverges.

Scope of Theorem 4. Theorem 4 covers the idealized recursion of Eq. 24 with unbiased estimators and a fixed Hider. PPO clipping, minibatch epochs, and stale rollouts lie outside it, and Algorithm 1 also departs from Eq. 24 in its objective and step sizes. Its per-turn budget cost $\lambda_B \mathbb{1}[t \geq B]$ sums to $\lambda_B(T-B)_+$, the budget term of Eq. 22 without the early-stop bonus $\lambda_B(B-T)_+$, while the dual ascent of Algorithm 1 keeps the linear residual $\bar{T} - B$, so in Phase A λ_B rises, up to $\lambda_{\max}$, if the batch-mean episode length exceeds B. The penalties are discounted inside GAE, a KL penalty to the SFT reference is added, and the step sizes are constant, so the implemented fixed points need not be saddles of Eq. 22.

Table 8: In-domain accuracy and training stability across the minimum-dwell sweep. The oscillation amplitude is the peak-to-peak range of λ_B over the final 500 outer iterations.

$K_{\min}$	In-domain	Phase switches	Dual oscillation amplitude	Divergence
50 (default)	53.5	3.5	0.38	none
25	53.0	6.7	0.49	none
10	51.6	12.5	0.79	none

B HSSP Algorithm

Algorithm 1 gives the update scheme used throughout the experimental section. Its three timescales are the PPO-Lagrangian primal updates (fast), the dual updates on $\lambda_B, \lambda_\kappa$ (slower), and the Hider–Seeker Phase A/B hysteresis (slowest). Step sizes appear in **Table 15**.

B.1 Hider Policy Specification

The Hider in HSSP is an adaptive self-play policy that reweights the fixed V-TRACE skeleton pool $\mathcal{T}$. It selects among existing skeletons and does not generate new instances.

Hider MDP. Let $\mathcal{T} = \bigcup_{(\text{src},\text{type},\text{bin})} \mathcal{T}_{\text{src},\text{type},\text{bin}}$ be the partition of the accepted V-TRACE pool by source benchmark $\text{src} \in \{\text{O3}, \text{V*}, \text{Tree}, \text{VP-H}, \text{HR-4K}, \text{MME}\}$, trajectory type $\text{type} \in \{\text{A (backtrack)}, \text{B (distractor)}, \text{C (absent)}\}$, and difficulty bin $\text{bin} \in \{\text{easy}, \text{med}, \text{hard}\}$. The difficulty bin is the per-skeleton tertile of the cold-start Seeker's pre-terminal maximum potential $\max_{t<T} \Phi(s_t)$, evaluated under the policy obtained from the 10% V-TRACE cold-start SFT stage, with the hard bin holding the lowest tertile. By Theorem 2 this maximum equals the last pre-terminal potential $\Phi(s_{T-1})$. The Hider's sliding-window MDP $\langle \mathcal{S}_H, \mathcal{A}_H, P_H, r_H \rangle$ has four components.

- **State** s^H is the moving-window per-bucket Seeker success-rate vector $\bar{p} \in [0,1]^{|\mathcal{B}|}$ over the most recent $W = 512$ rollouts, with $\bar{p}_b = 0$ for a bucket absent from the window, the moving-window mean pre-terminal maximum potential, the running CMDP constraint-residual vector $(\mathbb{E}[T] - B,\ \mathbb{E}[\sum_t \text{Redun}_t] - \kappa)$, and the per-bucket visitation count. The sliding window of rollout summaries is the underlying Hider state, and the policy observes these aggregate features.
- **Action** a^H is a categorical distribution over the $|\mathcal{B}|=6\times3\times3=54$ buckets. Each skeleton of an outer-loop batch is drawn by sampling a bucket from this distribution and then a skeleton uniformly at random from $\mathcal{T}_{\text{src},\text{type},\text{bin}}$. The action space is therefore the simplex $\mathcal{A}_H = \Delta^{|\mathcal{B}|-1}$.
- **Transition** P_H adds the new rollout summaries to the sliding window and removes the oldest ones, so it is deterministic given the Seeker's rollouts.
- **Reward** $r_H = -r^{\text{primary}}$ is the zero-sum complement of the Seeker's primary reward. The Hider receives no shaping reward, so that Theorem 1 applies.

Constraints. Two soft constraints are applied to the Hider's action distribution to guard against curriculum collapse. The first is a per-bucket lower bound $\Pr[\text{bucket}_i] \geq \epsilon_{\min}/|\mathcal{B}| = 1/(2|\mathcal{B}|^2)$, enforced by mixing the policy output with a uniform distribution at weight $\epsilon_{\min} = 1/(2|\mathcal{B}|)$ before sampling. The second is an entropy bonus, so that the Hider maximizes $\mathbb{E}[-r^{\text{primary}}] + \beta_H \mathcal{H}(q)$ with $\beta_H=0.05$, as in Algorithm 2. With these two terms, the effective support size of q stays above $0.6\,|\mathcal{B}|$ throughout training (**Fig. 5**), and no narrow subset such as Type-C-only dominates.

Algorithm 1 HSSP training loop.

Require: skeleton pool $\mathcal{T}$, potential configuration (α, c), constraint bounds (B, κ), hysteresis $(\tau_{\text{hi}}, \tau_{\text{lo}}, K_{\min})$, dual bound $\lambda_{\max}$, and learning rates $\eta_\theta, \eta_\lambda$ (Appendix D.3).

Initialize: π_S by cold-start SFT on 10% of $\mathcal{T}$, π_H at initial parameters $\theta_H^{(0)}$, duals $\lambda_B, \lambda_\kappa \leftarrow 0$, phase $m \leftarrow$ A, and dwell $d \leftarrow 0$.
for $n = 1, \ldots, T_{\text{iter}}$ **do**
  Sample skeletons $\{\tau_i\} \sim \pi_H$, roll out $\{h_i\} \sim \pi_S(\tau_i)$, and compute potentials $\{\Phi(s_t^{(i)})\}$.
  Shape reward: $r_t^{\text{shape}} \leftarrow \gamma\,\Phi(s_{t+1}) - \Phi(s_t)$, with $\Phi(s_T) := 0$.
  Penalized per-turn reward: $\tilde{r}_t \leftarrow r_t^{\text{primary}} + r_t^{\text{shape}} - \lambda_B \mathbb{1}[t \geq B] - \lambda_\kappa\ \text{Redun}_t$.
  Advantages: $\{A_t^{(i)}\} \leftarrow$ Bi-Level GAE [8] on $\{\tilde{r}_t^{(i)}\}$.
  if $m =$ A **then** ▷ PHASE A, Seeker update
    PPO-clip gradient step on π_S with advantages A and learning rate η_θ.
    Dual ascent: $\lambda_B \leftarrow \Pi_{[0,\lambda_{\max}]}\big[\lambda_B + \eta_\lambda(\bar{T} - B)\big]$,
      $\lambda_\kappa \leftarrow \Pi_{[0,\lambda_{\max}]}\big[\lambda_\kappa + \eta_\lambda(\overline{\textstyle\sum_t \text{Redun}_t} - \kappa)\big]$.
  else ▷ PHASE B, Hider update
    Hider update on π_H (Algorithm 2) with reward $-\mathbb{1}[\hat{y} = a^*]$ and entropy bonus $\beta_H \mathcal{H}(q)$.
  end if
  Evaluate held-out Seeker accuracy p on skeletons drawn from π_H, and set $d \leftarrow d + 1$.
  Phase switch (hysteretic, dwell-gated).
  if $m =$ A and $d \geq K_{\min}$ and $p \geq \tau_{\text{hi}}$ **then** $m \leftarrow$ B, $d \leftarrow 0$.
  else if $m =$ B and $d \geq K_{\min}$ and $p \leq \tau_{\text{lo}}$ **then** $m \leftarrow$ A, $d \leftarrow 0$.
  end if
end for
return Seeker π_S (the Hider is not used at inference).

Algorithm 2 Hider Phase-B update step (one outer-loop iteration).

Require: skeleton pool partitioned as $\mathcal{T} = \bigcup_{(\text{src,type,bin})} \mathcal{T}_{\text{src,type,bin}}$, with $|\mathcal{B}|{=}54$ buckets, window size W, uniform-mixing weight $\epsilon_{\min}$, entropy weight β_H, learning rate η_H.

Form Hider state s^H from the most recent W rollouts: per-bucket success-rate vector $\bar{p}$ ($\bar{p}_b = 0$ for a bucket absent from the window), mean pre-terminal maximum potential $\overline{\max_{t<T} \Phi(s_t)}$, CMDP constraint residuals $(\mathbb{E}[T] - B,\ \mathbb{E}[\sum_t \text{Redun}_t] - \kappa)$, and per-bucket visitation ν.
Compute bucket logits $\ell_b \leftarrow f_{\theta_H}(s^H)_b$ for $b \in \{1, \ldots, |\mathcal{B}|\}$.
Mix with per-bucket lower bound: $q_b \leftarrow (1 - \epsilon_{\min})\,\text{softmax}(\ell)_b + \epsilon_{\min}/|\mathcal{B}|$.
Sample $b^* \sim q$ and skeleton $\tau \sim \text{Unif}(\mathcal{T}_{b^*})$, then pass τ to the Seeker for rollout.
After the Seeker emits h, compute role-advantage reward $r_H \leftarrow -\mathbb{1}[\hat{y}(h) = a^*(\tau)]$ and baseline b^H via the role-advantage estimator of [10].
Gradient step: $\theta_H \leftarrow \theta_H + \eta_H\,(r_H - b^H)\,\nabla_{\theta_H} \log q_{b^*} + \eta_H\,\beta_H\,\nabla_{\theta_H} \mathcal{H}(q)$.

Update. The Hider is trained with the role-advantage REINFORCE estimator of Liu et al. [10], which offsets the role-conditional information asymmetry between the two players, within the same outer loop as the Seeker. Algorithm 2 gives one Hider update step.

B.2 Hider Training Dynamics

Fig. 5 reports four diagnostic quantities along the same training schedule for the adaptive Hider, the matched-volume random Hider of **Table 24**, and the uniform Hider. The adaptive Hider shifts the difficulty distribution over training while keeping broad support over $\mathcal{T}$.

The type-mass panel shows that the adaptive Hider shifts from a near-uniform type mix toward a mild Type-C upweighting, and Type-C mass stays below the 0.55 collapse threshold throughout training. The effective support size of the bucket distribution $\text{ESS}(q)/|\mathcal{B}|$ contracts during the first Phase B interval and remains above 0.6 through iteration 3000. The hard-bucket mass increases under the adaptive Hider while the matched-volume random Hider stays near the uniform-tertile reference $|\mathcal{B}_{\text{hard}}|/|\mathcal{B}| = 1/3$ throughout, indicating that the rise reflects adaptive curriculum pressure rather than sampling drift. The median pre-terminal maximum potential $\max_{t<T} \Phi(s_t)$ of Hider-sampled

skeletons rises during Phase B intervals, partially relaxes in the following Phase A windows, and plateaus near 0.49, well below the target-branch maximum of 1.

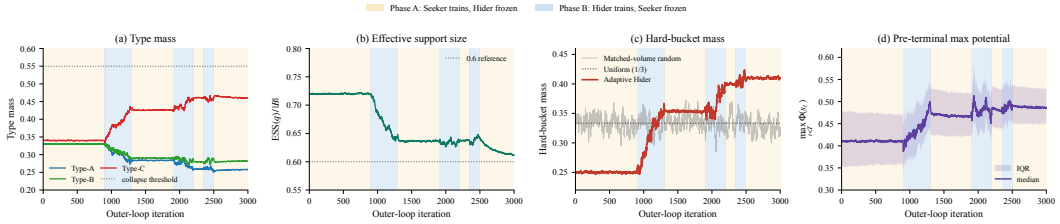


Figure 5: Hider training dynamics. *(a)* Type mass with the 0.55 collapse threshold. *(b)* Effective support size of the bucket distribution $\mathrm{ESS}(q)/|\mathcal{B}|$. *(c)* Hard-bucket mass for the adaptive, matched-volume random, and uniform Hiders. *(d)* Pre-terminal peak potential $\max_{t<T} \Phi(s_t)$ (median, IQR) of sampled skeletons.

C V-TRACE Construction Details

This appendix specifies the construction of $\mathcal{T}$ and the reproducibility-relevant choices omitted from the main method description. The construction has three stages, namely source aggregation, rule-based skeleton synthesis, and multi-agent enrichment with structural validation.

C.1 Source Datasets

V-TRACE aggregates six publicly released visual-search benchmarks that differ along annotation availability, image domain, resolution, and post-construction acceptance rate (**Table 9**). Three of the six carry native target-bounding-box annotations, namely O3-Bench [2] with layout-informed bounding boxes, V*Bench [16] with per-image target (x, y, w, h), and TreeBench [27] with per-instance target annotations. The remaining three sources, VisualProbe-Hard [3], HR-Bench 4K [28], and MME-RealWorld-Lite [29], lack native boxes.

Table 9: Source benchmarks aggregated into V-TRACE. Sources without native boxes are grounded by Qwen3.5 397B A17B. **Accept rate** is the fraction of synthesized trajectories clearing the mechanism-designer acceptance threshold τ_{acc}=0.85 (Algorithm 3). **# Trajectories** is the per-source accepted count after expansion into the three trajectory types (Type-A, Type-B, Type-C) and multi-agent filtering, and the total accept rate is weighted by source size.

Benchmark	**# QAs**	**Image domains**	**Native boxes**	**Avg. resolution**	**Accept rate**	**# Traj.**
V*Bench [16]	191	natural (100%)	✓	2246×1583	97.5%	559
TreeBench [27]	405	natural (100%)	✓	2152×1615	97.5%	1,185
VisualProbe-Hard [3]	106	natural (100%)	✗	4944×3980	91.7%	292
HR-Bench 4K [28]	800	natural 89%, chart 5%, map 6%	✗	4024×3503	100%	2,400
MME-RealWorld-Lite [29]	1,919	natural >60%, chart 25%, map 6%, etc.	✗	2708×1844	100%	5,757
O3-Bench [2]	204	chart 47%, map 53%	✓	4602×3967	95%	580
Total	**3,625**				99.1%	**10,773**

For these three sources, boxes are inferred by deterministic Qwen3.5 397B A17B grounding conditioned on the ground-truth answer. Every inferred box is then verified and, if needed, corrected by human annotators, and the verified box is recorded as the target region of the potential.

Data Leakage. The leakage analysis tests three conditions, namely (i) no image in the held-out test partition of any V-TRACE source appears in that source's training partition, (ii) no image in any V-TRACE source appears in any of the five OOD evaluation benchmarks, and (iii) VLM-inferred boxes do not leak into native-box sources. Conditions (i) and (ii) are verified by a perceptual-hash [45]

comparison. Every image is hashed with a 64-bit DCT hash, and no pairwise Hamming distance falls under the standard near-duplicate threshold of 8 bits. Condition (iii) holds by construction, because native-box and inferred-box sources are processed in isolation.

C.2 Skeleton Synthesis

Three rule-based generators expand each seed $\langle I, q, a^*, \mathbf{b}^* \rangle$ into one or more *trajectory templates*, each of which specifies an action sequence without per-step thought text. Each generator emits action tokens from $\{\langle \texttt{PLAN} \rangle, \langle \texttt{CROP} \ldots \rangle, \langle \texttt{BACKTRACK k} \rangle, \langle \texttt{VERIFY} \rangle, \langle \texttt{STOP} \ldots \rangle\}$ with per-step process labels (`good_step`, `backtrack_needed`, `verify_needed`), stored with each released trajectory.

Type-A: Teacher-Forced Backtracking. Type-A skeletons teach the policy to admit a wrong decision and recover. The generator pairs the target layout $\mathbf{b}^*$ with a *wrong* crop b_{wrong} drawn from the same image, selected to be disjoint from every target box and to lie outside a margin of the target center. The emitted trajectory is $\texttt{PLAN}\,\texttt{CROP}(b_{\text{wrong}})\,\texttt{BACKTRACK 1}\,\texttt{PLAN}\,\texttt{CROP}(\mathbf{b}^*)\,\texttt{STOP}(a^*)$, with the first `CROP` carrying the `backtrack_needed` label and the remaining steps carrying `good_step`.

Type-B: Distractor Contrast. Type-B skeletons defeat confirmation bias by pairing the target with the most visually similar wrong layout in the same image. Writing t for the target box and c for a candidate distractor, distractor similarity aggregates four factors,

$$\begin{aligned} s(b_t, b_c) \;=\; & 3\mathbb{1}[\ell_t{=}\ell_c] + 2\,\tfrac{\min(A_t,A_c)}{\max(A_t,A_c)} \\ & + \tfrac{\min(r_t,r_c)}{\max(r_t,r_c)} + \max\big(0,\, 1 - \mathrm{dist}(b_t, b_c)\big), \end{aligned} \tag{30}$$

where ℓ is the layout label, A the bounding-box area, r the aspect ratio, and dist the normalized center-to-center distance. The most-similar admissible candidate $b_{\text{distract}} = \arg\max_b s(b_t, b)$ is used to emit $\texttt{PLAN}\,\texttt{CROP}(b_{\text{distract}})\,\texttt{VERIFY}\,\texttt{BACKTRACK 1}\,\texttt{CROP}(\mathbf{b}^*)\,\texttt{VERIFY}\,\texttt{STOP}(a^*)$, with the distractor `CROP` labeled `verify_needed` and `backtrack_needed`, so that the Seeker learns to compare the two regions via `VERIFY` and to leave a plausible wrong region before committing to the target.

Type-C: Systematic Absence. Type-C skeletons teach the policy to declare "not found" only after systematic coverage. A counterfactual question is sampled from the template pool {*"Is there a {category} visible anywhere in this image? If not, answer None"*, *"Locate the {category} in the image. If it does not exist, respond with None"*, *"Find the {category}. If the image does not contain one, output None"*, *"Does this image contain a {category}? Answer with its location or None"*} with {category} drawn from a source-specific pool of ten counterfactual categories (e.g., chart images use *pie chart*, *waterfall chart*, *radar chart* absent from the actual layout). A 2×2 grid of overlapping crops with 5% overlap tiles the image, and the emitted trajectory is $\texttt{PLAN}\,[\texttt{CROP}(g_1)\,\texttt{CROP}(g_2)\,\texttt{CROP}(g_3)\,\texttt{CROP}(g_4)]\,\texttt{STOP}(\texttt{None})$, with all `CROP` steps labeled `good_step` because systematic coverage is the desired behavior under Type-C.

Examples. **Fig. 6** displays accepted V-TRACE skeletons, one backtrack (Type-A) and one absent (Type-C) trajectory per source. The examples illustrate the breadth of visual content spanned by the six sources and the structural consistency of V-TRACE trajectories across them.

C.3 Multi-Agent Enrichment and Validation

Each synthesized skeleton is completed by a Qwen3.5 397B A17B cascade of four agents, namely a *Generator* and three parallel verifiers (*Spatial*, *Semantic*, *Consistency*), followed by a *Mechanism Designer* that aggregates the verifier scores. Their system prompts are given verbatim below.

Generator. The generator writes per-step `thought` text from the cropped observation. Its system prompt requires each thought to be grounded in the current viewport and specific to the action type.

■ **PROMPT | Generator (system role)**

You are a visual reasoning agent examining a high-resolution composite image. You will be given a sequence of actions (`PLAN`, `CROP`, `BACKTRACK`, `VERIFY`, `STOP`) that form a search trajectory. For each action step, you must write a visually grounded thought that



Figure 6: V-TRACE skeletons, one Type-A and one Type-C per source. Each panel shows the source image with the target box (dashed red), the views along the trajectory, and the question.

1. describes exactly what you see in the current viewport (specific objects, text, colors, and spatial relationships rather than vague descriptions),
2. explains *why* this action is taken based on visual evidence, and
3. connects the step to the overall search goal.

Step-specific rules:

- `PLAN`: state the search strategy with reference to visible layout structure.
- `CROP`: describe what you expect or observe in the target region, mentioning specific visual elements (e.g., *"a bar chart with blue bars labeled Q1-Q4"* rather than *"a chart"*).
- `BACKTRACK`: explain what was wrong about the previous crop using specific visual evidence (e.g., *"this region contains a scatter plot, not the revenue table I need"*).
- `VERIFY`: compare the current region against the question criteria with explicit evidence for *and* against.
- `STOP`: summarize the evidence chain that leads to the answer.

Never hallucinate content not visible in the image, and never use generic phrases such as *"examining the region"* without specifics. Return a JSON object with key `"thoughts"`: a list of objects, each with `"step_idx"` (int) and `"thought"` (str).

Spatial Verifier. The spatial verifier checks bounding-box geometry and view-stack dynamics of the entire trajectory from coordinates and thoughts alone, without access to image pixels, which makes its signal complementary to the pixel-level semantic verifier.

■ **PROMPT | Spatial Verifier (system role)**

You are a spatial consistency verifier for visual search trajectories. You verify that actions are geometrically sound and internally consistent. Check each step against the following criteria:

1. `CROP` **coordinates** lie within $[0, 1]$, the crop area is non-degenerate (width and height > 0.02), and the accompanying thought mentions a region consistent with the coordinates.
2. `BACKTRACK` thoughts reference a specific failure in the previous crop, and the backtrack depth k matches the view-stack state.
3. `VERIFY` thoughts compare at least two pieces of evidence.
4. **View-stack consistency**: after `BACKTRACK k` the viewport reverts to `stack[-k]`.
5. **Coverage**: the trajectory avoids redundant crops, defined as IoU > 0.8 between any two crops without an intervening `BACKTRACK`.
6. **Answer reachability**: the final `STOP` viewport overlaps at least one target layout with IoU > 0.1.

Return a JSON object:

```
{
  "dimensions": {
    "coordinate_validity": {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []},
    "stack_consistency":   {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []},
    "redundancy_free":     {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []},
    "answer_reachable":    {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []}
  }
}
```

The spatial verifier receives task-type-aware hints that relax the `redundancy_free` and `answer_reachable` thresholds on intentional Type-A wrong crops and Type-C absent rollouts, so design features are not mistaken for defects.

Semantic Verifier. The semantic verifier is the only verifier with pixel access. It receives one crop and the generator's thought for it, and checks the grounding of the thought in visible content.

■ **PROMPT | Semantic Verifier (system role)**

You are a visual grounding verifier. You receive an image crop and the agent's thought about that crop, and your task is to verify that the thought is factually grounded in what is actually visible. For each `CROP` step, evaluate:

1. **Object accuracy**: are objects or elements mentioned in the thought actually visible in the crop?
2. **Text accuracy**: does any text quoted in the thought match visible text in the crop?
3. **Spatial accuracy**: are spatial claims (left/right/above/below) correct relative to the crop?
4. **Absence claims**: if the thought says something is *not* present, verify it is truly absent.
5. **Detail level**: is the description specific enough to distinguish this crop from other regions?

Return a JSON object:

```
{
  "dimensions": {
    "object_grounding":  {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []},
```

```
    "text_grounding":    {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []},
    "spatial_grounding": {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []},
    "specificity":       {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []}
  }
}
```

Consistency Checker. The consistency checker performs a trajectory-level text-only check that catches contradictions between steps and confirms that each task-type template is honored.

■ **PROMPT | Consistency Checker (system role)**

You are a logical coherence verifier for multi-step reasoning trajectories. You check whether the chain of thoughts tells a consistent, non-contradictory story. Evaluate:

1. **Narrative flow**: does each thought logically follow from the previous?
2. **Contradiction detection**: does the agent claim X in step i but contradict it in step j?
3. **Goal alignment**: does each thought connect to the original question?
4. **Backtrack coherence**: when the agent backtracks, does the stated reason match what was observed in the failed step?
5. **Evidence accumulation**: does the `STOP` thought reference evidence gathered in prior steps, or does it pull an answer from nowhere?
6. **Task-type adherence**: *backtrack* rollouts must contain at least one wrong crop that is acknowledged and corrected, *distractor* rollouts must explicitly compare target against distractor, and *absent* rollouts must demonstrate systematic coverage before concluding *"None"*.

Return a JSON object:

```
{
  "dimensions": {
    "narrative_flow":     {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []},
    "contradiction_free": {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []},
    "goal_alignment":     {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []},
    "evidence_chain":     {"score": 0.0-1.0, "confidence": 0.0-1.0, "issues": []}
  }
}
```

Mechanism Designer. The mechanism designer aggregates the twelve dimension scores emitted by the three verifiers into a single quality score $\mathrm{Q} \in [0, 1]$. Writing $\mathcal{J}$ for the set of non-null dimensions and (s_d, w_d) for the score and confidence of dimension d, Q is a confidence-weighted mean penalized by cross-verifier disagreement on overlapping concerns,

$$\mathrm{Q} \;=\; \frac{\sum_{d\in\mathcal{J}} w_d\, s_d}{\sum_{d\in\mathcal{J}} w_d} \;\cdot\; \big(1 - \beta\, \Delta_{\text{disagree}}\big), \tag{31}$$

where Δ_{disagree} is the mean absolute difference of paired scores on the two overlapping dimension pairs (`coordinate_validity` with `spatial_grounding`, and `answer_reachable` with `evidence_chain`). Skeletons with $\mathrm{Q} \geq \tau_{\text{acc}}$ are accepted, those in the intermediate range $[\tau_{\text{disc}}, \tau_{\text{acc}})$ are revised on every dimension whose per-dimension score s_d falls below τ_{rep}, and those with $\mathrm{Q} < \tau_{\text{disc}}$ are rejected. The thresholds τ_{acc}=0.85, τ_{disc}=0.50, τ_{rep}=0.80 and the mixing weight β=0.10 are set as in Algorithm 3.

The mean final quality score across the six sources is 0.961, with per-source acceptance in **Table 9**. Algorithm 3 summarizes skeleton synthesis, multi-agent enrichment, and structural validation.

Rule-Based Validation. After multi-agent enrichment, every accepted trajectory is checked against five deterministic structural rules. Trajectories that fail any rule are discarded without repair.

♦ **RULE | Structural Rules**

1. **Thought non-emptiness.** Every step carries a non-empty `thought` string.
2. **Valid `CROP` geometry.** Each `<CROP x1 y1 x2 y2>` action satisfies $x_1, y_1, x_2, y_2 \in [0, 1]$ and yields a non-degenerate box with width and height strictly greater than 0.01.
3. **Terminal `STOP`.** The final step's action token matches the `<STOP answer>` pattern.
4. **Sequential indices.** Step indices are exactly $0, 1, 2, \ldots, T-1$ with no gaps or permutations.
5. **Action-token grammar.** Every action token matches one of the five patterns `<PLAN>...</PLAN>`, `<CROP x1 y1 x2 y2>`, `<BACKTRACK k>`, `<VERIFY>...</VERIFY>`, or `<STOP answer>`.

Algorithm 3 V-TRACE Skeleton Synthesis and Multi-Agent Enrichment (High-level)

Require: Source pool $\mathcal{S} = \{(I, q, a^*, \mathbf{b}^*)\}$ from the six benchmarks of Table 9, type set $\mathcal{Y} \subseteq \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$ admissible per seed, repair rounds $R{=}2$, acceptance threshold $\tau_{\mathrm{acc}}{=}0.85$ and discard threshold $\tau_{\mathrm{disc}}{=}0.50$, per-dimension rewrite cutoff $\tau_{\mathrm{rep}}{=}0.80$, and disagreement weight $\beta{=}0.10$.
Ensure: Released skeleton pool $\mathcal{T}$ with thoughts, action sequences, and target boxes.

Native-or-grounded boxes. For each $(I, q, a^*) \in \mathcal{S}$, if native $\mathbf{b}^*$ exists use it, else infer $\mathbf{b}^*$ by deterministic Qwen3.5 397B A17B grounding conditioned on a^* and have humans verify it.
for each seed $\sigma_0 = (I, q, a^*, \mathbf{b}^*)$ in $\mathcal{S}$ **do**
  for each admissible trajectory type $y \in \mathcal{Y}(\sigma_0)$ **do** ▷ one seed may emit one or more skeletons
    Step 1 – Skeleton synthesis.
    if $y = \mathrm{A}$ **then** pick wrong box b_{wrong} disjoint from $\mathbf{b}^*$ and outside a margin of the target center, and emit $\texttt{PLAN}\,\texttt{CROP}(b_{\mathrm{wrong}})\,\texttt{BACKTRACK 1}\,\texttt{PLAN}\,\texttt{CROP}(\mathbf{b}^*)\,\texttt{STOP}(a^*)$.
    else if $y = \mathrm{B}$ **then** choose distractor $b_{\mathrm{distract}} = \arg\max_b s(\mathbf{b}^*, b)$ via Eq. 30, and emit $\texttt{PLAN}\,\texttt{CROP}(b_{\mathrm{distract}})\,\texttt{VERIFY}\,\texttt{BACKTRACK 1}\,\texttt{CROP}(\mathbf{b}^*)\,\texttt{VERIFY}\,\texttt{STOP}(a^*)$.
    else if $y = \mathrm{C}$ **then** sample counterfactual question q' with absent category from a fixed template pool, tile the image with a $2{\times}2$ overlapping grid $\{g_1, \ldots, g_4\}$, and emit $\texttt{PLAN}\,[\texttt{CROP}(g_1) \ldots \texttt{CROP}(g_4)]\,\texttt{STOP}(\texttt{None})$.
    end if
    Step 2 – Multi-agent enrichment.
    for repair round $r = 0, \ldots, R$ **do**
      GENERATOR: emit per-step `thought` text grounded in the current viewport.
      Apply SPATIALVERIFIER, SEMANTICVERIFIER, and CONSISTENCYCHECKER in parallel, each returning four scored verification dimensions.
      MECHANISMDESIGNER aggregates twelve dimension scores into a trajectory-level $\mathrm{Q} \leftarrow \left(\sum_d w_d s_d / \sum_d w_d\right) \cdot (1 - \beta \Delta_{\mathrm{disagree}})$, with $\Delta_{\mathrm{disagree}}$ the cross-verifier mean absolute difference of paired scores on overlapping concerns (Eq. 31).
      if $\mathrm{Q} \geq \tau_{\mathrm{acc}}$ **then break** (accept). ▷ enriched trajectory finalized
      else if $\mathrm{Q} < \tau_{\mathrm{disc}}$ **then break** and discard. ▷ below the discard threshold
      else return dimensions $\{d : s_d{<}\tau_{\mathrm{rep}}\}$ to the generator for a targeted rewrite.
      end if
    end for
    Step 3 – Structural validation. Reject violations of the five rules of Appendix C (non-empty thoughts, `CROP` geometry, terminal `STOP`, sequential indices, action-token grammar).
    Append the accepted, validated skeleton to $\mathcal{T}$.
  end for
end for
return $\mathcal{T}$. ▷ Released pool: 10,773 trajectories at 99.1% weighted acceptance.

These rules are applied uniformly across the six sources and the three trajectory types, so the released pool inherits a single, source-agnostic structural guarantee.

C.4 Human Evaluation and Annotation

To establish the quality of the released pool beyond the multi-agent verification above, we conduct a *full-coverage* human evaluation over every accepted V-TRACE trajectory. Ten graduate-level annotators from the authors' research group with prior vision-language research experience rate each of the 10,773 trajectories on four dimensions, with three independent ratings per item and three sampled steps per trajectory for the step-level dimension, yielding 32,319 trajectory-level ratings per dimension and 96,957 step-level ratings. Annotators receive a written rubric (**Table 10**), a pilot round of 40 items used to calibrate the guidelines, and a training session on adjudicated examples before the full evaluation begins. 10% of items are seeded gold checks with known-valid or known-invalid ground truth, and annotators whose gold-check accuracy drops below 95% are retrained. Ratings are aggregated by majority vote on binary dimensions and median on ordinal dimensions, with pre-adjudication inter-annotator agreement reported in **Table 11**.

♦ RULE | Predefined Pass Thresholds

1. **Target Validity** (binary, trajectory-level): $\geq 90\%$ valid.
2. **Trajectory Executability** (binary, trajectory-level): $\geq 90\%$ valid.
3. **Thought–Action Faithfulness** (3-point ordinal, step-level): $\geq 80\%$ faithful and $\geq 95\%$ faithful+partial.
4. **Mechanism Fidelity** (5-point Likert, trajectory-level): mean ≥ 4.0 and $\geq 75\%$ of ratings in $\{4, 5\}$.

Table 10: Full annotation rubric used by raters. Every annotator receives the exact prompt wording, positive criterion, and failure cues reproduced below.

Positive criterion	**Failure cues**
Target Validity (trajectory, binary)	
Prompt. "Given the question and the highlighted target region on the full image, is this a well-posed search task whose answer is uniquely determined by that region?"	
The dashed-red box contains a single unambiguous entity that answers the question, and an attentive human could answer without ambiguity from the image alone.	Target box off-object or empty, multiple regions equally answer the question, or question is under-determined.
Trajectory Executability (trajectory, binary)	
Prompt. "Is the displayed sequence of actions a legal multi-step search under the action grammar `{PLAN, CROP, BACKTRACK, VERIFY, STOP}`?"	
Every `CROP` lies in $[0, 1]$ and is non-degenerate, `BACKTRACK` depth never exceeds the number of open views, the rollout terminates with `STOP`, and step indices are sequential.	Degenerate crop (w or $h \leq 0.01$), backtracking beyond the first view, missing `STOP`, or out-of-order step indices.
Thought–Action Faithfulness (step, 3-point ordinal)	
Prompt. "Does this step's thought text describe what is actually visible in the current crop and justify the next action with specific visual evidence?"	
(*faithful*) Thought names specific objects, text, or spatial relations visible in the crop and links them to the next action. (*partial*) Only one claim is grounded.	(*unfaithful*) Thought hallucinates content not in the crop, uses generic phrases (*"examining the region"*), or contradicts visible evidence.
Mechanism Fidelity (trajectory, 5-point Likert)	
Prompt. "On a 1–5 scale, how clearly does this trajectory exhibit its declared Type-A/B/C search pattern (backtrack-and-recover, distractor-contrast-with-VERIFY, or systematic-coverage-then-absent)?"	
(*5*) Trajectory unambiguously follows the declared pattern, (*4*) pattern is clear with one minor weakness, or (*3*) pattern is partially recognizable.	(*2*) Pattern is weak or inconsistent, or (*1*) rollout is a single-shot attempt or the declared type does not match observed behavior.

Table 11: Human-evaluation results on the full V-TRACE pool ($n_{\text{traj}} = 10{,}773$, ten annotators, three ratings per item). Thought–Action Faithfulness is aggregated over three sampled steps per trajectory ($n_{\text{step}} = 32{,}319$). Confidence intervals are Wilson-score 95% (binary) and bootstrap 95% (ordinal), and pre-adjudication agreement uses Fleiss κ (binary) and Krippendorff α (ordinal).

Dimension	**Unit**	n	**Mean / Pass rate**	**95% CI**	**Agreement**
Target Validity	trajectory	10,773	96.8% valid	[96.5, 97.1]	Fleiss $\kappa = 0.82$
Trajectory Executability	trajectory	10,773	95.4% valid	[95.0, 95.8]	Fleiss $\kappa = 0.79$
Thought–Action Faithfulness	step	32,319	88.2% faithful, 8.6% partial, 3.2% unfaithful	[87.8, 88.6]	Krippendorff $\alpha = 0.74$
Mechanism Fidelity	trajectory	10,773	mean 4.31 ± 0.84, 84.7% in $\{4, 5\}$	[4.29, 4.33]	Krippendorff $\alpha = 0.71$

All four dimensions clear their predefined pass thresholds, with consistent results across the six sources and across the native-box and VLM-grounded subsets (per-source and per-level breakdowns in **Fig. 7**, and a complementary verification of the VLM-grounded bounding boxes alone in Appendix C.5).

C.5 Reliability of Training-Time Bounding Boxes

Three studies examine the target boxes behind Φ, namely human verification of the VLM-grounded boxes, a propagation study of box noise, and a substitution study of the box source.

Human Verification of VLM-Grounded Boxes. We verify the Qwen3.5 397B A17B boxes of the three VLM-grounded sources against human judgment. Two hundred boxes are sampled stratified

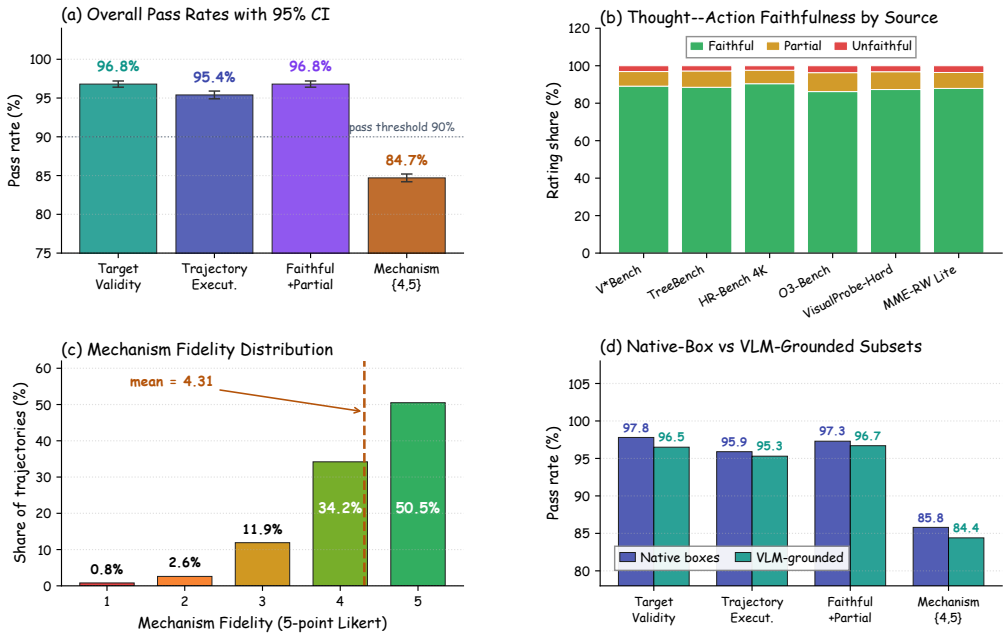


Figure 7: Human-evaluation quality of V-TRACE. *(a)* Overall pass rates on the four rubric dimensions with 95% confidence intervals (Wilson-score for binary and bootstrap for ordinal dimensions), with the dotted line at the 90% predefined threshold of the two binary dimensions. *(b)* Thought–Action Faithfulness stacked by source. *(c)* Mechanism Fidelity distribution over the 5-point Likert scale with a dashed line at the class mean. *(d)* Native-box and VLM-grounded subsets per dimension.

across them (70 VisualProbe-Hard, 70 HR-Bench 4K, 60 MME-RealWorld-Lite) and shown to three independent annotators alongside the original image and question. Each annotator draws the smallest bounding box they would themselves assign for the question and rates their agreement with the VLM-generated box on an ordinal three-level scale (*full*, *partial*, *disagree*). The full-agreement rate is 85.2% (full+partial rate 95.9%), the mean and median IoU against the median human-drawn reference box are 0.78 and 0.82, the fraction of VLM boxes with IoU $\geq$ 0.5, the effective grounding accuracy, is 92.5%, Krippendorff's α across all 200 items is 0.73, and the mean pairwise IoU among the three annotators is 0.85. **Fig. 8** shows the distribution and per-source breakdown.

To trace the path from residual box error to the reward, we recompute the potential and the discounted shaping sum on the 200 verified items under either box source, with every trajectory held fixed (**Table 12**). The two sources select the same active branch of the potential in 96.5% of steps. Branch mismatches concentrate in the 15 items with IoU below 0.5, which account for 57% of mismatches while comprising 7.5% of the verification sample, and on the remaining 185 items agreement matches inter-annotator variation. The per-step rank correlations agree within 0.03 across the three VLM-grounded sources, so the VLM boxes yield the same shaping signal as human boxes.

Table 12: Geometric quantities under the VLM-generated box and under the median human-drawn box on the 200 verified items. The last column compares two human annotators.

Quantity	**All** 200	**IoU** ≥ 0.5 (185)	**IoU** < 0.5 (15)	**Human pair**
Per-step potential, Spearman ρ	0.94	0.97	0.61	0.98
Per-step potential, mean absolute difference	0.043	0.031	0.19	0.026
Discounted shaping sum, Spearman ρ	0.91	0.95	0.58	0.96
Active max-pool branch, agreement	96.5%	98.4%	73.6%	98.7%

Box-Noise Propagation. This study quantifies the propagation of residual box error through the potential Φ and through subsequent training. Three calibrated noise regimes are applied to every target box during training-time V-TRACE sampling and are layered cumulatively from *mild* to *heavy*.

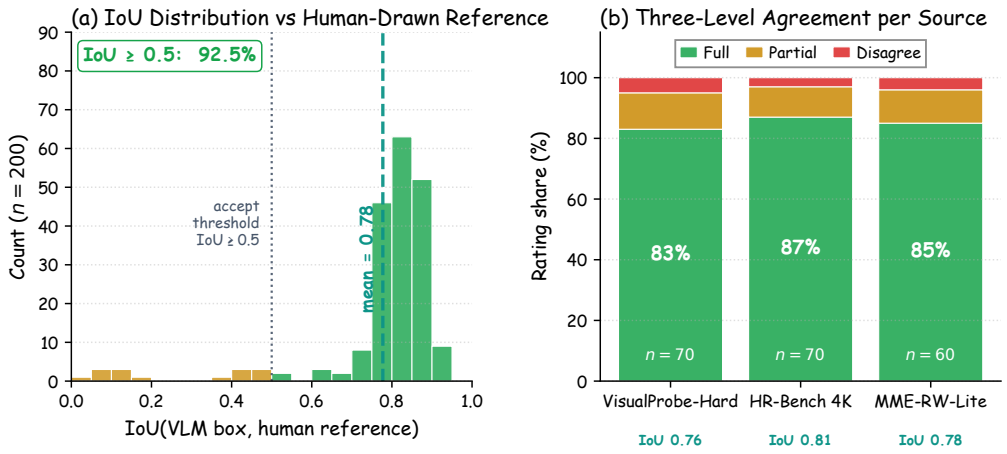


Figure 8: Human verification of VLM-grounded bounding boxes ($n = 200$). *(a)* IoU of VLM-inferred boxes against the median human-drawn reference, with the IoU ≥ 0.5 cutoff and mean marked. *(b)* Three-level agreement per source, with *full*-agreement percentages in the bars and per-source mean IoU below each name.

The mild regime applies an i.i.d. shift of $\pm 5\%$ of box width and height plus an aspect-ratio jitter of $\pm 5\%$. The moderate regime adds an i.i.d. scale jitter of $\pm 10\%$ together with a 5%-probability per-step random rectangular occlusion of up to 20% of the box area. The heavy regime additionally replaces the box with the nearest distractor (Eq. 30 of Appendix C) at a 10% rate. **Table 13** reports the correlation between clean and noisy potentials and the retrained accuracy under each regime.

Table 13: Box-noise propagation. *Spearman* is the rank correlation between Φ_{clean} and Φ_{noisy} over $5{,}000$ held-out trajectories. *In-domain* and *OOD* are the six-source and five-OOD mean accuracies of HSSP retrained from scratch under each regime, three-seed mean.

Box-noise regime	**Spearman**$(\Phi_{\text{clean}}, \Phi_{\text{noisy}})$	**In-domain mean**	**OOD mean**	**Δ vs. clean**
Annotated boxes (no noise)	1.00	53.5	55.1	—
Mild ($\pm 5\%$ jitter)	0.93	52.8	54.6	-0.7 / -0.5
Moderate ($\pm 10\%$ + occlusion)	0.89	52.0	53.7	-1.5 / -1.4
Heavy (moderate + distractor)	0.85	51.0	53.3	-2.5 / -1.8

The rank correlation declines monotonically with noise but stays at 0.85 under the heavy regime, while accuracy stays within 2.5 points in-domain and 1.8 points OOD even at that noise level. The OOD drop is smaller than the in-domain drop under every regime, in line with OOD accuracy depending on learned spatial-search behavior more than on exact target boxes.

Verification Budget. We vary the number of human-verified boxes across $\{50, 200, 500\}$, drawn under the same stratified sampling protocol as the human verification above, with all unverified boxes retained in their VLM-inferred form. In-domain mean accuracy is 51.9 at 50 verified boxes, 53.0 at 200, and 53.4 at 500, approaching the 53.5 obtained with every VLM-grounded box verified, as in our main results. The curve flattens by 500 verified boxes, indicating that the marginal value of additional human review is small once verification coverage approaches that level.

Box-Source Substitution. The boxes used to compute Φ are training-time supervision only. At inference the deployed Seeker takes neither bounding boxes nor any other geometry annotation. To measure the dependence of training on the source of those boxes, **Table 14** replaces the native-or-grounded boxes with two alternative sources and retrains HSSP from scratch under each. The first alternative replaces the native boxes on the three native-box sources with the same Qwen3.5 397B A17B grounding used on the other three sources, so all six sources are then VLM-grounded without human verification. The second alternative replaces every box with the prediction of an off-the-shelf open-vocabulary detector (Grounding-DINO [46]) conditioned on the question.

Under both substitutions HSSP stays above the strongest retrained multi-turn baseline (Mini-o3 at 50.7 in-domain and 51.7 OOD), even with the entire training-time geometry signal taken from

Table 14: Box-source substitution. Training IoU is against human boxes, and inference is unchanged.

Training-time box source	**Training IoU vs. human**	**In-domain mean**	**OOD mean**	**Δ vs. default**
Annotated boxes (default)	0.92	53.5	55.1	—
VLM grounding on all six sources	0.78	51.7	53.6	$-1.8 \,/\, -1.5$
Open-vocabulary detector (Grounding-DINO)	0.69	50.9	52.8	$-2.6 \,/\, -2.3$

unverified VLM grounding or from an off-the-shelf detector. HSSP is therefore robust to the origin of the training-time boxes as long as they preserve the rank-ordering of target relevance.

D Supplementary Experimental Details

This appendix specifies the backbones, benchmarks, baselines, and hyperparameters used in Section 5.

D.1 Evaluation Datasets

The evaluation suite comprises six in-domain V-TRACE sources (**Table 2**) and five zero-shot OOD benchmarks (**Table 3**). For each in-domain source, training uses the V-TRACE trajectories of a held-in portion of the questions, and in-domain evaluation uses the remaining held-out questions.

In-domain Benchmarks. *O3-Bench* [2] pairs 204 layout-informed questions on high-resolution charts and maps with native bounding-box annotations, so intermediate spatial decisions can be checked against ground truth. *V*Bench* [16] contains 191 SA-1B [47] natural images that probe attribute recognition and spatial-relationship reasoning over fine-grained details, each with a single (x, y, w, h) target. *TreeBench* [27] extends the V*Bench protocol to 405 object-dense scenes with bounding-box evidence for every answer, enabling per-step geometric process reward to be checked against the annotated evidence region. *VisualProbe-Hard* [3] is the 106-question hardest subset of VisualProbe, with small targets and many distractors that force iterative trial-and-error search. *HR-Bench 4K* [28] contains 800 ultra-high-resolution questions split evenly between fine-grained single-instance and cross-instance perception on downsampled 4K images. *MME-RealWorld-Lite* [29] is the official $1{,}919$-sample subset of MME-RealWorld [29], spanning 43 task types across autonomous driving, video surveillance, remote sensing, and entertainment scenarios. V*Bench, HR-Bench 4K, and MME-RealWorld-Lite are scored by MCQ accuracy, the other three by exact match.

Out-of-distribution Benchmarks. The five OOD benchmarks never contribute questions or images to V-TRACE, and the perceptual-hash analysis of Appendix C.1 verifies the absence of cross-contamination. Each benchmark probes a sub-skill that visual search relies on and that an effective search policy should transfer to without fine-tuning. *BLINK* [37] reformats fourteen classical computer-vision tasks (including relative depth, visual correspondence, multi-view reasoning, and visual similarity) into $1{,}901$ validation MCQs that isolate the perceptual primitives humans solve "within a blink". *CV-Bench* [38] is the Cambrian-1 vision-centric benchmark of $2{,}638$ MCQs repurposed from ADE20K, COCO, and OMNI3D that test 2D spatial relationships and object counting together with 3D depth ordering and relative distance. *MMMU* [39] contains approximately 11.5K college-level questions across six disciplines and 30 heterogeneous image types, probing robustness under heavy task-and-image distribution shift, and we use the validation split, scored by the official MMMU evaluation protocol (letter match for MCQ items, normalized string match for open items). *RefCOCOg* [40] contains 85K referring expressions over 25K MS-COCO images, scored as correct if the predicted box has $\mathrm{IoU} \geq 0.5$ with the gold box, and it reads the agent's final bounding box rather than its textual answer. *MathVista* [41] aggregates $6{,}141$ multimodal-math problems spanning seven reasoning types, on the standard $1{,}000$-example `testmini` split, in which many items hinge on reading a small numerical or symbolic token from a figure before solving.

D.2 Baselines

Eighteen baselines are organized into seven families, namely proprietary reference, base VLMs, single-turn RLVR, multi-turn visual search, game-based VLM self-play, process-reward multi-turn, and text-domain self-play adapted to vision.

Adaptation Protocol for Non-Native Baselines. All multi-turn baselines reported in **Tables 2–3** are reproduced or adapted under a matched training protocol that aligns HSSP and the baselines on backbone, cold-start supervision, training data, and compute budget. Concretely, every multi-turn baseline (Mini-o3, Insight-o3, DeepEyes, Pixel-Reasoner, Chain-of-Focus, VAGEN, SSP-Visual, Vision-Zero, VisPlay) is (i) initialized from the Qwen2.5-VL-7B checkpoint used for HSSP-Qwen-7B, (ii) given the same 10% V-TRACE cold-start SFT stage that teaches the action vocabulary, and (iii) retrained on V-TRACE under the baseline's own RL objective and trajectory sampling, with a matched number of outer-loop iterations and the same $32\times$A100-80GB compute allocation. The matched protocol fixes the hardware allocation and the number of outer-loop iterations, and per-iteration compute follows each method's own design, such as an external judge call or a multi-agent loop. The five *Multi-Turn Visual Search* methods are retrained from their published procedures, namely Mini-o3

(GRPO with verifiable rewards), DeepEyes (end-to-end agentic RL), Pixel-Reasoner (two-phase GRPO with a curiosity bonus), Chain-of-Focus (two-stage SFT + RL), and Insight-o3 (vSearcher RL). Crop-and-backtrack action vocabulary is shared across all multi-turn methods. Method-specific reward signals (the conditional tool-use bonus in DeepEyes, the curiosity bonus in Pixel-Reasoner, format compliance in Chain-of-Focus, LLM-judge in VAGEN) are preserved unchanged, so that any two multi-turn methods share data, backbone, and compute and differ by design in the reward and update mechanism. Game-based VLM self-play methods (Vision-Zero, VisPlay) retain their original roles, self-play loop, and reward but operate in the same crop-based observation space, and SSP-Visual follows [12] with the text `<SEARCH>` action replaced by `<CROP>`. Base VLMs and single-turn RLVR methods (R1-OneVision-7B, VLAA-Thinker, and OpenVLThinker on Qwen2.5-VL-7B, together with MM-Eureka-8B on its native InternVL2.5-Instruct-8B) retain their original single-pass formulation and therefore emit no crop or backtrack actions, and MM-Eureka-8B additionally serves as a cross-backbone RLVR control. All methods share the evaluation prompt and held-out splits.

Proprietary Reference. Commercial VLMs evaluated as a closed-model reference on the in-domain benchmarks under the same prompt budget.

- GPT-4o [42]. A multimodal instruction-tuned OpenAI model evaluated under the same system and user prompts as the open-weight models. Inference is zero-shot with no in-context examples, and the model internalizes its chain of thought.
- Gemini-2.5-Flash. A Google multimodal instruction-tuned model evaluated under the same evaluation prompt and zero-shot protocol.

Base VLMs. The three backbones of Appendix D.3, Qwen2.5-VL-7B [32], InternVL3-8B [33], and InternVL-14B [34], evaluated without V-TRACE post-training, isolating its effect.

Single-Turn RLVR Post-Training. Methods that apply rule-based verifiable rewards on a single forward pass and therefore isolate the contribution of multi-turn structure.

- R1-OneVision-7B [23]. Applies a DeepSeek-R1-style verifiable-reward training procedure to a Qwen2.5-VL backbone under GRPO with correctness-only rewards on mixed math and visual-reasoning tasks. Training mixes curated math, chart, and natural-image problems, each evaluated by a rule-based verifier.
- MM-Eureka-8B [24]. Extends the Eureka accuracy-only RLVR procedure from text reasoning to multimodal inputs under GRPO on multi-step math and perception problems. The reward combines a lightweight format-compliance check with a rule-based correctness verification on the final answer. Its released checkpoint uses InternVL2.5-Instruct-8B.
- VLAA-Thinker [25]. A two-stage method that first distills an R1-style reasoning trace into the student VLM under supervised learning and then fine-tunes the student under RLVR on visual-language tasks. The distillation stage transfers long-form chain-of-thought patterns, and the subsequent RL stage uses rule-based correctness rewards under GRPO.
- OpenVLThinker [26]. An open-source RLVR baseline that alternates supervised fine-tuning with RLVR over visual-reasoning tasks under correctness-only rewards. Each round pairs supervised updates on model-generated traces with on-policy RLVR steps.

Multi-Turn Visual Search. The family closest to HSSP, sharing the crop-and-reason interaction loop so that comparison isolates the contribution of self-play and of the geometric reward.

- Mini-o3 [3]. Trains a VLM to emit crop actions and reasoning over long interaction horizons under outcome-only RLVR. A curriculum over trajectory length extends the interaction budget beyond ten turns and stabilizes long-horizon credit assignment.
- Insight-o3 [2]. Pairs a commodity verifier VLM with a fine-tuned searcher that issues zoom and crop actions under RLVR. The verifier supplies a rule-based format check, and the searcher is optimized against the resulting joint reward.
- DeepEyes [5]. Trains multi-turn visual search end to end with reinforcement learning under an outcome reward of answer accuracy and format, together with a tool-use bonus granted only if the answer is correct and at least one zoom-in step is taken.

- Pixel-Reasoner [6]. Performs pixel-space reasoning through visual operations such as zoom-in, trained with a curiosity-driven reward that encourages these operations over purely textual reasoning. Crop actions operate on raw image patches, and the intermediate state is fed back to the VLM as visual tokens.
- Chain-of-Focus [19]. Iteratively refines a bounding box by following a natural-language chain of focus steps emitted by the VLM. Each focus step emits a textual instruction and an updated crop, and the VLM is trained with format-compliance rewards on the focus steps.

Game-Based VLM Self-Play. Non-spatial self-play baselines that isolate the gain of the geometric process reward over a generic self-play loop.

- Vision-Zero [13]. Originally a multi-VLM clue-guessing game on VQA-style images, Vision-Zero assigns different roles to VLM copies and awards a zero-sum reward for the correctness of each copy's vote. The clue-generation and decision phases, the hysteretic stage switch, and the original zero-sum reward are transferred onto the V-TRACE skeleton pool and the crop-based action vocabulary of HSSP, so the two differ in game structure, roles, and per-turn reward.
- VisPlay [14]. Originally a questioner–reasoner self-play formulation for single-turn VQA, VisPlay co-evolves a question generator and an answering policy over a pool of images. The REINFORCE objective, the questioner–reasoner role split, and the original answer-correctness reward are transferred onto the same V-TRACE skeleton pool and action vocabulary as HSSP, with the questioner selecting skeletons instead of synthesizing images.

Process-Reward Multi-Turn. Directly compares geometric and judge-based process rewards.

- VAGEN [8]. Trains a multi-turn VLM with PPO and the Bi-Level GAE estimator that HSSP also builds on, taking a natural-language judgment score as the per-turn reward. VAGEN is retrained on V-TRACE under the same data, backbone, cold-start, and compute allocation as HSSP, so the two methods differ in the per-turn reward and in HSSP's Hider.

Text-Domain Self-Play Adapted to Visual Search. Isolates self-play as a generic training mechanism from self-play structured around image geometry.

- SSP-Visual [12]. Adapts SSP self-play by replacing `<SEARCH>` tokens with `<CROP>` over visual regions while keeping the questioner–answerer self-play formulation, the answerability reward, and the policy-gradient objective intact. The questioner proposes cropped regions and the answerer produces a response, with both roles optimized jointly under REINFORCE.

D.3 Training and Hyperparameter Settings

Backbones. Following [13], we use two instruction-tuned and one foundation-pretrained VLM.

- Qwen2.5-VL-7B-Instruct [32]. Primary backbone. A 7B instruction-tuned VLM with native bounding-box grounding. As for every backbone, its crops are expressed in the normalized $[0, 1]^2$ coordinates of the action grammar (Appendix D.4).
- InternVL3-8B [33]. Cross-architecture control at comparable scale. InternVL3-8B pairs an InternViT-300M vision encoder with a Qwen2.5-7B language core through an MLP projector, and its coordinate convention and visual encoder both differ from Qwen's, isolating backbone-specific from reward-design effects.
- InternVL-14B [34]. Larger foundation-pretrained control. An earlier 14B vision-language foundation model that is less aligned to the multi-turn search action vocabulary, testing the transfer of HSSP's gains beyond instruction-tuned backbones.

All three backbones are initialized from their public checkpoints without further pretraining, and the post-training behavior reported in the main body derives from the cold-start SFT stage below followed by the HSSP updates of Algorithm 1.

Cold-Start Supervised Fine-Tuning. Ten percent of the V-TRACE trajectories are used for one epoch of supervised fine-tuning that teaches the backbone the action vocabulary of Section 4. The loss is next-token cross-entropy on the action-and-thought tokens with image tokens kept frozen, and the optimizer is AdamW with linear warm-up and cosine decay (**Table 15**).

Reinforcement Learning Stage. HSSP self-play follows standard PPO-Lagrangian conventions and is implemented with verl [35] and VAGEN [8]. The Seeker performs clipped PPO updates against Bi-Level GAE advantages with a KL penalty to the SFT reference policy, and the Hider performs a REINFORCE update with the role-advantage estimator [10]. Trajectory sampling and optimization share 32×A100-80GB GPUs, with rollouts generated by vLLM [36] under tensor-parallel decoding and parameters sharded with FSDP. All learning rates, clipping parameters, GAE constants, batch sizes, and self-play constants appear in **Table 15** and are held fixed across backbones. Each outer iteration takes one dual step at $\eta_\lambda{=}10^{-2}$ and K_{PPO} primal epochs at learning rate 10^{-6}.

Table 15: HSSP hyperparameter configuration.

Symbol	Value	Description
Reward design		
α	0.5	Weight on maxIoU inside the target-approach potential
c	0.5	Saturation level of the exploration potential
γ	0.99	Discount factor used by PBRS and GAE
B	10	Step-budget constraint in Eq. 3
κ	0.337	Redundancy constraint in Eq. 3
η_λ	0.01	Dual step size for both $\lambda_B, \lambda_\kappa$
$\lambda_{\max}$	10	Upper bound of the dual variables
Self-play loop		
τ_{hi}	0.75	Held-out accuracy threshold, Phase A → B
τ_{lo}	0.45	Held-out accuracy threshold, Phase B → A
$K_{\min}$	50	Minimum dwell time before a phase switch is allowed
T_{iter}	3000	Total outer-loop iterations
PPO-Lagrangian inner loop		
η_{θ_S}	10^{-6}	Seeker actor (policy) learning rate
η_{ϕ_S}	$5{\times}10^{-6}$	Seeker critic (value) learning rate
η_{θ_H}	$5{\times}10^{-6}$	Hider REINFORCE learning rate
ϵ	0.2	PPO clip ratio
λ_{GAE}	0.95	GAE λ
β_{KL}	10^{-3}	KL penalty against the SFT reference policy
Rollout batch	64	Rollouts per PPO iteration
PPO epochs	4	Inner-loop gradient epochs per batch
Rollout turns	≤ 20	Maximum rollout length in training and inference

Training Compute. Under identical hardware (32×A100-80GB) and identical $T_{\mathrm{iter}}{=}3000$ outer-loop iterations with the same trajectory sampler (**Table 16**), HSSP trains in 124.3 A100-hours against 168.4 for VAGEN, a 1.35× trajectory throughput. HSSP requires no judge queries, which account for 39.1% of VAGEN's compute, and its Hider takes 7.2% of its own.

Table 16: Training-time compute breakdown. Totals (A100-hours) are three-seed means, and percentage shares are pooled across seeds.

	Total (A100-h)	**Rollout**	**Judge**	**Policy and critic update**	**Hider**	**Other**
VAGEN	168.4	36.2%	39.1%	22.7%	–	2.0%
HSSP-Qwen-7B	**124.3**	55.7%	–	33.4%	7.2%	3.7%
Relative to VAGEN	1.35× throughput	+19.5pp	−39.1pp	+10.7pp	+7.2pp	+1.7pp

Evaluation-Time Inference. At evaluation time only the Seeker is retained, with the Hider discarded. Decoding is greedy at temperature 0 with at most 20 actions per trajectory. Each per-seed

accuracy is the bootstrap mean over 3000 trajectory-level resamples of that seed's held-out partition, and the reported value is the mean of the three per-seed bootstrap means. Bootstrap 95% confidence intervals over the three-seed population enter only the paired-bootstrap significance tests of **Table 17**, which compare HSSP-Qwen-7B with the strongest retrained multi-turn baseline on each benchmark and are cited in Section 5.

Table 17: HSSP-Qwen-7B against the strongest retrained multi-turn baseline on each benchmark under the paired bootstrap above. ✓ marks $p < 0.05$ and n.s. a non-significant gap.

Benchmark	**Strongest retrained multi-turn**	**HSSP**	Δ	**Sig.**
O3-Bench	30.0 (VAGEN)	37.4	+7.4	✓
V*Bench	85.0 (Mini-o3)	83.6	−1.4	n.s.
TreeBench	39.5 (Insight-o3)	42.7	+3.2	✓
VisualProbe-Hard	32.0 (Mini-o3)	31.3	−0.7	n.s.
HR-Bench 4K	77.0 (Insight-o3)	76.4	−0.6	n.s.
MME-RealWorld-Lite	47.1 (Vision-Zero)	49.6	+2.5	✓
In-domain mean	50.7 (Mini-o3)	53.5	+2.8	✓
BLINK	46.0 (Mini-o3)	50.0	+4.0	✓
CV-Bench	57.5 (Mini-o3)	60.5	+3.0	✓
MMMU	45.0 (Mini-o3)	48.8	+3.8	✓
RefCOCOg	66.2 (VAGEN)	67.5	+1.3	✓
MathVista	47.0 (VAGEN)	48.7	+1.7	✓
OOD mean	51.7 (Mini-o3)	55.1	+3.4	✓

Test-Time CMDP Adherence. The CMDP duals $\lambda_B, \lambda_\kappa$ are training-only quantities and are not applied at inference. The deployed policy is the final Seeker iterate, the empirical counterpart of the best response θ_S^{**} at the saddle of Eq. 22, evaluated under greedy decoding with the turn limit $T_{\max}$=20 and no penalty term. **Table 18** reports the resulting trajectory-length and redundancy distributions on the in-domain and OOD test partitions. Mean turns are 8.1 in-domain and 8.4 OOD, both well below the budget B=10, and the violation rate $\Pr[T > B]$ remains under 5% on either partition. Mean cumulative redundancy $\mathbb{E}[\sum_t \mathrm{Redun}_t]$ is 0.184 in-domain and 0.211 OOD, both below the CMDP constraint κ=0.337. This cumulative quantity is distinct from the repeated-crop *rate* (0.122 in-domain) reported in Appendix D.7 and **Fig. 2**. The rate counts the per-trajectory fraction of crops with overlap above 0.5 with prior crops, while the cumulative quantity integrates the per-step overlap fraction Redun_t over the rollout. The deployed Seeker thus meets the two CMDP constraints in expectation without dual penalties.

Table 18: Test-time distribution of trajectory length T and redundancy under the deployed Seeker, three-seed mean. *Overlapping crops* counts the crops with $\mathrm{Redun}_t > 0$ per trajectory, a looser criterion than the repeated-crop rate, which counts crops with overlap above 0.5. Budget B=10 and redundancy bound κ=0.337 are training-time constraints, and the test-time turn limit is $T_{\max}$=20.

Partition	**Mean** T	**Median** T	$p_{95}(T)$	$\Pr[T>B]$	**Overlapping crops**	**Mean** $\sum_t \mathrm{Redun}_t$
In-domain (six sources)	8.1	8	10	3.2%	1.7	0.184
OOD (five benchmarks)	8.4	8	10	4.1%	1.9	0.211

D.4 Inference-Time Seeker Prompt

All multi-turn accuracy results in Tables 2–3 are produced under a single Seeker system prompt that fixes the action grammar of Section 4 and the per-trajectory turn limit. The prompt is deliberately minimal, so search behavior reflects the trained policy. The same prompt applies to every backbone and multi-turn baseline on all eleven benchmarks. Base VLMs and single-turn RLVR baselines retain their original single-pass formulation and never see this prompt.

PROMPT | Inference-Time Seeker (system role, all backbones)

You are a long-horizon visual-search agent. Given a high-resolution image and a question, you may emit a sequence of actions, one per turn, drawn from the following grammar:

- `<PLAN>...</PLAN>`: a short natural-language plan describing the next region you intend to inspect and why.
- `<CROP` x_1 y_1 x_2 y_2`>`: zoom into the rectangle $[x_1, y_1, x_2, y_2] \subseteq [0, 1]^2$ in normalized image coordinates. The crop must be non-degenerate.
- `<BACKTRACK` k`>`: pop the top k crops from the active view stack to revisit an earlier region.
- `<VERIFY>...</VERIFY>`: compare visible evidence against the question using both supporting and contradicting cues before committing.
- `<STOP` ⟨`answer`⟩`>`: emit the final answer in the format required by the benchmark and terminate the trajectory.

You may issue up to 20 actions per question. If the question presupposes a target that is not present in the image, emit `<STOP None>`. Reply only in the action grammar above.

Per-Benchmark Output Contracts. The only per-benchmark variation is the *output contract*, the answer format that the final `<STOP>` action must follow under the official scoring protocol of the benchmark. **Table 19** lists the answer type, the exact format directive appended to the user message at evaluation time, and the scoring rule applied to the answer. Before scoring, the answer in the final `<STOP>` action is converted to the input format of the official scorer (a text answer for MCQ-letter and free-form benchmarks, a box quadruple for RefCOCOg, and a type-specific answer for MathVista), which then normalizes and matches it. Scorer versions are documented in the supplementary material.

Table 19: Per-benchmark output contracts. The Seeker prompt and action grammar are identical across benchmarks. The *format directive* is appended to the user message and specifies the answer format of the final `<STOP>` action. The *scorer* is the official evaluation protocol of each benchmark, applied after the answer is converted to its input format under the rules stated in the boxes below.

Benchmark	Answer type	Format directive (verbatim)	Scoring rule
In-domain (six V-TRACE *sources)*			
O3-Bench [2]	free-form	"Output the layout-region answer as a single phrase."	case/whitespace-normalized exact match
V*Bench [16]	MCQ	"Answer with a single letter A, B, C, or D."	letter match
TreeBench [27]	free-form	"Output the target category label as a single phrase."	case-insensitive exact match
VisualProbe-Hard [3]	free-form / absent	"Output a single phrase, or `None` if the target is absent." (full form in the box below)	case-insensitive exact match (with absent handling)
HR-Bench 4K [28]	MCQ	"Answer with a single letter."	letter match
MME-RealWorld-Lite [29]	MCQ	"Answer with a single letter."	letter match
Out-of-distribution (five held-out benchmarks)			
BLINK [37]	MCQ	"Answer with a single letter."	letter match
CV-Bench [38]	MCQ (up to 6 opts)	"Answer with the option letter in parentheses, e.g., `(C)`."	parenthesized-letter match
MMMU [39]	MCQ + open-ended	MCQ items: "Answer with a single letter.", open items: "Output a short phrase."	official MMMU scorer (letter match for MCQ, normalized string match for open)
RefCOCOg [40]	bounding box	"Output the box as $[x_{\min}, y_{\min}, x_{\max}, y_{\max}]$ in $[0, 1]^4$." (full form in the box below)	IoU ≥ 0.5 against gold (Acc@0.5)
MathVista [41]	MCQ + free-form	item-type-conditional, see the box below	official MathVista scorer (answer extraction + per-item-type normalization)

Worked-Example Output Contracts. The three benchmarks below have output contracts that diverge from the MCQ-letter or short-phrase format and therefore warrant verbatim format directives.

PROMPT | RefCOCOg (Bounding-Box Output Contract)

The benchmark question is appended verbatim, then the following directive is added to the user message:

> *"Localize the referent described above. Once you have committed to a region, emit <STOP [$x_{\min}, y_{\min}, x_{\max}, y_{\max}$]> where the four coordinates are floats in $[0, 1]$ in normalized image space (top-left origin), with two-decimal precision. Do not output natural-language descriptions of the box."*

The bracketed quadruple in the final <STOP> action is rescaled to absolute pixels and scored with the standard RefCOCO/+/g IoU protocol. A prediction is counted as correct if IoU ≥ 0.5 against the gold pixel box (Acc@0.5). Answers that miss the bracket pattern, or boxes with coordinates outside $[0, 1]$, score 0.

PROMPT | VisualProbe-Hard (Absent-Target Output Contract)

Every VisualProbe-Hard question is followed by the directive below, the full form of its **Table 19** entry:

> *"If the question presupposes an entity that is not present anywhere in the image, you must demonstrate systematic coverage (at least four crops that jointly cover the full image) before emitting <STOP None>. Do not output None until the coverage requirement is satisfied. If the entity is present, emit <STOP ⟨phrase⟩> as usual."*

An answer is correct if (i) the entity is absent and the rollout meets the coverage criterion and emits <STOP None>, or (ii) the entity is present and the phrase exactly matches the gold answer (ignoring case and whitespace).

PROMPT | MathVista (Item-Type-Conditional Output Contract)

MathVista items are annotated with an answer_type $\in$ {integer, float, list, text, multi_choice} together with an optional precision. The directive appended to each question is conditioned on this metadata:

> *"If this item is multiple-choice, emit <STOP ⟨letter⟩>. If it is integer or float, emit <STOP ⟨number⟩> in plain digits matching the requested precision, with no units, no thousands separator, and no natural-language explanation. If it is a list, emit <STOP ⟨comma-separated values⟩>. If it is free-form text, emit <STOP ⟨short phrase⟩>. Do not include explanation text inside the <STOP> payload."*

The answer in the final <STOP> action is then scored by the official MathVista protocol, which performs answer extraction and per-item-type normalization before the final match.

D.5 Per-Trajectory-Type Decomposition

Across the three V-TRACE trajectory types (**Table 20**), the largest gap between HSSP and VAGEN under matched data and compute appears on Type-A (backtrack-and-recover, $+4.7$ against $+3.7$ on Type-B and $+2.7$ on Type-C), since the max-pool potential retains the progress of the corrected crop through BACKTRACK (Theorem 2). The smaller Type-C gap is consistent with absent declarations being tractable under both reward shapes once the coverage skeleton is internalized.

Table 20: Per-trajectory-type accuracy decomposition on the held-out V-TRACE partition. *Type-A*: backtrack-and-recover. *Type-B*: distractor contrast. *Type-C*: systematic absence. *Mean* is the trajectory-level type-weighted mean of each row, which differs from the source-weighted mean of **Table 2**. Δ is the gap of HSSP-Qwen-7B over VAGEN.

Method	Type-A	Type-B	Type-C	Mean
Qwen2.5-VL-7B (base)	42.0	46.0	51.2	46.4
Insight-o3 (retrained)	47.5	50.0	54.0	50.5
Mini-o3 (retrained)	46.0	49.0	52.8	49.3
VAGEN (retrained)	44.3	47.8	52.5	48.2
HSSP-Qwen-7B	**49.0**	**51.5**	**55.2**	**51.9**
Δ over VAGEN	$+4.7$	$+3.7$	$+2.7$	$+3.7$

Per-Type Linear-Scalar Merge Ablation. The aggregate ablation of **Table 4** shows a -2.1 drop on the source-weighted six-source mean if the max-pool potential is replaced by a linear-scalar merge of the same target-approach and exploration components. **Table 21** decomposes that drop on the

same held-out partition as **Table 20** (trajectory-level type-weighted mean 51.9, distinct from the source-weighted 53.5 used in **Table 2**). The drop is mildest on Type-A (-0.9), moderate on Type-B (-1.9), and largest on Type-C (-2.7), with a type-weighted mean drop of -1.8. The smaller drop on the trajectory-level mean compared to the source-weighted -2.1 reflects source-level reweighting toward regimes (O3-Bench, VisualProbe-Hard) on which the linear-scalar merge lowers accuracy most. The per-type pattern is consistent with Theorem 3, since a linear-scalar merge collapses target-approach and exhaustive-coverage along a single axis, eroding the closed-form $\gamma^L(1-c)$ separation. Type-A, in which the target box is reached again after a deliberate wrong crop, tolerates the loss because the recovery crop contributes through both branches. Type-C, in which the trajectory must declare absence after systematic coverage, is most exposed because the exploration branch is the only signal of progress and its mass is diluted by a target-approach term that stays at zero.

Table 21: Per-type accuracy ($\%$) under the max-pool potential of Eq. 2 (HSSP) versus a linear-scalar merge ablation that replaces the pointwise maximum by the equal-weight convex combination of the two branches (Appendix D.9). Held-out V-TRACE test partition, three-seed mean.

Trajectory type	**HSSP (max-pool)**	**Linear-scalar merge**	Δ
Type-A (backtrack-and-recover)	49.0	48.1	-0.9
Type-B (distractor contrast)	51.5	49.6	-1.9
Type-C (systematic absence)	55.2	52.5	-2.7
Type-weighted mean	51.9	50.1	-1.8

D.6 Failure-Mode Taxonomy

Residual errors of HSSP-Qwen-7B and VAGEN fall into five mutually exclusive categories, obtained by manual labeling of 400 failure cases per model and shown in **Fig. 9**. *Localization-then-misread* denotes a wrong answer after the Seeker locates the right region. *Never-localize* denotes termination without inspecting the target. *Over-explore* denotes exhausting the turn budget on redundant or irrelevant crops. *Hallucinated-absent* denotes a `<STOP None>` answer for a target that is in fact present. *Other* covers answers that violate the output format. Under VAGEN the residual errors are dominated by *never-localize* (37.8%) and *over-explore* (24.1%), the two failure modes most directly addressed by the geometric process reward. Under HSSP the largest residual category is *localization-then-misread* (41.2%), an error that arises after the Seeker has reached the target region, so the remaining errors shift from search to backbone-level reading.

The adaptive Hider also targets these search failures. Its per-bucket sampling mass tracks the Seeker's rate of never-localize and over-explore failures with $\rho = 0.71$, and it correlates only weakly with the base VLM's single-pass error rate ($\rho = 0.18$). The association holds after controlling for that error rate and for the source, type, and difficulty bin of each bucket, and the inverse-success-rate sampler of Appendix D.10 shows the reverse ordering. The Hider therefore targets skeletons on which search fails rather than those the backbone finds hard to perceive.

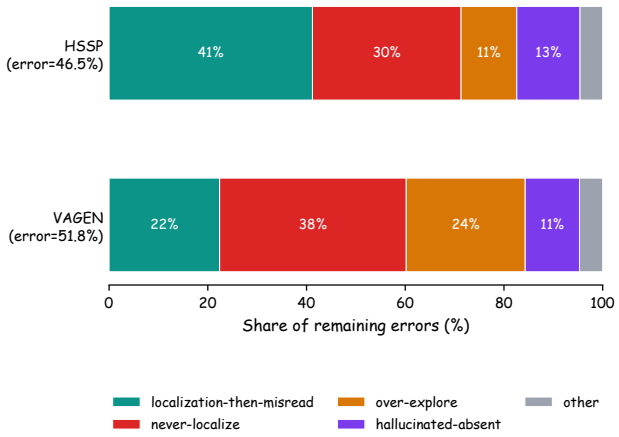


Figure 9: Failure-mode breakdown for HSSP-Qwen-7B versus VAGEN. Bars are stacked over the residual error budget of each method (HSSP 46.5%, VAGEN 51.8%).

D.7 Reward-Integrity Analysis

The analysis checks the Seeker trained under HSSP for reward hacking, namely optimizing the potential Φ at the expense of final-answer correctness. Six diagnostics, computed on the three seeds of the final HSSP-Qwen-7B policy with a bootstrap 95% confidence interval for the Pearson correlation, show that the geometric potential stays aligned with answer correctness. The Pearson correlation between per-trajectory mean Φ and answer correctness is 0.62 ± 0.03, the high-Φ wrong-answer tail is 7.8%, and the low-Φ correct-answer tail is 5.3%, all in line with the policy invariance of PBRS shaping (Theorem 1). The repeated-crop rate falls over training to 0.122 (**Fig. 2**d), and the premature-`VERIFY` rate is 0.041, indicating conservative search under the joint effect of the geometric potential and the κ-redundancy constraint. On Type-C the absent-declaration calibration is 0.84, consistent with the cumulative-coverage branch supporting systematic-absence behavior.

Correlation $\mathrm{Corr}(\bar{\Phi}, \mathbb{1}[\hat{y} = a^*])$**.** For each held-out test trajectory we compute the mean $\bar{\Phi} = \frac{1}{T}\sum_t \Phi(s_t)$ and the binary correctness indicator, and take the Pearson correlation across trajectories.

High-Φ Wrong-Answer Tail and Low-Φ Correct-Answer Tail. Two tail quantities measure dissociation between Φ and final correctness. With $\Phi_{\text{final}} := \Phi(s_{T-1})$, they are the fraction of trajectories with $\Phi_{\text{final}} > 0.5$ yet $\hat{y} \neq a^*$, and the fraction with $\Phi_{\text{final}} < 0.1$ yet $\hat{y} = a^*$.

Repeated-Crop and Premature-`VERIFY` Rates. A repeated crop is one whose overlap with the union of prior crops exceeds 0.5, and we report the mean rate of repeated crops per trajectory. A premature `<VERIFY>` is a `<VERIFY>` emitted while $\Phi(s_t) < 0.3$, and we report its per-trajectory rate.

Absent Calibration on Type-C Trajectories. For Type-C test trajectories (target absent) we report the absent-declaration calibration $\mathbb{P}[\hat{y} = \texttt{None} \mid \text{target genuinely absent}]$.

D.8 Judge-Free Non-Geometric Progress Control

The VAGEN baseline compares LLM-judge supervision against HSSP's geometric supervision as training regimes. A second control isolates the orthogonal axis of *geometric versus generic dense progress signals*. It replaces Φ with a CLIP-based semantic-progress potential $\Phi_{\text{CLIP}}(s_t) = \cos\big(\text{CLIP}_{\text{img}}(\text{view}(s_t)),\ \text{CLIP}_{\text{txt}}(q)\big)$ that scores the agent's current viewport against the question text using a frozen CLIP-ViT-L/14 encoder [48], while keeping the canonical PBRS shape, the CMDP constraints, the Bi-Level GAE estimator, and the self-play loop unchanged. The CLIP control therefore shares everything with HSSP except the source of the per-step progress signal, which is judge-free and dense but not verifiable against the recorded target region. **Table 22** reports accuracy alongside reward-integrity metrics that probe the exploitation of each reward shape.

Table 22: Judge-free non-geometric control. *Reward shape* is the per-step dense supervision used during training. The CLIP control replaces only the potential, VAGEN is the retrained baseline of **Table 2**, and outcome-only RLVR is the sparse-supervision endpoint of the comparison, not a single-component ablation. *Pearson* correlates the trajectory-mean dense signal with final correctness. *High-reward wrong* is, for HSSP, the high-Φ wrong-answer tail of Appendix D.7 and, for the CLIP control and VAGEN, the wrong-answer rate among trajectories whose mean reward lies in the top decile. *Repeat crop* is the mean repeated-crop rate. Dashes mark metrics that are undefined for outcome-only RLVR, whose reward is answer correctness itself.

Reward shape	**In-domain mean**	**OOD mean**	**Pearson**	**High-reward wrong ↓**	**Repeat crop ↓**
HSSP (geometric Φ, judge-free)	53.5	55.1	0.62	7.8%	0.122
CLIP-progress potential (judge-free)	49.8	50.5	0.43	14.7%	0.187
LLM-judge process reward (VAGEN)	48.2	51.4	0.51	11.3%	0.187
Outcome-only RLVR (no dense reward)	46.4	50.7	–	–	0.241

On in-domain accuracy the CLIP control exceeds both VAGEN and outcome-only RLVR (49.8 against 48.2 and 46.4), while on OOD it is comparable to outcome-only (50.5 against 50.7) and below VAGEN (51.4). A dense semantic signal is therefore informative on training-distribution skeletons but does not by itself improve over outcome supervision under distribution shift. Among its top-decile trajectories the CLIP control answers 14.7% incorrectly, against 11.3% for VAGEN, because a viewport that globally matches the question's textual semantics can be unrelated to the actual target box. The repeated-crop rate of the CLIP control matches VAGEN's at 0.187, higher than HSSP's 0.122, because neither non-geometric reward feeds the redundancy-aware spatial signal that the cumulative-multiset $\mathrm{cov}_{\mathrm{img}}$ branch provides. The geometric reward therefore improves in-domain accuracy, OOD accuracy, and reward-integrity metrics jointly.

Outcome-only RLVR isolates the contribution of answer reward alone (**Table 23**). It leaves the in-domain mean at the base level, with the largest drops on O3-Bench and VisualProbe-Hard, the two sources that require iterative localization, while it raises all five OOD benchmarks. Sparse reward thus retains the crop-and-zoom interface from cold-start SFT, which benefits shorter-horizon perceptual tasks, but does not teach long-horizon search, the capability that the dense geometric reward supplies.

Table 23: Per-benchmark accuracy of outcome-only RLVR (**Table 22**) and of the base model.

In-domain	O3-B	V*	TreeB	VP-Hard	HR-4K	MME-RW	**Mean**
Qwen2.5-VL-7B (base)	27.0	74.8	36.7	24.3	68.6	47.2	46.43
Outcome-only RLVR	26.1	75.6	36.4	23.1	68.9	48.2	46.38
OOD	BLINK	CV-B	MMMU	RefC	MathV		**Mean**
Qwen2.5-VL-7B (base)	39.7	52.9	43.8	61.3	45.2		48.6
Outcome-only RLVR	43.6	55.8	44.5	63.9	45.6		50.7

D.9 Full Six-Source Ablation Matrix

Across all six in-domain sources, the ablation pattern is consistent across the four groups (**Table 24** expands **Table 4**). The largest drops cluster on O3-Bench (layout-grounded) and VisualProbe-Hard (long-horizon, distractor-heavy), while the other four sources see drops of at most 3.6 points. MME-RealWorld-Lite is the only source on which the $\mathrm{cov}_{\mathrm{img}}$-only ablation rises above the full configuration, consistent with many MME questions being answerable through coverage alone. The two pooling alternatives isolate the role of the exploration cap. With Φ_1 the target-approach branch of Eq. 2, the sum-pool variant uses the clipped sum $\min(1, \Phi_1 + c\,\mathrm{cov}_{\mathrm{img}})$, which keeps the margin $\gamma^L(1-c)$ of Theorem 3 but also credits coverage beyond the target, and it loses 0.5. The linear-scalar merge uses the convex combination $\frac{1}{2}\Phi_1 + \frac{1}{2}c\,\mathrm{cov}_{\mathrm{img}}$, which halves both the margin and the exploration signal on absent skeletons, and it loses 2.1. The max-pool potential keeps the full margin and credits coverage only while it exceeds target approach.

Table 24: Full six-source ablation matrix for HSSP-Qwen-7B. Convention follows Table 2.

Configuration	O3-B	V*	TreeB	VP-Hard	HR-4K	MME-RW	**Mean**	Δ
Full HSSP	37.4	83.6	42.7	31.3	76.4	49.6	**53.5**	–
Self-Play (Hider) variants								
uniform Hider	34.8	81.5	41.7	29.6	75.5	48.9	52.0	−1.5
random Hider (matched volume)	35.1	81.7	41.8	29.9	75.8	49.0	52.2	−1.3
frozen Hider	36.1	82.0	42.3	30.5	76.3	49.2	52.7	−0.8
Reward structure								
w/o PBRS shaping ($r^{\text{shape}}\equiv 0$)	33.0	81.0	41.0	28.3	74.3	49.0	51.1	−2.4
linear-scalar merge	33.8	81.2	41.3	28.9	74.8	48.3	51.4	−2.1
task-type-switch Φ	35.8	81.8	42.2	30.2	76.2	49.4	52.6	−0.9
CMDP constraints								
w/o duals ($\lambda_B{=}\lambda_\kappa{=}0$)	34.5	81.3	41.7	29.6	75.2	50.2	52.1	−1.4
w/o step budget ($\lambda_B{=}0$)	35.6	82.0	42.1	30.2	76.2	49.5	52.6	−0.9
w/o redundancy ($\lambda_\kappa{=}0$)	35.8	82.0	42.2	29.9	76.3	49.9	52.7	−0.8
Potential Φ composition								
$\Phi{=}\,\text{maxIoU}$ only	33.8	81.5	41.3	29.2	75.4	49.5	51.8	−1.7
$\Phi{=}\,\text{cov}_{\text{tgt}}$ only	33.3	81.3	40.9	28.9	74.9	48.4	51.3	−2.2
$\Phi{=}c{\cdot}\text{cov}_{\text{img}}$ only	30.4	80.3	39.8	25.5	72.8	51.1	50.0	−3.5
sum-pool variant	36.6	82.2	42.6	31.1	76.7	48.9	53.0	−0.5
$c{=}0$ (no exploration)	34.2	81.2	41.2	28.6	74.9	49.7	51.6	−1.9
$c{=}1$ (saturation symmetry)	34.8	81.2	41.5	29.6	75.3	50.2	52.1	−1.4

D.10 Alternative Hider Samplers

The Self-Play group of **Table 24** replaces the learned Hider with uniform, frozen, and matched-volume random sampling. To separate the adversarial objective from adaptive reweighting, we retrain the Seeker under four heuristic curricula and a learned sampler over the same 54 buckets (**Table 25**). The Hider leads every alternative on both means, and a paired bootstrap separates it at $p < 0.05$ from the strongest, the learned sampler. It shares the Hider's architecture and optimizer but optimizes learning progress, so the remaining margin isolates the adversarial objective, not the capacity to learn a sampler. Heuristic samplers recover at most half of the Hider's margin over uniform.

Table 25: In-domain and OOD mean accuracy of HSSP-Qwen-7B under each sampler.

Sampler	**In-domain**	**OOD**
Uniform (**Table 4**)	52.0	53.9
Static hard-bucket weighting	51.3	53.3
Low-process-reward sampling	52.2	54.1
Temperature-based sampling over recent failures	52.5	54.5
Inverse-success-rate sampling	52.6	54.2
Learned sampler (learning-progress objective)	52.9	54.6
Learned Hider (HSSP)	**53.5**	**55.1**

Hider Gain by Pool Size. The Hider's gain over uniform sampling widens with pool size (**Table 26**), from $+0.6$ at 25% to $+1.5$ at 100%, so adaptive selection matters more in larger pools. On the full pool it concentrates on long-horizon regimes, with $+2.6$ on O3-Bench and $+1.7$ on VisualProbe-Hard against $+0.7$ on MME-RealWorld-Lite (**Table 24**), $+3.2$ on the hardest tertile, and $+2.4$ on Type-A.

D.11 Fixed Penalties in Place of the Duals

The CMDP group of **Table 4** removes length and redundancy control. A stronger alternative keeps it and replaces the duals with fixed per-turn penalties c_B and $c_\kappa\,\text{Redun}_t$. We search (c_B, c_κ) over a $4{\times}4$ grid with one seed per configuration, selecting by the held-out accuracy that drives phase switching (Algorithm 1), and retrain the best configuration ($c_B = 0.01$, $c_\kappa = 0.1$) with three seeds (**Table 27**).

Table 26: In-domain mean accuracy of the learned and uniform Hider by pool size.

Pool size	Uniform	Learned Hider	Gain
25%	50.6	51.2	+0.6
50%	51.5	52.3	+0.8
100% (full pool)	52.0	53.5	+1.5

The best fixed penalty trails the duals by 0.7 points in-domain ($p < 0.05$, paired bootstrap) and 1.1 points OOD. Only 3 of 16 configurations come within 0.5 points of it, and the grid spans 2.4 points. Stronger coefficients cut search below the budget and cost accuracy, while zero penalty violates both constraints. On InternVL3-8B the Qwen-tuned coefficients reach 49.5 in-domain against 51.1 for the duals. The targets $B = 10$ and $\kappa = 0.337$ transfer unchanged across backbones. The duals thus replace a backbone-specific search with two interpretable targets.

Table 27: In-domain and OOD mean accuracy and test-time constraint statistics under fixed per-step penalties and under the PPO-Lagrangian duals. Constraint statistics are defined as in **Table 18**.

Configuration	In-domain	OOD	$\mathbb{E}[T]$	$\Pr[T > 10]$	$\mathbb{E}[\sum_t \text{Redun}_t]$
w/o duals (λ_B=λ_κ=0)	52.1	53.7	11.8	51.4%	0.412
Fixed penalty, best of 16	52.8	54.0	8.6	5.7%	0.238
Fixed penalty, worst of 16	50.4	51.9	4.9	0.3%	0.101
Lagrangian duals (HSSP)	**53.5**	**55.1**	8.1	3.2%	0.184

D.12 Transfer to Unseen Target Categories

The max-IoU and coverage terms score crop geometry and never reference target identity, so the trained search policy should carry over to unseen target categories. We group the target categories of the six V-TRACE sources into 12 super-categories and remove animals, vehicles, and text or signage from both the training pool and the cold-start SFT data, which removes 22% of the skeletons. A size-matched control removes a random 22%, and both models are tested on the three removed categories (**Table 28**). The model trained without them still gains +6.4 over the base model, 86% of the +7.4 gain of the size-matched control. The learned behavior is thus a category-agnostic search policy, not a category-specific appearance prior.

Table 28: Accuracy on test questions from the three held-out target categories.

Model	Held-out-category accuracy	Δ vs. base
Qwen2.5-VL-7B (base)	41.5	–
HSSP (full training pool)	49.3	+7.8
HSSP (*w/o* size-matched random subset)	48.9	+7.4
HSSP (*w/o* target categories)	47.9	+6.4

E Case Studies

This section presents HSSP rollouts on all six V-TRACE sources and on MMMU, and comparisons with single-pass RLVR and VAGEN on three in-domain cases. Each case shows the input image, its resolution, the question, and the ground-truth answer. Each step shows the active view (dashed red: the target box, or the abandoned crop in backtrack views; dashed teal: the next crop), action type, crop coordinates, and agent output. Highlights distinguish three levels of relevance: red marks load-bearing visual evidence, amber supporting observations, and cyan contextual or planning-related content. Incorrect baseline outcomes in the comparison boxes are gray.

E.1 HSSP Rollouts on V-TRACE Sources

Nine cases cover the six in-domain sources, from small-target natural images to charts, documents, and ultra-high-resolution scenes, with one Type-A case per source and three Type-C cases.

■ CASE STUDY | V*Bench / backtrack-and-recover



Glove material under street-stall clutter

Resolution: 2000 × 1500

Question: What is the material of the glove?

(A) rubber | **(B)** cotton | **(C)** kevlar | **(D)** leather

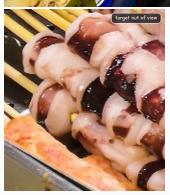
Ground truth: (A) rubber

Model output (HSSP): (A) rubber



t=0 **[PLAN]**

The image displays a street food stall with skewers of seafood and meat. In the upper left quadrant, there are two individuals wearing aprons. One person is holding a smartphone and wearing blue gloves. The question asks about the material of the glove. I need to locate the glove and examine its visual properties (texture, sheen, fit) to determine if it looks like latex, nitrile, or vinyl.



t=1 **[CROP]** bbox=[0.2044, 0.6431, 0.2887, 0.7699]

This crop lands on the lower-left tray of the display: skewers of raw octopus and squid tentacles on bamboo sticks, with a tray of marinated seafood along the bottom edge. No hands or gloves are visible in this view. The glove named in the question must belong to one of the vendors standing behind the counter, who are outside this crop, so this region does not contain the target and the search should move to the people in the upper left.



t=2 **[BACKTRACK]**

The previous crop (Step 1) targeted the bottom-left area of the image, which contains trays of raw octopus and squid skewers. It completely missed the people standing behind the counter. Since the question asks about the glove material, and the gloves are worn by the vendor in the upper left, the previous search trajectory was incorrect because it focused on the food display rather than the vendors.



t=3 **[PLAN]**

I need to correct the search strategy. The relevant visual evidence for 'glove material' is located in the upper-left quadrant of the image, specifically on the hands of the person wearing the white apron and gray shirt underneath. I will plan to crop that specific region to get a clear view of the blue gloves.



t=4 **[CROP]** bbox=[0.2632, 0.0569, 0.3774, 0.2102]

This crop targets the upper-left section of the image, specifically focusing on the vendor wearing the white apron. I can clearly see both hands covered in bright blue disposable gloves. The gloves appear to be made of a synthetic material like nitrile or latex due to their snug fit, tight wrinkling around the knuckles, and glossy finish with specular highlights. There are no visible cotton fibers or heavy rubber textures. The way they conform to the fingers suggests [...]



t=5 **[STOP]**

Based on the visual evidence in the final crop, the gloves are bright blue, disposable, and fit tightly around the fingers and wrists. The texture appears smooth and slightly elastic, consistent with materials like nitrile or latex commonly used in food service. Both nitrile and latex are rubber compounds rather than cotton, kevlar, or leather, so the answer is (A) rubber.

■ **CASE STUDY | HR-Bench 4K / backtrack-and-recover**



Reading the door-plate number on a 4K image

Resolution: 4032 × 4032

Question: What is the number displayed above the entrance where the woman is standing?

(A) 27B | **(B)** 37B | **(C)** 27D | **(D)** 27E

Ground truth: (A) 27B

Model output (HSSP): (A) 27B



t=0 **[PLAN]**

PLAN: The image shows a woman standing in the entrance of a building with stairs going up. Above the doorway, there is a small rectangular plate. I need to locate this specific area to read the number displayed above the entrance.



t=1 **[CROP]** bbox=[0.2051, 0.6418, 0.2894, 0.7686]

CROP: This region (Step 1) seems to focus on the lower part of the image, the dark stairwell wall to the left of the doorway. It does not contain the area above the entrance where the number should be located. This crop is misplaced for answering the question.



t=2 **[BACKTRACK]**

BACKTRACK: The previous crop (Step 1) was incorrect because it focused on the lower section of the image, missing the crucial detail above the doorway. A small plate is visible at the top center of the entrance frame, but its characters are not legible at this scale, and it was not captured in the first attempt.



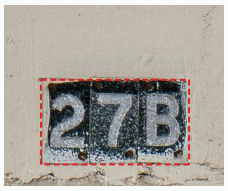
t=3 **[PLAN]**

PLAN: I will now search the upper central region of the image, specifically focusing on the area directly above the doorway where the woman is standing. This is where the house number is typically placed.



t=4 **[CROP]** bbox=[0.4442, 0.1662, 0.5559, 0.2579]

CROP: This region (Step 4) correctly captures the area above the entrance. A small black and white sign is visible with the text '27B' clearly displayed. This matches the expected location for the number mentioned in the question.



t=5 **[STOP]**

STOP A: The visual evidence confirms that the number displayed above the entrance is '27B'. The final crop provides clear visibility of this detail, allowing me to answer the question accurately.

■ **CASE STUDY | TreeBench / backtrack-and-recover**



Object-dense relational scene

Resolution: 2258 × 1500

Question: From the perspective of the building with the text "Camara Municipal", in which direction is the road sign with the text "Titan" to the entrance of the building?

(A) To the left | **(B)** To the right | **(C)** Front right | **(D)** Front left

Ground truth: (C) Front right

Model output (HSSP): (C) Front right



t=0 **[PLAN]**

PLAN: The image shows a street scene with two main buildings. On the left is a white building labeled 'CÂMARA MUNICIPAL'. On the right is a church-like building with a cross. The question asks about the relative position of a road sign 'Titan' to the entrance of the 'Camara Municipal' building. I need to locate both elements and determine their spatial relationship from the perspective of the building.



t=1 **[CROP]** bbox=[0.4624, 0.0234, 0.5474, 0.1406]

CROP 0.4624 0.0234 0.5474 0.1406: This crop lands on the sky above the church, catching only the right arm of the rooftop cross at its lower edge. There is no text 'Camara Municipal' or 'Titan' visible here. This region is irrelevant to the question's target objects.



t=2 **[BACKTRACK]**

BACKTRACK 1: The previous crop (Step 1) targeted the sky above the church roof, which contains neither the 'Camara Municipal' building nor the 'Titan' sign. The search must shift to the left side of the image where the municipal building is located.



t=3 **[PLAN]**

PLAN: Search the left half of the image for the 'CÂMARA MUNICIPAL' building and any nearby signage. The goal is to find the entrance of that building and then locate the 'Titan' sign relative to it.



t=4 **[CROP]** bbox=[0.0472, 0.5275, 0.3934, 0.8287]

CROP 0.0472 0.5275 0.3934 0.8287: This crop captures the lower-left section of the image, including the sidewalk, the curb, and the facade of the white building. Visible are advertisements ('Ragazzi', 'Dog Boy'), a yellow post in front of the awning entrance, the 'CÂMARA MUNICIPAL' lettering, and a small red-and-white 'Titan' sign on a post to the right of the entrance. The sign is too small at this scale to fix its placement relative to the door, so I zoom in on it.



t=5 **[CROP]** bbox=[0.1729, 0.6626, 0.2572, 0.7148]

CROP 0.1729 0.6626 0.2572 0.7148: This zoomed-in crop shows the first-floor balcony of the 'CÂMARA MUNICIPAL' building. Hanging just below the balustrade, on a post, is a red-and-white shield-shaped sign with the word 'TITAN' visible, above a blue street-name plate. The entrance itself lies outside this crop to the left; in the previous view the sign stands on the sidewalk in front of the facade, to the right of the awning entrance when facing the building.



t=6 **[STOP]**

STOP C: Evidence confirms that the 'Titan' sign stands on a post in front of the facade, to the right of the main entrance of the 'CÂMARA MUNICIPAL' building when facing the building. Therefore, relative to the building's entrance, the sign is in front and to the right: option (C) Front right.

■ **CASE STUDY | VisualProbe-Hard / backtrack-and-recover**



Tiny target under distractors

Resolution: 1200 × 1800

Question: What color is the body of the butterfly with black wings?

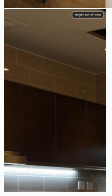
Ground truth: Brown.

Model output (HSSP): Brown.



t=0 **[PLAN]**

PLAN: The image displays a dining room with a table, chairs, and decorative items. To find the butterfly with black wings, I will scan the wall decorations and furniture details. I'll start by examining the middle-left area which might contain wall art or objects.



t=1 **[CROP]** bbox=[0.097, 0.3039, 0.2654, 0.5032]

CROP: This region shows the kitchen area through an opening, including the ceiling vent, a band of wall tiles, dark upper cabinets, and an under-cabinet light strip. There are no butterflies visible here, so this is not the correct location.



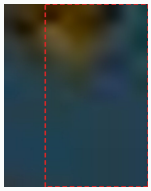
t=2 **[BACKTRACK]**

BACKTRACK: The previous crop focused on the kitchen area, which does not contain any butterfly imagery. I need to shift my focus to other parts of the room where decorative elements like plates or figurines might be displayed.



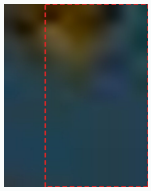
t=3 **[PLAN]**

PLAN: Since the kitchen area did not yield results, I will now examine the wall behind the dining table where several decorative plates are hung. Additionally, the blue-lacquered sideboard below the plates carries small painted motifs on its front panel that could include a butterfly.



t=4 **[CROP]** bbox=[0.6875, 0.572, 0.6997, 0.583]

CROP: This close-up lands on the front panel of the blue sideboard, where a small painted butterfly motif sits; at this magnification the crop itself is only a gold-brown patch over blue, so the motif is read together with the surrounding view. The butterfly has black wing tips and a brown body, matching the description in the question.



t=5 **[STOP]**

STOP: The visual evidence confirms that the butterfly with black wings has a brown body. This conclusion is based on direct observation of the painted motif on the front of the sideboard in the dining area.

■ **CASE STUDY | O3-Bench / backtrack-and-recover**



Layout-grounded chart question

Resolution: 2533 × 987

Question: If the Avg. Surfboard Utilization Rate is at its maximum sensitivity value and the # of Surfboards is at its minimum sensitivity value, what is the Levered IRR percentage according to the sensitivity table?

(A) 73.0% | **(B)** 117.1% | **(C)** 10.0% | **(D)** 59.0% | **(E)** 52.6% | **(F)** No right choice

Ground truth: (F) No right choice

Model output (HSSP): (F) No right choice



t=0 **[PLAN]**

The image is a 'Sensitivity Analysis' dashboard. I need to find the Levered IRR value under specific conditions: maximum Avg. Surfboard Utilization Rate and minimum # of Surfboards. The layout contains a tornado bar chart on the left and several tables with green headers on the right. The table headers mention '# of Surfboards' and 'Avg. Surfboard Utilization Rate', which matches the variables in the question. My strategy is to locate the intersection of these [...]



t=1 **[CROP]** bbox=[0.3366, 0.6165, 0.6775, 0.7972]

I am examining the cropped region showing the 'Key Value Drivers (Sorted)' table (bottom middle). This table has columns for 'Position', 'Key Value Driver', 'Min', 'Difference', and 'Max', and its rows rank the drivers from Avg. Surfboard Utilization Rate down to Levered IRR. This does not appear to be the correct table because the question asks for the *intersection* of 'Avg. Surfboard Utilization Rate' and '# of Surfboards', but this table only lists each driver's range. I need to backtrack and look at other tables.



t=2 **[BACKTRACK]**

The previous crop was incorrect. It showed the 'Key Value Drivers (Sorted)' ranking table with Min / Difference / Max columns, not a two-way sensitivity grid. The question specifically asks about 'Avg. Surfboard Utilization Rate' and '# of Surfboards'. I need to find a table that has these as its primary row or column headers to find the corresponding Levered IRR values.



t=3 **[PLAN]**

I am planning to search the right-hand section of the image where there are tables with green headers. I see a two-way table headed '# of Surfboards' across the top with 'Avg. Surfboard Utilization Rate' as its row label. These match the variables in the question. I will examine the table labeled '# of Surfboards' first to see if it contains the sensitivity analysis data needed.



t=4 **[CROP]** bbox=[0.7346, 0.3468, 0.9795, 0.4893]

I have found the correct table on the right side. The table header is '# of Surfboards' with columns 37, 36, 34, 32, 31, whose first data row gives the percentage steps 110%, 105%, 100%, 95%, 90%. The row labels on the left belong to 'Avg. Surfboard Utilization Rate' with values 94.1%, 89.1%, 84.1%, 79.1%. The question asks for the Levered IRR when 'Avg. Surfboard Utilization Rate' is at its maximum sensitivity value and '# of Surfboards' is at its minimum. Looking at [...]



t=5 **[STOP]**

Based on the visual evidence in the table titled '# of Surfboards' (right side), I identified the relevant data points. The row for 'Avg. Surfboard Utilization Rate' at 94.1% (the maximum value in the row labels) reads 127.5%, 133.6%, 117.1%, 59.0%, 44.9% across the columns; it intersects the column for '# of Surfboards' at 31 boards, the 90% step and minimum of the sweep, in the last column, where the value is 44.9%. None of the listed options equals 44.9%: 117.1% (B) is the 100% base column and 59.0% (D) the 95% column, so the correct choice is (F) No right choice.

■ **CASE STUDY | MME-RealWorld-Lite / backtrack-and-recover**



Aerial DOTA scene with distractors

Resolution: 7360 × 4912

Question: Where is the white tower in the picture?

(A) To the left above the picture, amidst the green trees. | **(B)** To the left corner of the picture, amidst the red trees. | **(C)** To the right at the bottom in the picture, amidst the green trees. | **(D)** To the right above the picture, amidst the green trees. | **(E)** This image doesn't feature the position.

Ground truth: (D) To the right above the picture, amidst the green trees.

Model output (HSSP): (D) To the right above the picture, amidst the green trees.



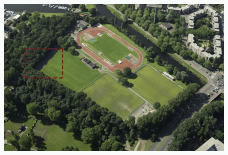
t=0 **[PLAN]**

The image shows an aerial view of a sports complex with a running track, football fields, surrounding trees, and adjacent buildings. The question asks for the location of a 'white tower.' I need to scan the image for any tall, vertical structures that are white. Initial scanning suggests residential buildings on the right and green spaces/trees elsewhere. I will plan to search specific regions where towers or tall structures might be located.



t=1 **[CROP]** bbox=[0.0982, 0.3027, 0.2666, 0.5020]

I am examining the region at coordinates [0.0982, 0.3027, 0.2666, 0.5020]. The left part of this crop is dense green woodland; the middle is a shaded asphalt car park with painted bays and a chain-link fence running diagonally, and the right part is the bright green edge of a football pitch with a floodlight pole, with a white van and a dark car at the top. There is no tall white structure anywhere in this specific crop.



t=2 **[BACKTRACK]**

This region (Step 1) contains woodland, a car park, and the edge of a pitch, which does not match the target 'white tower'. Since there is no tall vertical structure here, this was a false lead. I must backtrack and search other parts of the image where buildings are more likely to appear.



t=3 **[PLAN]**

Based on the full image layout, the upper right quadrant contains several multi-story apartment buildings and urban infrastructure. A 'tower' would most likely be found among these built-up areas rather than in the park or water sections. I will focus my search on the upper right section near the residential blocks.



t=4 **[CROP]** bbox=[0.6715, 0.0258, 0.6857, 0.1342]

In the cropped region [0.6715, 0.0258, 0.6857, 0.1342], located in the upper middle-right area in front of the apartment block, there is a distinct, tall, thin vertical structure. It appears to be a white lattice-style tower or mast standing amongst the trees and near the buildings. This matches the description of a 'white tower'.



t=5 **[STOP]**

The visual evidence confirms the presence of a white tower-like structure in the upper right portion of the image, specifically among the green trees in front of the residential buildings. The previous search in the tree-heavy left side was incorrect. The final answer identifies the location based on the successful finding in Step 4: option (D), to the right above the picture, amidst the green trees.

■ **CASE STUDY | V*Bench / systematic-absence**



Counterfactual probe with no target present

Resolution: 2250 × 1500

Question (counterfactual): Locate the satellite imagery layer in the image. If it does not exist, respond with None.

Original benchmark question: What is the color of the tissue box?

(A) gray | **(B)** white | **(C)** black | **(D)** blue

Original-question ground truth: (D) blue

Counterfactual GT: None

Model output (HSSP): None



t=0 **[PLAN]**

The image displays a white MAN truck on a road with trees and vegetation in the background. There are no visible overlays, maps, or graphical layers that resemble satellite imagery. I will systematically divide the image into four quadrants to ensure thorough coverage before concluding.



t=1 **[CROP]** bbox=[0.0, 0.45, 0.55, 1.0]

Cropped the bottom-left quadrant (x: 0-55%, y: 45-100%). This region shows the lower part of the truck's trailer, wheels, and some grassy roadside vegetation. There is no satellite imagery layer here; it is purely photographic content of the vehicle and environment.



t=2 **[CROP]** bbox=[0.45, 0.45, 1.0, 1.0]

Cropped the bottom-right quadrant (x: 45-100%, y: 45-100%). This area contains the front bumper of the truck, part of the road guardrail, and dense green bushes. No satellite imagery or map overlay is present in this section either.



t=3 **[CROP]** bbox=[0.0, 0.0, 0.55, 0.55]

Cropped the top-left quadrant (x: 0-55%, y: 0-55%). This view includes tree branches, leaves, and distant hills/mountains through the foliage. The scene remains a standard photograph with no digital satellite data or map layers superimposed.



t=4 **[CROP]** bbox=[0.45, 0.0, 1.0, 0.55]

Cropped the top-right quadrant (x: 45-100%, y: 0-55%). This region shows the upper cab of the truck, more trees, and a hazy mountainous background. A traffic sign post is visible on the far right. Still, there is no evidence of a satellite imagery layer anywhere in this final quadrant.



t=5 **[STOP]**

After examining all four quadrants of the image, no satellite imagery layer was found. The entire image consists of a real-world photograph of a truck on a road surrounded by nature. Since every quadrant has now been inspected and no such layer exists, the correct response is None.

■ **CASE STUDY | O3-Bench / systematic-absence**



Confirming absence on a financial dashboard

Resolution: 2560 × 1941

Question (counterfactual): Locate the radar chart in the image. If it does not exist, respond with None.

Original benchmark question: In which forecast year does the 'LTV' value first drop below 15%, and what is the corresponding Gross Fixed Assets value in that year?

(A) 2029 F and 4,292,050 USD | **(B)** 2030 F and 5,038,550 USD | **(C)** 2031 F and 5,169,400 USD | **(D)** 2031 F and 5,038,550 USD | **(E)** 2030 F and 4,907,700 USD | **(F)** No right choice

Original-question ground truth: (E) 2030 F and 4,907,700 USD

Counterfactual GT: None

Model output (HSSP): None



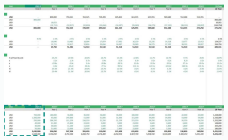
t=0 **[PLAN]**

The image is a composite financial dashboard containing multiple sections: Cash Flow Statement tables and charts, Debt Schedule with bar/line charts, CAPEX charts, and Fixed Asset Schedule. I will systematically scan these regions to look for any circular or spider-web style radar chart.



t=1 **[CROP]** bbox=[0.0, 0.0, 0.55, 0.3833]

Cropping the top-left region (0.0000, 0.0000 to 0.5500, 0.3833). This area contains the 'Cash Flow Statement' section. Visually, there is a large stacked bar chart on the left showing cash flow components over years, and a detailed data table on the right. There are no circular or radial charts here.



t=2 **[CROP]** bbox=[0.45, 0.2833, 1.0, 0.7167]

Cropping the middle-right region (0.4500, 0.2833 to 1.0000, 0.7167). This section contains the value columns of the 'Debt Schedule' data table, whose row labels sit just left of this crop, with the first rows of the CAPEX table at the bottom. The view is purely tabular text and numbers; no graphical charts are present in this specific crop.



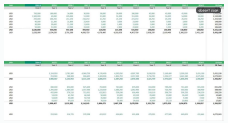
t=3 **[CROP]** bbox=[0.0, 0.6167, 0.55, 1.0]

Cropping the bottom-left region (0.0000, 0.6167 to 0.5500, 1.0000). The upper half shows the CAPEX band, stacked bars with the green CAPEX line and its table, and the lower half the 'Fixed Asset Schedule' with its row labels and unit column. Neither element resembles a radar chart.



t=4 **[CROP]** bbox=[0.0, 0.2833, 0.55, 0.7167]

Cropping the middle-left region (0.0000, 0.2833 to 0.5500, 0.7167). This contains the 'Debt Schedule' visualizations. Specifically, there is a bar chart comparing 'Financial Debt' and 'EBITDA', and below it a bar/line chart for 'Debt Service Coverage' (DSCR). These are standard vertical bar and line plots, not radar charts.



t=5 **[CROP]** bbox=[0.45, 0.6167, 1.0, 1.0]

Cropping the bottom-right region (0.4500, 0.6167 to 1.0000, 1.0000). This displays the value columns of the 'CAPEX' data table across years, with the Fixed Asset Schedule values below; the row labels lie left of this crop. It is strictly tabular data with no associated chart visible in this corner.



t=6 **[CROP]** bbox=[0.45, 0.0, 1.0, 0.3833]

Cropping the top-right region (0.4500, 0.0000 to 1.0000, 0.3833). This area shows the value columns of the 'Cash Flow Statement' data table, with the top of the Debt Schedule values below; the row labels lie left of this crop. It is a spreadsheet-style layout with no graphical elements like a radar chart.



t=7 **[STOP]**

After scanning all major sections of the image (Cash Flow, Debt Schedule, CAPEX, Fixed Assets), only stacked bar charts, line charts, and data tables were found. No radar chart exists in the image.

■ **CASE STUDY | TreeBench / systematic-absence**



Coverage-based negative declaration

Resolution: 1500 × 2247

Question (counterfactual): Find the elevation contour. If the image does not contain one, output None.

Original benchmark question: From the perspective of the girl on the left side of the image, wearing a white short-sleeved top and sunglasses, in which direction is the man wearing a black robe and a hat located?

(A) Front left | **(B)** Front right | **(C)** Left rear | **(D)** Right rear

Original-question ground truth: (A) Front left

Counterfactual GT: None

Model output (HSSP): None



t=0 **[PLAN]**

The image displays a European-style town square with historic buildings (a tall red one and an orange one), cobblestone ground, and tourists. I need to search for an 'elevation contour'—typically a line on a map indicating equal altitude. The current full view shows no maps or technical diagrams. Strategy: Divide the image into quadrants to systematically check for any graphical elements resembling a contour plot.



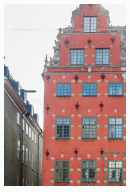
t=1 **[CROP]** bbox=[0.0, 0.45, 0.55, 1.0]

Cropping the bottom-left quadrant of the image (x=0 to 0.55, y=0.45 to 1.0). This region contains the lower parts of the buildings, the cobblestone street, and people sitting/standing. There are no lines, shading, or symbols that resemble elevation contours. The focus is on real-world scenery, not data visualization.



t=2 **[CROP]** bbox=[0.45, 0.0, 1.0, 0.55]

Cropping the top-right quadrant (x=0.45 to 1.0, y=0.0 to 0.55). This area shows the upper stories of the orange building and sky. No text, graphs, or contour lines are visible. The visual content is purely architectural.



t=3 **[CROP]** bbox=[0.0, 0.0, 0.55, 0.55]

Cropping the top-left quadrant (x=0.0 to 0.55, y=0.0 to 0.55). This covers the upper facade of the red building. Visible details include windows, decorative stonework, and gables. There is no evidence of an elevation contour here; it is a photograph of a building exterior.



t=4 **[CROP]** bbox=[0.45, 0.45, 1.0, 1.0]

Cropping the bottom-right quadrant (x=0.45 to 1.0, y=0.45 to 1.0). This includes the base of the orange building, some signage ('CHOKLADKOPPEN'), and more pedestrians. Again, no cartographic elements or contour lines are present. All four quadrants have been checked and contain only photographic imagery of a street scene.



t=5 **[STOP]**

After examining all regions of the image, there is no sign of an elevation contour. The image is a photograph of a city square with people and buildings, lacking any maps, charts, or scientific diagrams. Therefore, the answer is None.

E.2 HSSP Rollouts on Out-of-Distribution Benchmarks

Two examples illustrate HSSP on multi-discipline reasoning (MMMU), an anatomy figure and a tabulated materials-science problem, with the in-domain action grammar and Seeker prompt.

■ **CASE STUDY | OOD / MMMU / multi-discipline**



College-level anatomy question

Resolution: 330 × 260

Question: Pallidothalamic axons cross the <image 1>

(A) Anterior limb of the internal capsule. | **(B)** Genu of the internal capsule. | **(C)** Posterior limb of the internal capsule.

Ground truth: (C) Posterior limb of the internal capsule.

Model output (HSSP): (C) Posterior limb of the internal capsule.



t=0 **[CROP]** bbox=[0.35, 0.4, 0.65, 0.6]

The image appears to be a coronal section of the brain, showing structures such as the internal capsule. To determine where pallidothalamic axons cross, I need to identify the specific region of the internal capsule. The anterior limb, genu, and posterior limb are distinct parts, and the pallidothalamic fibers typically cross in the posterior limb.



t=1 **[CROP]** bbox=[0.45, 0.45, 0.55, 0.55]

This zoom lands on the midline of the section: the pale vertical slit is the third ventricle, flanked by medial thalamic tissue. The internal capsule is not inside this crop; in the full view it is the dark curved band lateral to the thalamus on each side, and its posterior limb is the segment between the thalamus (medial) and the lentiform nucleus (lateral). I will verify against the full section.



t=2 **[VERIFY]**

Further zoom on the midline adds no new detail, so I verify against the full section: the dark band lateral to the thalamus on each side is the internal capsule, and the portion level with the thalamus is its posterior limb. Pallidothalamic axons are known to cross through the posterior limb of the internal capsule, which matches this layout.



t=3 **[STOP]**

The verification places the posterior limb between the thalamus and the lentiform nucleus in the full section, and the pallidothalamic axons cross through the posterior limb of the internal capsule there.

■ CASE STUDY | OOD / MMMU / table reading

Temperature (°C)	*Impact Energy (J)*
−25	124
−50	123
−75	115
−85	100
−100	73
−110	52
−125	26
−150	9
−175	6

Charpy impact table in materials science

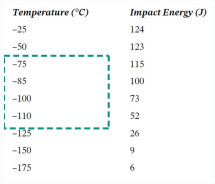
Resolution: 493 × 429

Question: Following are tabulated data that were gathered from a series of Charpy impact tests on a ductile cast iron. <image 1> Determine a ductile-to-brittle transition temperature as that temperature corresponding to the average of the maximum and minimum impact energies.

(A) $-100°C$ | **(B)** $-110°C$ | **(C)** $-105°C$ | **(D)** $-125°C$

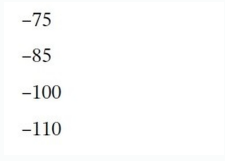
Ground truth: (C) $-105°C$

Model output (HSSP): (C) $-105°C$



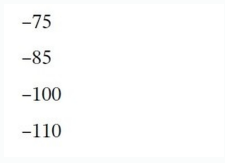
t=0 **[PLAN]**

I need to identify the maximum and minimum impact energies from the table to calculate the ductile-to-brittle transition temperature as their average. The maximum impact energy is 124 J at $-25°C$, and the minimum is 6 J at $-175°C$. The average of these two values will give the transition temperature.



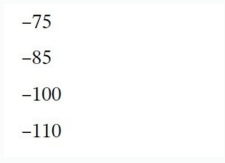
t=1 **[CROP]** bbox=[0.0, 0.3, 0.5, 0.7]

The maximum impact energy is 124 J and the minimum is 6 J. The average of these two values is (124 + 6) / 2 = 65 J. Now, I need to find the temperature corresponding to an impact energy of approximately 65 J from the table.



t=2 **[VERIFY]**

The crop isolates the temperature rows from $-75°C$ to $-110°C$; the matching energies are read from the full table. I need to locate the temperature corresponding to an impact energy close to 65 J. From the table, the impact energy at $-100°C$ is 73 J, and at $-110°C$ it is 52 J. The value 65 J lies between these two, so the transition temperature should be between $-100°C$ and $-110°C$.



t=3 **[STOP]**

The transition temperature corresponds to the average of the maximum and minimum impact energies, which is 65 J. This value lies between $-100°C$ (73 J) and $-110°C$ (52 J). Linear interpolation gives $-100 - 10 \times (73 - 65)/(73 - 52) \approx -104°C$, so the closest option is (C) $-105°C$.

E.3 HSSP versus Single-Pass RLVR and VAGEN

The following cases compare HSSP rollouts with the single-pass RLVR baseline and VAGEN (LLM-judge process reward) on three in-domain inputs. Each comparison shows the HSSP rollout and summarizes the outcome of each baseline on the same input.

■ CASE STUDY | HSSP vs. baselines on HR-Bench 4K (door-plate number)

Question: What is the number displayed above the entrance where the woman is standing?
(A) 27B | **(B)** 37B | **(C)** 27D | **(D)** 27E

Ground truth: (A) 27B

Model output (HSSP): (A) 27B

Failure-mode summary: The RLVR baseline answers in a single pass. The LLM-judge baseline commits to a wrong-region crop early. HSSP backtracks after an off-target first crop and recovers the door plate.

HSSP (Ours)


t=0 **[PLAN]**


t=1 **[CROP]**


t=2 **[BACKTRACK]**


t=3 **[PLAN]**



t=4 **[CROP]**



t=5 **[STOP]**

HSSP reasoning (key steps):

t=0 **[PLAN]** PLAN: The image shows a woman standing in the entrance of a building with stairs going up. Above the doorway, there is a small rectangular plate. I need to locate this specific area to read the number displayed above the entrance.

t=1 **[CROP]** CROP: This region (Step 1) seems to focus on the lower part of the image, the dark stairwell wall to the left of the doorway. It does not contain the area above the entrance where the number should be located. This crop is misplaced for answering the question.

t=2 **[BACKTRACK]** BACKTRACK: The previous crop (Step 1) was incorrect because it focused on the lower section of the image, missing the crucial detail above the doorway. A small plate is visible at the top center of the entrance frame, but its characters are not legible at this scale, and it was not captured in the first attempt.

t=3 **[PLAN]** PLAN: I will now search the upper central region of the image, specifically focusing on the area directly above the doorway where the woman is standing. This is where the house number is typically placed.

Method	Final answer / failure mode
HSSP (Ours)	Correct. After an off-target first crop, the policy backtracks, re-plans, and reads the door plate, the recovery behavior that the geometric process reward credits during training.
Single-pass RLVR	Wrong: (D) 27E. It answers in one pass from the full image and never zooms in on the door plate, misreading the last character.
VAGEN (LLM-judge)	Wrong: (C) 27D. It commits to a wrong-region crop early and never reaches the door plate.

■ CASE STUDY | HSSP vs. baselines on V* Bench (glove material)

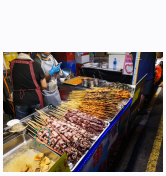
Question: What is the material of the glove?
(A) rubber | **(B)** cotton | **(C)** kevlar | **(D)** leather

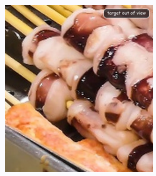
Ground truth: (A) rubber

Model output (HSSP): (A) rubber

Failure-mode summary: The RLVR baseline cannot resolve the glove material in a single pass. The LLM-judge baseline settles on a nearby irrelevant crop. HSSP backtracks after an off-target first crop and re-targets the upper-left vendor.

HSSP (Ours)

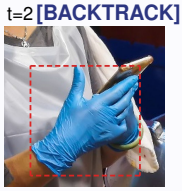

t=0 **[PLAN]**


t=1 **[CROP]**


t=2 **[BACKTRACK]**


t=3 **[PLAN]**


t=4 **[CROP]**


t=5 **[STOP]**

HSSP reasoning (key steps):

t=0 **[PLAN]** The image displays a street food stall with skewers of seafood and meat. In the upper left quadrant, there are two individuals wearing aprons. One person is holding a smartphone and wearing blue gloves. The question asks about the material of the glove. I need to locate the [...]

t=1 **[CROP]** This crop lands on the lower-left tray of the display: skewers of raw octopus and squid tentacles on bamboo sticks. No hands or gloves are visible in this view. The glove must belong to one of the vendors behind the counter, so this region does not contain the target [...]

t=2 **[BACKTRACK]** The previous crop (Step 1) targeted the bottom-left area of the image, which contains trays of raw octopus and squid skewers. It completely missed the people standing behind the counter. Since the question asks about the glove material, and the gloves are worn by the vendor [...]

t=3 **[PLAN]** I need to correct the search strategy. The relevant visual evidence for 'glove material' is located in the upper-left quadrant of the image, specifically on the hands of the person wearing the white apron and gray shirt underneath. I will plan to crop that specific region to [...]

Method	Final answer / failure mode
HSSP (Ours)	Correct. After an off-target first crop, the policy backtracks and re-targets the upper-left vendor to read the glove material, the recovery behavior that the geometric process reward credits during training.
Single-pass RLVR	Wrong: (B) cotton. It answers in one pass from the full image and never zooms in on the glove.
VAGEN (LLM-judge)	Wrong: (D) leather. It settles on a nearby but irrelevant crop and never reaches the vendors.

■ CASE STUDY | HSSP vs. baselines on TreeBench (object-dense relational)

Question: From the perspective of the building with the text "Camara Municipal", in which direction is the road sign with the text "Titan" to the entrance of the building?

(A) To the left | **(B)** To the right | **(C)** Front right | **(D)** Front left

Ground truth: (C) Front right

Model output (HSSP): (C) Front right

Failure-mode summary: Both baselines commit to a wrong distractor. HSSP backtracks from an off-target crop on the church, re-plans toward the municipal building, and locates the Titan sign relative to its entrance.

HSSP (Ours)


t=0 **[PLAN]**


t=1 **[CROP]**


t=2 **[BACKTRACK]**


t=3 **[PLAN]**


t=4 **[CROP]**


t=5 **[CROP]**

HSSP reasoning (key steps):

t=0 **[PLAN]** PLAN: The image shows a street scene with two main buildings. On the left is a white building labeled 'CÂMARA MUNICIPAL'. On the right is a church-like building with a cross. The question asks about the relative position of a road sign 'Titan' to the entrance of the [...]

t=1 **[CROP]** CROP 0.4624 0.0234 0.5474 0.1406: This crop lands on the sky above the church, catching only the right arm of the rooftop cross at its lower edge. There is no text 'Camara Municipal' or 'Titan' visible here. This region is irrelevant to the question's target objects.

t=2 **[BACKTRACK]** BACKTRACK 1: The previous crop (Step 1) targeted the sky above the church roof, which contains neither the 'Camara Municipal' building nor the 'Titan' sign. The search must shift to the left side of the image where the municipal building is located.

t=3 **[PLAN]** PLAN: Search the left half of the image for the 'CÂMARA MUNICIPAL' building and any nearby signage. The goal is to find the entrance of that building and then locate the 'Titan' sign relative to it.

Method	Final answer / failure mode
HSSP (Ours)	Correct. After an off-target crop on the church, the policy backtracks, re-plans toward the municipal building, and locates the Titan sign relative to its entrance, the recovery behavior that the geometric process reward credits during training.
Single-pass RLVR	Wrong: (A) To the left. It answers in one pass from the full image and never zooms in on the sign.
VAGEN (LLM-judge)	Wrong: (D) Front left. It commits to a distractor and does not recover.

F Limitations and Broader Impacts

Limitations. Our process reward is geometry-verifiable, in that each step's progress is measured against the recorded target region. Per-step shaping is therefore most informative for bounding-box-localizable targets, and tasks dominated by diffuse context, free-form text, or non-rectangular evidence benefit less. The Hider performs adaptive self-play over a fixed skeleton pool, reweighting existing skeletons without synthesizing new instances, so trajectory diversity is bounded by the pool. Extending the Hider to instance generation is left to future work. CMDP constraints are training-time only, so out-of-distribution inputs whose intrinsic difficulty exceeds the trained budget can still approach the inference-time turn limit. Theorem 4 analyzes an idealized two-timescale recursion with a fixed Hider and unbiased estimators, so it does not certify the realized PPO optimization.

Broader Impacts. The work targets long-horizon visual search, supporting document understanding, scientific image analysis, accessibility tools for users with low vision, and educational image-based question answering. Replacing language-model-judge process rewards with geometry-verifiable supervision lowers training-time dependence on auxiliary judge models and the corresponding energy and inference cost. The same localization capability can also improve fine-grained recognition in surveillance or person-identification settings, which raises privacy and consent concerns for high-resolution visual analysis. Deployment should therefore be paired with consent-respecting dataset curation, verification of presence and absence declarations on sensitive attributes, for which false positives are most harmful, and disclosure of the system's confidence and inspected regions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 summarize four contributions (the HSSP framework, the geometry-verifiable process reward with closed-form guarantees, the V-TRACE companion dataset, and eleven-benchmark empirical validation), each reflected in Sections 3–5 with matching scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix F discusses rectangular-box supervision, the curriculum-weighting scope of the Hider, train-time-only CMDP enforcement, and the idealized status of Theorem 4.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All four theorems in Section 4.3 carry full restatements with explicit scope and assumptions (Assumptions 5, 4, 7, 8, and 1), with complete proofs in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 5 specifies backbone, training stack (verl + vLLM + FSDP + Bi-Level GAE), dataset, hyperparameters, three-seed protocol, and three evaluation regimes; further detail appears in Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our code is publicly available at *this link* under Apache 2.0, together with a representative sample of the V-TRACE dataset under CC BY 4.0. Full reproduction details (architecture, hyperparameters, dataset construction, and matched baseline protocol) appear in Section 5 and Appendices C–D.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5 lists default hyperparameters, optimizer, seed protocol, and three evaluation regimes; the full hyperparameter table is Table 15 in Appendix D, and the matched-protocol baseline reproduction is described in the same appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Section 5 reports three-seed means with the maximum seed standard deviation and paired-bootstrap significance tests for the main comparisons (Table 17); Wilson-score 95% CIs are used for binary human-evaluation dimensions in Appendix C.4.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Training uses 32×A100-80GB (Section 5 and Appendix D), with a per-method compute breakdown in Table 16 (Appendix D.3).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics: no human-subjects experiments, no scraped data collection, and all source benchmarks are public research datasets used under their original terms.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix F discusses positive applications (document and scientific imaging, accessibility tools) and negative applications (surveillance and identification privacy risks), together with mitigation strategies (consent, absent-declaration auditing, user disclosure).

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released artifacts are a domain-specialized visual-search policy and a trajectory dataset built from existing public benchmarks; neither poses a high misuse risk requiring gated release.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Source benchmarks (O3-Bench, V*Bench, TreeBench, VisualProbe, HR-Bench, MME-RealWorld-Lite), OOD benchmarks (BLINK, CV-Bench, MMMU, RefCOCOg, MathVista), base VLMs (Qwen2.5-VL, InternVL3), and frameworks (verl, VAGEN, vLLM) are cited with their original licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: V-TRACE is the new asset, documented end-to-end in Appendix C (sources, skeleton synthesis, multi-agent enrichment, structural filtering, and full-coverage human evaluation). Our code is publicly available at *this link* under Apache 2.0, with the accompanying V-TRACE data under CC BY 4.0.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The human evaluation reported in Appendix C.4 was conducted internally by the authors' own research team, not via crowdsourcing or research with external human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The human evaluation in Appendix C.4 is an internal quality-assessment activity over public-benchmark images, not a human-subjects experiment, so IRB review was not required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use a VLM, Qwen3.5-397B-A17B, as an annotation tool when building V-TRACE. It writes the step-level text of each synthesized trajectory skeleton, which three verifier agents then score, and it grounds target boxes for the three sources without native boxes. Every grounded box is verified by human annotators, and the full procedure is given in Appendix C. LLMs were otherwise used only for editorial polish.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.