REFINE: ACTIVE VISUAL REASONING WITH SPATIALLY INDEXED EVIDENCE REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Active visual reasoning requires models to acquire fine-grained observations and retain them across successive decisions. Appending every crop grows the visual context even when an inspection revisits a previously observed region. We introduce REFINE, a reinforcement-learning framework that organizes acquired evidence into a persistent spatial state. Each region owns a fixed set of slots initialized from a coarse observation; higher-resolution inspection replaces those slots while preserving evidence elsewhere. A frozen answer-only probe measures the change in reference-answer support between initial and final states. This bounded, signed utility augments correctness and format rewards within Group Relative Policy Optimization, favoring useful refinement without rewarding inspection count. Evaluations span visual search, mathematical reasoning, chart understanding, and real-world perception. REFINE-7B achieves 85.26 on V* positional search and 69.2 on RealWorldQA, compared with 82.89 and 66.4 for the GRPO baseline. Its MME-RealWorld perception and reasoning scores are 68.60 and 50.37. Component ablations show complementary benefits from spatial storage and utility-aware learning. REFINE connects selective observation to persistent evidence through local replacement and outcome-grounded policy optimization.

1 INTRODUCTION

Vision-language models can reason over scenes, documents, and diagrams, but their conclusions depend on the visual evidence available to the language model. Visual instruction tuning and dynamic-resolution encoders have expanded these capabilities (Liu et al., 2023; Wang et al., 2024b; Bai et al., 2025). Longer reasoning chains provide more intermediate computation (Wei et al., 2022); they cannot recover a small label or an object relation omitted from the visual representation. Such questions require the model to choose where to look, acquire a more informative observation, and retain it for subsequent reasoning.

Active perception makes observation part of the reasoning process. Visual CoT incorporates localized visual evidence, ZoomEye searches images hierarchically, and DeepEyes learns to interleave visual actions with text through reinforcement learning (Shao et al., 2024a; Shen et al., 2025b; Zheng et al., 2026). These approaches motivate a representation question: after an inspection, should the model retain another observation in its history, or update its current understanding of the inspected region?

Appending observations organizes storage by acquisition time. Revisiting a region can therefore create several representations of the same content. With a finite visual-token budget, keeping every observation becomes costly, while discarding older observations can remove evidence needed for a later comparison. Spatial coordinates identify an observation's origin but do not determine how its contents should replace earlier evidence. A history of observations and a current evidence state serve different purposes.

We propose **REFINE**, which treats inspection as a local update to a persistent spatial state. A fixed image partition assigns each region its own visual slots. The policy selects a region and a refinement level; the resulting crop replaces that region's slots. Evidence in other regions remains unchanged, and a global overview preserves the scene layout. This operation keeps visual-state capacity independent of inspection count while allowing the contents to improve over time.

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

Learning to refine also requires a signal for the value of an observation. Correctness rewards identify successful answers but do not distinguish informative inspections from redundant ones along successful trajectories. Tool-aware and perception-aware training provide relevant precedents (Hou et al., 2026; Wang et al., 2026c). REFINE measures the change in reference-answer support between two evidence states using the same frozen answer-only probe. Because the probe excludes generated reasoning and tool history, the comparison is tied to the evidence supplied to it. Signed gains cancel when a later update reverses an earlier improvement.

We incorporate the net support gain into GRPO (Shao et al., 2024b). The utility term applies to correct, format-valid trajectories, and its bounded weight preserves their reward ordering relative to incorrect format-valid trajectories. The visual encoder and pooling remain fixed; the policy learns inspection and answering from image–question–answer records. Figure 1 summarizes the reported 7B benchmark profile, while the domain breakdown separates perception from reasoning.

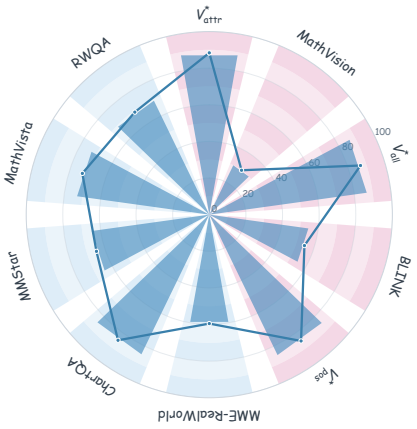


Figure 1: **REFINE-7B benchmark profile.** Scores use the 0–100 scale. The MME-RealWorld spoke is the equal-weight mean of the perception and reasoning overall scores in Table 2, rounded to 59.49.

The main contributions are summarized as follows:

- We introduce a spatially indexed evidence state for active visual reasoning. Higher-resolution observations locally replace the corresponding regional slots while preserving evidence elsewhere, keeping visual-state capacity independent of inspection count.
- We develop utility-aware GRPO, which augments outcome supervision with the net reference-answer support gained between initial and final evidence states. A frozen answer-only probe provides a bounded, signed signal that rewards useful refinement without rewarding additional inspections by themselves.
- We evaluate REFINE across visual search, mathematical reasoning, and real-world perception. Component ablations show complementary benefits from spatial storage and utility-aware learning, while the utility-weight sweep identifies a moderate coefficient that performs best across the tested benchmarks.

2 Related Work

2.1 Multimodal representations and active perception

Pretrained image–language interfaces, including BLIP-2, LLaVA, InternVL, and Qwen2.5-VL, provide the representations on which visual reasoning operates (Li et al., 2023a; Liu et al., 2023; Chen et al., 2024b; Bai et al., 2025). Sequential attention and targeted visual search introduce selective observation (Mnih et al., 2014; Wu & Xie, 2024), while Visual CoT, ZoomEye, and ReFocus integrate localization or visual editing into reasoning (Shao et al., 2024a; Shen et al., 2025b; Fu et al., 2025b). ReAct formalizes interleaved reasoning and action, and visual program synthesis provides executable interfaces to image operations (Yao et al., 2023; Gupta & Kembhavi, 2023; Surís et al., 2023). DeepEyes learns visual tool use from rollout rewards (Zheng et al., 2026). REFINE addresses the complementary storage decision: each acquired observation updates a persistent regional representation rather than becoming another visual block in the history.

2.2 Visual memory and compact reasoning states

Grounded reasoning and visual memory connect intermediate computation to selected image regions (Man et al., 2025; Wang et al., 2026a). Compact visual interfaces instead reduce or transform the representations available to the reasoner: Monet optimizes continuous visual thoughts, ILVR interleaves latent visual signals, and LIVR develops implicit visual reasoning tokens (Wang et al., 2026b; Dong et al., 2026; Li et al., 2026). These approaches differ in how intermediate visual

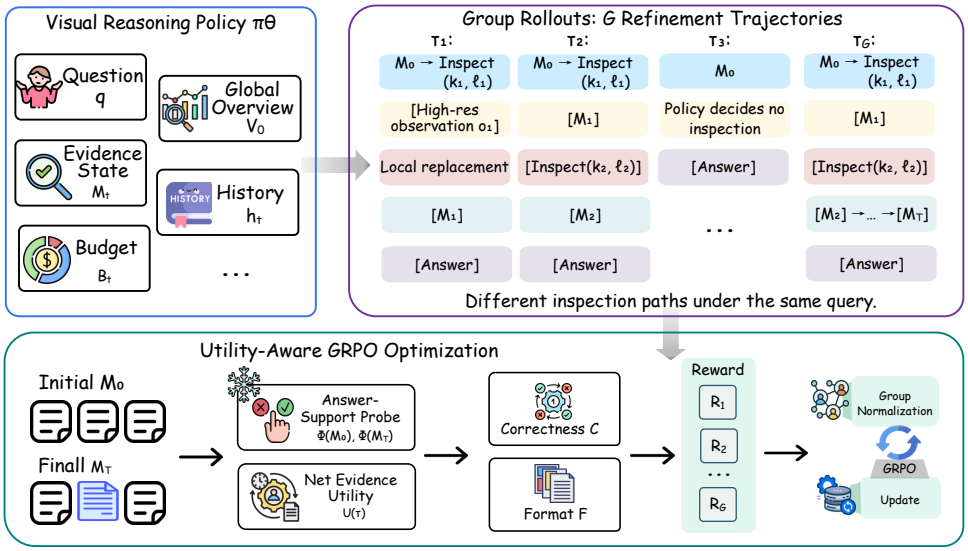


Figure 2: **REFINE training pipeline.** The policy samples trajectories that interleave inspection and answering. An inspection replaces the selected region's evidence. A frozen answer-only probe compares the initial and final states, and the resulting utility augments correctness and format rewards before group-relative policy optimization. The probe is removed at inference.

content is produced and interpreted. REFINE assigns storage to fixed image regions and derives each slot block from observed pixels through deterministic pooling. A revisit changes the contents of an existing region while preserving its spatial identity. Fixed-capacity storage and decoder-cache reuse are separate implementation concerns: replacing visual embeddings can require rebuilding the decoder prefix even though the number of stored visual tokens remains constant.

2.3 Visual Rewards and Policy Optimization

PPO and GRPO optimize sampled trajectories through clipped likelihood-ratio objectives (Schulman et al., 2017; Shao et al., 2024b). Multimodal training adds signals for visual reasoning: CodeV studies tool-aware rewards, PAPO regularizes perceptual behavior, and Perceval introduces perception-centric process rewards (Hou et al., 2026; Wang et al., 2026c; Min et al., 2026). REFINE changes the trajectory reward by comparing answer support under initial and final evidence states. The same frozen probe, reference answer, and answer-only prompt are used for both scores. Their signed difference measures net evidence gain, so an improvement subsequently reversed contributes zero net utility. This state-based signal complements answer correctness without requiring annotated intermediate reasoning or a separate teacher model.

3 Method

Overview. REFINE couples a persistent spatial representation with an inspection policy. The representation determines where acquired evidence is stored; utility-aware GRPO determines how its improvement contributes to learning. Figure 2 summarizes rollout sampling, evidence replacement, and the training-time support probe.

3.1 Spatially Indexed Evidence Refinement

Reasoning with an evolving state. Given an image I and question q, REFINE alternates text generation with regional inspection before producing an answer. Following the action–observation interface of ReAct and active visual reasoning (Yao et al., 2023; Zheng et al., 2026), the policy receives $s_t = (q, V_0, M_t, h_t, B_t)$: a global overview, the current evidence state, generated history,

and remaining inspection budget. An action either answers or requests a cell k at resolution level ℓ. The reference answer y^* is used for training rewards and never enters the rollout input.

Persistent spatial ownership. We partition original-image coordinates into K fixed cells $\{b_k\}_{k=1}^K$. Each cell owns m projected visual vectors and a resolution label:

$$M_t = \{(Z_{t,k}, \ell_{t,k})\}_{k=1}^K, \qquad Z_{t,k} \in \mathbb{R}^{m \times d}. \tag{1}$$

The frozen visual encoder and projection E are inherited from Qwen2.5-VL (Bai et al., 2025). A coarse image I_0 initializes $V_0 = E(I_0)$ and $Z_{0,k} = \mathcal{P}_m(V_0[b_k])$, with $\ell_{0,k} = 0$. Here $\mathcal{P}_m$ is deterministic two-dimensional adaptive average pooling into a prescribed slot grid. The overview remains available throughout reasoning. Regional embeddings are serialized in row-major order with explicit region and resolution identifiers; the initialization resolution must provide features for every cell.

Local replacement. A valid inspection increases the resolution of the selected cell, $\ell > \ell_{t,k}$, while keeping its original-image extent fixed. The environment extracts the regional crop, encodes it, and replaces its slot block:

$$\widehat{Z}_k = \mathcal{P}_m(E(\mathrm{crop}_\ell(I, b_k))), \qquad (Z_{t+1,k}, \ell_{t+1,k}) = (\widehat{Z}_k, \ell). \tag{2}$$

For every $j \neq k$, $(Z_{t+1,j}, \ell_{t+1,j}) = (Z_{t,j}, \ell_{t,j})$. Thus an inspection refines an existing representation rather than adding another visual block. Repeated levels are excluded, and both crop-pixel and inspection-count budgets bound acquisition. The original image determines the maximum available detail.

Storage and execution. With s metadata tokens per cell, the visual prefix contains $|V_0| + K(m + s)$ tokens, independently of inspection count. This bound concerns visual storage, not generated text or execution time. Replacing early embeddings invalidates downstream decoder caches, so the reference execution rebuilds the prefix after an update. Encoding, re-prefilling, and decoding contribute separately to execution cost. In contrast to continuous visual thoughts (Wang et al., 2026b), each regional block is deterministically derived from acquired pixels and retains an explicit spatial owner.

3.2 Net answer-support utility

A fixed support probe. For each rollout group, a frozen snapshot $\bar{\theta} = \theta_{\mathrm{old}}$ scores a canonical reference answer under an answer-only prompt p_{probe}. Its inputs contain q, V_0, M but exclude generated reasoning, tool history, and the sampled answer. We define

$$\Phi_{\bar{\theta}}(M; q, y^*) = \exp\left[\frac{1}{|y^*|}\sum_{v=1}^{|y^*|} \log \pi_{\bar{\theta}}(y_v^* \mid q, V_0, M, y_{<v}^*; p_{\mathrm{probe}})\right]. \tag{3}$$

This geometric mean lies in $[0, 1]$. Initial and final evaluations use identical answer tokenization, prompt, overview, and policy snapshot, isolating the change in the evidence-state input. Accepted answer aliases are canonicalized before scoring. The probe measures model-dependent answer support rather than calibrated correctness probability.

Signed evidence gain. The utility of an update is $u_t = \Phi_{\bar{\theta}}(M_{t+1}) - \Phi_{\bar{\theta}}(M_t)$, with the shared question and answer suppressed for clarity. Summing the signed changes yields

$$U(\tau) = \sum_{t=0}^{T-1} u_t = \Phi_{\bar{\theta}}(M_T) - \Phi_{\bar{\theta}}(M_0) \in [-1, 1]. \tag{4}$$

An improvement followed by an equal deterioration cancels, so repeatedly recording favorable intermediate changes cannot increase the net bonus. Only the initial and final probes are needed for training; intermediate evaluations are optional diagnostics. The initial score is shared across trajectories for the same input, and all scores are detached during policy optimization.

3.3 Utility-aware GRPO

Outcome-gated reward. Let $C(\tau), F(\tau) \in \{0, 1\}$ denote answer correctness and output-format validity. The trajectory reward combines outcome supervision with evidence utility:

$$R(\tau) = C(\tau) + \omega F(\tau) + \lambda C(\tau) F(\tau) U(\tau), \quad 0 \leq \lambda < 1, \quad 0 \leq \omega < 1. \tag{5}$$

Among format-valid trajectories, a correct answer receives at least $1 + \omega - \lambda$, whereas an incorrect answer receives ω. Hence the auxiliary term preserves correctness ordering while distinguishing the evidence acquired by successful trajectories. Negative utility can lower a correct trajectory's reward; $\lambda = 0$ recovers the same-interface outcome baseline.

Group-relative optimization. For G rollouts per question, GRPO (Shao et al., 2024b) assigns the normalized advantage $A_n = (R_n - \overline{R})/(\mathrm{std}(R_1, \ldots, R_G) + \epsilon)$. With the clipped likelihood ratio used in PPO (Schulman et al., 2017), the objective is

$$\mathcal{J}(\theta) = \frac{1}{G} \sum_{n=1}^{G} \frac{1}{|\mathcal{Y}_n|} \sum_{v \in \mathcal{Y}_n} \left[\min(r_{n,v} A_n, \mathrm{clip}(r_{n,v}, 1-\delta, 1+\delta) A_n) - \beta D_{\mathrm{KL},n,v} \right], \tag{6}$$

where $r_{n,v} = \pi_\theta(y_{n,v} \mid s_{n,v}) / \pi_{\mathrm{old}}(y_{n,v} \mid s_{n,v})$ and $D_{\mathrm{KL},n,v}$ regularizes against the initialization policy. $\mathcal{Y}_n$ contains generated text and command tokens; environment-inserted features and metadata are masked. Likelihoods are evaluated under the evidence state present when each segment was generated, including the same constrained action grammar. Utility remains a trajectory-level reward, not a token-level causal attribution.

Training and inference. REFINE starts from an instruction-tuned checkpoint and trains the language policy using image–question–answer records. The visual encoder, projection, pooling, and batchwise probe snapshot remain fixed. No additional teacher model or supervised intermediate trajectory is required. The reference configuration uses $G = 4$, $K = 16$, $m = 16$, two refinement levels, $\lambda = 0.2$, and $\omega = 0.1$. Appendix A gives the optimizer settings and execution details. Inference removes the support probe and reference answer, leaving the policy to inspect and answer directly.

3.4 Interaction between storage and acquisition

Separating location from acquisition order. The cell index is an address, not a timestamp. A regional block remains in the same position in the serialized visual state after inspection, and its resolution label records how that block was obtained. This stable correspondence lets later reasoning combine a refined region with coarse evidence elsewhere. For example, resolving an object's appearance can require a close observation of one cell, whereas comparing its position with another object still benefits from the global overview and the other cells. Local replacement supports both operations without retaining multiple resolutions of the inspected cell as separate visual observations.

Learning when refinement is worthwhile. Spatial storage alone does not determine a useful inspection policy. A higher-resolution crop can reveal a missing detail, duplicate evidence already sufficient for answering, or displace features that supported the answer. The signed utility distinguishes these outcomes through the final state's answer support. Its correctness gate then favors acquisitions that accompany a correct response. Conversely, utility learning depends on what the storage rule preserves: an informative observation has lasting value only when subsequent updates leave the relevant evidence accessible. The storage–reward ablation in Section 4.3 examines this interaction directly.

Stopping and bounded acquisition. Answering is available throughout a rollout. An inspection consumes budget and advances the selected cell's resolution; when no valid inspection remains, the action interface permits answering only. The state therefore evolves through a finite sequence of refinements. With L refinements available per cell and a count budget B_0, there are at most $\min(B_0, KL)$ inspections. This bound controls acquisition, while Equation 1 controls retained visual capacity. Algorithm B makes this separation explicit, and Algorithm B shows how the same interaction produces training trajectories.

4 Experiments

4.1 Evaluation setup

Tasks and metrics. The benchmark comparison covers visual search with V*Bench (Wu & Xie, 2024), mathematical reasoning with MathVision and MathVista (Wang et al., 2024a; Lu et al., 2024), chart understanding with ChartQA (Masry et al., 2022), and perception with BLINK, MMStar, and RealWorldQA (Fu et al., 2025a; Chen et al., 2024a; xAI, 2024). MME-RealWorld is reported

Table 1: **Multimodal benchmark comparisons.** Local evaluation results; row citations identify the original methods. RWQA: RealWorldQA. Bold/underlined entries mark the highest/second-highest displayed scores within each size group. †: equal-weight mean of REFINE's V* subtask accuracies. Thyme uses MathVision-Mini. Dashes indicate omitted evaluations.

Models	#Param	Visual Search			Math & Chart			Perception-intensive		
		V^*_{attr}	V^*_{pos}	V^*_{all}	MathVision	MathVista	ChartQA	BLINK	MMStar	RWQA
VLM-R1 (Shen et al., 2025a)	3B	75.65	67.11	72.25	21.71	65.1	83.48	46.25	**56.7**	61.5
LMM-R1 (Peng et al., 2025)	3B	46.09	53.95	49.21	24.47	63.5	85.04	46.60	**56.7**	58.7
R1-VL (Zhang et al., 2025a)	2B	59.13	57.89	58.64	10.20	47.0	68.08	42.81	38.9	49.2
Perception-R1 (Yu et al., 2025)	3B	57.39	48.68	53.92	22.03	58.1	81.60	46.44	54.8	55.8
Jigsaw-R1 (Wang et al., 2025c)	3B	72.17	65.79	69.63	19.40	61.0	84.60	45.01	54.4	57.9
Qwen2.5-VL (Bai et al., 2025)	3B	57.39	65.79	60.73	21.40	61.6	83.12	46.94	52.1	63.4
+ GRPO (Min et al., 2026)	3B	86.95	69.73	80.10	23.36	65.1	83.32	49.13	55.3	62.1
REFINE (ours)	3B	**91.17**	**72.79**	**81.98**†	**26.36**	**66.2**	**85.73**	**50.10**	55.4	**64.9**
Pixel-Reasoner (Su et al., 2025)	7B	–	–	84.30	22.03	69.4	77.36	51.10	63.1	64.0
Vision-R1 (Huang et al., 2026)	7B	–	–	–	**36.18**	71.3	83.36	49.72	55.4	64.1
VL-Rethinker (Wang et al., 2025a)	7B	54.78	59.21	56.54	31.91	72.6	85.60	49.91	64.0	64.0
VLAA-Thinker (Chen et al., 2025)	7B	43.47	52.63	47.12	27.96	70.3	85.36	49.38	64.0	62.0
R1-VL (Zhang et al., 2025a)	7B	47.83	67.11	55.50	22.37	64.1	82.80	47.19	55.5	59.5
OpenVLThinker (Deng et al., 2025)	7B	76.52	80.26	78.01	32.57	71.1	**89.00**	51.36	62.8	66.5
MM-Eureka (Meng et al., 2025)	7B	42.61	56.58	48.17	32.23	72.4	82.36	50.23	63.4	62.3
Thyme (Zhang et al., 2026)	7B	83.50	80.30	82.20	27.60	70.0	86.10	**56.10**	64.3	67.3
DeepEyes (Zheng et al., 2026)	7B	**90.43**	72.27	84.21	12.50	69.9	75.84	50.98	62.7	67.0
Qwen2.5-VL (Bai et al., 2025)	7B	73.91	65.79	70.68	26.97	70.2	84.28	48.56	62.3	60.6
+ GRPO (Min et al., 2026)	7B	85.22	82.89	84.29	27.96	71.7	85.16	53.55	62.0	66.4
REFINE (ours)	7B	88.37	**85.26**	**86.82**†	29.97	**72.9**	84.80	54.63	**64.8**	**69.2**

Table 2: **Performance comparison on MME-RealWorld.** Local evaluation results for perception and reasoning. Best and second-best scores in each section are bold and underlined. Improvement is the absolute gain over Qwen2.5-VL-7B in percentage points. Dashes indicate omitted evaluations.

Perception						
Model	*Monitoring*	*Autonomous Driving*	*OCR*	*Diagram and Table*	*Remote Sensing*	*Overall*
Qwen2.5-VL-7B (Bai et al., 2025)	38.75	22.70	87.00	78.83	45.40	60.94
Thyme-VL-7B (Zhang et al., 2026)	49.27	37.46	**88.36**	80.86	53.93	67.10
REFINE (ours)	**52.55**	**40.73**	88.10	**82.46**	**55.40**	**68.60**
Improvement ↑	*13.80*	*18.03*	*1.10*	*3.63*	*10.00*	*7.66*
Reasoning						
Model	*Monitoring*	*Autonomous Driving*	*OCR*	*Diagram and Table*	*Remote Sensing*	*Overall*
Qwen2.5-VL-7B (Bai et al., 2025)	26.10	24.25	64.80	63.40	–	38.59
Thyme-VL-7B (Zhang et al., 2026)	47.39	32.29	**70.60**	70.40	–	48.38
REFINE (ours)	**49.75**	**35.15**	69.50	**72.40**	–	**50.37**
Improvement ↑	*23.65*	*10.90*	*4.70*	*9.00*	*–*	*11.78*

separately by domain and by perception versus reasoning (Zhang et al., 2025b). HRBench4K/8K and POPE appear in the component and capacity analyses (Wang et al., 2025b; Li et al., 2023b). All scores are percentages; POPE uses F1, and the other listed tasks use their benchmark correctness metrics.

Models and reporting. Table 1 reports local evaluation results for REFINE and the comparison models at 3B and 7B scales. Row citations identify the original methods, rather than the numerical source of a reproduced score. Qwen2.5-VL provides the backbone comparison, and its GRPO variant provides the outcome-based learning comparison. Table 2 gives the full MME-RealWorld domain breakdown. Bold and underlining indicate the highest and second-highest displayed scores within each group, including ties. REFINE's V* aggregate is the equal-weight mean of attribute and positional accuracy; the two subtask columns provide the direct comparison. Thyme's MathVision entry uses the Mini split.

4.2 BENCHMARK COMPARISONS

Multimodal performance. As shown in Table 1, REFINE-3B reaches 72.79 on V* positional search, 85.73 on ChartQA, and 50.10 on BLINK, exceeding the 3B GRPO scores by 3.06, 2.41, and 0.97 percentage points. Its attribute-search score improves by 4.22 points to 91.17. At 7B, REFINE reaches 88.37 and 85.26 on the two search subtasks, improving over GRPO by 3.15 and 2.37 points. The positional result is also 19.47 points above the Qwen2.5-VL backbone. These subtask improvements accompany REFINE macro-averages of 81.98 at 3B and 86.82 at 7B.

Task-dependent benefits. The 7B results extend beyond visual search: REFINE obtains 29.97 on MathVision, 72.9 on MathVista, 64.8 on MMStar, and 69.2 on RealWorldQA. Relative to GRPO, the respective gains are 2.01, 1.2, 2.8, and 2.8 points. REFINE leads the displayed 7B group on MathVista, MMStar, and RealWorldQA, while other methods remain stronger on particular tasks. ChartQA is 0.36 points below GRPO and 4.20 points below OpenVLThinker; Thyme retains the strongest BLINK score, 56.10 versus REFINE's 54.63. Thus the benefit of persistent refinement varies with the evidence required by each benchmark rather than producing a uniform ranking across tasks.

Real-world perception and reasoning. Table 2 shows overall scores of 68.60 for perception and 50.37 for reasoning, improvements of 7.66 and 11.78 points over Qwen2.5-VL-7B. Perception gains are largest in autonomous driving (18.03) and monitoring (13.80); reasoning gains are largest in monitoring (23.65). Relative to Thyme, REFINE improves the two overall scores by 1.50 and 1.99 points. OCR follows a different pattern: REFINE is 0.26 points lower in perception and 1.10 points lower in reasoning. This breakdown separates broad aggregate gains from the recognition tasks on which those gains are smaller.

4.3 COMPONENT AND UTILITY-WEIGHT ANALYSES

Storage and utility. Table 3 compares FIFO and spatial storage under outcome-only and utility-aware rewards. Adding utility increases V* by 1.05 points with FIFO and 3.57 points with spatial storage; POPE increases by 0.66 and 1.19 points, respectively. Conversely, replacing FIFO with spatial storage improves V* by 2.62 points under outcome-only learning and 5.14 points under utility-aware learning. The complete model exceeds FIFO with outcome rewards by 6.19 V* points and 2.63 POPE points. Both components contribute, and the larger utility gain with spatial storage supports their complementarity: acquisition learning benefits from retaining observations at stable spatial addresses.

Complementary component effects. The difference between the two utility gains is 2.52 points on V* and 0.53 on POPE. This interaction contrast captures the additional benefit of utility-aware learning when storage is spatially organized. The improvement on POPE also shows that the component effect is not confined to spatial search: the full model reaches 89.10 F1, compared with 87.91 for spatial storage without utility. Equation 16 in the appendix records the contrast explicitly, making the storage and reward contributions distinguishable from the full-model comparison.

Utility weight. Table 4 varies the utility coefficient with spatial storage. The zero-weight V* score is 83.25, matching the spatial outcome-only condition in Table 3. Among the tested weights, $\lambda = 0.2$ achieves the highest score on all three benchmarks: 86.82 on V*, 75.03 on HRBench4K, and 72.18 on HRBench8K. These are gains of 3.57, 6.11, and 6.77 points over $\lambda = 0$. Raising the coefficient to 0.4 reduces the scores by 2.53, 0.15, and 1.06 points. The response is not monotonic: $\lambda = 0.1$ improves both HRBench scores but lowers V* by 2.62 points. A moderate utility weight therefore provides the strongest balance across the tested tasks. HR4K and HR8K abbreviate the two HRBench resolutions.

4.4 TRAINING AND EVIDENCE DIAGNOSTICS

Training dynamics. Figure 3 shows the supplied traces over 2,000 steps. Response length drops early and then partly recovers, while result reward and the separately supplied consistency-reward series increase before reaching fluctuating plateaus. These trajectories describe different aspects of optimization: response length tracks generated computation, whereas the reward traces track their logged objectives. The consistency series is distinct from the answer-support utility U in Equation 4. Ten-step ranges visualize local variation, not uncertainty across independent training runs.

Table 3: **Storage–utility ablation.** 7B component study comparing storage and reward choices. V*: accuracy (%); POPE: F1 (%).

Storage	Reward	V*	POPE
FIFO	Outcome	80.63	86.47
FIFO	Utility	81.68	87.13
Spatial	Outcome	83.25	87.91
Spatial	Utility	**86.82**	**89.10**

Table 4: **Utility-weight ablation.** Spatial storage with varying λ. V*, HRBench4K, and HRBench8K report accuracy (%).

λ	V*	HR4K	HR8K
0.0	83.25	68.92	65.41
0.1	80.63	71.08	69.35
0.2	**86.82**	**75.03**	**72.18**
0.4	84.29	74.88	71.12



Figure 3: **Training traces.** Response length, result reward, and a separately supplied consistency-reward series over 2,000 steps. Solid curves use an exponential moving average with span 50; vertical segments with circular endpoints show ten-step minimum–maximum ranges.

Answer-support ranking. Figure 4 illustrates how correctness lift characterizes the ranking induced by utility. At threshold q, lift is the accuracy among trajectories above utility quantile q divided by the unconditional accuracy. The horizontal baseline of one denotes no enrichment. The curves are close at lower thresholds and separate toward the upper tail: at $q = 0.95$, early, middle, and late values are approximately 1.13, 1.17, and 1.20. The illustration makes the diagnostic's interpretation concrete: a lift of 1.20 denotes a selected population with 20% higher correctness relative to the full population, rather than a 20-percentage-point accuracy increase. Appendix D gives its formal definition.

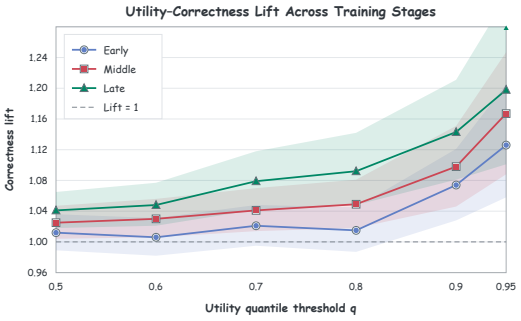


Figure 4: **Utility–correctness lift (illustrative).** Three training stages with supplied nominal 95% bands, clipped at 1.28.

Reading the utility tail. Increasing the threshold retains trajectories with higher utility but reduces the selected population. The comparison therefore concerns the concentration of correct answers, not the total number of questions answered correctly. A late-stage curve above an early-stage curve at the same quantile represents stronger enrichment within its own stage; absolute accuracy also depends on that stage's unconditional accuracy. Local plateaus reflect thresholds that change the selected set without substantially changing its correctness fraction. In this diagnostic, the separation near the upper tail is the relevant pattern, whereas the ablations in Tables 3 and 4 quantify the effect of including utility during training.

4.5 Capacity and Qualitative Analysis

Capacity-sensitive refinement. Figure 5 illustrates distinct capacity-response patterns through local peaks, crossings, and declines. V* and MME-RealWorld-Lite show larger changes than POPE, while HRBench8K exhibits a late-capacity decline. The utility-aware curve is generally strong but not uniformly highest. Here K counts spatial cells with slots per cell fixed, so changing K changes both

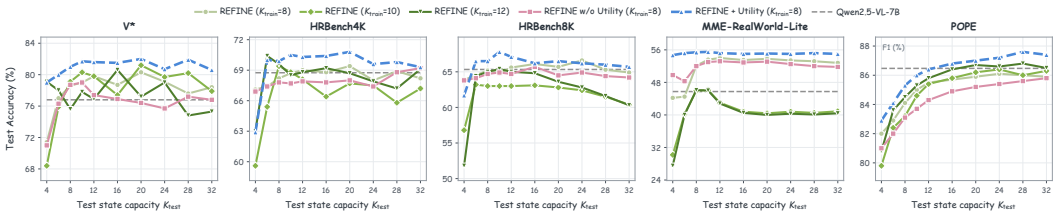


Figure 5: **Illustrative test-capacity sensitivity.** REFINE curves for different training capacities and objectives, with fixed Qwen2.5-VL-7B reference lines. The supplied development curves retain their local variation without additional smoothing. POPE reports F1; the other panels report accuracy.

spatial granularity and regional storage. Finer partitions allocate more regional blocks but reduce each cell's spatial extent; the global overview supplies context across their boundaries. This differs from increasing m, which adds representational slots without changing the image partition.

Relational grounding. Figure 6 contrasts two responses to the same above/below question. Both provide a fluent explanation, but they select opposite relations: REFINE answers "below," matching the annotation, whereas GRPO answers "above." This example distinguishes verbal plausibility from relational correctness. The target object's appearance and its location relative to the comparison object must support the same answer. Persistent regional evidence and the global overview are designed to make those two information sources available together. Appendix E presents three further cases covering relative position, shape placement, and geometric area.

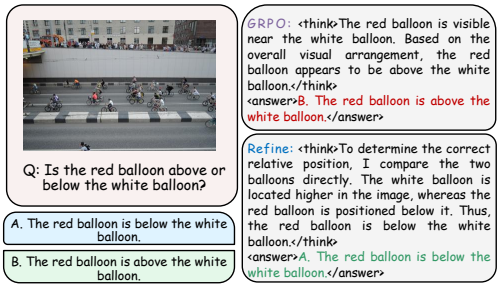


Figure 6: **Spatial-relation example.** REFINE's "below" answer matches the supplied annotation; GRPO selects "above."

What the experiments establish. The benchmark and ablation results address complementary questions. The benchmark tables show where the complete model improves over the backbone and comparison methods, with particularly large gains in real-world monitoring and positional search. The storage–utility ablation separates two design choices within REFINE: preserving regional evidence and rewarding the support it provides. Neither choice substitutes for the other, as the full combination exceeds both single-component variants on V* and POPE. The weight sweep further shows that increasing the auxiliary signal is not uniformly beneficial. Together, these results support a moderate utility coefficient coupled with persistent spatial storage, rather than relying on additional acquisition alone. Chart and OCR results identify tasks where the complete model's advantage is smaller or absent, giving a task-specific picture of its behavior.

5 Conclusion

REFINE turns selective inspection into local refinement of a persistent spatial evidence state. Its answer-support utility links the quality of retained evidence to outcome supervision through GRPO. Benchmark results show gains in positional search and real-world reasoning, and component ablations demonstrate that spatial storage and utility-aware learning work better together than either component alone. The strongest tested utility weight balances evidence support with answer correctness. REFINE's central principle is to preserve useful observations while refining the regions that warrant closer inspection.

AI USE STATEMENT

LLMs were used solely for grammar and language polishing; all ideas, analyses, and writing were produced entirely by the authors.

REPRODUCIBILITY STATEMENT

Section 3 defines the state representation, local replacement, support probe, reward, and optimization objective. Appendix A provides the reference configuration and state-consistent execution details; Algorithms B and B specify inference and training. Appendix C derives the utility and storage properties, and Appendix D specifies score aggregation and diagnostic calculations. The manuscript source includes the figure PDFs and bibliography required to reproduce the document.

ETHICS STATEMENT

This work studies visual reasoning on existing image benchmarks. Fine-grained inspection can expose identifying details, making dataset licenses, access controls, and privacy protections relevant to downstream use. Monitoring and autonomous-driving results concern offline question answering rather than safety-critical deployment. The support probe consumes reference answers only during training and is absent from inference. Responsible use requires keeping evaluation data separate from training rewards and assessing failure modes in the intended application context.

REFERENCES

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. arXiv preprint arXiv:2502.13923, 2025. URL https://arxiv.org/abs/2502.13923.

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. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=wZI5qkQeDF.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? In *Advances in Neural Information Processing Systems*, volume 37, pp. 27056–27087, 2024a. doi: 10.52202/079017-0850. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f8ee6a3d766b426d2618e555b5aeb39-Abstract-Conference.html.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 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*, pp. 24185–24198, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Chen_InternVL_Scaling_up_Vision_Foundation_Models_and_Aligning_for_Generic_CVPR_2024_paper.html.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. OpenVLThinker: Complex Vision-Language Reasoning via Iterative SFT-RL Cycles. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b33286edd8602ead8cb966358b01d115-Abstract-Conference.html.

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou, and Zhongyu Wei. Interleaved Latent Visual Reasoning with Selective Perceptual Modeling. In *Proceedings of the 64th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 29316–29335. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1351. URL https://aclanthology.org/2026.acl-long.1351/.

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 *Computer Vision – ECCV 2024*, volume 15081 of *Lecture Notes in Computer Science*, pp. 148–166. Springer Nature Switzerland, 2025a. doi: 10.1007/978-3-031-73337-6_9. URL https://link.springer.com/chapter/10.1007/978-3-031-73337-6_9.

Xingyu Fu, Minqian Liu, Zhengyuan Yang, John Richard Corring, Yijuan Lu, Jianwei Yang, Dan Roth, Dinei Florencio, and Cha Zhang. ReFocus: Visual Editing as a Chain of Thought for Structured Image Understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 17783–17805. PMLR, 2025b. URL https://proceedings.mlr.press/v267/fu25d.html.

Tanmay Gupta and Aniruddha Kembhavi. Visual Programming: Compositional Visual Reasoning Without Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Gupta_Visual_Programming_Compositional_Visual_Reasoning_Without_Training_CVPR_2023_paper.html.

Xinhai Hou, Shaoyuan Xu, Manan Biyani, Moyan Li, Jia Liu, Todd C Hollon, and Bryan Wang. CodeV: Code with Images for Faithful Visual Reasoning via Tool-Aware Policy Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21500–21510, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Hou_CodeV_Code_with_Images_for_Faithful_Visual_Reasoning_via_Tool-Aware_CVPR_2026_paper.html.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/pdf/ff5de4e779d104fc077d7afe5d46f4b419fcdab4.pdf.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a. URL https://proceedings.mlr.press/v202/li23q.html.

Kelvin Li, Chuyi Shang, Leonid Karlinsky, Rogerio Feris, Trevor Darrell, and Roei Herzig. Latent Implicit Visual Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33457–33466, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Latent_Implicit_Visual_Reasoning_CVPR_2026_paper.html.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

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*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/663bce02a0050c4a11f1eb8a7f1429d3-Abstract-Conference.html.

Yunze Man, De-An Huang, Guilin Liu, Shiwei Sheng, Shilong Liu, Liang-Yan Gui, Jan Kautz, Yu-Xiong Wang, and Zhiding Yu. Argus: Vision-Centric Reasoning with Grounded Chain-of-Thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14268–14280, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Man_Argus_Vision-Centric_Reasoning_with_Grounded_Chain-of-Thought_CVPR_2025_paper.html.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring Visual Aha Moment with Rule-based Large-scale Reinforcement Learning, 2025. URL https://arxiv.org/abs/2503.07365v1.

Yingqian Min, Kun Zhou, Yifan Li, Yuhuan Wu, Han Peng, Yifan Du, Wayne Xin Zhao, Min Yang, and Ji-Rong Wen. Improving Vision-language Models with Perception-centric Process Reward Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33099–33109, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Min_Improving_Vision-language_Models_with_Perception-centric_Process_Reward_Models_CVPR_2026_paper.html.

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent Models of Visual Attention. In *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/hash/3e456b31302cf8210edd4029292a40ad-Abstract.html.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. LMM-R1: Empowering 3B LMMs with Strong Reasoning Abilities Through Two-Stage Rule-Based RL, 2025. URL https://arxiv.org/abs/2503.07536.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. arXiv preprint arXiv:1707.06347, 2017. URL https://arxiv.org/abs/1707.06347.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark for Chain-of-Thought Reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642, 2024a. doi: 10.52202/079017-0275. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/0ff38d72a2e0aa6dbe42de83a17b2223-Abstract-Datasets_and_Benchmarks_Track.html.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv preprint arXiv:2402.03300, 2024b. URL https://arxiv.org/abs/2402.03300.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model, 2025a. URL https://arxiv.org/abs/2504.07615.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. ZoomEye: Enhancing Multimodal LLMs with Human-Like Zooming Capabilities through Tree-Based Image Exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6602–6618. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.emnlp-main.335. URL https://aclanthology.org/2025.emnlp-main.335/.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel Reasoner: Incentivizing Pixel Space Reasoning via Curiosity-Driven Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/0c38f54740062529aa4117a04b583f3c-Abstract-Conference.html.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual Inference via Python Execution for Reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Suris_ViperGPT_Visual_Inference_via_Python_Execution_for_Reasoning_ICCV_2023_paper.html.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2c84844a559e4f962752570bff456ae4-Abstract-Conference.html.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and Jun Xiao. VGR: Visual Grounded Reasoning. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/90aeee6dfe75ab6b5d6958bce40d3e16-Abstract-Conference.html.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad0edc7d5fa1a783f063646968b7315b-Abstract-Datasets_and_Benchmarks_Track.html.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv preprint arXiv:2409.12191, 2024b. URL https://arxiv.org/abs/2409.12191.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in Latent Visual Space Beyond Image and Language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12030–12040, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_Monet_Reasoning_in_Latent_Visual_Space_Beyond_Image_and_Language_CVPR_2026_paper.html.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):7907–7915, 2025b. doi: 10.1609/aaai.v39i8.32852. URL https://ojs.aaai.org/index.php/AAAI/article/view/32852.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. Perception-Aware Policy Optimization for Multimodal Reasoning. In *International Conference on Learning Representations*, 2026c. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/82846e19e6d42ebfd4ace4361def29ae-Abstract-Conference.html.

Zifu Wang, Junyi Zhu, Bo Tang, Zhiyu Li, Feiyu Xiong, Jiaqian Yu, and Matthew B. Blaschko. Jigsaw-R1: A Study of Rule-based Visual Reinforcement Learning with Jigsaw Puzzles. *Transactions on Machine Learning Research*, 2025c. URL https://openreview.net/forum?id=XqQCsuyPve.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Penghao Wu and Saining Xie. V*: Guided Visual Search as a Core Mechanism in Multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13084–13094, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.html.

xAI. Grok-1.5 Vision Preview. Official release introducing RealWorldQA, April 2024. URL https://x.ai/news/grok-1.5v.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. Perception-R1: Pioneering Perception Policy with Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/880ee6e23b6ee548ebcccafe0ef0f513-Abstract-Conference.html.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1859–1869, 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_R1-VL_Learning_to_Reason_with_Multimodal_Large_Language_Models_via_ICCV_2025_paper.html.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MME-RealWorld: Could Your Multimodal LLM Challenge High-Resolution Real-World Scenarios that are Difficult for Humans? In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/df29d63af05cb91d705cf06ba5945b9d-Abstract-Conference.html.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think Beyond Images. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8080912b919ad0344c0d34e68983dfcb-Abstract-Conference.html.

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. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cdb347d7516d52a9280ea9d6708911f6-Abstract-Conference.html.

A STATE REPRESENTATION AND EXECUTION

Reference configuration. REFINE uses Qwen2.5-VL as the backbone, with a frozen visual encoder and projection and a trainable language policy (Bai et al., 2025). The reference RL configuration follows the supplied Thyme settings (Zhang et al., 2026): learning rate 5×10^{-7}, one epoch, batch size 256, four rollouts per question, KL coefficient 0.001, and repetition penalty 1.05. The base evidence state has $K = 16$ spatial cells, $m = 16$ slots per cell, and two refinement levels. The reward coefficients are $\lambda = 0.2$ and $\omega = 0.1$. These parameters specify distinct controls: K determines the partition, m determines regional representational capacity, and the acquisition budget limits how often the state can be refined.

Coordinates and spatial identity. Let the original image have width W and height H. For a rectangular partition with $n_x n_y = K$, cell (i, j) has normalized boundaries

$$b_{i,j} = \left[\frac{j}{n_x}, \frac{j+1}{n_x}\right) \times \left[\frac{i}{n_y}, \frac{i+1}{n_y}\right), \quad 0 \le i < n_y, \quad 0 \le j < n_x. \tag{7}$$

The last row and column include the image boundary. A crop maps these boundaries to the original pixel coordinates, so changing the requested resolution does not change the cell's spatial owner. A row-major index $k = 1 + in_x + j$ fixes the location of each regional block in the serialized state. The overview and cell identifiers preserve the relation between a local observation and the complete image.

Pooling and slot dimensions. After visual encoding and projection, a regional feature map has shape $h \times w \times d$. Adaptive average pooling maps it to a prescribed $r \times c$ grid with $rc = m$. Each output vector averages a deterministic spatial bin of the input feature map. Flattening the pooled grid in row-major order produces $Z_k \in \mathbb{R}^{m \times d}$. Consequently, a refined crop can contain more input pixels and encoder features without increasing the number of retained regional vectors. Pooling parameters are fixed, and no gradient is propagated through acquisition or state construction.

Valid actions. For current resolution labels $\ell_{t,k}$, an inspection is admissible when the requested level exceeds the stored level and its cost fits the remaining budget. Writing $c(k, \ell)$ for its pixel cost gives

$$\mathcal{A}_t^{\text{inspect}} = \{(k, \ell) : 1 \le k \le K,\ \ell_{t,k} < \ell \le L,\ B_t^{\text{count}} > 0,\ c(k, \ell) \le B_t^{\text{pixel}}\}. \tag{8}$$

Answering remains admissible at every step. The action grammar rejects malformed commands and masks inadmissible inspections before execution. An accepted inspection decrements the count budget by one, subtracts its pixel cost, and updates only the selected cell. This interface avoids repeated acquisition at a resolution already represented in the state.

State-consistent likelihoods. The rollout record associates each generated segment with its input evidence state, action mask, and old-policy log probabilities. Recomputing a likelihood uses that same state rather than the trajectory's final state. Otherwise, an early inspection decision would be evaluated using evidence obtained only after the decision. Environment-inserted features and metadata are excluded from the optimized token set; generated reasoning, inspection commands, and answer tokens are included. Immutable regional blocks can be shared across stored snapshots because replacement allocates a new selected block and leaves all others unchanged.

Decoder execution. The visual prefix contains the global overview and the current regional blocks. Text history retains generated content and action metadata without accumulating additional crop embeddings. Updating an early visual block changes the prefix seen by subsequent tokens, so the reference execution rebuilds the corresponding decoder context. The distinction between fixed visual storage and decoder computation is important: a bounded state limits retained evidence, while encoding and prefix reconstruction still contribute to runtime.

B INFERENCE AND TRAINING ALGORITHMS

Inference loop. Algorithm B implements Figure 2: each inspection returns an updated state to the policy, and answering terminates the interaction. Inference uses neither the reference answer nor the probe. The optional rollout record retains state snapshots for training.

[htbp] [1] Image I, question q, policy π_θ, frozen encoder E, pooling $\mathcal{P}_m$, cells $\{b_k\}$, levels L, initial budget B_0 Encode coarse observation: $V_0 \leftarrow E(I_0)$ Initialize $M_0 \leftarrow \{(\mathcal{P}_m(V_0[b_k]), 0)\}_{k=1}^{K}$

$M \leftarrow M_0$; $B \leftarrow B_0$; $h \leftarrow \varnothing$; $\tau \leftarrow \varnothing$ Construct $s \leftarrow (q, V_0, M, h, B)$ and valid actions from Equation 8 Generate a reasoning segment and action with $\pi_\theta(\cdot \mid s)$ Record generated tokens, their state, mask, and policy likelihoods in τ the action is Answer(a) a, M_0, M, τ Valid Inspect(k, ℓ) $Z \leftarrow \mathcal{P}_m(E(\mathrm{crop}_\ell(I, b_k)))$ Replace (Z_k, ℓ_k) in M by (Z, ℓ) Other cells unchanged Decrement inspection-count and pixel budgets Append generated text and action metadata to h Rebuild the visual prefix from V_0 and the updated M

Groupwise training. Algorithm B uses one frozen behavior snapshot for group sampling and support evaluation. One initial score is shared across the group, and each final state receives an answer-only score. Standardized trajectory rewards enter the clipped objective; the KL reference remains fixed at initialization.

[htbp] [1] Training records (I, q, y^*), initial policy π_{θ_0}, rollout count G, reward weights ω, λ, KL coefficient β $\theta \leftarrow \theta_0$; freeze reference policy $\pi_{\mathrm{ref}} \leftarrow \pi_{\theta_0}$ each training minibatch Freeze behavior policy $\pi_{\mathrm{old}} \leftarrow \pi_\theta$ and probe $\bar{\theta} \leftarrow \theta$ each (I, q, y^*) in the minibatch Canonicalize the reference answer tokens y^* Sample G trajectories using Algorithm B with π_{old} Evaluate $\phi_0 \leftarrow \Phi_{\bar{\theta}}(M_0; q, y^*)$ using Equation 3 $n = 1, \ldots, G$ Evaluate $\phi_n \leftarrow \Phi_{\bar{\theta}}(M_{T_n}^{(n)}; q, y^*)$ $U_n \leftarrow \phi_n - \phi_0$ Compute answer correctness C_n and format validity F_n $R_n \leftarrow C_n + \omega F_n + \lambda C_n F_n U_n$ $\overline{R} \leftarrow G^{-1} \sum_n R_n$; $\sigma_R \leftarrow \mathrm{std}(R_1, \ldots, R_G)$ $A_n \leftarrow (R_n - \overline{R})/(\sigma_R + \epsilon)$ for each trajectory Recompute generated-token likelihoods under their recorded evidence states Update θ by maximizing Equation 6; detach rewards and probe scores π_θ with the frozen encoder and state-update interface

Answer-only scoring interface. The probe receives the question, overview, current spatial state, and a direct-answer instruction, but no sampled reasoning or answer. Teacher forcing supplies canonical reference prefixes only for likelihood evaluation. Initial and final scores share the same answer tokens and prompt; reference tokens never enter trajectory generation.

Two policy snapshots. The refreshed behavior snapshot supplies rollout probabilities and the within-group probe; the initialization snapshot anchors KL regularization. This separation keeps initial–final comparisons consistent while the policy evolves. Probe scores are detached forward evaluations, not an independently trained reward model.

C Properties of the state and utility

Locality invariant. Let $\mathcal{R}_{k,\ell}$ denote the replacement operator for a valid inspection. By Equation 2, its projection onto every other cell is the identity:

$$\big[\mathcal{R}_{k,\ell}(M)\big]_j = M_j \quad (j \neq k). \tag{9}$$

Induction over an inspection sequence shows that a cell never inspected retains its initialized block. A cell inspected multiple times retains the block from its latest valid inspection. Thus changes to the state can be traced to specific acquired regions rather than shifts in a recency queue.

Order of independent replacements. For two distinct cells $j \neq k$, fixed crop encodings, and admissible levels, replacement commutes at the state level:

$$\mathcal{R}_{j,\ell_j}(\mathcal{R}_{k,\ell_k}(M)) = \mathcal{R}_{k,\ell_k}(\mathcal{R}_{j,\ell_j}(M)). \tag{10}$$

Each operation writes a disjoint block and preserves the other. The resulting visual state is therefore determined by the retained regional observations, although the policy's text history and remaining budget can still depend on acquisition order. This property separates the evidence representation from the sequence of decisions that produced it.

Visual storage bound. With s metadata tokens per cell, the retained visual prefix has

$$N_{\mathrm{visual}} = |V_0| + K(m + s), \qquad \frac{\partial N_{\mathrm{visual}}}{\partial T} = 0. \tag{11}$$

An append-based representation retaining a_t visual tokens from inspection t instead has $|V_0| + \sum_{t=1}^{T} a_t$ tokens, before any truncation or compression. REFINE's bound concerns the model-visible state at one decision. Rollout records, text history, crop encoding, and repeated decoder prefilling have their own costs.

Signed telescoping utility. For a fixed probe and reference answer, adjacent potential terms cancel:

$$\sum_{t=0}^{T-1} u_t = \sum_{t=0}^{T-1} \big[\Phi(M_{t+1}) - \Phi(M_t)\big] = \Phi(M_T) - \Phi(M_0). \tag{12}$$

Because $0 \le \Phi \le 1$, the net utility lies in $[-1, 1]$. A no-inspection trajectory has $M_T = M_0$ and zero utility. More generally, any sequence ending with the same probe input as it started has zero net gain. Clipping individual negative increments would break cancellation; REFINE instead uses the signed endpoint difference. The auxiliary reward therefore contains no bonus proportional to inspection count.

Correctness-preserving reward gap. For two format-valid trajectories, let τ_c be correct and τ_i incorrect. The gate gives

$$R(\tau_c) - R(\tau_i) = 1 + \lambda U(\tau_c) \ge 1 - \lambda > 0. \tag{13}$$

Among two correct, format-valid trajectories, the reward difference is $\lambda[U(\tau_a) - U(\tau_b)]$. Utility consequently orders their evidence gains without reversing their priority relative to an incorrect format-valid answer. This result follows from the bounded potential and the gate, not from an assumption that every inspection improves support.

Within-group normalization. For fixed group rewards, subtracting their mean and dividing by a positive standard deviation preserves reward ordering. When all trajectories are correct and format-valid, their common correctness and format terms cancel, as does the shared initial potential:

$$R_n - \overline{R} = \lambda \left[\Phi(M_{T_n}^{(n)}) - \frac{1}{G} \sum_{j=1}^{G} \Phi(M_{T_j}^{(j)}) \right]. \tag{14}$$

Such a group can supply a policy gradient through differences in final evidence support even though binary correctness is constant. If rewards are identical, normalized advantages are zero. In mixed-correctness groups the outcome gate remains active, and Equation 13 determines the ordering before normalization.

Support versus answer probability. The potential is the geometric mean of reference-token probabilities. Averaging log probabilities reduces the direct dependence on answer length, but the score remains conditional on the probe, prompt, tokenization, and reference wording. Using an identical canonical answer within a comparison holds these factors fixed. The utility quantifies a change in this particular support measure; it is not a proof that the acquired pixels causally determine the answer or that the probe is calibrated.

D METRICS AND DIAGNOSTIC CALCULATIONS

Benchmark aggregation. REFINE's V* aggregate is the subtask macro-average,

$$A_{\text{macro}} = \tfrac{1}{2}(A_{\text{attr}} + A_{\text{pos}}). \tag{15}$$

It yields 81.98 for the 3B model and 86.82 after rounding for the 7B model. A pooled accuracy instead weights subtasks by question counts; V*Bench contains 115 attribute and 76 spatial-relation questions (Wu & Xie, 2024). Table 1 preserves the recorded aggregates and exposes both subtask scores. The radar's MME-RealWorld spoke is the descriptive equal-weight mean $(68.60 + 50.37)/2 = 59.485$, rounded to 59.49. Table 2 retains perception and reasoning separately.

Absolute gains. The improvement rows in Table 2 subtract the Qwen2.5-VL-7B score from the REFINE score in each domain. They are percentage-point differences, not percentage changes. For example, the perception overall gain is $68.60 - 60.94 = 7.66$, and the reasoning overall gain is $50.37 - 38.59 = 11.78$. Domain scores expose which question categories contribute to the comparison without treating the overall benchmark metric as an unweighted mean of displayed domains.

Factorial contrast. Let $A_{s,u}$ denote the score with storage $s \in \{\text{FIFO}, \text{Spatial}\}$ and utility indicator $u \in \{0, 1\}$. The interaction contrast is

$$\Delta_{\text{int}} = (A_{\text{Spatial},1} - A_{\text{Spatial},0}) - (A_{\text{FIFO},1} - A_{\text{FIFO},0}). \tag{16}$$

Table 3 gives $\Delta_{\text{int}} = 2.52$ on V* and 0.53 on POPE. These differences summarize the dependence of the utility gain on storage choice. Table 4 complements this binary comparison by varying the strength of utility under spatial storage; the highest tested coefficient is not the highest-scoring configuration.

Correctness lift. For trajectories indexed by i, let u_i be utility, $c_i \in \{0, 1\}$ correctness, and Q_q the empirical utility quantile. Define $S_q = \{i : u_i \geq Q_q\}$. The lift at threshold q is

$$\text{Lift}(q) = \frac{|S_q|^{-1} \sum_{i \in S_q} c_i}{N^{-1} \sum_{i=1}^{N} c_i}, \tag{17}$$

for a population with nonzero unconditional correctness. Ties at the threshold remain in the selected set. Lift measures enrichment within a trajectory population, whereas the ablations compare training configurations. Figure 4 is an illustrative visualization of this diagnostic with nominal bands; those bands are not resampled confidence estimates from question-level outcomes.

Reading the visual diagnostics. The training figure displays an exponential moving average and ten-step minimum–maximum ranges; the latter summarize temporal variation rather than independent-run uncertainty. The utility and capacity figures are illustrative development plots. In particular, the capacity plot retains its fixed V* reference line of 76.8, while the benchmark table records 70.68 for Qwen2.5-VL-7B. These plots are interpreted within their displayed settings and do not enter the numerical benchmark comparisons. The qualitative example compares annotated final answers rather than a recorded sequence of inspection actions.

Acquisition and computation accounting. Rollout cost comprises coarse encoding, crop encoding, prefix reconstruction, and decoding. Training adds one shared initial and up to G final probe evaluations per question, independently of inspection count; identical states can reuse their scores. Endpoint utility thus avoids probing every intermediate update.

E Comprehensive Case Studies

E.1 Case 1: relative-position reasoning on V*

Case 1. Relative-Position Reasoning on V*: Dog versus Scooter.



Question: Is the dog on the left or right side of the scooter?
Ground truth: A. The dog is on the left side of the

AI Think with image...

<think>
The global overview V_0 is not sufficient for a confident spatial comparison. I will refine the evidence for the dog and scooter regions and compare their retained spatial positions.
</think>
<inspect>Inspect(k_dog, ℓ1)</inspect>



<reasoning>
The high-resolution observation refines the dog-related slots in Mt, placing the dog in the right-central foreground.
</reasoning>
<inspect>Inspect(k_scooter, ℓ1)</inspect>



<reasoning>
The scooter-related slots are refined while the previously acquired dog evidence is preserved. In the updated evidence state, the scooter lies to the right of the dog.
</reasoning>
<answer>The dog is on the left side of the scooter.</answer>

Figure 7: **Case 1: dog–scooter spatial comparison.** The trace inspects the two objects in sequence, retains their spatial correspondence, and answers that the dog is to the left of the scooter, matching the annotation.

Comparing retained object locations. Figure 7 separates localization from relational reasoning. The first inspection establishes the dog's position; the second refines the scooter evidence while retaining the dog-related slots. Locating the scooter to the right of the dog then determines the requested relation: the dog is on the scooter's left. The reference frame matters because "right-central foreground" describes the dog's absolute image position, whereas "left of the scooter" describes a relation between objects. These statements are compatible. The case illustrates the role of persistent spatial indexing: a later observation is interpreted together with an earlier object's retained location, rather than replacing that object in a recency-based visual history.

E.2 CASE 2: SHAPE-PLACEMENT REASONING ON MATHVISION

Case 2 . Shape-Placement Reasoning on MathVision: Two Identical Pieces in a \(4\times3\) Grid.

<image1>:



<image2>:



Question: Question: Ben wants to cut out two identical pieces out of the \(4\times3\) grid. For which of the following shapes can he not achieve that?

Ground truth: A.

Think with image...

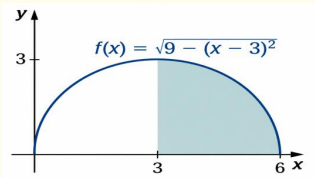
<think>
The task is to determine whether two identical copies of each candidate shape can be placed within the 4×3 grid without overlap. Full coverage of the grid is not required, so I should inspect the candidate geometry rather than infer the answer from the total grid area.
</think>
<inspect>4×3 grid and candidate shapes, high resolution</inspect>
<reasoning>
The target region is a 4×3 grid, and each candidate consists of five unit cells. The relevant constraint is whether two congruent copies of a candidate can both fit inside the grid without overlapping or extending beyond its boundary.
</reasoning>
<inspect>shape A, high resolution</inspect>
<reasoning>
After refining the geometry of Shape A, a second identical copy cannot be placed within the remaining grid cells without overlap or crossing the grid boundary.
</reasoning>
<inspect>shapes B–E, high resolution</inspect>
<reasoning>
The refined geometries of Shapes B–E each admit two congruent, non-overlapping placements within the 4×3 grid. Shape A is therefore the only candidate that does not satisfy the placement constraint.
</reasoning>
<answer>A. Shape A.</answer>

Figure 8: **Case 2: fitting two congruent pieces.** The trace compares candidate geometries with a 4×3 grid and selects A, matching the annotation. The task requires non-overlapping placement within the grid, not complete coverage.

Preserving constraints during candidate comparison. Figure 8 maintains the grid boundary while inspecting the candidate shapes. Each candidate has five cells, so two copies occupy ten of the twelve cells; an area check alone cannot distinguish the options. The decisive constraint is geometric compatibility: both congruent copies must fit without overlap or boundary crossing. The displayed reasoning rejects A and identifies feasible placements for B–E. This sequence connects local shape inspection with a persistent global constraint. REFINE's state retains the grid structure while candidate evidence is updated, allowing successive comparisons to use the same spatial frame. The selected answer agrees with the provided ground truth.

E.3 CASE 3: LINKING A CURVE TO ITS SHADED AREA

Case 3. Geometric Area Reasoning on MMStar: The Shaded Quarter-Circle.



Question: What is the size of the shaded area under the curve? Round the answer to 2 decimal places. Options: (A) 7.07, (B) 6.97, (C) 7.14, (D) 9.00.
Ground truth: A.

Think with image...

<think>
The global overview suggests that the plotted curve is part of a circle, but the exact geometry and shaded interval should be verified before computing the area. I will refine the evidence around the function expression and the shaded region.
</think>
<inspect>curve equation and coordinate axes</inspect>
<reasoning>
The refined evidence shows
f(x) = sqrt(9 - (x - 3)^2),
which is the upper semicircle of
(x - 3)^2 + y^2 = 9.
Thus, the circle is centered at (3,0) with radius 3.
</reasoning>
<inspect>shaded region from x = 3 to x = 6</inspect>
<reasoning>
The shaded portion occupies the right half of the upper semicircle, corresponding to one quarter of the full circle. Retaining the radius r = 3 from the previous evidence, its area is (1/4)πr^2 = 9π/4 ≈ 7.07.
</reasoning>
<answer>A. 7.07</answer>

Figure 9: **Case 3: geometric area reasoning on MMStar.** The supplied trace reads the circle equation and the shaded interval, then computes the quarter-circle area. The answer, A: 7.07, agrees with the annotation.

Retaining geometry across observations. Figure 9 separates recognition of the curve from identification of the shaded domain. Reading $f(x) = \sqrt{9 - (x-3)^2}$ establishes a radius of three and center $(3, 0)$. The second observation locates the shaded interval at $x \in [3, 6]$, selecting half of the upper semicircle. Combining these two facts gives $A = \pi(3)^2/4 = 9\pi/4 \approx 7.07$, consistent with option A. The trace illustrates why evidence acquired at one step must remain available after attention moves elsewhere: the radius comes from the equation, whereas the fraction of the circle comes from the shading. Neither observation alone completes the calculation.