POLICY-ADAPTIVE CREDIT ASSIGNMENT FOR SPATIAL REASONING WITH VERIFIABLE REWARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Image-based spatial reasoning requires perception, decomposition, and inference, but the stage that limits accuracy can change during training. Answer-level rewards neither localize failures to individual stages nor provide discriminative group-relative feedback when all rollouts receive the same reward. Process supervision provides intermediate feedback, but fixed weights cannot adapt reward allocation to these changing bottlenecks. To address this gap, we propose CALIPER, an adaptive process-reward framework supported by SpatialTrace, a corpus of stage-decomposed spatial reasoning traces. CALIPER jointly estimates the associations between stage-wise reference coverage and answer correctness through within-group ridge regression, using these associations as utility proxies to gradually update reward weights. The resulting process reward supervises all rollout groups, including uniformly failing ones. Because utility estimates can become unreliable when answer-discriminative groups are scarce, a closed-loop harness monitors estimator reliability and validation performance to guide checkpoint rollback, rollout diversification, and early stopping. Extensive experiments across eight benchmarks and three backbones show that CALIPER achieves the highest mean under both one-shot and harness training, with consistent gains over fixed allocation across RL optimizers. Further analyses confirm the benefits of adaptive reward allocation and harness control, with favorable scaling in model/data size.

1 INTRODUCTION

Image-based spatial reasoning remains a persistent weakness of vision-language models, spanning metric and relational reasoning in 3D scenes (Ma et al., 2025), open-domain scene understanding (Zhang et al., 2025) and region-level search in high-resolution images (Wu & Xie, 2024). Although typically evaluated as a direct mapping from an image and question to an answer, solving these tasks often involves distinct stages of perception, decomposition and inference, whose relative importance varies across instances. Reinforcement learning from verifiable rewards (RLVR) (Guo et al., 2025), often using group-relative advantages over multiple rollouts of the same question (Shao et al., 2024), has been extended to multimodal reasoning by Vision-R1 (Huang et al., 2026) and DeepEyes (Zheng et al., 2026). However, its answer-level supervision has two limitations. It cannot distinguish failures in perception, decomposition or inference, and it provides no discriminative signal when all rollouts receive the same reward, especially on difficult spatial instances.

Reference-assisted verification enriches answer grading but retains a scalar outcome reward, leaving stage-level credit assignment and uniform-reward degeneracy unresolved, as illustrated in Fig. 1a,b. Process supervision (Uesato et al., 2022) provides finer feedback through human-annotated step labels (Lightman et al., 2024) or rollout-derived labels in Math-Shepherd (Wang et al., 2024b). However, a stage score quantifies alignment with a reference or grading criterion. It does not determine the supervision weight appropriate to that stage. Combining such scores with predefined weights leaves reward allocation insensitive to changes in the relationship between stage performance and answer correctness. Fig. 1c illustrates this gap in adaptive credit allocation across perception, decomposition and inference despite access to intermediate feedback.

Stage utility guides allocation through the association between stage performance and answer correctness under the current rollout distribution. Low reference coverage does not establish the expected benefit of increasing a stage's reward weight. Allocation therefore requires an estimate of

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

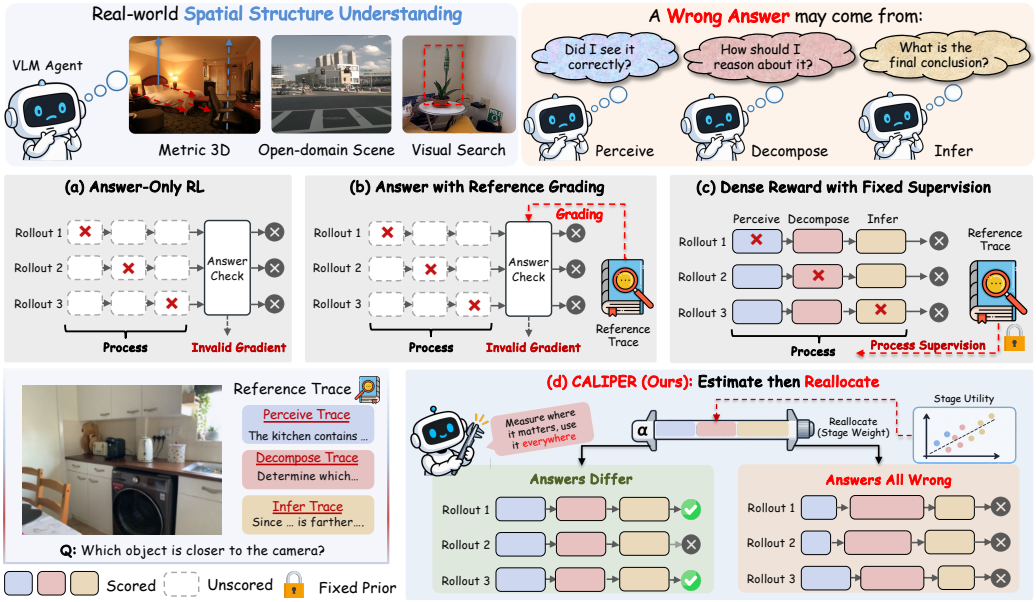


Figure 1: Supervision for spatial reasoning. Outcome-only rewards (a), as in GRPO (Shao et al., 2024) and DeepSeek-R1 (Guo et al., 2025), supervise final answers. Reference-assisted verification (b) improves grading but retains uniform-reward degeneracy. Process supervision (c) combines human-labeled feedback (Lightman et al., 2024) or Math-Shepherd scores (Wang et al., 2024b) with fixed weights. CALIPER (d) adapts these weights from estimated utility.

this association, conditioned on the policy's evolving behavior across spatial competencies. Rollout groups with variation in answer correctness supply the contrast needed for estimation, whereas uniformly correct or incorrect groups provide no such contrast. Adaptive supervision must estimate utility from informative groups and maintain reliable allocation as they become scarce.

To address this gap, we introduce SpatialTrace and CALIPER, as illustrated in Fig. 1d. SpatialTrace represents heterogeneous spatial tasks through a common schema for perception, decomposition and inference while retaining available reasoning from incomplete traces. CALIPER uses this representation to allocate process supervision according to estimated stage utility. It estimates the association between stage coverage and answer correctness from answer-discriminative rollout groups, then applies the resulting allocation to all groups, including uniformly failing ones. A closed-loop harness monitors estimation reliability and validation performance to guide checkpoint rollback, rollout diversification and stopping. Our contributions are as follows.

- We formulate process-reward allocation as a policy-dependent estimation problem, distinguishing stage quality from its association with answer correctness. Utility estimation is restricted to informative groups, while process supervision applies to all groups.
- We construct SpatialTrace, a reasoning-augmented spatial dataset spanning heterogeneous spatial tasks, in which approximately 84% of items carry complete three-stage traces and the remainder retain available reasoning without imputed stage labels.
- We propose CALIPER, a closed-loop framework that estimates stage utility from answer-discriminative rollout groups, applies the resulting allocation to all groups, including degenerate groups, and audits that estimate through a rule-based harness.
- Across eight benchmarks and three backbones, CALIPER attains the highest mean under both protocols and improves on fixed allocation across RL optimizers. Ablations support adaptive allocation and harness control, with gains at larger model and data scales.

2 Related Work

Spatial Reasoning in Vision-Language Models. Spatial reasoning in vision-language models advances through spatial supervision, visual search and outcome optimization. SpatialVLM (Chen

et al., 2024a) and SpatialRGPT (Cheng et al., 2024) derive metric and relational supervision from 3D reconstruction or depth. Benchmarks include 3DSRBench (Ma et al., 2025) for 3D relations, SpatialEval (Wang et al., 2024a) for synthetic reasoning, MME-RealWorld (Zhang et al., 2025) for real-world scenes and RoboSpatial (Song et al., 2025) for embodied settings. V* (Wu & Xie, 2024), ZoomEye (Shen et al., 2025) and DeepEyes (Zheng et al., 2026) improve evidence acquisition through region selection, cropping and zooming. RLVR extends to multimodal reasoning in Vision-R1 (Huang et al., 2026), R1-Zero-style spatial training (Liao et al., 2025) and video spatial reasoning in SpaceR (Ouyang et al., 2025). These efforts improve supervision, evidence acquisition and outcome optimization while leaving intermediate-stage credit allocation unresolved.

Process Supervision in RLVR. Chain-of-thought prompting (Wei et al., 2022) makes intermediate reasoning explicit, while DeepSeek-R1 (Guo et al., 2025) optimizes reasoning trajectories with verifiable outcome rewards. GRPO (Shao et al., 2024) uses group-relative advantages, DAPO (Yu et al., 2025) modifies clipping and rollout filtering, and GSPO (Zheng et al., 2025a) uses sequence-level importance ratios. These methods govern how rewards are translated into policy updates, but do not change their granularity. Process supervision instead assigns feedback to intermediate reasoning steps, using direct annotations (Lightman et al., 2024) or labels derived from rollouts (Wang et al., 2024b). PRIME (Cui et al., 2025) integrates implicit process rewards into online RL, and PURE (Cheng et al., 2025) replaces summation with min-form credit assignment. Such feedback alleviates outcome-reward sparsity, but its aggregation follows a fixed rule, leaving reward allocation insensitive to the policy's changing bottlenecks. CALIPER addresses this orthogonal problem by estimating policy-dependent stage utility and adapting process-reward weights accordingly.

3 SpatialTrace Dataset

Task Formulation. SpatialTrace targets spatial question answering with an explicit reasoning trace. Given an image I and a question x, a vision-language policy receives $q = (I, x)$ and generates a trace τ followed by an answer a. The desired behavior comprises extracting relevant visual evidence, decomposing the question into spatial subproblems, and deriving an answer consistent with that evidence. We represent these components as $\tau = (\tau_{\mathrm{p}}, \tau_{\mathrm{d}}, \tau_{\mathrm{i}})$, corresponding to perception, decomposition and inference. A reference answer $a^{\star}$ supports outcome verification, while stage annotations support separate assessment of these reasoning components. This distinction permits analysis of intermediate reasoning even when the final answer is incorrect.

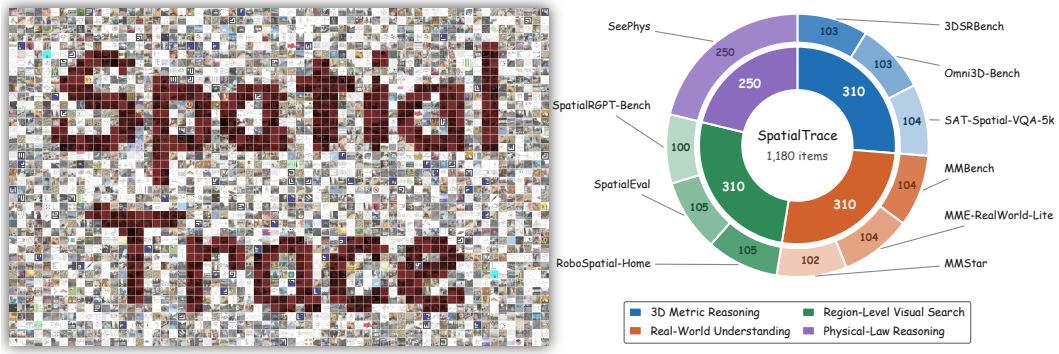


Figure 2: SpatialTrace examples (left) and source composition (right).

Construction. Learning this behavior requires coverage of spatial problems with different perceptual and inferential demands. SpatialTrace draws metric 3D questions from 3DSRBench (Ma et al., 2025), Omni3D-Bench (Brazil et al., 2023) and SAT-Spatial-VQA-5k (Ray et al., 2025). Open-domain questions are sampled from MMBench (Liu et al., 2024), MME-RealWorld-Lite (Zhang et al., 2025) and MMStar (Chen et al., 2024b). Region-level search instances are sampled from RoboSpatial-Home (Song et al., 2025), SpatialEval (Wang et al., 2024a) and SpatialRGPT-Bench (Cheng et al., 2024), with SeePhys (Xiang et al., 2026) supplying auxiliary physical reasoning. Fig. 2 summarizes SpatialTrace's 1,180 single-image questions with verifiable answers. Family quotas preserve coverage, with source-stratified splits of 1,000 training and 180 validation items.

Traces. The annotations encode the relationship between visual evidence, intermediate reasoning and the reference answer. Perception records relevant visual observations. Decomposition identifies the subproblems required by the question. Inference connects the observations to the answer. Each record also includes answer reconciliation and the final answer, allowing discrepancies between visual evidence and the source label to be recorded. An image-grounded annotator drafts the trace, and a text-only editor refines its reasoning and records unsupported inferences or ambiguity. Approximately 84% of the corpus contains all three stage fields. Other records retain their answers and available reasoning without imputed stage labels. Appendix A details the annotation protocol.

4 METHODOLOGY

Problem Formulation. Given an image–question pair $q = (I, x)$ and a verifiable answer $a^\star$, a policy π_θ generates a group g of n rollouts $\{o_i\}_{i\in g} \sim \pi_\theta(\cdot \mid q)$. Let r_i denote the reward of rollout o_i. Group-relative policy optimization (Shao et al., 2024) computes its advantage from the group mean μ_g and standard deviation σ_g,

$$A_i = \begin{cases} (r_i - \mu_g)/\sigma_g, & \sigma_g > 0, \\ 0, & \sigma_g = 0, \end{cases} \qquad \mu_g = \tfrac{1}{n}\textstyle\sum_{j\in g} r_j, \;\; \sigma_g^2 = \tfrac{1}{n}\textstyle\sum_{j\in g}(r_j - \mu_g)^2, \tag{1}$$

The same A_i applies to every token of o_i. If all rewards in g agree, then $\sigma_g = 0$ and the advantages are set to zero. Such groups contribute no advantage-weighted policy gradient, although other objective terms may remain. Let $\mathrm{acc}_i \in \{0, 1\}$ denote the verifier judgment of the answer in o_i against $a^\star$. Under a binary outcome reward, groups that are uniformly correct or uniformly incorrect both yield zero advantages. The problem is to assign credit among perception, decomposition and inference even when answer-level rewards cannot distinguish the rollouts.

Overview. Fig. 3 shows CALIPER's two coupled loops. During a policy update, a stage judge scores each rollout against the corresponding Spatial-Trace reference for perception, decomposition and inference. Groups with both correct and incorrect answers support the stage-utility estimate. The estimated utilities determine the weights used to reward every group. Across training rounds, the harness checks validation performance and estimator reliability. It then continues training, restores a checkpoint, increases rollout diversity or stops.

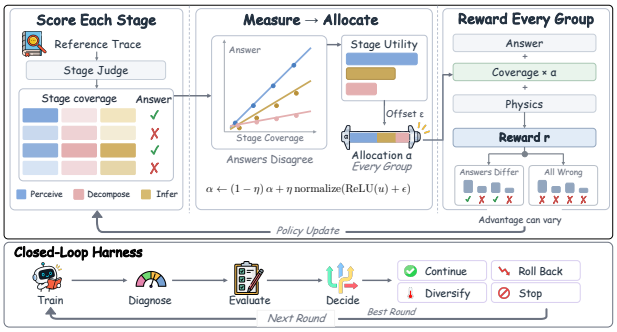


Figure 3: CALIPER scores stages against reference traces, estimates utility from groups with mixed answer correctness, and updates the policy using rewards for every group. Validation and estimator diagnostics guide continuation, rollback, rollout diversification or stopping.

Stage-Decomposed Process Reward. Let $\mathcal{S}$ index the perception, decomposition and inference stages. The reference trace for q is $\tau^\star = (\tau_s^\star)_{s\in\mathcal{S}}$. A judge J assigns each rollout a reference-coverage score $\mathrm{cov}_s(o_i) = J(o_i, \tau_s^\star) \in [0, 1]$ for stage s. Let $\mathrm{phys}_i \in [0, 1]$ be a physical-plausibility score and $\mathrm{len}_i \in [0, 1]$ a response-length penalty. The nonnegative weights w_a, w_m, w_p and $w_\ell \in [0, 1]$ control answer correctness, process coverage, physical plausibility and length, respectively. The allocation $\alpha \in \Delta^{|\mathcal{S}|}$ lies on the probability simplex and distributes the process weight w_m across stages. We define the rollout reward as

$$r(o_i) \;=\; \Big[\, w_a\,\mathrm{acc}_i \;+\; w_m \sum_{s\in\mathcal{S}} \alpha_s\,\mathrm{cov}_s(o_i) \;+\; w_p\,\mathrm{phys}_i \,\Big] \cdot \big(1 - w_\ell\,\mathrm{len}_i\big), \tag{2}$$

The score phys_i quantifies consistency with physical constraints on magnitudes, units and relations. For response length $|o_i|$ in tokens, target length L_0 and cap $L_{\max} > L_0$, the penalty is

$$\mathrm{len}_i = \min\left\{1, \max\left\{0, \frac{|o_i| - L_0}{L_{\max} - L_0}\right\}\right\}.$$

It increases linearly from zero at L_0 to one at $L_{\max}$. For incomplete references, the reward sums over available stages and renormalizes their weights.

Stage Utility. Stage utility is the conditional association between coverage and answer correctness under the current policy. Among groups with complete references, only answer-discriminative groups $\mathcal{M} = \{g : 0 < \sum_{i \in g} \mathrm{acc}_i < n\}$ contain the outcome variation needed to estimate it. For rollout o_i in such a group, we center each stage score and the outcome within the group,

$$\tilde{x}_{i,s} = \mathrm{cov}_s(o_i) - \frac{1}{n} \sum_{j \in g} \mathrm{cov}_s(o_j), \qquad \tilde{y}_i = \mathrm{acc}_i - \frac{1}{n} \sum_{j \in g} \mathrm{acc}_j.$$

The vector $\tilde{x}_i$ collects the centered stage scores. The design matrix $\tilde{X}$ stacks these vectors as rows, and $\tilde{y}$ stacks the corresponding outcomes. With $\bar{\sigma}^2 = \mathrm{tr}(\tilde{X}^\top \tilde{X})/|\mathcal{S}|$ and ridge coefficient $\lambda > 0$, we estimate utility $u \in \mathbb{R}^{|\mathcal{S}|}$ by

$$u = \arg\min_{\beta \in \mathbb{R}^{|\mathcal{S}|}} \left\{ \sum_{g \in \mathcal{M}} \sum_{i \in g} \big(\tilde{y}_i - \beta^\top \tilde{x}_i\big)^2 + \lambda \bar{\sigma}^2 \|\beta\|_2^2 \right\} = \big(\tilde{X}^\top \tilde{X} + \lambda \bar{\sigma}^2 \mathrm{I}\big)^{-1} \tilde{X}^\top \tilde{y}, \tag{3}$$

where I is the identity matrix and the closed form applies when $\bar{\sigma}^2 > 0$. Centering compares rollouts of the same question, and joint regression accounts for correlations among stage scores. The coefficients quantify each stage's association with correctness after accounting for the other stages. The scale $\bar{\sigma}^2$ is the mean diagonal of the Gram matrix, so the penalty follows the feature scale. Across batches, exponential decay $\gamma \in (0, 1)$ gives recent samples greater weight in the sufficient statistics $\tilde{X}^\top \tilde{X}$ and $\tilde{X}^\top \tilde{y}$. Let N_{eff} be the corresponding decayed sample mass and $N_{\min}$ its update threshold. We retain the current allocation unless $N_{\text{eff}} \geq N_{\min}$, feature variation is nonzero and at least one coefficient is positive.

Utility-Adaptive Allocation. Define $\mathrm{ReLU}(u) = \max(u, 0)$ elementwise and $\mathrm{normalize}(v) = v / \sum_s v_s$. An offset $\epsilon > 0$ keeps the normalized target positive, and a step size $\eta \in (0, 1]$ limits the change between updates. CALIPER interpolates between the current allocation and this target,

$$\alpha \;\leftarrow\; (1 - \eta)\, \alpha \;+\; \eta\, \mathrm{normalize}\big(\mathrm{ReLU}(u) + \epsilon\big), \tag{4}$$

The offset is added componentwise. It preserves positive weights but does not impose a lower bound of ϵ after normalization. Utilities are estimated from answer-discriminative groups, and the resulting allocation supervises all groups.

Closed-Loop Harness. As training progresses, answer-discriminative groups may become scarce, weakening the utility estimate even when the allocation appears stable. Validation performance can also deteriorate during training. The harness diagnoses both failure modes to determine whether and how training should continue. At round k, let α_k be the current allocation, m_k the validation accuracy, σ its standard error and $m^\star$ the best accuracy from earlier rounds. We record allocation drift $\delta_k = \|\alpha_k - \alpha_{k-1}\|_1$ and the mean number $n_k^{\mathcal{M}}$ of answer-discriminative groups per batch. The allocation target is $\bar{u}_k = \mathrm{normalize}(\mathrm{ReLU}(u_k) + \epsilon)$, with separation $\Delta u_k = \max_s \bar{u}_{k,s} - \min_s \bar{u}_{k,s}$. The decision $\mathcal{D}_k$ applies the fixed regression tolerance ρ_{r}, drift threshold ϵ_δ, group-count threshold $\nu_{\min}$ and separation threshold ϵ_u in the following order,

$$\mathcal{D}_k = \begin{cases} \textsc{rollback} & m_k < m^\star - \max(\rho_{\mathrm{r}}\, m^\star,\, 2\sigma), \\ \textsc{diversify} & \delta_k < \epsilon_\delta \;\wedge\; n_k^{\mathcal{M}} < \nu_{\min}, \\ \textsc{stop} & \delta_k < \epsilon_\delta \;\wedge\; n_k^{\mathcal{M}} \geq \nu_{\min} \;\wedge\; \max(\Delta u_k, \Delta u_{k-1}) < \epsilon_u, \\ \textsc{continue} & \text{otherwise}, \end{cases} \tag{5}$$

Rollback requires a saved checkpoint. We set $\Delta u_0 = +\infty$ and use $\Delta u_k = +\infty$ when the utility estimate is unavailable, disabling utility-based stopping in those rounds. ROLLBACK restores the policy, allocation and rollout temperature of the best validation round. DIVERSIFY raises temperature when allocation drift is low and answer-discriminative groups are scarce. STOP terminates training when current drift is low, sufficient answer-discriminative groups are available and utility separation is small in two consecutive rounds. Evaluation uses the best validation checkpoint.

Policy Objective. The policy update computes A_i from the stage-weighted reward in Eq. 2. Let $o_{i,t}$ be token t of rollout o_i, $o_{i,<t}$ its prefix and θ_{old} the parameters of the sampling policy. The

token-level importance ratio is

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}.$$

With lower and upper clipping tolerances $\varepsilon_{\text{lo}}, \varepsilon_{\text{hi}} > 0$, the clipped policy-loss term averaged over sampled questions and rollout groups is

$$\mathcal{L}_{\text{policy}}(\theta) \;=\; -\mathbb{E}\Big[\tfrac{1}{n}\textstyle\sum_i \tfrac{1}{|o_i|}\sum_t \min\big(\rho_{i,t}A_i,\; \text{clip}(\rho_{i,t}, 1-\varepsilon_{\text{lo}}, 1+\varepsilon_{\text{hi}})A_i\big)\Big], \tag{6}$$

For GRPO, we augment this objective with a KL penalty that constrains deviation from the frozen reference policy. The complete CALIPER training procedure is given in Algorithm 1.

Proposition 1 (Stage-Contrast Preservation) *Fix the policy and the conditional distributions of complete-reference groups. Let $d = |\mathcal{S}|$, $p \in (0,1)$ be the rollout fraction in answer-discriminative groups, D the centered feature covariance in the remaining groups, and c the centered feature–outcome cross-moment in answer-discriminative groups. Assume exact moments, positive feature variance in both group types, and $\lambda, \epsilon > 0$. Let $\mathcal{T}_\epsilon$ denote the target map in Eq. 4 and α_{unif} the uniform allocation. The mixed-group utility u_{M} and its target are independent of p, whereas applying the same estimator to all groups gives a utility u_p satisfying*

$$\big\|\mathcal{T}_\epsilon(u_p) - \alpha_{\text{unif}}\big\|_1 \le \frac{2\sqrt{d}\,p\|c\|_2}{(1-p)\lambda\epsilon\,\text{tr}\,D} \longrightarrow 0 \quad \textit{as } p \downarrow 0. \tag{7}$$

Moreover, a bounded three-stage construction preserves a positive association between normalized advantages and a correctness-associated stage contrast under mixed-group estimation, while this association vanishes under pooled estimation as $p \downarrow 0$. The proof is provided in Appendix C.

5 Experiments

Implementation Details. The primary experiments use the GRPO implementation of EasyR1 (Zheng et al., 2025b) on verl (Sheng et al., 2025). The backbone is frozen throughout and a LoRA adapter (Hu et al., 2022) of rank 64 is trained with $n = 8$ rollouts per question at temperature 1.2, AdamW at learning rate 1×10^{-5}, and 512 questions per optimizer step. Backbones are Qwen3-VL-4B (Bai et al., 2025a), Qwen2.5-VL-7B (Bai et al., 2025b) and InternVL3-8B (Zhu et al., 2025). One-shot training comprises ten optimizer steps, approximately five passes over $1{,}000$ items. Harness training comprises at most six rounds of five steps. Each run uses eight NVIDIA H200 GPUs (141 GB each) on a single node. Unless varied explicitly, the backbones share the reward, estimator, allocation and harness settings specified in Appendix B.

Baselines. We compare two untrained configurations and three reward-based training baselines. Direct prompting requests an answer without a reasoning trace. Chain-of-thought prompting (Wei et al., 2022) uses the same three-stage format as trained methods to isolate explicit reasoning before training. Within each training protocol, all methods share the backbone, adapter and rollout settings. Outcome reward uses final-answer supervision, following GRPO (Shao et al., 2024) and Vision-R1 (Huang et al., 2026). Reference-assisted reward is a controlled variant in which the verifier receives the SpatialTrace reference as grading context. The process reward model (PRM) baseline follows stage-level labeling in Lightman et al. (2024) and Wang et al. (2024b), and combines its three predicted scores under fixed weights. Appendix B.1 details reward scorers, baseline and prompts.

Benchmarks & Metrics. We train each backbone on our SpatialTrace dataset (Section 3) and evaluate it on eight primary benchmarks. Four target 3D and relational spatial understanding, namely 3DSRBench (Ma et al., 2025), RoboSpatial-Home (Song et al., 2025), and spatial question-answering splits built on Omni3D (Brazil et al., 2023) and SAT (Ray et al., 2025). Two are the general multimodal benchmarks MMBench (Liu et al., 2024) and MMStar (Chen et al., 2024b), and the remaining two are the high-resolution real-world benchmark MME-RealWorld-Lite (Zhang et al., 2025) and the synthetic spatial puzzles of SpatialEval (Wang et al., 2024a). Training and evaluation images are disjoint (Appendix A). We report benchmark scores (%) and their unweighted mean (Avg) on fixed 200-item subsets. The metrics include binary accuracy, a mixed numerical score for Omni3D, and a proxy context score within RoboSpatial. Appendix B specifies these differences.

Table 1: Main results (%). **Avg** is the mean, and **1st** counts wins (ties included). **Bold**/underline mark best/second-best scores. Δ compares protocols.

Method	Harness	3DSR	MMB	MME-RW	MMStar	Omni3D	RoboSp	SAT	SpatEval	Avg	Δ	1st
Qwen3-VL-4B (Bai et al., 2025a)												
Direct (No RL)	✗	55.00	89.00	49.00	61.00	**38.90**	43.92	75.00	56.00	58.48	✗	1
CoT (No RL)	✗	59.00	92.00	47.50	67.00	31.85	44.72	73.50	68.50	60.51	✗	0
Outcome Reward	✗	60.50	91.00	46.00	66.50	33.70	46.15	76.00	70.00	61.23	+0.13	0
	✓	58.00	91.00	50.50	64.50	33.65	46.72	76.00	70.50	61.36		0
Reference-Assisted	✗	61.00	92.00	47.50	68.00	33.95	47.13	**77.00**	68.00	61.82	+0.13	1
	✓	61.50	92.00	50.00	65.00	32.55	48.03	75.00	71.50	61.95		0
Process Reward Model	✗	63.00	**93.50**	49.50	68.50	37.20	49.20	74.00	69.50	63.05	−0.15	1
	✓	63.50	92.50	48.00	69.50	37.05	46.11	75.00	71.50	62.90		0
CALIPER (ours)	✗	**66.00**	**93.50**	52.00	69.00	37.05	46.95	76.50	72.50	64.19		2
	✓	65.50	93.00	**52.50**	**70.00**	36.35	**50.37**	72.50	**75.00**	**64.40**	+0.21	**4**
Qwen2.5-VL-7B (Bai et al., 2025b)												
Direct (No RL)	✗	49.50	88.00	44.50	62.00	23.15	43.73	51.50	53.00	51.92	✗	0
CoT (No RL)	✗	51.50	88.50	45.50	65.00	24.65	47.23	54.50	57.00	54.23	✗	0
Outcome Reward	✗	55.00	89.00	51.00	62.50	22.75	44.05	58.50	54.50	54.66	+0.98	0
	✓	**55.50**	89.50	49.50	**65.50**	24.05	43.06	**62.50**	55.50	55.64		3
Reference-Assisted	✗	53.50	89.50	**52.00**	**65.50**	20.45	44.10	57.50	56.50	54.88	+0.11	2
	✓	53.00	88.50	50.00	63.50	22.45	43.47	62.00	57.00	54.99		0
Process Reward Model	✗	52.50	85.50	47.00	59.50	**24.75**	45.25	53.50	54.50	52.81	−0.33	1
	✓	48.50	85.50	44.50	60.00	24.00	45.87	56.00	55.50	52.48		0
CALIPER (ours)	✗	54.50	**90.50**	50.50	63.50	23.05	46.73	57.00	62.00	55.97		1
	✓	53.50	**90.50**	49.50	**65.50**	**24.75**	**48.27**	55.50	**62.50**	**56.25**	+0.28	**5**
InternVL3-8B (Zhu et al., 2025)												
Direct (No RL)	✗	51.00	**89.00**	54.00	71.00	36.95	39.90	74.50	50.00	58.29	✗	1
CoT (No RL)	✗	56.50	88.50	53.00	**73.00**	38.65	44.32	72.50	50.50	59.62	✗	1
Outcome Reward	✗	55.50	88.00	53.00	67.50	37.40	41.63	**75.00**	52.50	58.82	+0.71	1
	✓	55.50	87.50	54.50	66.50	37.00	43.75	74.00	57.50	59.53		0
Reference-Assisted	✗	56.50	88.00	52.50	71.50	37.90	44.28	74.00	54.00	59.84	+0.45	0
	✓	55.00	88.00	54.00	68.50	38.10	47.19	74.00	57.50	60.29		0
Process Reward Model	✗	57.50	88.50	**56.50**	70.00	38.15	49.08	73.50	56.00	61.15	−0.67	1
	✓	57.00	**89.00**	52.50	70.00	38.35	50.04	72.50	54.50	60.49		1
CALIPER (ours)	✗	**58.00**	**89.00**	55.00	72.50	**41.65**	46.10	**75.00**	57.50	61.84		**4**
	✓	57.50	**89.00**	55.50	71.00	40.25	**52.14**	**75.00**	**60.00**	**62.55**	+0.71	**4**

5.1 Main Results

Performance Comparison. Table 1 shows that CALIPER achieves the highest mean score across the eight benchmarks for all three backbones under both training protocols. Under harness training, its margin over the strongest trained baseline reaches 2.06 percentage points on InternVL3-8B. PRM improves on outcome reward for two backbones but falls behind on Qwen2.5-VL-7B, indicating that intermediate supervision alone is insufficient. The consistent advantage over frozen allocation supports adapting stage weights to the current policy.

OOD Generalization. Table 2 shows CALIPER's strongest relative gain on viewpoint-dependent spatial reasoning. Its mean exceeds frozen allocation by 3.00 percentage points, with the largest individual gain on ViewSpatial-Bench. The evaluation covers BLINK-Spatial (Fu et al., 2024), RealWorldQA (xAI, 2024), EmbSpatial-Bench (Du et al., 2024) and ViewSpatial-Bench (Li et al., 2025) under fixed prompts and decoding after SpatialTrace training. Harness PRM uses the separate harness protocol. The pattern is consistent with better transfer of stage-specific supervision to viewpoint reasoning while retaining performance on the other OOD tasks.

Table 2: Qwen3-VL-4B OOD accuracy (%).

Method	BLINK	RealWorld QA	Emb Spatial	View Spatial	Avg
Direct (no RL)	80.50	73.00	76.50	37.50	66.88
CoT (no RL)	83.00	75.00	69.50	43.00	67.63
Outcome Reward	84.00	71.50	72.50	41.50	67.38
Reference-Assisted	85.00	71.50	73.00	43.00	68.13
Process Reward Model	80.50	72.00	72.00	45.00	67.38
Frozen α	83.00	74.50	73.00	41.50	68.00
CALIPER (ours)	85.50	76.00	76.50	46.00	71.00
Harness PRM	81.50	76.00	75.00	45.50	69.50

5.2 Framework Analysis

Generality Across Optimizers. Table 3 shows that adaptive allocation improves performance across distinct policy optimizers. Under both DAPO (Yu et al., 2025) and GSPO (Zheng et al., 2025a), CALIPER outperforms frozen allocation, with gains of up to 1.92 percentage points despite their different update rules. On Qwen2.5-VL-7B, outcome reward remains stronger than process-based variants, indicating that adaptive allocation improves process supervision without requiring process rewards to universally dominate outcome rewards. These comparisons support utility-based allocation across the tested policy objectives.

Table 3: Optimizer transfer under harness training, with Δ relative to CALIPER.

Optim.	Method	Q3-VL-4B		Q2.5-VL-7B		IVL3-8B	
		Avg	Δ	Avg	Δ	Avg	Δ
GRPO	Outcome	61.36	-3.04	55.64	-0.61	59.53	-3.02
	Ref.-assisted	61.95	-2.45	54.99	-1.26	60.29	-2.26
	Frozen α	62.47	-1.93	55.22	-1.03	60.96	-1.59
	CALIPER	64.40	—	56.25	—	62.55	—
DAPO	Outcome	63.31	−0.04	56.31	+2.34	61.34	−1.13
	Ref.-assisted	63.36	+0.01	56.26	+2.29	61.21	−1.27
	Frozen α	62.56	−0.79	53.38	−0.59	61.09	−1.39
	CALIPER	63.35	—	53.97	—	62.48	—
GSPO	Outcome	61.43	−1.71	55.46	+0.78	59.97	−0.61
	Ref.-assisted	62.95	−0.20	53.91	−0.76	60.17	−0.42
	Frozen α	62.78	−0.37	52.75	−1.92	60.44	−0.14
	CALIPER	63.15	—	54.67	—	60.58	—

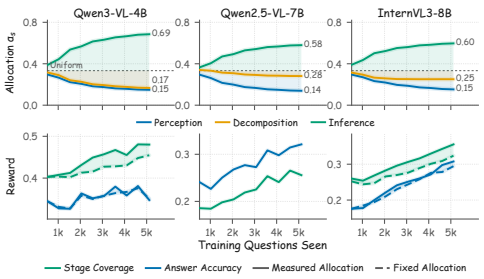


Figure 4: Stage allocation (top) and rewards (bottom) over optimizer steps. Dashed curves use fixed allocation.

Reward Allocation Dynamics. Fig. 4 shows a consistent reallocation toward inference across backbones. Starting from uniform weights, inference accounts for 0.58–0.69 of the final process-reward allocation. The lower panels further reveal that higher stage coverage does not uniformly translate into higher answer accuracy: coverage increases on Qwen3-VL-4B despite fluctuating accuracy, whereas the other two backbones exhibit more aligned improvements in both. The fixed-allocation baselines also follow backbone-specific trajectories on Qwen3-VL-4B and InternVL3-8B. These trajectories support using within-question correctness contrasts to estimate stage utility.

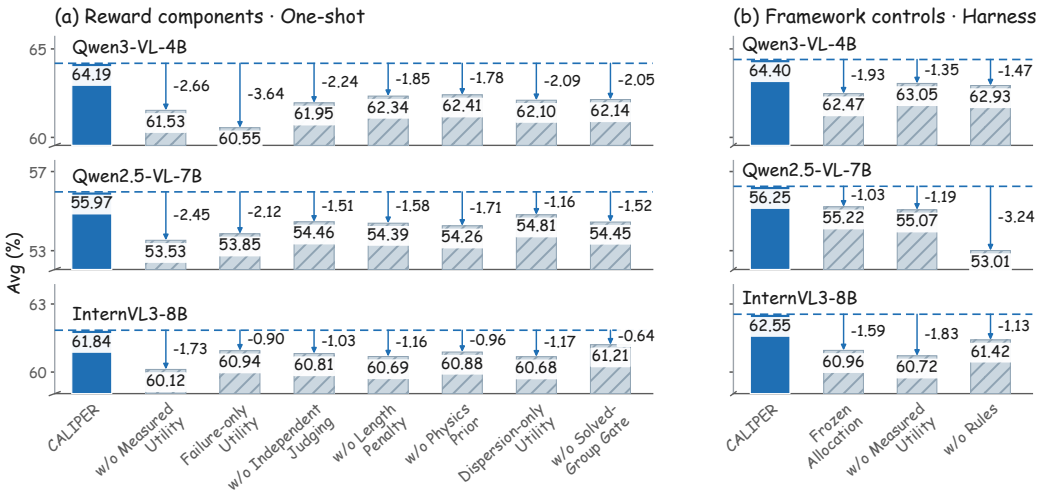


Figure 5: Reward-component (a) and framework (b) ablations. Gaps are in percentage points.

Ablation Studies. Fig. 5 shows consistent contributions from both adaptive reward allocation and harness control. Removing measured utility degrades all three backbones, with failure-only utility causing the largest drop on Qwen3-VL-4B (3.64 points), while freezing the learned allocation also consistently reduces performance. Removing harness rules produces further degradation across backbones, reaching 3.24 points on Qwen2.5-VL-7B. Together, these results show that CALIPER benefits from both policy-dependent reward adaptation and closed-loop training control.

Hyperparameter Sensitivity. Fig. 6 shows that the shared default gives the highest mean on all three backbones. Group-size effects differ by backbone: Qwen3-VL-4B and InternVL3-8B recover at $n = 16$ after a dip, whereas Qwen2.5-VL-7B declines beyond $n = 8$. Non-default step size, temperature and decay also reduce performance, consistent with a balance between diverse rollouts and stable allocation updates. Increasing the ridge penalty from 0.01 to 0.1 improves the mean on all three backbones.

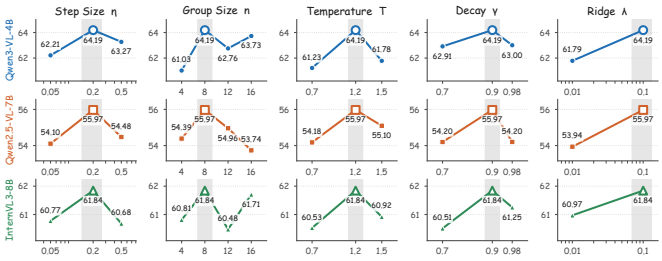


Figure 6: Hyperparameter sensitivity.

Case Study. Fig. 7 shows two failures to translate recognized objects into spatial answers. CALIPER resolves the camera-depth relation in the first item and distinguishes possible placement from current location in the second. Both cases illustrate gains in inference or decomposition despite similar object recognition. Appendix E provides cross-backbone examples and a regression.

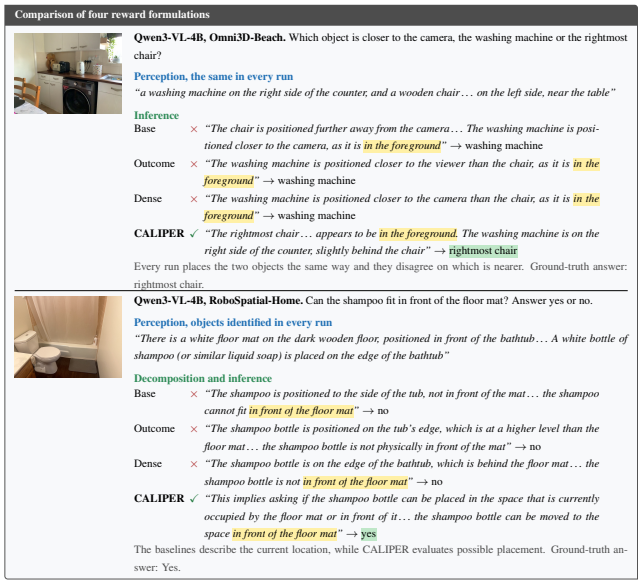


Figure 7: Qualitative effects of adaptive stage supervision.

Scaling Behaviour. Fig. 8 shows diminishing gains with model capacity, unique training items, optimizer steps and complete-trace coverage. The spatial group gains more than the general and real-world group across these settings. The fixed-model panels vary data exposure, optimization and annotation coverage, while the model-size panel compares members of the InternVL3 family. Appendix D.2 gives the corresponding settings.

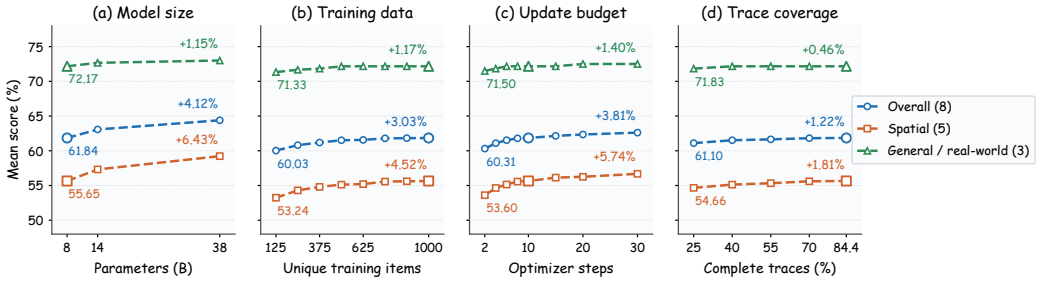


Figure 8: Scaling with model size, training data, update budget and trace coverage.

6 Conclusion

This paper proposes CALIPER, an adaptive process-reward framework that allocates spatial-reasoning supervision according to policy-dependent stage utility. CALIPER estimates utility from answer-discriminative rollout groups and applies the resulting stage weights to all groups, including those with uniformly incorrect answers. A closed-loop harness monitors estimator reliability and validation performance to guide training decisions. Extensive experiments across eight benchmarks, three backbones and three optimizers show gains over matched fixed weights.

AI USE STATEMENT

In this work, we used AI tools for manuscript preparation, including language editing, selected icons in Fig. 1, and code revision assistance. We also used a vision-language model to synthesize the initial SpatialTrace reasoning traces, which were subsequently refined by a text-only model. The authors take responsibility for the final manuscript, figures, dataset and code, including AI-assisted content. These tools did not independently determine the methods or technical conclusions.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A describe SpatialTrace sources, sampling and reference annotation. Section 4 specifies the reward, utility estimator and allocation update, with pseudocode in Algorithm 1. Appendix B records training protocols, hyperparameters, evaluation metrics and reward-scoring prompts. Appendix C provides the assumptions and proof of Proposition 1, while Appendix D details experimental controls and scaling settings.

REFERENCES

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

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

Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13154–13164. IEEE, 2023.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14455–14465, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Jie Cheng, Gang Xiong, Ruixi Qiao, Lijun Li, Chao Guo, Junle Wang, Yisheng Lv, and Fei-Yue Wang. Stop summation: Min-form credit assignment is all process reward model needs for reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards, 2025.

DeepSeek-AI, Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chengyu Hou, Chenhao Xu, Chenze Shao, Chong Ruan, Conner Sun, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Donghao Li, Dongjie Ji, Erhang Li, Fang Wei, Fangyun Lin, Fangzhou Yuan, Feiyu Xia, Fucong Dai, Guangbo Hao, Guanting Chen, Guoai Cao, Guolai Meng, Guowei Li, Han Yu, Han Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoling Zhang, Haoming Luo, Haoran Wei, Haotian Yuan, Haowei Zhang, Haowen Luo, Haoyu Chen, Haozhe Ji, Hengqing Zhang, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, J Yang, JQ Zhu, Jia Luo, Jia Song, Jia Yu, Jialiang Huang, Jialu Cai, Jian Liang, Jiangting Zhou, Jiasheng Ye, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jieyu Yang, Jin Chen, Jin Yan, Jingchang Chen, Jingli Zhou, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jingzi Zhou, Jinhua Zhu, Jiping Yu, Joseph Sun, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junmin Zheng, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Leyi Xia, Li Zhang, Liang Zhao, Lihua Guo, Lingxiao Luo, Linwang Ma, Linyan Zhu, Litong Wang, Liyu Cai, Liyue Zhang, Longhao Chen, MS Di, MY Xu, Max Mei, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Mingxu Zhou, Minmin Han, Ning Wang, Panpan Huang, Panpan Wang, Peixin Cong, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Qiwei Jiang, Rui Tian, Ruifan Xu, Ruijie Lu, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqian Chen, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, Ruyi Chen, SH Liu, Shanghao Lu, Shangmian Sun, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoheng Nie, Shaoqing Wu, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Shuying Yu, Songyang Zhou, Tao Ni, Tao Yun, Tian Jin, Tian Pei, Tian Ye, Tianle Lin, Tianran Ji, Tianyi Cui, Tianyuan Yue, Tingting Yu, Tun Wang, W Zhang, WL Xiao, Wangding Zeng, Wei An, Weilin Zhao, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjing Yao, Wenjun Gao, Wenkai Yang, Wenlve Huang, Wenqing Hou, Wentao Zhang, Wenting Ma, Xi Gao, Xiang He, Xiangwen Wang, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingchen Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyu Zhang, Xu Chen, Xuanyu Wang, Xuecheng Su, Xueyin Chen, Xuheng Lin, Xuwei Fu, YC Yan, YQ Wang, YW Ma, Yanfeng Luo, Yang Zhang, Yanhong Xu, Yanru Ma, Yanwen Huang, Yao Li, Yao Li, Yao Xu, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Shao, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yijia Wu, Yiliang Xiong, Yiling Ma, Ying He, Ying Tang, Ying Zhou, Yingjia Luo, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiang Zhang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, YuKun Li, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuanhao Li, Yuduan Wang, Yuehan Yang, Yuer Xu, Yuhan Wu, Yuhao Meng, Yuheng Zou, Yukun Zha, Yunfan Xiong, Yupeng Chen, Yuping Lin, Yuqian Cao, Yuqian Wang, Yushun Zhang, Yuting Yan, Yutong Lin, Yuxian Gu, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuxuan Zhou, Yuyang Zhou, Yuzhen Huang, ZF Wu, Zehao Wang, Zehua Zhao, Zehui Ren, Zekai Zhang, Zhangli Sha, Zhe Fu, Zhe Ju, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zheren Gao, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhixuan Chen, Zhiyu Wu, Zhizhou Ren, Zhongyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihua Qu, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Ziyi Wan, Zizheng Pan, and Zongqing Yao. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models, 2024. URL https://arxiv.org/abs/2406.05756.

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, 2024. URL https://arxiv.org/abs/2404.12390.

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

Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Zhe Xu, Yao Hu, Shaohui Lin, et al. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 63794–63812, 2026.

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*, pp. 611–626, 2023.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models, 2025. URL https://arxiv.org/abs/2505.21500.

Zhenyi Liao, Qingsong Xie, Yanhao Zhang, Zijian Kong, Haonan Lu, Zhenyu Yang, and Zhijie Deng. Improved visual-spatial reasoning via r1-zero-like training. *arXiv preprint arXiv:2504.00883*, 2025.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. URL https://arxiv.org/abs/2307.06281.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso De Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934. IEEE, 2025.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. Sat: Dynamic spatial aptitude training for multimodal language models, 2025. URL https://arxiv.org/abs/2412.07755.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

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

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*, pp. 1279–1297, 2025.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15768–15780. IEEE, 2025.

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

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024a.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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

xAI. RealWorldQA. Dataset, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, et al. Seephys: Does seeing help thinking?–benchmarking vision-based physics reasoning. *Advances in Neural Information Processing Systems*, 38, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2a36EMSSTp.

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025a. URL https://arxiv.org/abs/2507.18071.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, Yuwen Xiong, and Richong Zhang. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025b.

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

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

Roadmap Appendix A details SpatialTrace construction. Appendix B specifies implementation and evaluation protocols. Appendix C proves Proposition 1. Appendix D provides detailed experiments and scaling settings, and Appendix E presents additional qualitative examples.

A Dataset Construction Details

Sources and Sampling. SpatialTrace combines ten single-image question sources grouped by reasoning demand in Table 4. The 3D family comprises 3DSRBench (Ma et al., 2025), Omni3D-Bench (Brazil et al., 2023) and SAT-Spatial-VQA-5k (Ray et al., 2025). General multimodal questions come from MMBench (Liu et al., 2024), MME-RealWorld-Lite (Zhang et al., 2025) and MMStar (Chen et al., 2024b). Region-level spatial questions come from RoboSpatial-Home (Song et al., 2025), SpatialEval (Wang et al., 2024a) and SpatialRGPT-Bench (Cheng et al., 2024), while SeePhys (Xiang et al., 2026) adds physical reasoning. Records with empty question or answer fields or Chinese text were removed to match the English training prompt. Fixed quotas assign 310 items to each core family and 250 to physical reasoning. Within each core family, the largest-remainder method allocates items in proportion to filtered source sizes. Sampling uses seed 42, followed by a source- and family-stratified split of 1,000 training and 180 validation items. Each 200-item evaluation subset in Section 5 excludes every image in these splits, so SpatialTrace shares source benchmarks with the evaluation but no evaluation images.

Table 4: SpatialTrace source families and reasoning demands.

Family	Source datasets	Items	Primary reasoning demand
3D metric	3DSRBench, Omni3D-Bench, SAT-Spatial-VQA-5k	310	Depth, orientation and metric relations
General multimodal	MMBench, MME-RealWorld-Lite, MMStar	310	Object grounding and open-domain relations
Region-level spatial	RoboSpatial-Home, SpatialEval, SpatialRGPT-Bench	310	Placement and local spatial relations
Physical reasoning	SeePhys	250	Physical quantities and plausibility
Total	Ten source datasets	1,180	Single-image questions with verifiable answers

Reference Annotation and Coverage. Qwen2.5-VL-32B (Bai et al., 2025b), served locally through Ollama, generates visual observations, question decomposition and inference from each image, question and reference answer. The annotation prompt requests verifiable evidence and explicit acknowledgment of ambiguity. Decomposition identifies detection, counting and spatial subproblems, while inference connects observations to an answer. A text-only editing pass preserves visual observations, clarifies inferences and flags unsupported estimates or evidence–label conflicts. The three resulting fields form the stage references in Eq. 2. Reflection and verification record annotation concerns without adding reward stages. Complete three-stage records cover 844 training and 152 validation items. The remaining 156 training and 28 validation items retain answer supervision and available reasoning. Policy training uses all 1,000 training items, while utility estimation uses records with complete stage references.

Human Validation. Three unpaid members of the research group independently assess 100 randomly sampled SpatialTrace items using the rubric in Table 5. They judge question validity and answer alignment from the image, question and reference answer before reading the stage references. Items without complete stage references are rated only on available stages. Table 6 reports positive rater-item judgments and Krippendorff's nominal α on the original response categories, which counts disagreement between adjacent ordinal levels as complete disagreement. Positive denotes a valid question, faithful perception, adequate decomposition, inference rated at least four, or agreement with the recorded answer alignment, respectively. Agreement is weakest for decomposition, where judging the completeness of a subproblem set requires the most interpretation.

Table 5: Rubric used by the SpatialTrace raters.

QUESTION VALIDITY — item, binary

"Given the image and the question, can an attentive reader determine a single answer from visible evidence?"

POSITIVE The image contains the objects and relations the question asks about, and one answer follows from them.

FAILURE The evidence is occluded or outside the frame, the reference frame is unspecified, or several answers remain defensible.

PERCEPTION FIDELITY — perception, three-point ordinal

"Does each observation report objects, attributes or positions that are visible in the image and needed for the question?"

POSITIVE *Faithful* reports only visible content and covers the evidence the answer depends on.

FAILURE *Partial* is correct but omits one needed object or relation. *Unfaithful* describes absent objects, misreads a position or orientation, or states the answer in place of evidence.

DECOMPOSITION ADEQUACY — decomposition, three-point ordinal

"Would answering these subquestions in order suffice to answer the original question?"

POSITIVE *Adequate* covers every object, quantity and reference frame the answer depends on.

FAILURE *Partial* misses one of them. *Inadequate* restates the question, targets irrelevant objects, or answers the subquestions itself.

INFERENCE SOUNDNESS — inference, five-point Likert

"How well does the inference derive the reference answer from the recorded observations?"

POSITIVE *5* every step follows from the observations and reaches the reference answer, *4* it is sound apart from one minor gap.

FAILURE *3* it reaches the answer through one weakly supported step, *2* it relies on facts absent from the perception stage or on unjustified numerical precision, *1* it contradicts the observations or reaches a different answer.

ALIGNMENT AGREEMENT — item, binary

"Judging from the image, is the reference answer supported at its stated precision?"

POSITIVE The judgment, made before the recorded answer alignment is shown, matches that record.

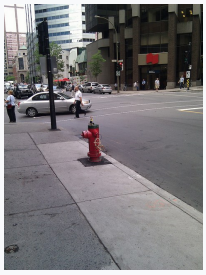
FAILURE The rater finds an unsupported label or unjustified precision that the record accepts, or rejects a recorded concern.

Table 6: Human assessment of SpatialTrace references.

Criterion	Items	Positive ratings	Positive (%)	α
Question validity	100	269/300	89.7	0.785
Perception fidelity	95	226/285	79.3	0.715
Decomposition adequacy	94	217/282	77.0	0.640
Inference soundness	95	212/285	74.4	0.808
Alignment agreement	100	255/300	85.0	0.765

Reference Traces. The reference examples distinguish visual evidence, decomposition and inference. The Omni3D example separates visible evidence from unsupported numerical precision, illustrating the complementary roles of reference coverage and physical plausibility.

3DSRBench, 3D metric reasoning



Question. Consider the real-world 3D locations and orientations of the objects. If I stand at the grey car's position facing where it is facing, is the fire hydrant on the left or right of me?

Ground-truth answer. B. on the right

Perception

- A grey car is parked on the left side of the image, facing towards the right.
- A red fire hydrant is located on the sidewalk near the intersection, closer to the right side of the image.

Decomposition

- What is the position and orientation of the grey car?
- Where is the fire hydrant located relative to the car?
- If standing at the car's position and facing its direction, which side is the fire hydrant on?

Inference

- Step 1: The grey car is observed to be parked on the left side of the image and is facing towards the right. This establishes its orientation as directed to the right.
- Step 2: The fire hydrant is located on the sidewalk near the intersection and is closer to the right side of the image. This places it spatially to the right of the car's orientation.
- Step 3: If a person stands at the grey car's position and faces the direction the car is facing (towards the right), the fire hydrant, being positioned to the right of the car, would also be on the right side of the person.

Omni3D-Bench, 3D metric reasoning



Question. If the combined height of the TV and TV stand is 2.7m, how tall is the fireplace?

Ground-truth answer. 2.989

Perception

- The TV is mounted on a stand, and both are visible in the image.
- The fireplace is visible on the right side of the image.
- The TV and TV stand are positioned at the same height as the fireplace.
- The height of the TV and stand is given as 2.7 meters.
- The fireplace has a visible height that extends above the TV and stand.

Decomposition

- What is the height of the TV and stand? (Given as 2.7 meters)
- How much taller is the fireplace compared to the TV and stand?
- What is the total height of the fireplace?

Inference

- Step 1: The height of the TV and stand is explicitly given as 2.7 meters.
- Step 2: The fireplace is visually observed to extend above the top of the TV and stand. Based on the image, the additional height of the fireplace above the TV and stand appears to be slightly less than 0.3 meters.
- Step 3: Using a coarse visual estimate, the additional height of the fireplace is approximated as 0.289 meters. This estimate is based on proportional comparison in the image, but no exact measurement tools are available.
- Step 4: Adding the height of the TV and stand (2.7 meters) to the estimated additional height of the fireplace (0.289 meters) gives a total height of 2.989 meters for the fireplace.

Answer alignment, recorded as not supported

- The stated precision of 2.989 meters exceeds what can be justified from the available visual evidence. The answer should be treated as a plausible estimate rather than an exact measurement.

B Implementation Details

Framework and Optimization. EasyR1 implements GRPO on verl, with vLLM (Kwon et al., 2023) for rollout generation and FSDP (Zhao et al., 2023) for policy updates. Rank-64 LoRA adapters with scaling 64 update the language model's linear layers. The vision tower remains frozen. AdamW uses bfloat16, weight decay 10^{-2}, gradient-norm clipping at 1.0 and linear warmup over the first 5% of training. Gradient checkpointing, unpadded packing and on-device states reduce memory overhead.

Rollouts and Batching. Each optimizer step processes 512 training questions in four gradient updates of 128. Rollouts use top-$p = 1.0$, and prompts and responses are each limited to 3072 tokens. Overlength prompts are excluded, and images are rescaled to 2.6×10^5–4.2×10^6 pixels. GRPO uses a KL penalty of 10^{-2} against the frozen reference policy. DAPO and GSPO disable this

Algorithm 1 CALIPER. Each harness round comprises H steps of stage-adaptive GRPO.

Require: SpatialTrace $\mathcal{D}$ with reference traces $\tau^\star$, policy π_θ, stages $\mathcal{S}$
Settings: rounds K, steps per round H, patience P, step size η, offset ϵ, decay γ, gate $N_{\min}$
Return: the policy of the best-validating round

$\alpha_0 \leftarrow$ uniform, statistics and $N_{\text{eff}} \leftarrow 0$, $m^\star \leftarrow -\infty$, $\Delta u_0 \leftarrow +\infty$
$\alpha \leftarrow \alpha_0$
for round $k = 1, \ldots, K$ **do**
  Inner loop: stage-adaptive GRPO
  for step $t = 1, \ldots, H$ **do**
    draw a batch from $\mathcal{D}$ and sample n rollouts $\{o_i\}$ per question from π_θ
    $\text{cov}_s(o_i) \leftarrow J(o_i, \tau_s^\star)$ for each $s \in \mathcal{S}$ ▷ stages judged independently
    score every rollout by Eq. 2 under the allocation α of the previous update
    form A_i by Eq. 1 and update θ by Eq. 6
    $\mathcal{M} \leftarrow \{g : 0 < \sum_{i \in g} \text{acc}_i < n\}$ ▷ answer-discriminative groups
    decay statistics and N_{eff} by γ, then add complete-trace samples from $\mathcal{M}$
    if $N_{\text{eff}} \geq N_{\min}$ and feature variation is nonzero **then**
      $u \leftarrow$ Eq. 3
      if $\max_s u_s > 0$ **then**
        $\alpha \leftarrow$ Eq. 4 ▷ otherwise hold the allocation
      end if
    end if
  end for
  Outer loop: closed-loop harness
  $m_k \leftarrow$ validation accuracy of θ, and $\mathcal{D}_k \leftarrow$ Eq. 5
  if $m_k > m^\star$ **then**
    $m^\star \leftarrow m_k$ and retain the current checkpoint
  end if
  if $\mathcal{D}_k =$ ROLLBACK **then**
    restore the policy, allocation and rollout temperature of the best round
  else if $\mathcal{D}_k =$ DIVERSIFY **then**
    raise the rollout temperature
  else if $\mathcal{D}_k =$ STOP **then**
    break ▷ stages indistinguishable in utility
  end if
  if $m^\star$ has not improved for P rounds **then**
    break
  end if
end for

penalty and use the optimizer-specific settings in Table 7. Data and reward definitions remain fixed across optimizers.

Protocols. One-shot training evaluates the final checkpoint after ten optimizer steps. Harness training uses at most six rounds of five steps, with early stopping after two rounds without validation improvement. Each round is evaluated on the 180 validation items using greedy decoding, and the best validation checkpoint is selected for benchmark evaluation. Eq. 5 combines validation performance with allocation drift, group diversity and utility separation. Rollback requires a decline exceeding both the relative tolerance and twice the binomial standard error. Checkpoints retain the policy, allocation and rollout temperature.

Evaluation. Adapters are merged into the backbone for greedy decoding with a 2048-token response budget. All methods use the same three-stage prompt except direct prompting, which retains the answer format but omits reasoning. Scoring is deterministic and task-specific. Multiple-choice items use option matching. Omni3D combines exact or categorical matching with mean relative accuracy over ten error thresholds for numerical answers. RoboSpatial-Home averages compatibility, configuration and context scores. Its context implementation uses the convex hull of reference points with a nearest-point fallback instead of the official mask. The reported mean weights the eight benchmark scores equally, preserving their task-specific scales rather than pooling them as binary outcomes.

Table 7: Default hyperparameters and optimizer variants. AdamW uses bfloat16. Clipping values specify lower / upper thresholds.

Hyperparameter	Symbol	Value
Policy optimization		
LoRA rank / scaling		64 / 64
Rollouts per question	n	8
Rollout temperature	T	1.2
Questions / optimizer step		512
Gradient updates / step		4
AdamW learning rate		1×10^{-5}
Maximum response length		3072
Process reward (Eq. 2)		
Outcome weight	w_a	0.60
Process weight	w_m	0.25
Physics weight	w_p	0.08
Length weight	w_ℓ	0.10
Length target	L_0	1500
Length cap	$L_{\max}$	3072
Utility estimator (Eq. 3)		
Ridge penalty / mean diagonal	λ	0.1
Statistics decay	γ	0.9
Identifiability threshold	$N_{\min}$	24

Hyperparameter	Symbol	Value
Allocation update (Eq. 4)		
Step size	η	0.2
Additive target offset	ϵ	0.15
Initial allocation	α_0	uniform
Closed-loop harness (Eq. 5)		
Steps per round		5
Maximum rounds / patience		6 / 2
Regression tolerance	ρ_{r}	0.02
Allocation drift threshold	ϵ_δ	0.02
Answer-discriminative group threshold	$\nu_{\min}$	1
Flat-utility threshold	ϵ_u	0.05
Temperature step	Δ_T	0.15
Temperature cap	$T_{\max}$	1.6
Round-selection items		180 items
Optimizer variants (Table 3)		
DAPO clipping range		0.2 / 0.28
DAPO online filtering		enabled
GSPO clipping range		3×10^{-4} / 4×10^{-4}
GSPO importance ratio		sequence level
DAPO and GSPO KL penalty		disabled

B.1 Reward Scoring and Baseline Details

Reward Computation. The policy generates three numbered reasoning stages followed by an answer. DeepSeek-V4-Flash (DeepSeek-AI et al., 2026), at temperature 0, scores all three stages in one call by comparing each stage with its corresponding reference. Independent grading prevents evidence from one stage from inflating another stage's score. For incomplete references, the scorer excludes missing stages and renormalizes the remaining allocation. The same judge evaluates free-form answers if option matching is inapplicable. The physics term checks units, magnitudes and plausibility in the answer region and is active only if at least one check applies. The length penalty increases linearly from zero at 1500 tokens to one at the 3072-token cap, scaling the weighted reward as in Eq. 2.

Process Reward Model Scorer. The PRM baseline combines a LoRA adapter with a three-output regression head on the policy backbone. It is fitted to stage labels obtained by scoring the untrained policy's rollouts against SpatialTrace references. During policy training, its predictions replace reference-based stage scores in Eq. 2 and receive equal weights. CALIPER additionally accesses reference traces at scoring time. The frozen-allocation control preserves that access and the judge, isolating the contribution of weight adaptation.

Reward Prompts. The judge computes the stage-coverage terms in Eq. 2 from one call per rollout. Inputs contain up to 500 question characters, 6000 reasoning characters and ten reference steps per stage. Stage labels distinguish visual observations, question decomposition and reasoning. The split clause indicates a multi-stage reference, and the strict clause requires evidence to match its designated stage. The loose-judging ablation replaces only that requirement, allowing a rollout step to match any reference stage.

Stage-Matching Judge

You are checking whether a model's reasoning covers the steps of a reference solution.{split_clause}

Question: {question}

{stage_blocks}

Model's reasoning:
{model_reasoning}

For EACH stage, count how many of its numbered reference steps have a semantically equivalent counterpart in the model's reasoning (same idea, not necessarily same words).{strict_clause}

Return ONLY a JSON object: {"visual_covered": <int>, "decomp_covered": <int>, "reasoning_covered": <int>}

Independence Instruction, Default and Loose Variant

strict: Judge each stage independently: count a reference step as covered ONLY if the model's reasoning explicitly states its content, not merely because content from another stage implies it.

loose: A step in the model's reasoning may cover steps from any stage.

Baseline Access and Evaluation Prompts. The outcome-only judge receives the question, reference answer and predicted answer. Reference-assisted grading adds up to 800 reference-trace characters. The PRM receives up to 400 question characters and 2500 reasoning characters without the reference trace. The frozen-allocation control holds reference access fixed to isolate weight adaptation. Trained methods and untrained chain-of-thought prompting share the three-stage response format. Direct prompting omits the reasoning request while retaining the answer format.

C Theoretical Analysis

Assumptions and Notation. Fix a policy and the conditional distributions of complete-reference rollout groups. Let $d = |\mathcal{S}| \geq 2$, $x \in [0,1]^d$ collect stage-coverage scores, and $y \in \{0,1\}$ denote answer correctness. Within each group g, define $\tilde{x} = x - \bar{x}_g$ and $\tilde{y} = y - \bar{y}_g$ using its rollout means. Let $\mathcal{M}$ denote answer-discriminative groups and let $\mathcal{M}^c$ denote the remaining, outcome-degenerate groups, whose answers are either all correct or all wrong. The fraction of rollout mass in $\mathcal{M}$ is $p \in (0,1]$. Changing p leaves both conditional distributions fixed. Define the conditional moments

$$\Sigma = \mathbb{E}[\tilde{x}\tilde{x}^\top \mid \mathcal{M}], \qquad D = \mathbb{E}[\tilde{x}\tilde{x}^\top \mid \mathcal{M}^c], \qquad c = \mathbb{E}[\tilde{x}\tilde{y} \mid \mathcal{M}]. \tag{8}$$

These are exact expected moments or exact weighted empirical moments. Assume $\operatorname{tr}\Sigma > 0$, $\operatorname{tr} D > 0$, and fixed $\lambda, \epsilon > 0$. Full rank is unnecessary. For a covariance matrix H, define

$$R(H) = H + \frac{\lambda}{d}\operatorname{tr}(H) I_d, \qquad \mathcal{T}_\epsilon(u) = \frac{[u]_+ + \epsilon \mathbf{1}}{\|[u]_+\|_1 + d\epsilon}, \qquad \alpha_{\text{unif}} = \frac{\mathbf{1}}{d}, \tag{9}$$

where I_d is the identity, $\mathbf{1}$ is the all-ones vector, and $[u]_+$ is the componentwise positive part. Thus R uses the trace-scaled penalty in Eq. 3, and $\mathcal{T}_\epsilon$ is the target before interpolation in Eq. 4.

Formal Statement. Under these assumptions, the mixed-group and pooled estimators satisfy

$$u_{\text{M}} = R(\Sigma)^{-1} c, \qquad u_p = R\big(p\Sigma + (1-p)D\big)^{-1} pc. \tag{10}$$

The utility u_{M} and target $\mathcal{T}_\epsilon(u_{\text{M}})$ are independent of p. For $0 < p < 1$, the pooled target satisfies Eq. 7 and converges to α_{unif} as $p \downarrow 0$. Consequently, the limiting targets differ whenever the mixed-group target is nonuniform. This assertion concerns the computed target, before the algorithm's update gates.

For the normalized-advantage claim, take $d = 3$, $v = (1, -1, 0)^\top$, and any $a, b > 0$ such that $a + b \le 1/2$. Each group contains the four equally represented sign pairs $(z, \xi) \in \{-1, 1\}^2$, repeated m times for any integer $m \ge 1$, so its size is $n = 4m$. Define

$$x = \tfrac{1}{2}\mathbf{1} + azv + b\xi\mathbf{1}, \qquad y = \begin{cases} (z+1)/2 & \text{in mixed groups,} \\ 0 & \text{in outcome-degenerate groups.} \end{cases} \tag{11}$$

The conditional feature distributions are identical, while only mixed groups associate the contrast z with correctness. Let $\eta \in (0, 1]$, initialize both allocations at α_{unif}, and apply $j \ge 1$ eligible updates with fixed moments. Eligibility means that the effective-sample and feature-variation gates admit every update. Define

$$\begin{aligned} r_\lambda &= \frac{\lambda}{3}(2a^2 + 3b^2), \qquad h = \frac{a}{2(2a^2 + r_\lambda)}, \quad \kappa_j = 1 - (1-\eta)^j, \\ \delta_{\mathrm{M},j} &= \kappa_j \frac{h}{h + 3\epsilon}, \qquad \delta_{p,j} = \kappa_j \frac{ph}{ph + 3\epsilon}. \end{aligned} \tag{12}$$

On an all-wrong group from this family, assume a common physics score, a common positive length multiplier, and $w_m > 0$. The advantages in Eq. 1, using the group standard deviation with divisor n, are

$$A_{\mathrm{M},j} = \frac{a\delta_{\mathrm{M},j} z + b\xi}{\sqrt{a^2\delta_{\mathrm{M},j}^2 + b^2}}, \qquad A_{p,j} = \frac{a\delta_{p,j} z + b\xi}{\sqrt{a^2\delta_{p,j}^2 + b^2}}. \tag{13}$$

For every $0 < p < 1$, their uniform within-group averages obey

$$0 < \mathbb{E}[A_{p,j} z] < \mathbb{E}[A_{\mathrm{M},j} z] = \frac{a\delta_{\mathrm{M},j}}{\sqrt{a^2\delta_{\mathrm{M},j}^2 + b^2}}, \qquad \lim_{p \downarrow 0} \mathbb{E}[A_{p,j} z] = 0. \tag{14}$$

The mixed-group association remains positive and independent of p. Here z identifies the same stage-coverage contrast across both group types. It predicts correctness in mixed groups, while all evaluation-group answers in this construction are wrong.

Proof. Within any outcome-degenerate group, $y = \bar{y}_g$ for every rollout. Hence $\tilde{y} = 0$ and $\mathbb{E}[\tilde{x}\tilde{y} \mid \mathcal{M}^c] = 0$. The law of total expectation gives

$$G_p = \mathbb{E}[\tilde{x}\tilde{x}^\top] = p\Sigma + (1-p)D, \qquad \mathbb{E}[\tilde{x}\tilde{y}] = pc.$$

Thus these groups contribute feature covariance but no feature–outcome cross-moment. For either conditional distribution and any vector w, the quadratic form of its covariance is $\mathbb{E}[(w^\top \tilde{x})^2] \ge 0$. Both Σ and D, and therefore G_p, are positive semidefinite.

For a feature covariance H and cross-moment t, expanding the ridge objective gives

$$\mathbb{E}[\tilde{y}^2] - 2\beta^\top t + \beta^\top H\beta + \frac{\lambda}{d}\operatorname{tr}(H)\beta^\top\beta.$$

Its gradient is $2R(H)\beta - 2t$. For every nonzero w and positive-trace H considered here,

$$w^\top R(H) w \ge \frac{\lambda}{d}\operatorname{tr}(H)\|w\|_2^2 > 0.$$

The objective is strictly convex, so its unique minimizer is $R(H)^{-1}t$. Substitution of $(H, t) = (\Sigma, c)$ and (G_p, pc) proves Eq. 10. Neither Σ nor c depends on p, establishing invariance of the mixed-group utility and target. For weighted empirical data, dividing the squared-error sum and the Gram-matrix trace by the total sample weight scales the entire objective by the same positive constant. The minimizer is unchanged, giving exactly these moment formulas for Eq. 3.

For $0 < p < 1$, positive semidefiniteness and the trace penalty imply

$$R(G_p) \succeq \frac{(1-p)\lambda \operatorname{tr} D}{d} I_d.$$

The spectral decomposition of this real symmetric positive-definite matrix bounds the Euclidean operator norm of its inverse by $d/((1-p)\lambda \operatorname{tr} D)$. Applying this inverse to pc yields

$$\|u_p\|_2 \le \frac{pd\|c\|_2}{(1-p)\lambda \operatorname{tr} D}.$$

Let $w = [u_p]_+$ and $s = \|w\|_1$. Subtracting the uniform vector from the allocation target gives

$$\mathcal{T}_\epsilon(u_p) - \alpha_{\mathrm{unif}} = \frac{w - (s/d)\mathbf{1}}{s + d\epsilon}.$$

Since w is nonnegative, the triangle inequality bounds the numerator's ℓ_1 norm by $s + s = 2s$. Cauchy–Schwarz and $\|[u_p]_+\|_2 \le \|u_p\|_2$ give $s \le \sqrt{d}\|u_p\|_2$. Consequently,

$$\big\|\mathcal{T}_\epsilon(u_p) - \alpha_{\mathrm{unif}}\big\|_1 \le \frac{2s}{s + d\epsilon} \le \frac{2\sqrt{d}}{d\epsilon}\|u_p\|_2 \le \frac{2\sqrt{d}\,p\|c\|_2}{(1-p)\lambda\epsilon\,\mathrm{tr}\,D}.$$

The last expression tends to zero as $p \downarrow 0$. The mixed-group target is fixed, so it remains separated from this limit whenever it is nonuniform. This proves the target bound and limiting separation.

In Eq. 11, the first two coordinates satisfy $|azv_s + b\xi| \le a + b \le 1/2$, and the third satisfies $|b\xi| = b \le 1/2$. Every coverage score therefore lies in $[0, 1]$, and the specified labels are binary. Balanced sign pairs satisfy

$$\mathbb{E}[z] = \mathbb{E}[\xi] = \mathbb{E}[z\xi] = 0, \qquad \mathbb{E}[z^2] = \mathbb{E}[\xi^2] = 1.$$

The group feature mean is $\mathbf{1}/2$, giving $\tilde{x} = azv + b\xi\mathbf{1}$. In mixed groups, $\bar{y}_g = 1/2$ and $\tilde{y} = z/2$. In all-wrong groups, $\tilde{y} = 0$. Expanding the products and using the zero cross-moment of the signs gives

$$\Sigma = D = a^2 vv^\top + b^2\mathbf{1}\mathbf{1}^\top, \qquad c = \frac{a}{2}v, \qquad \mathrm{tr}\,\Sigma = 2a^2 + 3b^2.$$

Since $v^\top v = 2$ and $\mathbf{1}^\top v = 0$, we obtain

$$R(\Sigma)v = (2a^2 + r_\lambda)v.$$

Both group types have the same covariance, so $G_p = \Sigma$. Eq. 10 then gives $u_{\mathrm{M}} = hv$ and $u_p = phv$. Their first coordinate is positive for $p > 0$, and the feature variance is nonzero. Thus the positivity and feature-variation gates admit both estimates. Replicating balanced groups increases the mixed-sample mass without changing the conditional moments, making the assumed effective-sample eligibility compatible with this construction.

For either fixed target α_{tar}, Eq. 4 becomes $\alpha_{\ell+1} = (1-\eta)\alpha_\ell + \eta\alpha_{\mathrm{tar}}$. Starting at $\alpha_0 = \alpha_{\mathrm{unif}}$, this recursion yields

$$\alpha_j = (1-\eta)^j\alpha_{\mathrm{unif}} + \big(1 - (1-\eta)^j\big)\alpha_{\mathrm{tar}}.$$

The formula holds at $j = 0$. If it holds at j, substitution into the recursion makes the coefficient on α_{unif} equal to $(1-\eta)^{j+1}$ and the coefficient on α_{tar} equal to $(1-\eta)(1-(1-\eta)^j)+\eta = 1-(1-\eta)^{j+1}$. Induction proves the formula for every $j \ge 0$. Each allocation stays in the simplex as a convex combination of two simplex vectors. Since $[hv]_+ = (h, 0, 0)^\top$,

$$\mathcal{T}_\epsilon(hv) = \frac{(h+\epsilon, \epsilon, \epsilon)^\top}{h + 3\epsilon}, \qquad v^\top\mathcal{T}_\epsilon(hv) = \frac{h}{h + 3\epsilon}.$$

Replacing h by ph gives the pooled target's contrast $ph/(ph + 3\epsilon)$. Also $v^\top\alpha_{\mathrm{unif}} = 0$. Taking the inner product of the recursion solution with v proves $v^\top\alpha_{\mathrm{M},j} = \delta_{\mathrm{M},j}$ and $v^\top\alpha_{p,j} = \delta_{p,j}$ in Eq. 12.

On an all-wrong group, the answer reward is zero. Let P be the common physics score and $L = 1 - w_\ell\mathrm{len} > 0$ the common length multiplier. For either allocation α, Eq. 2 and $\mathbf{1}^\top\alpha = 1$ give

$$r = L\big(w_m\alpha^\top x + w_p P\big), \qquad \alpha^\top x = \tfrac{1}{2} + a(v^\top\alpha)z + b\xi.$$

Write $\delta = v^\top\alpha$. Averaging over the balanced group removes both sign terms, so

$$\bar{r} = L(w_m/2 + w_p P), \qquad r - \bar{r} = Lw_m(a\delta z + b\xi).$$

Squaring the centered reward and averaging gives

$$\sigma_r^2 = L^2 w_m^2\big(a^2\delta^2\mathbb{E}[z^2] + 2a\delta b\mathbb{E}[z\xi] + b^2\mathbb{E}[\xi^2]\big) = L^2 w_m^2(a^2\delta^2 + b^2).$$

The assumptions $L, w_m, b > 0$ ensure $\sigma_r > 0$, so the nondegenerate branch of Eq. 1 applies. Division by σ_r cancels the positive common factor Lw_m and gives Eq. 13. Multiplying by z and averaging gives

$$\mathbb{E}[Az] = F(\delta), \qquad F(t) = \frac{at}{\sqrt{a^2t^2 + b^2}}, \qquad F'(t) = \frac{ab^2}{(a^2t^2 + b^2)^{3/2}} > 0.$$

For $j \ge 1$, $\kappa_j > 0$. Since $h, \epsilon > 0$ and $0 < p < 1$, cross-multiplication gives $ph/(ph+3\epsilon) < h/(h+3\epsilon)$, equivalently $3ph\epsilon < 3h\epsilon$. Hence $0 < \delta_{p,j} < \delta_{\mathrm{M},j}$, with $\delta_{p,j} \to 0$ as $p \downarrow 0$. Strict monotonicity and continuity of F, together with $F(0) = 0$, prove Eq. 14. The mixed-group association stays positive and independent of p, completing the proof. □

D DETAILED EXPERIMENTS

Table 8 shows task-specific sensitivity to utility proxies. On Qwen3-VL-4B, failure-only utility produces the weakest SAT result among the tested component variants despite retaining competitive performance on other benchmarks. Grading and reward constraints have different task-specific effects, indicating that they do not substitute for estimating which stages distinguish correct answers. This pattern supports task-level assessment of utility beyond aggregate coverage.

Table 9 shows backbone-dependent sensitivity to harness settings. Relaxing the mixed-sample threshold approaches CALIPER on Qwen3-VL-4B but widens the gap on Qwen2.5-VL-7B. Frozen allocation and disabled rules also produce distinct task-level changes, separating reward adaptation from training intervention. This pattern is consistent with complementary contributions from informative rollout groups and the harness response to their scarcity.

Table 10 shows that CALIPER's advantage over frozen allocation extends to DAPO and GSPO. The gain varies by task despite matched optimizer settings and reference access within each comparison. This variation supports utility-based allocation across policy objectives, with task-dependent gains.

Table 8: Component ablations by benchmark.

Method	3DSR	MMB	MME-RW	MMSTAR	OMNI3D	ROBOSP	SAT	SPATEVAL	**Avg**
Qwen3-VL-4B									
CALIPER (ours)	66.00	**93.50**	**52.00**	**69.00**	**37.05**	**46.95**	**76.50**	72.50	**64.19**
w/o Measured Utility	61.50	92.50	48.00	67.50	32.05	44.69	73.50	72.50	61.53
w/o Independent Judging	59.50	92.00	50.00	66.00	31.55	45.06	76.00	75.50	61.95
w/o Solved-Group Gate	62.50	91.00	49.50	67.00	35.00	43.13	74.00	75.00	62.14
w/o Length Penalty	64.00	91.00	50.50	66.50	36.70	40.48	74.50	75.00	62.34
w/o Physics Prior	**67.00**	92.00	50.50	65.50	33.10	43.15	74.00	74.00	62.41
w/o Utility, Dispersion	61.50	92.50	49.00	67.00	32.75	43.56	75.50	75.00	62.10
w/o Utility, Failure	61.50	92.00	50.50	67.00	36.90	44.97	54.50	**77.00**	60.55
Qwen2.5-VL-7B									
CALIPER (ours)	**54.50**	90.50	**50.50**	63.50	**23.05**	46.73	57.00	**62.00**	**55.97**
w/o Measured Utility	51.50	88.00	46.00	61.50	**23.05**	45.66	56.00	56.50	53.53
w/o Independent Judging	51.50	90.00	48.50	61.00	22.95	45.72	58.50	57.50	54.46
w/o Solved-Group Gate	51.50	90.00	46.50	64.50	22.50	45.09	56.00	59.50	54.45
w/o Length Penalty	51.00	88.00	49.50	63.00	21.55	47.07	55.00	60.00	54.39
w/o Physics Prior	52.00	89.50	49.50	63.50	20.45	45.61	56.00	57.50	54.26
w/o Utility, Dispersion	51.00	90.00	49.00	63.00	21.85	**47.17**	**59.00**	57.50	54.81
w/o Utility, Failure	49.50	**91.00**	48.50	**65.50**	21.25	43.58	55.00	56.50	53.85
InternVL3-8B									
CALIPER (ours)	**58.00**	89.00	55.00	**72.50**	**41.65**	46.10	75.00	**57.50**	**61.84**
w/o Measured Utility	55.50	88.50	54.00	70.50	37.70	46.73	75.50	52.50	60.12
w/o Independent Judging	56.50	**89.50**	54.00	72.00	37.30	48.22	73.50	55.50	60.81
w/o Solved-Group Gate	55.00	89.00	**56.00**	71.00	36.90	**49.75**	75.50	56.50	61.21
w/o Length Penalty	56.50	88.00	53.50	69.50	37.15	47.83	**77.00**	56.00	60.69
w/o Physics Prior	56.50	89.00	55.00	**72.50**	35.25	47.82	74.50	56.50	60.88
w/o Utility, Dispersion	56.50	87.50	54.00	70.00	36.20	48.71	76.00	56.50	60.68
w/o Utility, Failure	57.50	88.00	55.00	69.50	39.80	45.73	76.00	56.00	60.94

D.1 HYPERPARAMETER SENSITIVITY

Table 11 shows that the shared default yields the highest mean across the tested settings for each backbone. Larger rollout groups partly recover performance on Qwen3-VL-4B and InternVL3-8B but reduce it on Qwen2.5-VL-7B. The nonmonotonic response indicates that larger rollout groups alone do not consistently improve performance.

Table 9: Harness ablations by benchmark.

Method	3DSR	MMB	MME-RW	MMStar	Omni3D	RoboSp	SAT	SpatEval	Avg
Qwen3-VL-4B									
Outcome Reward	58.00	91.00	50.50	64.50	33.65	46.72	76.00	70.50	61.36
Reference-Assisted	61.50	92.00	50.00	65.00	32.55	48.03	75.00	71.50	61.95
Frozen α	64.50	94.00	50.50	65.50	32.70	44.09	75.50	73.00	62.47
w/o Measured Utility	62.50	93.00	50.00	69.00	35.20	48.19	73.50	73.00	63.05
α EMA $= 0.05$	60.00	92.50	52.50	65.50	35.85	49.58	77.00	73.50	63.30
Min. Mixed Samples $= 8$	65.00	93.00	50.50	68.50	34.15	46.46	76.50	79.00	64.14
Rollout Temperature $= 1.5$	60.50	91.50	49.00	70.50	32.35	48.61	76.50	79.00	63.49
w/o Rules	62.00	93.50	53.00	65.00	34.35	47.61	74.00	74.00	62.93
CALIPER (ours)	65.50	93.00	52.50	70.00	36.35	50.37	72.50	75.00	64.40
Qwen2.5-VL-7B									
Outcome Reward	55.50	89.50	49.50	65.50	24.05	43.06	62.50	55.50	55.64
Reference-Assisted	53.00	88.50	50.00	63.50	22.45	43.47	62.00	57.00	54.99
Frozen α	52.50	89.50	50.50	65.00	22.60	46.65	55.50	59.50	55.22
w/o Measured Utility	51.00	90.50	52.00	65.00	20.95	43.58	58.50	59.00	55.07
α EMA $= 0.05$	51.50	89.50	49.50	64.50	23.90	46.18	56.50	59.00	55.07
Min. Mixed Samples $= 8$	52.50	90.00	48.00	65.00	19.80	44.10	53.50	57.50	53.80
Rollout Temperature $= 1.5$	52.00	90.50	48.00	65.00	24.65	44.15	58.00	57.50	54.98
w/o Rules	49.00	87.00	46.50	63.00	21.40	45.66	55.50	56.00	53.01
CALIPER (ours)	53.50	90.50	49.50	65.50	24.75	48.27	55.50	62.50	56.25
InternVL3-8B									
Outcome Reward	55.50	87.50	54.50	66.50	37.00	43.75	74.00	57.50	59.53
Reference-Assisted	55.00	88.00	54.00	68.50	38.10	47.19	74.00	57.50	60.29
Frozen α	55.50	88.50	55.00	73.50	37.95	47.20	74.50	55.50	60.96
w/o Measured Utility	56.00	89.00	56.00	70.00	35.40	45.32	75.50	58.50	60.72
α EMA $= 0.05$	57.50	88.50	54.00	72.00	37.30	46.65	74.50	57.00	60.93
Min. Mixed Samples $= 8$	58.50	89.00	54.50	75.00	37.60	46.75	76.50	55.00	61.61
w/o Rules	57.50	88.50	54.50	69.50	38.60	50.24	75.50	57.00	61.42
CALIPER (ours)	57.50	89.00	55.50	71.00	40.25	52.14	75.00	60.00	62.55

D.2 Scaling Settings

Model Size. Fig. 8a compares the 8B, 14B and 38B variants of InternVL3 (Zhu et al., 2025) with $1{,}000$ training items, rank-64 LoRA, ten update steps and eight rollouts per question. The 38B variant also changes its vision encoder, so this panel describes family-level capacity scaling.

Training Data. The data-size panel fixes InternVL3-8B and uses nested subsets that approximately preserve source proportions and complete-trace coverage. Each update samples 512 questions with replacement, holding the update count fixed while varying exposure to unique items.

Update Budget. The update panel evaluates final checkpoints after 2, 4, 6, 8, 10, 15, 20 and 30 steps. All settings use $1{,}000$ training items, eight rollouts per question and 512 questions per step, isolating the number of updates.

Trace Coverage. The annotation panel fixes ten updates and varies complete-trace coverage over $\{25, 40, 55, 70, 84.4\}\%$. Nested, source-stratified subsets retain all stages for selected complete traces. Other records retain answers and available reasoning, with missing stages masked. The fixed $1{,}000$-item training set isolates stage evidence for utility estimation.

E Qualitative Evidence

The left–right example documents an inference error despite an accurate visual description, separating evidence recognition from its use in the answer. In the Omni3D example, the InternVL3 fixed-weight trace identifies the nearer chair but selects the washing machine, whereas CALIPER follows the depth evidence. In the RoboSpatial placement example, CALIPER answers correctly on both backbones while the other rewards exhibit backbone-dependent errors.

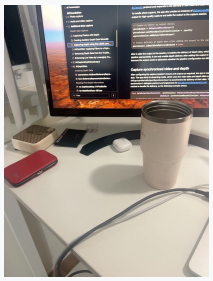
Table 10: Optimizer comparisons under harness training.

Method	3DSR	MMB	MME-RW	MMStar	Omni3D	RoboSp	SAT	SpatEval	**Avg**
Qwen3-VL-4B									
GRPO, Outcome Reward	58.00	91.00	50.50	64.50	33.65	46.72	76.00	70.50	61.36
GRPO, Reference-Assisted	61.50	92.00	50.00	65.00	32.55	48.03	75.00	71.50	61.95
GRPO, Frozen α	64.50	**94.00**	50.50	65.50	32.70	44.09	75.50	73.00	62.47
GRPO, CALIPER (ours)	**65.50**	93.00	**52.50**	**70.00**	36.35	50.37	72.50	75.00	**64.40**
DAPO, Outcome Reward	61.50	93.50	50.50	65.50	34.45	51.03	76.50	73.50	63.31
DAPO, Reference-Assisted	62.50	**94.00**	**52.50**	66.50	31.70	**51.18**	75.50	73.00	63.36
DAPO, Frozen α	64.50	91.00	51.00	69.00	33.90	43.11	73.50	74.50	62.56
DAPO, CALIPER (ours)	**65.50**	92.00	50.50	64.50	33.60	50.19	74.50	**76.00**	63.35
GSPO, Outcome Reward	60.00	92.00	52.00	66.00	32.45	45.03	74.00	70.00	61.43
GSPO, Reference-Assisted	65.00	91.50	51.00	66.50	32.00	47.62	**77.00**	73.00	62.95
GSPO, Frozen α	**65.50**	92.00	52.00	65.00	**37.65**	44.59	74.50	71.00	62.78
GSPO, CALIPER (ours)	64.00	91.50	50.00	67.00	35.05	50.15	75.00	72.50	63.15
Qwen2.5-VL-7B									
GRPO, Outcome Reward	**55.50**	89.50	49.50	**65.50**	24.05	43.06	62.50	55.50	55.64
GRPO, Reference-Assisted	53.00	88.50	50.00	63.50	22.45	43.47	62.00	57.00	54.99
GRPO, Frozen α	52.50	89.50	50.50	65.00	22.60	46.65	55.50	59.50	55.22
GRPO, CALIPER (ours)	53.50	**90.50**	49.50	**65.50**	24.75	**48.27**	55.50	**62.50**	56.25
DAPO, Outcome Reward	**55.50**	89.00	53.00	62.00	23.35	45.09	**63.00**	59.50	**56.31**
DAPO, Reference-Assisted	55.00	87.50	**53.50**	63.50	**25.00**	44.05	61.50	60.00	56.26
DAPO, Frozen α	49.50	86.50	47.00	63.00	24.35	44.67	55.50	56.50	53.38
DAPO, CALIPER (ours)	51.00	85.00	48.00	61.00	23.55	48.21	59.50	55.50	53.97
GSPO, Outcome Reward	**55.50**	89.50	52.00	64.00	22.60	45.04	56.50	58.50	55.46
GSPO, Reference-Assisted	**55.50**	86.50	48.50	62.00	24.75	42.02	57.50	54.50	53.91
GSPO, Frozen α	52.00	87.00	48.00	60.50	18.85	43.68	56.00	56.00	52.75
GSPO, CALIPER (ours)	51.00	90.00	50.00	64.50	22.80	45.08	57.00	57.00	54.67
InternVL3-8B									
GRPO, Outcome Reward	55.50	87.50	54.50	66.50	37.00	43.75	74.00	57.50	59.53
GRPO, Reference-Assisted	55.00	88.00	54.00	68.50	38.10	47.19	74.00	57.50	60.29
GRPO, Frozen α	55.50	88.50	55.00	**73.50**	37.95	47.20	74.50	55.50	60.96
GRPO, CALIPER (ours)	57.50	89.00	55.50	71.00	**40.25**	52.14	**75.00**	60.00	**62.55**
DAPO, Outcome Reward	54.50	89.00	52.50	71.00	38.25	43.50	74.00	**68.00**	61.34
DAPO, Reference-Assisted	55.00	**89.50**	54.00	69.50	36.10	48.10	71.00	66.50	61.21
DAPO, Frozen α	**59.00**	**89.50**	55.50	70.50	34.30	49.38	73.00	57.50	61.09
DAPO, CALIPER (ours)	57.00	88.50	**56.50**	73.00	39.15	**52.17**	**75.00**	58.50	62.48
GSPO, Outcome Reward	56.00	88.00	52.00	70.00	39.10	46.67	74.50	53.50	59.97
GSPO, Reference-Assisted	57.50	88.00	51.50	71.00	39.20	46.14	71.00	57.00	60.17
GSPO, Frozen α	57.00	89.00	53.00	72.50	38.30	45.26	72.00	56.50	60.44
GSPO, CALIPER (ours)	57.50	89.00	53.50	70.00	37.90	47.78	73.00	56.00	60.58

Table 11: One-shot sensitivity averaged over eight benchmarks.

Configuration	Qwen3-VL-4B	Qwen2.5-VL-7B	InternVL3-8B
Default configuration	64.19	55.97	61.84
$\eta = 0.05$	62.21	54.10	60.77
$\eta = 0.5$	63.27	54.48	60.68
$\lambda = 0.01$	61.79	53.94	60.97
$n = 4$	61.03	54.39	60.81
$n = 12$	62.76	54.96	60.48
$n = 16$	63.73	53.74	61.71
$T = 0.7$	61.23	54.18	60.53
$T = 1.5$	61.78	55.10	60.92
$\gamma = 0.7$	62.91	54.20	60.51
$\gamma = 0.98$	63.00	54.20	61.25

RoboSpatial-Home, left–right regression



Can the speaker fit left of the cup? Answer yes or no.

Perception and inference

IVL3-8B Base ✓ *"The speaker is located to the left of the cup in the image."* → yes

IVL3-8B Outcome ✓ *"the speaker is already positioned to the left of the cup"* → yes

IVL3-8B Dense ✓ *"the speaker is currently to the left of the cup"* → yes

IVL3-8B CALIPER × *"The speaker is located to the left of the computer monitor, and the cup is placed to the right of the monitor on the desk."* Yet it concludes *"the speaker is not to the left of the cup"* → no

Ground-truth answer: Yes.

Omni3D-Bench, both backbones on one item



Which object is closer to the camera: the washing machine or the rightmost chair?

Perception

"The image shows a kitchen area with a washing machine on the right side of the counter, and a wooden chair (partially visible) on the left side, near the . . . "

Inference

Q3-4B Base × *". . . positioned closer to the camera, as it is in the foreground and its front is directly facing the viewer."* → washing machine

Q3-4B Outcome × *". . . the foreground of the image and occupies a more central position relative to the camera's perspective."* → washing machine

Q3-4B Dense × *". . . a more central position in the frame. The chair is further away, behind the table and to the left."* → washing machine

Q3-4B CALIPER ✓ *". . . and the washing machine is further back (in the mid-ground), the chair is closer to the camera."* → rightmost chair

IVL3-8B Base ✓ *". . . to the wall. - The rightmost chair is positioned closer to the foreground, nearer to the camera. Answer:"* → rightmost chair

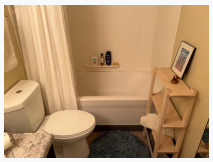
IVL3-8B Outcome × *". . . front of the camera, while the rightmost chair is further away, placed at the end of the table. Answer:"* → washing machine

IVL3-8B Dense × *". . . further back in the image compared to the rightmost chair, which is closer to the foreground. Answer:"* → washing machine

IVL3-8B CALIPER ✓ *". . . to the wall. - The rightmost chair is positioned closer to the foreground, nearer to the camera. Answer:"* → rightmost chair

Ground-truth answer: rightmost chair.

RoboSpatial-Home, both backbones on one item



Can the watch fit above the toilet? Answer yes or no.

Perception

"The image shows a bathroom with a toilet on the left, a bathtub with a shower curtain in the center, and a wooden shelving unit on the right. - On the top . . . "

Inference

Q3-4B Base	×	*". . . the watch cannot fit above the toilet as there is no such surface or space available in the image."* → no
Q3-4B Outcome	×	*". . . the shelf is not positioned to be directly above it. Therefore, the watch cannot fit above the toilet."* → no
Q3-4B Dense	×	*". . . above the toilet, and there is no indication that it could be moved there without obstruction."* → no
Q3-4B CALIPER	✓	*". . . and the watch is placed on that shelf. - Therefore, the watch is physically located above the toilet."* → yes
IVL3-8B Base	×	*". . . a watch could be placed without interference from the shower curtain or the bathtub fixtures. Answer:"* → no
IVL3-8B Outcome	×	*". . . above the toilet where a watch could be placed without interference from the shower curtain. Answer:"* → no
IVL3-8B Dense	✓	*". . . issues. Based on the observations, the watch is already placed above the toilet on the shelf. Answer:"* → yes
IVL3-8B CALIPER	✓	*". . . issues. Based on the observations, the watch is already placed above the toilet on the shelf. Answer:"* → yes

Ground-truth answer: Yes.