SGT-3D: 3D-GUIDED SELF-DISTILLATION AND GEOMETRIC VERIFICATION FOR SPATIAL REASONING

Anonymous authors

ABSTRACT

Spatial reasoning from images often involves estimating object geometry and using it to answer spatial questions. Some spatial reasoning approaches expose intermediate 3D attributes, but making them visible does not ensure accurate prediction or correct use. These two goals motivate complementary supervision: final-answer rewards do not directly assess attribute accuracy, while training on reference responses may leave reasoning from the model's own imperfect predictions insufficiently supervised. The resulting challenge is to improve object-level spatial understanding and turn better estimates into more accurate answers. We present SGT-3D, which couples learning to predict object centers, dimensions and orientations with learning to reason from the states the student actually generates. A training-only 3D expert guides self-distillation, with continuation supervision following the student's stated attributes within a fixed-state interface, and subsequent reinforcement learning supplies numerical attribute feedback alongside answer rewards. Inference requires only RGB images and questions. SGT-3D achieves QA accuracy of 69.2% on 3DSRBench-real and 96.2% on CV-Bench-3D, with 85.8% center/size/orientation pass rate against independent annotations on a separate diagnostic subset.

1 INTRODUCTION

Spatial reasoning from images involves interpreting spatial relationships to answer questions about a scene. Comparing object depths, physical sizes or box orientations, for example, involves both estimating relevant 3D attributes and correctly using those estimates. Attribute estimation captures an important aspect of object-level spatial understanding, while subsequent comparison and computation determine how these estimates inform the answer. However, vision-language models (VLMs) may leave these estimates implicit, making them difficult to supervise directly.

Explicit approaches expose spatial attributes before reasoning: SpatialReasoner separates perception from computation, while GS-Reasoner includes predicted 3D boxes in its reasoning traces (Ma et al.,

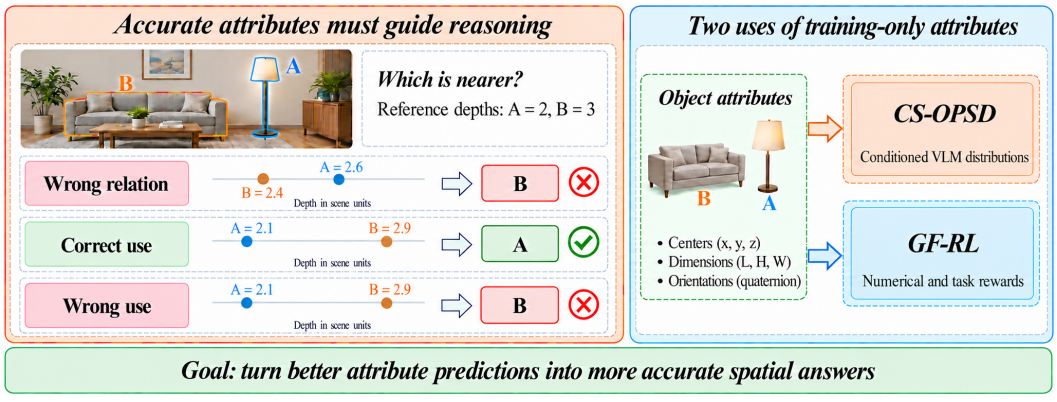


Figure 1: **Attribute prediction and use both matter for spatial answering.** Left: incorrect depths can reverse the relevant ordering, and correcting them can still leave an error in comparison. The repeated improved coordinates illustrate why attribute gains need not become answer gains. Right: centers, dimensions and orientations provide the shared object interface for full-response CS-OPSD and attribute-plus-answer GF-RL.

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.

2025b; Chen et al., 2025). Yet making attributes explicit leaves two obstacles to accurate spatial answering. Inaccurate predictions can distort task-relevant spatial relationships and provide a faulty basis for reasoning. Even accurate attributes can yield incorrect answers when the model makes errors in comparison or computation (Figure 1). Better attribute estimates therefore do not automatically translate into better spatial reasoning performance: the model must also learn to use the information they provide.

These obstacles motivate supervision of both attribute prediction and reasoning from predicted attributes. Training only on reference traces does not directly supervise reasoning from the imperfect attribute states encountered in student-generated responses, while final-answer rewards do not directly assess the numerical accuracy of those attributes. We therefore combine guidance on how attributes are expressed and used along student-generated responses with numerical feedback on their predicted values. Through these complementary signals, we aim to improve object-level spatial understanding and translate better spatial estimates into more accurate answers.

SGT-3D instantiates this principle using references from a training-only 3D expert to guide self-distillation on student-generated responses and provide numerical attribute feedback. Structured supervised fine-tuning initializes Scan–Gaze–Think chain-of-thought (SGT CoT) traces. We introduce Committed-State On-Policy Self-Distillation (CS-OPSD), a 3D-attribute-conditioned variant of OPSD (Zhao et al., 2026) that guides complete responses along the student's own generated prefixes. Its committed-state priority rule uses the expert for unfinished fields and preserves the student's stated values as the basis for continuation supervision, including when they differ from the reference. Its principle is to *improve future attribute predictions while preserving supervision for reasoning from the current state*.

A correct spatial answer can coexist with inaccurate attribute values, leaving their numerical quality unresolved by answer rewards alone. Our Geometric Feedback Reinforcement Learning (GF-RL) stage checks each predicted center, dimension and orientation field against the admitted expert reference. Each object–field reward is normalized across rollouts for that target and assigned through its completion boundary, alongside a separate terminal answer reward. Inference requires only RGB images and questions.

We present SGT-3D, a training framework for improving 3D attribute prediction and its use in spatial answering, while retaining RGB-only inference. Our technical contributions are:

1. We introduce **CS-OPSD**, an evidence-priority rule within on-policy self-distillation. Within the fixed-state interface, the expert guides unfinished attribute fields, while continuation supervision follows values already stated by the student. This design provides supervision for reasoning consistently from the student's own attribute predictions, including imperfect ones.
2. We define **GF-RL** with individual object–field targets as the units of numerical verification and reward normalization. Completion boundaries delimit each attribute reward's contribution to the generated trace, while a separately normalized final-answer reward retains the task objective. This design connects checks of predicted 3D values with feedback on their generation.

2 Related Work

2.1 Explicit interfaces for spatial reasoning

SpatialVLM and SpatialRGPT use generated spatial supervision and 3D scene graphs, respectively (Chen et al., 2024; Cheng et al., 2024). SpatialReasoner and GS-Reasoner expose geometry before reasoning (Ma et al., 2025b; Chen et al., 2025). GS-Reasoner and N3D-VLM also use depth in their reported setups (Chen et al., 2025; Wang et al., 2025). Visual Spatial Tuning (VST, ECCV 2026) uses large-scale SFT followed by RL without a specialized 3D encoder (Yang et al., 2026). Making geometry explicit does not by itself ensure accurate attribute predictions or correct reasoning from those predictions. Our shared object interface treats attribute prediction and reasoning from student-generated states as complementary training targets, combining response distillation with numerical feedback while retaining RGB-only inference.

2.2 Privileged supervision and on-policy distillation

Privileged distillation transfers depth, object and relation evidence to RGB-only VLMs (Hu et al., 2026). GKD supervises student-generated sequences (Agarwal et al., 2024). OPSD scores shared prefixes with a privileged shared-weight teacher (Zhao et al., 2026). PW-OPSD studies reliability weighting (Liu et al., 2026), EDGE-OPD uses guided rollouts and evidence masking (Lazaridis et al., 2026), and related work identifies self-correction risks (Kaur et al., 2026). ViGOS separates image-only perception teaching from reference-conditioned reasoning teaching on the same student prefix (Wang et al., 2026). When privileged attributes conflict with a student-generated prefix, supervision must resolve which values subsequent reasoning should follow. CS-OPSD makes this choice explicit: the expert guides unfinished fields, while reasoning follows completed student values. This retains continuation supervision on states that disagree with the reference.

2.3 Attribute verification and process credit

Process supervision uses step labels and reward-to-go (Lightman et al., 2023; Shao et al., 2024), with spatial feedback in SpatialReasoner, RoboTracer and DR-MV3D (Ma et al., 2025b; Zhou et al., 2026a; Choi et al., 2026). Our auxiliary reference refinement projects attributes, distinct from posterior regularization of distributional expectations (Ganchev et al., 2010). SCOUT separately normalizes reward types, including grounding, depth and reasoning consistency, and combines them into segment-specific advantages (Zhou et al., 2026b). GF-RL studies a finer unit of feedback: each object–field target has its own normalization population and completion boundary. Its numerical reward reaches tokens up to that boundary, while answer feedback reaches the whole response. Matched controls cross normalization populations with segment and boundary credit under identical raw rewards.

3 Method

Given an RGB image and question $x = (I, q)$, SGT-3D learns to predict 3D object attributes and reason from those predictions to answer spatial questions. A shared spatial trace connects two training signals: CS-OPSD guides attribute expression and subsequent comparison and computation on the student's own prefixes, while GF-RL directly scores attribute values and final answers. Together they target the quality of the spatial estimates and their use in answering. Training proceeds sequentially through structured SFT, CS-OPSD and GF-RL (Figure 2).

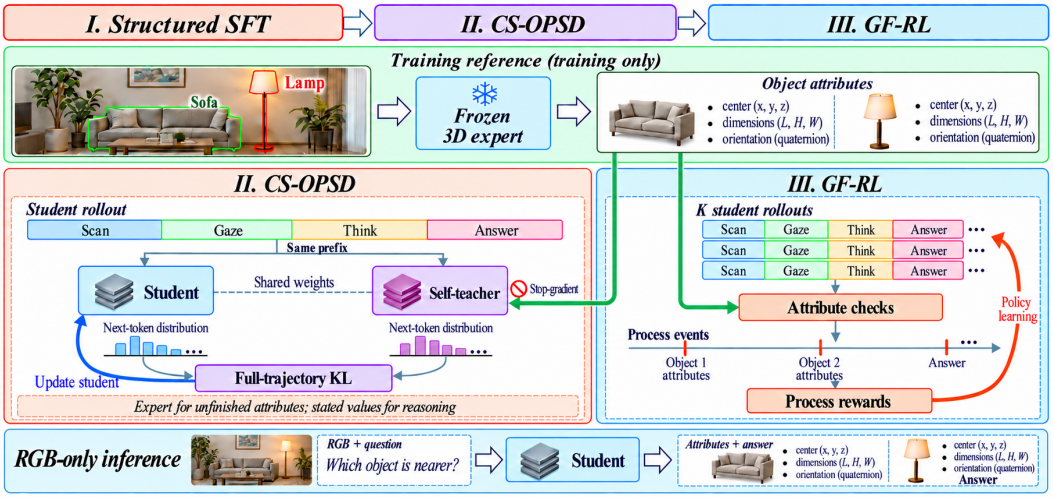


Figure 2: **Sequential training and information flow: SFT → CS-OPSD → GF-RL.** The expert branch supplies reference centers, dimensions and orientations in Z^R to both the detached self-teacher and checker. In the refinement variant, training answers enter a preprocessing step on this branch before either use. Student and self-teacher score identical student prefixes. KL covers every output token. Student rollouts and inference use RGB and the question.

3.1 TASK SETUP AND SCAN–GAZE–THINK TRACES

To supervise both attribute prediction and its use in answering, we represent object selection, 3D attributes and subsequent computation within a shared spatial trace. Scan–Gaze–Think chain-of-thought (SGT CoT) connects these steps while making the correspondence between question referents, predicted attributes and reference attributes explicit:

$$\hat{y} = \text{serialize}(\tau), \qquad \tau = \langle \underbrace{G}_{\text{Scan}}, \underbrace{Z}_{\text{Gaze}}, \underbrace{U, a}_{\text{Think and answer}} \rangle, \tag{1}$$

Here G binds referents to object identifiers and 2D regions, Z records box centers $\mathbf{c}_i$, dimensions $\mathbf{s}_i$ and orientations $\mathbf{r}_i$, and U records the computation leading to answer a. A scalar-first unit quaternion $\mathbf{r}_i$ represents the rotation from the box's local axes to the camera frame. Attributes are serialized text fields, learned without a separate regression head. Structured SFT initializes this interface using token cross-entropy on grounded traces, mixed with general conversation supervision. The objective and supervision provenance are in Appendix A.3.

Frozen WildDet3D (Huang et al., 2026), a specialist prompted 3D object detector, supplies training-reference centers, dimensions and orientations using trusted 2D object prompts and its declared calibration/depth route. These prompts fix object identities. Raw expert inference receives no answer. Camera-frame centers and dimensions use one common scene-depth scale, while normalized rotation quaternions are scale-independent, yielding $Z^E = \mathcal{C}_E(E(I, \mathcal{G}_{\text{train}}))$. We distinguish this raw expert state, the admitted training reference Z^R, and independent evaluation attributes Z^*. Orientation describes the 3D box rotation, with box symmetries accounted for, and does not by itself establish semantic front/back. Coordinate conventions and expert inputs are specified in Appendix A.1.

The default retains valid raw references, $Z^R = Z^E$, fixed across both objectives. Missing or invalid references supply no attribute supervision. Optional reference filtering or refinement is a separate robustness study (Appendix A.1). The refined targets remain training references.

3.2 COMMITTED-STATE ON-POLICY SELF-DISTILLATION (CS-OPSD)

This trace exposes the attributes the student actually predicts, which may differ from the expert reference. Our serialization fixes completed fields, so a conditioned teacher needs a rule for subsequent reasoning when these states disagree. CS-OPSD uses *committed-state priority*: the expert guides unfinished attribute fields, while subsequent computation and answering follow values already printed by the student. A teacher-only instruction u_{teach} specifies this rule and forbids silent replacement of completed values. It is an evidence-priority prompt, not an additional learned module.

For example, suppose expert depths are $z_A = 2, z_B = 5$, but the student has printed $z_A = 5, z_B = 2$. Expert priority teaches "A is nearer", while committed-state priority teaches "B is nearer" from the printed values. A consistent continuation can therefore be factually wrong. Earlier-token supervision seeks to improve attribute predictions. We hypothesize that comparison and computation learned on current states can transfer to future responses with more accurate attributes. Because distillation also covers answer tokens, it can reinforce factual errors. Complete-answer comparisons against masking under identical subsequent GF-RL test the net effect.

We implement this rule within full-trajectory OPSD (Zhao et al., 2026). At iteration n, the student samples $\hat{y} \sim s^{\text{CS}}_{\theta_n}(\cdot \mid x)$ without privileged input, where s^{CS} applies the configured decoder to p_θ. Student and shared-weight self-teacher then score each identical prefix over vocabulary $\mathcal{V}$:

$$\begin{aligned} p_{S,t}(v;\theta) &= p_\theta(v \mid x, \hat{y}_{<t}), \\ q^{(n)}_t(v) &= \text{stopgrad}\left[p_{\theta_n}(v \mid x, Z^R, u_{\text{teach}}, \hat{y}_{<t})\right], \quad v \in \mathcal{V}. \end{aligned} \tag{2}$$

The teacher receives Z^R and u_{teach} but generates no second trajectory. Its context contains accepted attribute fields without answer strings or audit metadata. Both branches start from θ_n. The teacher targets and sampled response are fixed during the student update.

Full-vocabulary forward KL transfers the teacher distributions across the complete response, with $T = |\hat{y}|$:

$$\mathcal{L}^{(n)}_{\mathrm{OPSD}}(\theta) = \mathbb{E}_{x,\, \hat{y}\sim s^{\mathrm{CS}}_{\theta_n}(\cdot|x)}[\ell_n(x,\hat{y})],$$
$$\ell_n(x,\hat{y}) = \begin{cases} \frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}\big(q_t^{(n)} \| p_{S,t}(\cdot;\theta)\big), & h_I = 1, \\ 0, & h_I = 0. \end{cases} \tag{3}$$

Here h_I is the reference-admission flag fixed before sampling. For each admitted response, all Scan, Gaze, Think, formatting and answer tokens contribute, with only input and padding positions excluded. Gradients pass through student scoring, and the teacher refreshes at the next iteration. Appendix A.3 specifies loss aggregation and sampling conventions.

CS-OPSD changes the teacher's evidence-priority instruction within the inherited OPSD objective. Plain conditioning, expert priority and conflict-selective masking isolate this choice (Appendices B and B.2). Matched prompts appear in Appendix B.1. In particular, masking uses the same committed teacher but withholds conflicting continuation losses, testing whether retaining that supervision helps.

3.3 Geometric Feedback Reinforcement Learning (GF-RL)

GF-RL directly evaluates the numerical quality of the attributes used for spatial answering. Different depth estimates can yield the same correct ordering despite different numerical errors, so answer correctness alone cannot determine attribute accuracy. CS-OPSD guides attribute prediction and use, but its token-distribution loss does not explicitly measure geometric discrepancies between decoded values. We therefore score predicted centers, dimensions and orientations against the same admitted Z^R, alongside a terminal answer reward. Independent Z^* evaluates attribute accuracy separately from agreement with this training reference.

Writing $\pi_\theta = p_\theta$, we initialize from the CS-OPSD checkpoint and sample K responses per question with rollout policy π_{old}, without privileged input. A deterministic checker matches parsed object fields to available references. For a valid center target, the reward is

$$r^{\mathrm{center}}_{k,i} = \exp\left(-\frac{\|\hat{\mathbf{c}}_{k,i} - \mathbf{c}^{\mathrm{ref}}_i\|_2}{\tau_c}\right), \qquad m_{i,\mathrm{center}} = 1, \tag{4}$$

where $m_{i,\mathrm{center}}$ marks reference availability and τ_c sets the reward scale. Dimensions use the same exponential transform of mean absolute log-dimension error, and orientation uses a symmetry-aware 3D rotation error. The terminal reward checks canonical answer correctness. To compare alternative predictions of the same spatial quantity, each fixed object–field target is normalized separately across the K rollouts. Invalid predictions receive zero raw reward and remain in that population. Missing references are omitted, and attribute weights sum to one. The answer has its own unit weight and normalization population. For identical rollouts, adding attributes leaves its normalized event unchanged, although clipping can change its contribution to the total gradient. Appendix A.2 gives the transforms, scales and parsing rules.

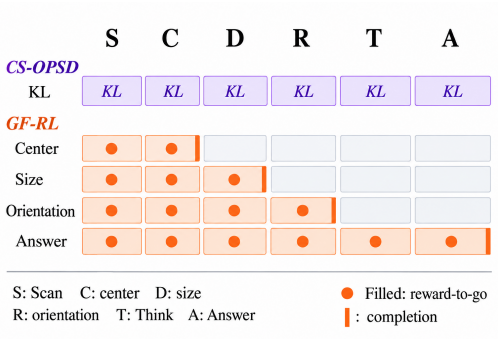


Figure 3: **Supervision coverage.** KL spans the full response. Attribute rewards stop at field completion; answer rewards cover all tokens.

A completed attribute cannot be changed by later text under the fixed serialization. We therefore place its reward at the field-completion boundary $b_{k,j}$, and the answer reward at response end T_k. Following process GRPO (Shao et al., 2024), weighted normalized event rewards $\widetilde{r}_{k,j}$ define

$$\widehat{A}_{k,t} = \sum_{j\in\mathcal{J}_k:\, b_{k,j}\geq t} \widetilde{r}_{k,j}, \tag{5}$$

where $\mathcal{J}_k$ contains every available attribute target and the terminal answer event. First assignments and raw rewards are immutable. Uncompleted targets receive zero at termination. A completed

attribute's reward thus reaches preceding tokens, including its boundary, but stops before subsequent fields and Think tokens. The answer reward reaches the whole response (Figure 3).

These advantages enter the standard clipped GRPO objective (Equation 11). GF-RL's design concerns the unit of numerical feedback and its completion boundary, using the existing optimizer without a learned verifier. Attribute rewards assess reference agreement, not Think arithmetic, and update future rollouts without repairing the current trace. The experiments compare this reward design with answer-only RL and direct attribute supervision, then cross normalization populations with credit boundaries under fixed raw rewards.

3.4 Training and inference

Each stage initializes from its predecessor: structured SFT $\rightarrow$ CS-OPSD $\rightarrow$ GF-RL. The two later stages share object targets but apply different supervision to the response-generating model. At inference, the final student generates G, Z, U, a from RGB and the question, without the expert, privileged self-teacher or training verifier. Table 7 summarizes the sequential training procedure and information access.

4 Experiments

We evaluate complete spatial answering on full 3DSRBench-real with CircularEval/FlipEval and use three complementary diagnostics. 3DSR-D1000 pairs independent attribute assessment with canonical-question QA on 1,000 questions. The attribute and joint readouts assess centers, dimensions and orientations; benchmark Orientation is a spatial-QA category. 3DSR-T400 audits teacher continuations on 400 sampled prefixes, which may repeat questions; 3DSR-I100 tests student use of supplied attributes on 100 eligible depth questions. These are evaluation cohorts with different scoring units, not training splits (Appendix C). CV-Bench-3D provides an additional main benchmark comparison.

All SGT-3D variants and ablations use Qwen2.5-VL-7B-Instruct as their base model and Spatial-Reasoner's training datasets with regenerated spatial labels. SFT uses the SR-CoT/LLaVA mixture; CS-OPSD and GF-RL share SR-QA. Dataset sizes, hardware, hyperparameters and stage budgets are detailed in Appendix A.5 and Table 8.

4.1 Benchmark comparisons

Table 1 compares generalist VLMs, spatial specialists and complete verification/tool methods. The shared evaluation protocol fixes benchmark releases, questions, option permutations and aggregation; native inference settings and training histories remain distinct (Appendix C.1). Original source settings are documented separately. CV-Bench-3D tests depth order and relative distance (Tong et al., 2024); HiSpatial-RGB's reported 95.6% leaves little headroom (Liang et al., 2026). We use it as a near-ceiling complementary test and 3DSRBench for broader spatial comparisons. The raw/refined CV-Bench-3D correct counts are 1,154/1,158 correct out of 1,200, compared with 1,150 (95.8%) for VST-7B-RL and 1,147 (95.6%) for HiSpatial-RGB in the shared-evaluator scenario. The small remaining error counts make CV-3D a near-ceiling complementary test. Full-benchmark ablations reuse the same E endpoint; D1000 supplies separate attribute and joint diagnostics.

Table 1 presents the shared-evaluator comparison. The comparison fixes test inputs and scoring, while preserving native reasoning, verification and tool use. Training histories and inference budgets retain their method-specific settings. Appendix C.1 specifies shared scoring and method-specific inference.

4.2 What do CS-OPSD and GF-RL contribute?

The stage comparison tests whether supervision of student-generated states and numerical attribute feedback improve spatial answering. A is structured SFT; B adds CS-OPSD, while C applies GF-RL directly after A. D, E and G share B and the same raw references: D uses answer-only RL, E uses GF-RL, and G adds attribute-token CE to answer-only RL. E–D tests the attribute-reward design as a whole, including its temporal assignment. E–G tests its value beyond supplying the same attribute labels through token supervision. Full-benchmark QA measures answering, while D1000 attributes and joint correctness test whether any benefit also includes better geometry.

Table 1: **Shared-evaluator comparison.** All models use the benchmark populations and aggregation specified below.

		3DSRBench-real					CV-Bench-3D		
Model	Venue	Height	Loc.	Orient.	Multi	Avg.	Depth	Dist.	Avg.
Qwen2.5-VL-7B-Instruct	NeurIPS (2025b)	44.1	62.7	40.6	40.5	48.4	82.5	83.2	82.8
SpatialReasoner	NeurIPS (2025b)	52.5	75.2	55.2	51.8	60.3	87.3	73.3	80.3
VST-7B-SFT	ECCV (2026)	50.3	68.8	46.9	46.8	54.6	96.8	94.3	95.6
VST-7B-RL	ECCV (2026)	55.8	74.3	52.4	52.1	60.0	97.0	94.7	95.8
Qwen3-VL-8B-Thinking (CoT)	ACL (2026)	56.8	72.5	50.2	56.7	60.6	95.0	90.5	92.8
VERDICT (default)	ECCV (2026)	54.1	73.0	49.0	52.7	58.9	86.0	78.7	82.3
HiSpatial-3B-RGB	CVPR (2026)	58.0	78.0	52.0	61.0	64.4	96.5	94.7	95.6
PolyV (8.4B active)	CVPR (2026)	60.0	78.0	57.0	58.0	64.7	98.0	58.0	78.0
Mini-o3 + entropy (7B)	ICML (2026)	56.5	77.0	52.0	59.0	63.2	85.0	72.7	78.8
SGT-3D (E)	Ours	60.6	85.3	56.1	64.6	69.2	96.7	95.7	96.2
SGT-3D + depth refinement	Ours	61.3	85.9	56.1	65.2	69.7	97.0	96.0	96.5

Accuracy (%), higher is better. Averages are computed from unrounded subset counts: 690/1,709/1,036/1,722 forms for 3DSRBench and 600/600 questions for CV-3D. VERDICT retains default verification; Mini-o3 retains tools but uses one episode per rendering here. Native inference costs and training/selection histories differ.

A crossed comparison uses continued structured SFT or CS-OPSD before answer-only RL or GF-RL, starting from the same SFT checkpoint with the same data and references. It measures each signal's effect under both alternatives for the other stage. The difference between these conditional effects measures interaction (Appendix C.3).

Table 2: Stage and shared-B controls.

Variant	QA_F	$Attr._D$	$Joint_D$
A: SFT	53.8	58.0	48.0
B: SFT+CS-OPSD	63.0	70.0	60.5
C: SFT+GF-RL	61.9	74.9	62.3
D: B+Ans.RL	65.4	72.0	63.5
E: B+GF-RL	69.2	85.8	78.4
G: B+Ans.RL+Attr.CE	66.6	82.0	72.5

F: full 3DSRBench-real (5,157 forms); D: independently annotated D1000 (1,000 canonical questions). A/B/C are stages; D/E/G share B. Attribute and joint metrics assess centers, dimensions and orientations.

Table 3: Crossed stages and reward objectives.

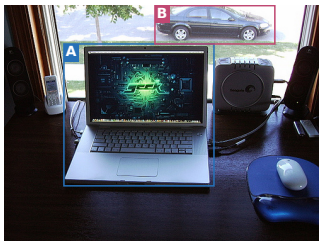
Stage	Objective	QA_F	$Attr._D$	$Joint_D$
SFT	Answer RL	61.5	61.4	49.9
SFT	GF-RL	63.9	74.4	62.8
CS-OPSD	Answer RL	65.4	72.0	63.5
CS-OPSD	GF-RL	69.2	85.8	78.4

F: full-benchmark QA; D: D1000 attribute/joint correctness. Shared SFT start, data and references; CS-OPSD rows reuse D/E, and continued-SFT/GF-RL reuses its endpoint. Attribute and joint metrics assess centers, dimensions and orientations.

E–D yields +3.8 full-benchmark QA points, with +13.8 attribute and +14.9 joint points on D1000. This contrast tests numerical attribute rewards together with their temporal assignment, while E–G compares this design with direct attribute-token supervision.

The full-benchmark attribute-reward QA increments are 2.4 and 3.8 points after continued SFT and CS-OPSD, respectively. The interaction is +1.4 points. These differences summarize the two conditional effects.

Qualitative analysis. Figure 4 illustrates the distinction between estimating spatial attributes and using them. The height and depth examples show how changing the estimated attribute order changes an otherwise consistent comparison. In (a), image-plane vertical order differs from the annotated physical-height order; in (b), a larger projected object is farther from the camera. By contrast, (c) fixes all centers: comparing only horizontal gaps selects a different object from using full 3D Euclidean distance. This comparison isolates whether a continuation uses the supplied geometry. Together, the examples connect GF-RL to the quality of predicted attributes and CS-OPSD to computation from committed states, making the two supervision roles visible at the response level.



(a) Height
GT: A. laptop on the table

Q: Consider the real-world 3D locations of the objects. Which object has a higher location?
A. laptop on the table; B. black car.

CS-OPSD + answer-only RL
The estimated heights are h(A) = 0.28 and h(B) = 0.70. Since 0.70 > 0.28, the car has the higher location.
Answer: B. black car

SGT-3D (ours)
The estimated heights are h(A) = 0.65 and h(B) = 0.20. Since 0.65 > 0.20, the laptop has the higher location.
Answer: A. laptop on the table

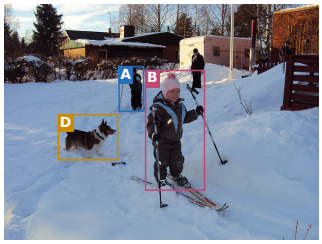


(b) Camera depth
GT: B. plane

Q: Consider the real-world 3D location of the objects. Which object is closer to the camera?
A. boat; B. plane.

CS-OPSD + answer-only RL
The estimated depths are z(A) = 0.28 and z(B) = 0.70. The smaller depth is 0.28, so the boat is closer to the camera.
Answer: A. boat

SGT-3D (ours)
The estimated depths are z(A) = 0.80 and z(B) = 0.35. The smaller depth is 0.35, so the plane is closer to the camera.
Answer: B. plane



(c) Relative distance
GT: B. boy with pink hat

Q: Consider the real-world 3D locations of the objects. Which is closer to the dog, the boy with black hat or the boy with pink hat?
A. boy with black hat; B. boy with pink hat.
Shared fixed prefix: D = (0.20, 0.15, 0.45); A = (0.45, 0.45, 0.95); B = (0.65, 0.45, 0.42). D denotes the dog.

Mask + GF-RL
The horizontal gaps from the dog are |Δx(D,A)| = 0.25 and |Δx(D,B)| = 0.45. Since 0.25 < 0.45, the boy with black hat is closer.
Answer: A. boy with black hat

SGT-3D (ours)
Using Euclidean distance in 3D, d(D,A) = 0.634 and d(D,B) = 0.542. Since 0.542 < 0.634, the boy with pink hat is closer.
Answer: B. boy with pink hat

Figure 4: **Qualitative comparisons on 3DSRBench-real.** (a,b) Height and camera-depth comparisons; h denotes height and z camera depth. (c) Continuations from identical supplied centers compare horizontal gaps with full 3D Euclidean distances. Boxes identify question referents; red/green answers disagree/agree with the official label. Photographs and QA come from 3DSRBench (Ma et al., 2025a); responses and continuous 3D values are author-designed illustrations, not measured outputs.

4.3 WHICH GF-RL DESIGN CHOICES MATTER?

We first compare boundary credit with sequence broadcast and segment credit from the same B checkpoint, keeping numerical reward events, weights and normalization fixed. A 2×2 comparison then crosses object–field versus attribute-type-pooled normalization with segment versus boundary credit (Appendix B.3). This separates each design choice and their interaction. The pooled control preserves individual reward events and the answer's separate statistics. These controls adapt normalization and segmentation within our rewards; they do not reproduce SCOUT's complete algorithm.

Table 4: Credit assignment at fixed rewards from B.

Credit assignment	QA_F	$Attr._D$	$Joint_D$	ΔQA_F
Sequence broadcast	66.1	80.7	71.1	+3.1
Segment credit	66.2	80.9	71.6	+3.0
E: GF-RL (boundary credit)	69.2	85.8	78.4	n/a

F: full-benchmark QA; D: D1000 attribute/joint correctness. Initialization and references are fixed. Broadcast, segment and E share reward events, weights and normalization. ΔQA is E minus the row. Attribute and joint metrics assess centers, dimensions and orientations.

Boundary credit has a +3.0 full-QA margin over segment credit at fixed field normalization. Field normalization adds +0.6/+1.4 points over type pooling under segment/boundary credit (Table 20). The crossed comparison separates the two design choices.

4.4 Does committed-state supervision improve attribute use?

We compare plain conditioning, expert priority, conflict masking and CS-OPSD under shared SFT, raw references and GF-RL. CS-OPSD follows completed student fields and uses the expert for unfinished fields. Masking removes reference-conflicting continuation targets; its rescaled control compensates for reduced supervision (Appendix B.2).

Table 5: Teacher policy and complete student answers.

Policy	OPSD QA_F	RL QA_F	Attr.$_D$	Joint$_D$	ΔQA_F
Plain conditioning	61.1	64.3	76.0	65.1	+4.9
Expert priority	61.0	64.1	75.5	64.4	+5.1
Conflict-selective mask	62.4	66.0	81.2	71.0	+3.2
CS-OPSD	63.0	69.2	85.8	78.4	n/a

F: full 3DSRBench-real (5,157 forms). D: the independently annotated D1000 cohort. Raw references, SFT and the GF-RL recipe are shared. ΔQA is CS-OPSD minus the row after RL. Attribute and joint metrics assess centers, dimensions and orientations.

CS-OPSD exceeds ordinary/rescaled masking by 3.2/2.5 full-QA points (Appendix D.5). Retaining Think while masking only conflicting Answer KL gives +2.8/+2.2 full-QA points over ordinary/rescaled continuation masking (Table 19). Full CS-OPSD adds 0.4/0.3 points. The main difference comes from preserving computation supervision.

Frozen-stratum, supervision-amount and prompt controls are detailed in Appendix C.3; identical supplied-state tests are in Appendix C.4. These diagnostics compare conditional attribute use on fixed supplied states.

4.5 Reference quality and robustness

We test the net value of retaining conflicting continuation supervision under added training-reference errors. Perturbations change both the teacher's guidance and GF-RL's numerical targets. Training-reference centers are perturbed at fixed scene-normalized levels, sharing each corrupted cache across the teacher, checker and compared policies while keeping evaluation labels clean (Appendix C.4).

Table 6: Full-benchmark QA under training-reference noise.

Noise SD	E QA	Mask QA	Rescaled QA	E–mask	E–rescaled
0.00	69.2	66.0	66.7	+3.2	+2.5
0.05	68.3	65.5	66.1	+2.8	+2.2
0.10	63.1	63.7	64.0	-0.6	-0.9

Full 3DSRBench-real, using the same 5,157-form protocol as Table 1. Each noise setting changes the shared training-reference centers; evaluation labels stay clean. Noise SD is scene-normalized and does not estimate the natural expert-error distribution.

The comparison retains gains at zero and moderate added noise and a reversal at the largest setting. This dependence concerns the shared reference used by both training stages.

5 Conclusion

SGT-3D trains an RGB-only student to predict 3D attributes and reason from them. CS-OPSD supervises complete responses on student-generated prefixes, preserving computation from completed attributes. GF-RL provides numerical attribute and answer feedback with credit at field completion. Matched teacher, reward and credit controls connect these two supervision choices to spatial answering and independent attribute assessment.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

Shuai Bai et al. Qwen2.5-vl technical report, 2025.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian 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 (CVPR)*, pp. 14455–14465, June 2024.

Yiming Chen, Zekun Qi, Wenyao Zhang, Xin Jin, Li Zhang, and Peidong Liu. Reasoning in space via grounding in the world. *arXiv preprint arXiv:2510.13800*, 2025.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jiho Choi, Seonho Lee, Seojeong Park, and Hyunjung Shim. Dense reward for multi-view 3D reasoning with global maps and local views. In *European Conference on Computer Vision (ECCV)*, 2026.

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11(67):2001–2049, 2010.

Yanbin Hu, Jin Cui, Jun Ye, Jiepeng Zhou, Jiangcheng Song, Boran Zhao, and Pengju Ren. Distill what RGB can recover: Privileged 3D evidence for RGB-only vision-language models, 2026.

Weikai Huang, Jieyu Zhang, Sijun Li, Taoyang Jia, Jiafei Duan, Yunqian Cheng, Jaemin Cho, Matthew Wallingford, Rustin Soraki, Chris Dongjoo Kim, Shuo Liu, Donovan Clay, Taira Anderson, Winson Han, Ali Farhadi, Bharath Hariharan, Zhongzheng Ren, and Ranjay Krishna. WildDet3D: Scaling promptable 3D detection in the wild, 2026. Version 2.

Sai Srinivas Kancheti, Aditya Sanjiv Kanade, Vineeth N. Balasubramanian, and Tanuja Ganu. Chain-of-thought degrades visual spatial reasoning capabilities of multimodal LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 862–876, 2026. doi: 10.18653/v1/2026.acl-short.71.

Simran Kaur, Narutatsu Ri, Yinghui He, Liam Fowl, and Sanjeev Arora. Rethinking on-policy self-distillation for thinking models, 2026. Version 1, July 6, 2026.

Dong-Hee Kim, Reuben Tan, and Donghyun Kim. Diversity over frequency: Rethinking tool use in visual chain-of-thought agents. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Aristotelis Lazaridis, Dylan Bates, Aman Sharma, Brian King, Vincent Lu, and Jack FitzGerald. EDGE-OPD: Internalizing privileged context with evidence guided on-policy distillation, 2026. Version 1, May 22, 2026.

Huizhi Liang, Yichao Shen, Yu Deng, Sicheng Xu, Zhiyuan Feng, Tong Zhang, Yaobo Liang, and Jiaolong Yang. HiSpatial: Taming hierarchical 3d spatial understanding in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023.

Xiaogeng Liu, Xinyan Wang, Yingzi Ma, Yechao Zhang, and Chaowei Xiao. When are teacher tokens reliable? position-weighted on-policy self-distillation for reasoning, 2026. Version 1, May 20, 2026.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6924–6934, October 2025a.

Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. SpatialReasoner: Towards explicit and generalizable 3D spatial reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025b.

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.

Rohit Sinha, Kunal Tilaganji, Tanuja Ganu, Nagarajan Natarajan, Amit Sharma, and Vineeth Balasubramanian. Verdict: Training-free step-wise verification of multimodal reasoning via disagreement-aware consensus. In *Computer Vision – ECCV 2026*, volume 17016 of *Lecture Notes in Computer Science*, pp. 500–517. Springer, 2026. doi: 10.1007/978-3-032-37383-0_29.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 87310–87356, 2024.

Sihan Wang, Xiyao Liu, Lianqing Liu, and Zhi Han. Seeing before reasoning: Decoupling perception and reasoning for shortcut-resilient multimodal on-policy self-distillation, 2026. Version 2, 19 June 2026.

Yuxin Wang, Lei Ke, Boqiang Zhang, Tianyuan Qu, Hanxun Yu, Zhenpeng Huang, Meng Yu, Dan Xu, and Dong Yu. N3d-vlm: Native 3d grounding enables accurate spatial reasoning in vision-language models. *arXiv preprint arXiv:2512.16561*, 2025.

Shengqiong Wu, Lanhu Wu, Mingyang Bao, Wenhao Xu, Hanwang Zhang, Shuicheng Yan, Hao Fei, and Tat-Seng Chua. Modeling cross-vision synergy for unified large vision model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22941–22952, June 2026.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. In *European Conference on Computer Vision (ECCV)*, 2026. Numerical results from arXiv version 1, 7 November 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026.

Enshen Zhou, Yibo Li, Jingkun An, Jiayuan Zhang, Shanyu Rong, Mengzhen Liu, Yi Han, Yuheng Ji, Huajie Tan, Jiawei He, Pengwei Wang, Zhongyuan Wang, Cheng Chi, Lu Sheng, and Shanghang Zhang. Towards spatial trace with reasoning in vision-language models for robotics, 2026a. Version 4.

Zile Zhou, Huining Yuan, Weichen Zhang, Xinlei Chen, and Xiao-ping Zhang. SCOUT: Unlocking enhanced spatial reasoning via structured chain-of-thought and multi-objective process reward, August 2026b. Version 1, 12 August 2026.

A Representation and Optimization

A.1 Object correspondence and coordinate conventions

Scan binds question-role identifiers to normalized 2D boxes; Gaze serializes the corresponding object-box centers $\mathbf{c}_i$ and local-axis dimensions $\mathbf{s}_i = (L_i, H_i, W_i)$ and scalar-first orientation quaternions $\mathbf{r}_i$. Each quaternion represents the rotation from local box axes to camera axes. A predicted binding must match the fixed identifier and have at least 0.5 IoU with its reference box. Matching never minimizes predicted 3D error. Gaze includes center, dimension and orientation predictions. Fixed trusted boxes prompt frozen WildDet3D (Huang et al., 2026); they are not student predictions. Its monocular route uses predicted depth and, when calibration is unavailable, predicted intrinsics, with measured-depth conditioning disabled. Raw expert inference receives no answer.

The inspected decoder at revision `1b8aa52b6ff3` returns $(c_x, c_y, c_z, W, L, H, q_w, q_x, q_y, q_z)$ in OpenCV camera axes (x right, y down, z forward). Canonical rotation must be enabled; it enforces $W \le L$ through a joint dimension/rotation transformation. The corner implementation uses local-axis lengths (L, H, W). The manuscript retains the normalized scalar-first quaternion as an orientation target. A single full-scene depth scale is fixed before object selection or sampling:

$$\alpha = \mathrm{median}_{u \in V}\, D_E(u), \qquad \mathbf{c}_i^E = \mathbf{c}_i^{E,\mathrm{raw}}/\alpha, \qquad \mathbf{s}_i^E = (L_i, H_i, W_i)/\alpha. \tag{6}$$

Here V excludes padding and nonpositive/nonfinite depths. All objects share this scale; none is centered or rescaled separately. Camera-up is not world-up, and canonical boxes do not define semantic front/back. World-height questions require an independent gravity direction.

Reference admission checks trusted prompt IDs, prompt-box IoU of at least 0.5, confidence floors of 0.3 (2D) and 0.1 (3D), finite positive dimensions/depth, a nonzero finite quaternion and canonical dimension order. Duplicate or unknown IDs are manifest errors. Missing or rejected fields remain unavailable across all rollouts. The data record retains prompt correspondence, intrinsics and resize/padding transforms. Independent labels and SFT targets use the same box semantics; visible-surface summaries cannot substitute for box centers or dimensions.

Optional reference refinement. The default keeps valid raw references, $Z^R = Z^E$. The separate refinement variant uses an explicit image-level bank of trusted training camera-depth relations with fixed identities. With raw center depths z_i^E, it solves

$$\widetilde{z} = \arg\min_z \frac{1}{2} \sum_i (z_i - z_i^E)^2 \quad \text{s.t.} \quad z_j - z_i \ge \delta \;\forall (i,j) \in \mathcal{E}_I, \quad z_i \ge \ell_i, \tag{7}$$

where $\ell_i = z_{\min} + \frac{1}{2}\sum_k |R_{i,3k}| s_{i,k}$ keeps every oriented-box corner in front of the near plane. Identities, lateral centers, dimensions, rotations and scale stay fixed. Accept only a feasible joint solution with $\max_i |\widetilde{z}_i - z_i^E| \le b$; contradictory constraints, missing objects and unsupported relation banks cause abstention. Compatible scenes remain unchanged. Filter instead rejects any scene violating the margin. Rejection removes attribute supervision from both stages but retains answer-based RL. Both teacher and checker use the same admitted reference, frozen before sampling.

The margin δ, clearance $z_{\min}$, shift bound b and solver tolerance are fixed on training/validation data. The tolerance obeys $0 < \epsilon_{\mathrm{sol}} < \min(\delta, z_{\min})/2$; feasibility, stationarity and complementarity are checked independently. "Nearer" denotes box-center camera depth, not radial distance or nearest visible surface. Adjusted coordinates are answer-conditioned pseudo-labels, even though answer strings and audit metadata never enter the teacher context. Positive margins can move already correctly ordered objects. Held-out answers never enter this preprocessing.

The small-gap example shows why a relation-compatible reference can still disagree with independent geometry.

A.2 Attribute rewards and parsing

Center rewards follow Equation 4. For positive dimensions and normalized nonzero quaternions,

$$\begin{aligned} d_{k,i}^{\mathrm{size}} &= \frac{1}{3} \| \log \hat{\mathbf{s}}_{k,i} - \log \mathbf{s}_i^{\mathrm{ref}} \|_1, \\ d_{k,i}^{\mathrm{ori}} &= \min_{S \in \mathcal{S}_i} \arccos\left(\mathrm{clip}\left(\frac{\mathrm{tr}(\hat{R}_{k,i}^\top R_i^{\mathrm{ref}} S) - 1}{2}, -1, 1 \right) \right). \end{aligned} \tag{8}$$

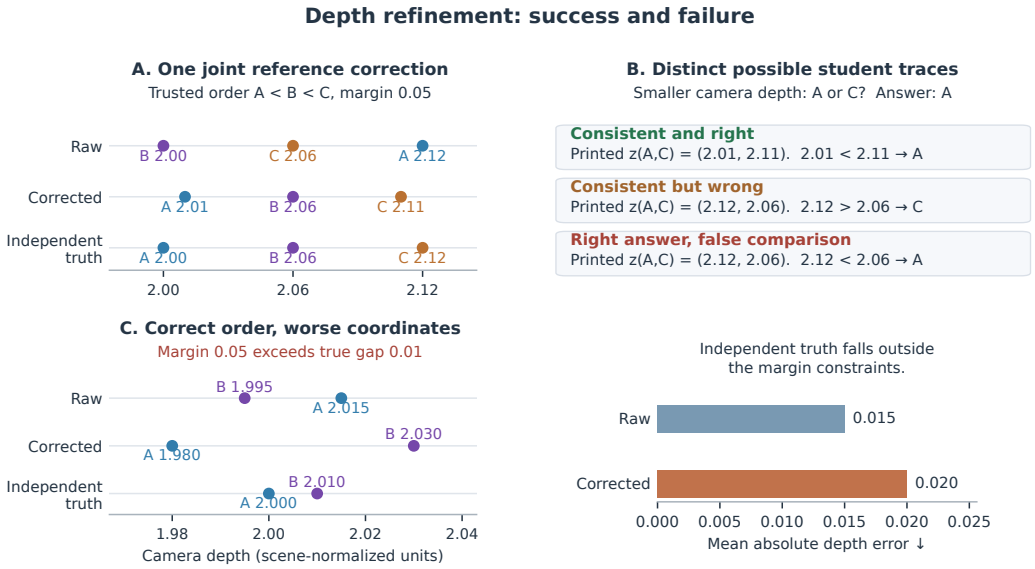


Figure 5: **Success and failure of depth refinement.** Fixed object identities and joint constraints define the geometric calculations. The failure uses a true depth gap smaller than the imposed margin, so the independent truth lies outside the constraint set. The examples use a shared camera-coordinate convention.

Here R_i maps local box axes to camera axes, and $\mathcal{S}_i$ contains proper local-axis rotations preserving the reference cuboid and dimension labels. This includes 180° axis flips and, for equal side lengths, corresponding permutations. The angular error is in radians and is invariant to quaternion sign. Size rewards use $\exp(-d^{\text{size}}/\tau_s)$. One complete quaternion is one Gaze field and one reward event. Zero or nonfinite quaternions receive zero orientation reward when a reference is available. The center/size configuration uses $\tau_c = 0.1$ and $\tau_s = 0.2$. Orientation uses $r^{\text{ori}} = \exp(-d^{\text{ori}}/\tau_R)$, with $\tau_R > 0$ fixed on training/validation data before comparison. The terminal reward is $r_k^{\text{answer}} = \mathbb{1}[\nu(\hat{a}_k) = \nu(a^*)]$, with benchmark option mapping before canonical string comparison and surrounding whitespace removed. Selection and computation receive no separate rewards; Think arithmetic is evaluated separately.

The parser accepts ordered JSON-lines records: Scan, Gaze, Think, Answer, with one object–field assignment per Gaze record. A boundary is the first output token containing that record's closing delimiter, excluding trailing whitespace. The first Scan binding and first field assignment are immutable. Unmatched bindings and invalid numerical values receive zero at their field boundaries. At the first duplicate assignment, malformed JSON, duplicate JSON key or stage-order violation, the checker stops accepting new records and assigns zero terminal reward. Previously completed attributes retain their rewards and boundaries; uncompleted reference targets receive zero at the actual response end. Generation is not truncated. Missing stages remain incomplete, and later text cannot rewrite committed fields. CS-OPSD still scores the original complete response.

For the same fixed target set across K responses, normalize each object–field and the answer independently:

$$\mu_j = K^{-1}\sum_k r_{k,j}, \qquad \sigma_j^2 = K^{-1}\sum_k (r_{k,j}-\mu_j)^2,$$
$$\widetilde{r}_{k,j} = \omega_j \frac{r_{k,j}-\mu_j}{\max(\sigma_j, 10^{-8})}. \tag{9}$$

For M available attributes, each has weight $1/M$ and the answer has weight one. If $M = 0$, only the answer remains. Missing references are omitted; invalid predictions remain as zero-valued events. Constant columns return exactly zero, including all-invalid groups. Adding attribute events leaves the normalized answer event unchanged for identical rollouts, although clipping can still couple

Table 7: Three-stage optimization and the training/inference boundary.

Step	Procedure
References	Freeze the expert, prompt identities, admitted Z^R and available fields before student sampling. Optional refinement uses training relations only.
SFT	Fit structured target responses from RGB/questions to obtain $\theta^{(1)}$.
CS-OPSD	Initialize from $\theta^{(1)}$, sample one student response and compare student/privileged teacher distributions at identical prefixes. Update only student scoring.
GF-RL	Initialize from CS-OPSD, generate K RGB-only responses, check attributes against the same Z^R and check the training answer. Apply separately normalized reward-to-go and clipped optimization.
Inference	Generate Scan, Gaze, Think and Answer from RGB and the question. The expert, privileged teacher and training verifier are removed.

their gradient contributions. Completed attribute rewards do not reach later Think tokens under reward-to-go.

A.3 Optimization and information access

Structured SFT fits validated grounded traces by response-token cross-entropy:

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{(x,y^{\mathrm{sup}})} \frac{1}{T} \sum_{t=1}^{T} \log p_\theta(y_t^{\mathrm{sup}} \mid x, y_{<t}^{\mathrm{sup}}). \tag{10}$$

Input and padding positions are excluded; text attributes require no regression head. CS-OPSD then samples one RGB-only response, scores identical prefixes with shared student/self-teacher weights, and detaches both the response and teacher targets. Full-vocabulary forward KL covers every output token, including the answer. The admission flag is fixed before sampling: rejected examples have zero distillation loss, and the batch mean retains all sampled QA examples. Teacher weights refresh at the next iteration; an EMA or lagged teacher would define a different update. This surrogate does not differentiate through sampling or equal KL under the teacher's trajectory distribution.

The rollout law s^{CS} uses the temperature and top-p/top-k settings in Table 8. Both teacher and student KL distributions use temperature one over the complete vocabulary, without truncation. GF-RL samples the untruncated temperature-one policy, also used for old/new likelihoods. LoRA dropout is enabled only during SFT. Rollout generation and CS-OPSD/GF-RL scoring disable dropout, while student scoring retains gradients. These are the execution conventions.

For GF-RL, $\rho_{k,t} = \pi_\theta(\hat{y}_{k,t} \mid x, \hat{y}_{k,<t}) / \pi_{\mathrm{old}}(\hat{y}_{k,t} \mid x, \hat{y}_{k,<t})$. The rollout denominator, rewards and advantages remain detached:

$$\mathcal{L}_{\mathrm{RL}} = -\frac{1}{K} \sum_{k=1}^{K} \frac{1}{T_k} \sum_{t=1}^{T_k} \min \big(\rho_{k,t} \widehat{A}_{k,t}, \\ \mathrm{clip}(\rho_{k,t}, 1-\varepsilon, 1+\varepsilon) \widehat{A}_{k,t} \big) + \beta \mathcal{R}_{\mathrm{KL}}. \tag{11}$$

Token losses are averaged within each response, then across responses. The reference objective uses $\varepsilon = 0.2$ and $\beta = 0$. An enabled KL term would require a separately designated reference policy, distinct from the rollout policy and self-teacher. There is no replay or additional alignment objective. The reproducibility record requires backbone revision, trainable parameters, SFT provenance, optimizer, batch/rollout sizes, decoding, updates, seeds and checkpoint selection.

A.4 Reference-component checks

The normalization experiment tests whether adding attribute rewards changes an otherwise identical answer signal. We generate 2,000 shared groups of eight Bernoulli(0.5) answer rewards and vary the number of attribute targets over $\{0, 1, 4, 16, 64\}$. Five attribute conditions use constant zero, constant one, or $\exp(-e)$ with exponential normalized error of mean 0.25, 1, or 4. Per-target normalization in Equation 9 preserves every normalized answer event exactly across all 25 conditions. Pooling attribute and answer events changes the answer signal: at 64 targets with exponential error of mean one, its mean absolute change is 0.7961, and constant-zero targets produce a change of 5.5083 (Figure 6). This numerical result removes a normalization confound from answer-only/GF-RL comparisons.

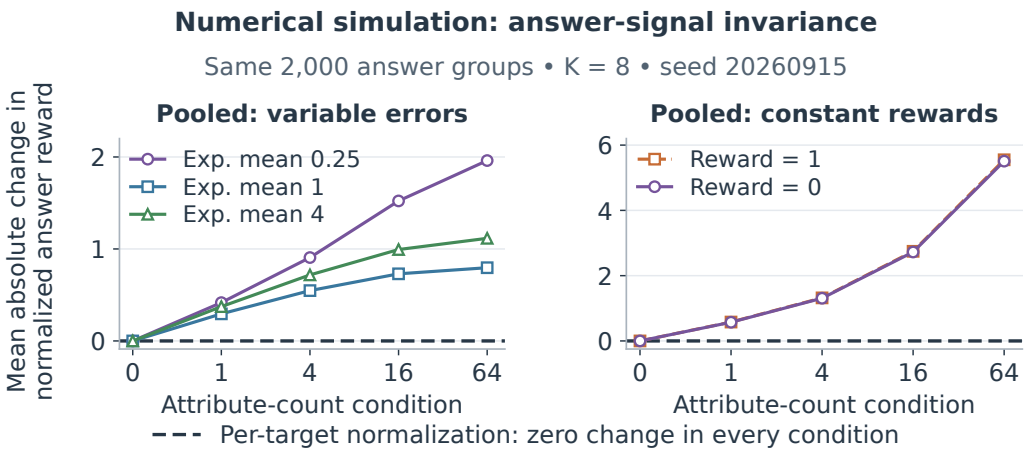


Figure 6: **Answer-signal stability in a controlled numerical experiment.** Mean absolute change from normalizing the same answer groups alone. Pooled normalization uses one population across answer and attribute events. Per-target normalization keeps them separate. Panels use different vertical scales. The five attribute conditions and five target counts share the same 2,000 answer groups. These are synthetic reward calculations, not VLM accuracy results.

The WildDet3D decoded-array adapter is checked on constructed asymmetric boxes, fixed prompt identities, quaternion-equivalent rotations, projected corners, and existing-checker integration. The depth preprocessor additionally checks analytic two-object projections, joint chains and branches, unchanged feasible scenes, rotated-box clearance, rejected cycles and excessive shifts, solver certificates, and consistency of updated metadata. Other checks cover parsing, normalization, arithmetic, and detached objective gradients. The checks use constructed inputs with known expected outputs.

The normalization experiment uses NumPy 1.26.4, generator PCG64, and seed 20260915. All 25 conditions share one 2000×8 binary answer array. Attribute rewards are constant zero, constant one, or $\exp(-e)$ for exponential e with the specified mean. The pooled comparator centers and scales all raw events in a group with population standard deviation and floor 10^{-8}, then retains the unit-weight answer term. Equation 9 instead uses only the answer column for that term. Eighteen groups have constant answers. Pooled conditions with exponential error of mean one assign nonzero answer terms to all 144 positions in these groups, whereas per-target normalization retains zero.

Table 8: **Literature-informed training recipe.** The settings draw on SpatialReasoner, ViGOS, OPSD and SCOUT (Ma et al., 2025b; Wang et al., 2026; Zhao et al., 2026; Zhou et al., 2026b).

Setting	Choice
Hardware	$8\times$ NVIDIA H200, 141 GB per GPU.
Initial model	`Qwen/Qwen2.5-VL-7B-Instruct`.
Trainable weights	LoRA adapters: rank 64, $\alpha = 128$, SFT dropout 0.05; base weights frozen. The same adapter configuration is used for all three stages.
Target modules	`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`.
Optimizer	AdamW with bf16 training.
SFT	Learning rate 5×10^{-6}; fixed budget 15,000 optimizer updates.
CS-OPSD	Learning rate 5×10^{-6}; fixed budget 1,500 updates; one rollout per question; temperature 1.1, top-$p = 0.95$, top-$k = 20$.
GF-RL	Learning rate 10^{-6}; fixed budget 3,750 updates; $K = 8$ rollouts per question; temperature 1.0 with no sampling truncation.
Training generation	At most 4,096 generated tokens in each rollout stage; the same limit applies to CS-OPSD and GF-RL.
SGT-3D evaluation	One greedy response per form/option order, at most 2,048 generated tokens; retain the official benchmark aggregation.

A.5 Training data construction and resource budget

We use SpatialReasoner's training datasets with independently regenerated spatial labels (Ma et al., 2025b). The official `ccvl/SpatialReasonerTrain-SFT` annotation split has

48,000 records: 24,000 SR-CoT examples and 24,000 general LLaVA conversations. The `ccvl/SpatialReasonerTrain-RL` split has 1,200 SR-QA records. The source releases provide training records, not a separate validation split. Our additional CS-OPSD stage shares the reannotated SR-QA bank with GF-RL. Counts below presume that relabeling preserves source membership; any later exclusion must be recorded and its effective pass count recomputed.

Source preservation and preprocessing. Retain source image/question IDs and the original 24K/24K SFT mixture. The two spatial training banks share some images; split auditing therefore uses image identities and reviewed duplicates. Audit image hashes and reviewed duplicates against evaluation for both the spatial and LLaVA components. Select validation-based settings using a separate non-test manifest. Decode available images as RGB and use the Qwen processor's aspect-preserving resize, capped at $1{,}280 \times 28^2$ pixels (Bai et al., 2025). Transform object regions consistently, use no random crop or flip, and preserve text-only LLaVA records.

Relabeling and trace provenance. Regenerate trusted object bindings and supported center, dimension and quaternion-orientation targets using the frozen WildDet3D route under our camera-frame and scene-scale convention. Record the original question and answer, newly acquired references, label source and any reviewed answer correction. Expert-generated geometry remains a pseudo-reference. Do not infer semantic orientation or gravity-relative labels from centers and dimensions alone. Missing fields and unsupported operators retain their declared masks and diagnostic exclusions.

For spatial SFT records, serialize Scan/Gaze/Think/Answer traces from the reannotated fields and supported operations, checking arithmetic after numerical serialization. Validate the available supervised fields and retain answer supervision when a geometric field is unavailable. Freeze any answer correction before all comparisons; reference-noise controls never recompute answer labels. General LLaVA conversations retain their original assistant targets and format, supervised only by assistant-token cross-entropy during SFT. These general conversations are used only in SFT. Each conversation remains one training record without packing or expansion into separate QA rows. The 1,200-question bank also requires validated supervised traces for matched continued-SFT and attribute-CE controls. The reannotation manifest must identify source revision, record ID, object mapping, reference and answer provenance, available-field mask, trace validation and exclusions.

Table 9: **Fixed training budget.** All stages use eight NVIDIA H200 GPUs (141 GB each) and 32 records per optimizer update. Record counts follow the official SpatialReasoner annotation datasets; passes assume reannotation preserves membership.

Stage	Records	Passes	Responses/record	Accumulation	Updates
SFT	48,000	10	1	4	15,000
CS-OPSD	1,200	40	1	4	1,500
GF-RL	1,200	100	8	32	3,750

SFT mixes 24,000 spatial records with 24,000 general LLaVA conversations; a conversation is one record, not one image or necessarily one QA turn. CS-OPSD and GF-RL share 1,200 reannotated SR-QA records. Accumulation is 4 for SFT/CS-OPSD and 32 for GF-RL at one record/response per GPU microbatch. GF-RL processes 256 responses per update and 960,000 in total. Carry partial batches across source passes without dropping or padding records.

Fixed training budget. With a global batch of 32 records, prescribe $48{,}000 \times 10/32 = 15{,}000$ SFT updates, $1{,}200 \times 40/32 = 1{,}500$ CS-OPSD updates and $1{,}200 \times 100/32 = 3{,}750$ GF-RL updates. SFT and RL adopt SpatialReasoner's 10- and 100-pass scales; the additional 40-pass CS-OPSD stage is our allocation. This intermediate allocation produces 48,000 question presentations before RL. The resulting total is 20,250 updates, including 5,250 after SFT. These batch and sampling conventions define the fixed SGT-3D schedule.

Carry partial question batches across shuffled source passes, without per-pass dropping or padding. GF-RL forms each question's complete eight-response group before microbatching and performs one update per rollout batch, giving 120,000 question presentations and 960,000 sampled responses. Each stage uses its terminal checkpoint for the next stage and primary comparisons. No pilot-dependent extension, early stopping or validation-selected checkpoint is part of this fixed schedule.

Memory and matched expenditure. Allocate one process to each of eight NVIDIA H200 GPUs with 141 GB per device, with no additional rollout GPU. Use activation checkpointing and one

record/response per backward microbatch. Score teacher and student sequentially; compute exact full-vocabulary KL in 256-position chunks at the output head and probability computation, with fp32 loss statistics. Avoid full sequence-by-vocabulary logits or concurrent retention of all chunk graphs. Reserve 4,096 completion tokens within an 8,192-token branch context cap after visual-token expansion. Apply one common admission check across all compared prompts and controls; any source-record exclusion changes the retained count and must be reported. Do not truncate reference fields or filter variants independently. Generated truncations follow the existing failure rules.

Teacher-policy controls share 1,500 intermediate and 3,750 subsequent GF-RL updates on the same 1,200-question bank. D/E/G and normalization–credit controls share B, question order, batch 32, $K = 8$ and 3,750 updates. G's extra attribute-CE pass adds compute. Continued-SFT and attribute-CE warm starts use 1,500 intermediate updates on that bank. Reverse order uses 3,750 GF-RL then 1,500 CS-OPSD updates; longer single-stage controls use 5,250 updates. All primary endpoints use terminal checkpoints. The matched budget fixes optimizer updates and question presentations; generated tokens and runtime are method-specific.

Limitations and future work. SGT-3D focuses on object-level spatial reasoning from single RGB images using normalized 3D box attributes. Training supervision depends on the quality and coverage of the geometric references. Extending the framework to dynamic scenes and finer geometric representations is a direction for future work.

B Matched Teacher Prompts and Controls

B.1 Exact prompt and context assembly

Plain conditioning uses the complete shared text below. CS-OPSD and expert priority append their respective blocks. The files are reproduced verbatim; student inputs exclude these teacher-only instructions and reference attributes. Image preprocessing, question bytes, serialized references, object identities and sampled prefixes remain identical across matched conditions.

Shared text.

```
You are scoring the next token of a spatial question-answering trace. You receive
an RGB image, a question, training-only reference object geometry, and the exact
student prefix. The geometry describes object-box centers, local-axis dimensions (L,H,W), and
      scalar-first orientation quaternions
in its declared common camera frame; centers and dimensions share one scene-depth scale.
     Orientations map local box axes to camera axes and are scale-independent. It contains
     neither a final answer nor a target reasoning trace.

Follow Scan -> Gaze -> Think -> Answer as JSON-lines records. Use the question's
object-role identifiers. Each Gaze record contains one object attribute. Think
states the supported operation, argument identifiers, and numerical values;
Answer states the chosen identifier. Do not silently rewrite prefix tokens or emit
duplicate attribute assignments. Unknown gravity does not authorize treating the
camera's vertical axis as world up. Reference geometry may contain errors.
```

CS-OPSD priority block.

```
Use reference geometry to guide fields not yet completed. After Gaze fields are
completed, compute and answer from those printed student values verbatim, even
when they disagree with the reference. Do not substitute reference values for
committed student values in the continuation.
```

Expert-priority block.

```
Use reference geometry to guide fields not yet completed. After Gaze fields are
completed, compute and answer from the reference values, even when they disagree
with the printed student values. This priority may produce a continuation that
conflicts with the committed student state.
```

Variable context payload.

```
RGB: image passed through the backbone image processor
QUESTION: {question}
REFERENCE_GEOMETRY: {reference_json}
ASSISTANT_PREFIX: {exact_student_prefix}
```

The fixed prompt precedes the payload. OPSD scores the sampled next-token position; it generates no second teacher trajectory. Independent frozen-prefix diagnostics decode continuations separately.

Center/size serialized illustration. For "Which object is closer to the camera, a or b?", expert depths are 1.0 and 1.4, while the student prefix reverses them. All fields share the declared camera frame and scene scale; `extent` means (L, H, W) box dimensions.

```
{"stage":"scan","id":"a","box":[0.05,0.1,0.4,0.7]}
{"stage":"scan","id":"b","box":[0.55,0.1,0.9,0.7]}
{"stage":"gaze","id":"a","field":"center","value":[0,0,1.4]}
{"stage":"gaze","id":"a","field":"extent","value":[0.2,0.3,0.2]}
{"stage":"gaze","id":"b","field":"center","value":[0,0,1.0]}
{"stage":"gaze","id":"b","field":"extent","value":[0.2,0.3,0.2]}
```

The committed-state continuation is:

```
{"stage":"think","op":"front","args":["a","b"],"values":[1.4,1.0]}
{"stage":"answer","value":"b"}
```

The expert-priority continuation is:

```
{"stage":"think","op":"front","args":["a","b"],"values":[1.0,1.4]}
{"stage":"answer","value":"a"}
```

The first follows the printed state but answers incorrectly under this example; the second answers correctly using values absent from that state. This isolates the distinction between factual accuracy and trace consistency.

B.2 Conflict masking and teacher comparisons

Conflict masking uses the same committed teacher, raw references and full-vocabulary KL. Let b complete the required Gaze state. For supported binary depth questions, $\chi = 1$ only when expert and printed attributes imply decidable opposing orders under a fixed tie tolerance. Trusted operators and bindings define this test without target answers or independent evaluation attributes. Unsupported, missing, tied or unparseable states have $\chi = 0$ with separate coverage counts:

$$m_t = 1 - \chi \mathbb{1}[t > b], \qquad \ell_{\text{mask}} = \frac{h_I}{T} \sum_{t=1}^{T} m_t D_{\text{KL}}(q_t^{(n)} \| p_{S,t}). \tag{12}$$

The original denominator T is retained. Rescaled masking multiplies this loss by $\min(T/R, c_{\max})$ for $R = \sum_t m_t > 0$ and returns zero otherwise; the cap is fixed on validation data. A separate supervision mask preserves the output-length denominator. Random masking matches removed-token counts within length/operation strata. Prompt paraphrases preserve the priority rule. All variants receive identical subsequent GF-RL; token counts alone do not match retained KL mass or gradient norms.

An unselected SFT stream uses at most one completed-Gaze prefix per question to estimate conflict frequency and exclusions. Independently scored relations distinguish a wrong student from a wrong expert. Frozen-prefix comparisons hold decoding fixed across no attributes, plain conditioning, committed priority and expert priority. They score factual QA, arithmetic, answer–trace consistency and invalids separately. The balanced 3DSR-T400 bank uses 400 prefixes, 100 per stratum, sampled from five SFT responses per original depth question. Repeated questions make this a conditional-behavior bank rather than a prevalence sample. Complete RGB-only students are evaluated after OPSD and after the same GF-RL recipe using full CircularEval/FlipEval; independent attributes and joint correctness retain the D1000 diagnostic scope.

B.3 Separating continuation targets and GF-RL design choices

Retain computation, mask only Answer KL. The continuation mask in Appendix B.2 removes both computation and answer supervision after a conflicting state. To separate their effects, retain the same conflict indicator χ, committed-state teacher and sampled response, but mask only the first parser-accepted Answer span. Let b_{prev} be the preceding accepted record's completion token and b_{ans} the Answer completion token. Set $\mathcal{A}(\hat{y}) = \{t : b_{\text{prev}} < t \le b_{\text{ans}}\}$. This preserves the preceding computation boundary when a token spans two records; any Answer-opening characters sharing that token retain KL. Define

$$m_t^{\text{ans}} = 1 - \chi \mathbb{1}[t \in \mathcal{A}(\hat{y})], \qquad \ell_{\text{ans}} = \frac{h_I}{T} \sum_{t=1}^{T} m_t^{\text{ans}} D_{\text{KL}}(q_t^{(n)} \| p_{S,t}). \tag{13}$$

Scan, Gaze and Think retain their original KL. If no Answer is accepted, or both boundaries share one token, $\mathcal{A}$ is empty and the loss is unchanged. A later parse failure does not erase an already accepted Answer span. An accepted null-valued Answer also supplies a mask span. Malformed responses remain evaluation failures. Conflicts are detected from the trusted operator and opposing student/reference relations, without ground-truth answers.

Compare ordinary and rescaled Answer-only masks with their corresponding continuation masks and full CS-OPSD. Rescaling uses the capped rescaling rule in Appendix B.2 with $R = \sum_t m_t^{\text{ans}}$ and the same validation-fixed cap. Keep the original response-length denominator, training data, optimizer schedule and subsequent GF-RL fixed. The Answer-only comparison isolates retention of computation supervision; the full-continuation comparison removes both Think and Answer targets.

Cross normalization with credit boundaries. Use a 2×2 design from the same CS-OPSD checkpoint: object–field or attribute-type-pooled normalization, crossed with segment or field-completion credit. The raw reward events, target identities, availability, weights and failure rules remain identical. Only the statistics population and token assignment vary. Reward definitions match for a fixed rollout; sampled values can diverge as policies change. For one question, let $\mathcal{J}_c$ contain its available targets of attribute type $c \in \{\text{center}, \text{dimension}, \text{orientation}\}$, with $M_c = |\mathcal{J}_c|$ and $M = \sum_c M_c$. Type pooling uses

$$\mu_c = \frac{1}{KM_c} \sum_{k=1}^{K} \sum_{j \in \mathcal{J}_c} r_{k,j}, \qquad \sigma_c^2 = \frac{1}{KM_c} \sum_{k=1}^{K} \sum_{j \in \mathcal{J}_c} (r_{k,j} - \mu_c)^2,$$
$$\widetilde{r}_{k,j}^{\text{type}} = \frac{1}{M} \frac{r_{k,j} - \mu_c}{\max(\sigma_c, 10^{-8})}, \quad j \in \mathcal{J}_c. \tag{14}$$

Normalize each original event before credit assignment. Do not collapse a response's attributes into one reward. Invalid predictions contribute zero raw reward and stay in the population; missing references do not. Empty types are omitted and $M = 0$ leaves answer-only RL. The answer keeps its own across-rollout normalization and unit weight in all four cells. A constant pooled type returns zero; a constant individual field need not do so under pooling. With one target of a type, its pooled and object–field statistics coincide.

Boundary credit uses Equation 5. For segment credit, add the sum of normalized attribute events to all tokens through $b_{k,\text{Gaze}}$, and give the normalized answer event to the whole response. Here $b_{k,\text{Gaze}}$ is the last token before the first accepted Think record, or termination if no Think is accepted. This is the existing matched segment control. The type-pooled alternative isolates a normalization population within our rewards; it is not a reproduction of SCOUT's complete reward construction (Zhou et al., 2026b).

Let $Q_{ps}, Q_{pb}, Q_{fs}, Q_{fb}$ denote full-benchmark QA for pooled/field normalization and segment/boundary credit. Normalization effects are $Q_{fs} - Q_{ps}$ and $Q_{fb} - Q_{pb}$; boundary effects are $Q_{pb} - Q_{ps}$ and $Q_{fb} - Q_{fs}$; their interaction is $(Q_{fb} - Q_{fs}) - (Q_{pb} - Q_{ps})$. Use the full 3DSRBench-real protocol for QA and D1000 for independent attribute and joint correctness. Apply Appendix C.4 to actual paired predictions.

B.4 Independent reference protocol

D1000 pairs canonical-question QA with independently sourced object-box annotations. The annotation protocol uses a fixed image/question manifest and object correspondence, with independent centers, dimensions and rotations required for the joint attribute cohort. Per-field reference coverage is recorded separately.

Reference eligibility and quality control. Require registered full-object boxes, camera calibration and scene depth for continuous-attribute scoring. Convert source units, box dimensions and local-to-camera rotations to the convention in Appendix A.1; check reconstructed corners and their projection against the queried object regions. Normalize centers and dimensions by the median over valid registered scene-depth pixels, accounting for axial depth versus ray distance. Two reviewers, blind to expert and student outputs, check correspondence, truncation, occlusion and full-box semantics, with disagreements adjudicated and exclusions recorded. Rotation evaluation additionally requires independent rotation labels and the declared box-symmetry convention.

Data separation and coverage. Audit source IDs, image hashes, reviewed duplicates and available scene/sequence groups against evaluation for both SFT components, the SR-QA bank and accessible expert-training sources. Reference-ineligible questions remain in full QA; reference coverage and prediction validity have separate denominators. Each diagnostic retains its own fixed question set and denominator.

C Evaluation Protocols and Key Comparisons

C.1 Benchmark and diagnostic populations

Primary native QA comparisons use the full 3DSRBench-real protocol. The pinned metadata contains 2,100 original questions, including 175 camera-depth questions, expanded to 5,157 processed forms: 690 Height, 1,709 Location, 1,036 Orientation and 1,722 Multi-Object forms. Original questions, processed forms, prefixes and scenes remain distinct counting units. CircularEval groups option permutations; FlipEval also includes prescribed horizontal flips with relation-dependent answers (Ma et al., 2025a). The release, form mapping, flip inclusion and extractor define the scoring configuration. CV-Bench-3D has 600 Depth Order and 600 Relative Distance questions, retaining its object-identifying colored boxes (Tong et al., 2024).

Table 1 follows the shared evaluation contract below. Benchmark averages use unrounded subset counts and are displayed to one decimal. Source-reported values retain their original aggregation and inference settings.

The shared evaluation protocol fixes the released RGB images, questions, option-order bank and benchmark aggregation for the Table 1 roster. Use the pinned 5,157-form 3DSRBench-real release and all 600 Depth plus 600 Distance CV-Bench questions. Circular option permutations are grouped within each form; original and horizontally flipped forms remain separate scoring units. Category means use their actual denominators, and the overall mean is computed from unrounded correct counts.

Source-specific evaluation. Published comparisons retain the identified model variant, test population, circular/flip grouping, aggregation and answer mapping. The shared evaluation protocol uses the settings below, while native prompts, budgets and tools remain method-specific.

Use the official VLMEvalKit choice/text matching procedure and one shared, gold-free `Qwen3-30B-A3B-Instruct-2507` option mapper for unresolved answers. Freeze its revision, prompt and deterministic settings; log failures and score unresolved outputs incorrect. Validate adapters on non-test examples before inference.

Retain native templates, processors and complete inference procedures, with one method execution per option rendering. Qwen3-VL-8B-Thinking retains its source 32,768-token cap (Kancheti et al., 2026); SGT-3D uses 2,048. VERDICT keeps default candidate verification, and Mini-o3 keeps crop/zoom from the same image. Their internal calls are not restricted to one model forward pass. Method-specific token budgets and internal calls accompany the shared scoring configuration; invalid or truncated answers follow the same failure rules.

Model predictions are paired on the same scoring units.

Method-specific inference and source settings. VERDICT (Sinha et al., 2026) uses a Qwen2.5-VL-7B reasoner and three frozen verification roles. Retain its default confidence and dispersion settings, candidate sampling and selection. The source uses Gemma answer extraction.

HiSpatial-3B-RGB (Liang et al., 2026) uses PaliGemma2-3B-Mix-448 preprocessing and the RGB-only interface. Its public evaluator merges original and `-flip` IDs; the shared protocol keeps these forms separate and applies the common answer mapper.

PolyV (Wu et al., 2026) retains its source-reported overall and category values in the published comparison. The shared protocol aggregates per-category correct counts under the pinned benchmark denominators.

Mini-o3 with entropy regularization (Kim et al., 2026) retains crop/zoom inference. Its source reports Avg@8 with a VLM answer judge and selects a checkpoint using online 3DSRBench validation. The shared protocol uses one complete episode per option rendering. These sampling and selection settings accompany the comparison.

Independent attribute and joint metrics target 3DSR-D1000, subject to Appendix B.4: 1,000 original, unflipped canonical questions requiring independent box-center, dimension and rotation annotations, comprising 100 Height, 325 Location (including all 175 camera-depth questions), 150 Orientation and 425 Multi-Object. Selection precedes student outputs. Its canonical-question QA supports joint and paired diagnostics, separately from full-benchmark CircularEval/FlipEval. Within the camera-depth pool, 3DSR-T400 assesses frozen-prefix behavior and 3DSR-I100 contains 100 eligible questions for supplied-state continuations. Annotations must not copy expert references or student outputs. Manifests retain image/scene/question mappings, annotation source and quality checks, calibration, common depth scale, exclusions and reference coverage. Training/validation require external image-/duplicate-group splits and separate student/expert exposure audits. Physical-scene separation requires actual source scene IDs.

Why separate these cohorts? The full benchmark supports broad answer comparisons without selecting questions for geometric diagnosability. D1000 concentrates the additional independent-box annotation needed to distinguish accurate geometry from a correct answer alone. Within its 175-question camera-depth pool, T400 balances student-relation correctness crossed with student–expert agreement (100 prefixes per cell), allowing teacher policies to be compared even in rare conflict cases. I100 instead fixes a common eligible question set across supplied-state substitutions and student variants, so changes reflect responses to the same states. T400 and I100 therefore branch from the depth pool; they do not form a sequence of progressively smaller datasets. The category allocation and sampling units remain fixed across comparisons.

C.2 Attribute, joint and answer metrics

Independent box centers and dimensions use the same camera-frame and (L, H, W) convention, divided by one independently established scene-depth median. No alignment is fitted to predictions. Center distance is dimensionless and dimension error follows Equation 8. Independent rotations use the same local-to-camera convention without scaling. Symmetry-aware orientation error is expressed in degrees, with reference and prediction coverage defined separately. The tabulated attribute and joint readouts assess centers, dimensions and orientations. Coverage distinguishes required targets, valid independent references and matched valid predictions. Conditional numerical errors use the last set with its denominator shown; coverage and unmatched/invalid predictions remain visible.

On the center/size/orientation reference-complete cohort $\mathcal{H}$, $g_x = 1$ requires a structurally valid trace, every required object match, and all required center, dimension and symmetry-aware orientation errors below their respective tolerances. Missing, invalid or unmatched predictions give $g_x = 0$. With answer correctness $a_x \in \{0, 1\}$,

$$P_{uv} = |\mathcal{H}|^{-1} \sum_{x \in \mathcal{H}} \mathbb{1}[g_x = u, a_x = v], \qquad u, v \in \{0, 1\}. \tag{15}$$

Attribute correctness is $P_{10} + P_{11}$; its joint correctness with QA is P_{11}. The retained tolerances are 0.1 for center distance and 0.2 for mean log-dimension error. Sensitivity uses center thresholds $\{0.05, 0.1, 0.2\}$ and dimension thresholds $\{0.1, 0.2, 0.4\}$, without test-set selection. Orientation correctness additionally requires each symmetry-aware angular error to be at most η_R, in radians, fixed on training/validation data before evaluation. Missing or invalid orientation predictions give $g_x = 0$. The four joint counts share the same reference-complete denominator and fixed angular tolerance. Full-cohort accuracy counts malformed/absent answers as incorrect.

A trusted question record fixes the operator, arguments and numerical tolerance. Re-executed arithmetic is evaluated against printed attributes; factual QA uses benchmark answers. Arithmetic consistency is conditional on executable operations, and answer–trace consistency on determinate implied answers. Each consistency rate uses its own eligible denominator and excludes unsupported or indeterminate operations. Unknown gravity, unavailable operands, nonfinite values and ties prevent an eligible relation check.

C.3 Controls for the two training signals

A is structured SFT; B adds CS-OPSD. C and F apply GF-RL and answer-only RL directly after A. D, E and G share B: D uses answer-only RL, E uses GF-RL, and G uses answer-only RL plus attribute-token CE. All share data, RGB-only deployment, reference policy and matched within-pair rollout/optimization settings. G's separate teacher-forced pass supervises center, dimension

and orientation value tokens from the same accepted references; Scan, Think and Answer have no auxiliary CE. Its coefficient is validation-selected. Geometry CE before GF-RL is a distinct warm-start control.

E–D and C–F test the attribute-reward package, while E–G compares it with direct supervision. To separate the two stages, the main factorial crosses continued SFT or CS-OPSD with answer-only RL or GF-RL from common SFT initialization. Its continued-SFT cells differ from C/F. Native QA uses full CircularEval/FlipEval; attributes and joint correctness use D1000. For full-protocol QA accuracies $Q_{sa}, Q_{sg}, Q_{oa}, Q_{og}$, with s/o denoting SFT/CS-OPSD and a/g denoting answer RL/GF-RL,

$$\Delta_{\text{teach,geom}} = (Q_{og} - Q_{oa}) - (Q_{sg} - Q_{sa}). \tag{16}$$

Longer single-stage, reversed-order and supervised warm-start routes use the same raw references and the fixed horizons and terminal checkpoints in Appendix A.5.

Credit assignment. Hold reward functions, available targets, weights, normalization, parser and optimization fixed. Broadcast uses $\widehat{A}_{k,t} = \sum_j \widetilde{r}_{k,j}$; completion-boundary reward-to-go includes only events with $b_{k,j} \geq t$ (Equation 5). The segment control assigns all attribute events to Scan/Gaze, with the answer event throughout:

$$\widehat{A}^{\text{segment}}_{k,t} = \widetilde{r}_{k,a} + \mathbb{1}[t \leq b_{k,\text{Gaze}}] \sum_{j \in \mathcal{J}^g_k} \widetilde{r}_{k,j}. \tag{17}$$

Here $b_{k,\text{Gaze}}$ precedes the first accepted Think record; without one, the segment extends to termination. This borrows functional segmentation from SCOUT while retaining our normalized rewards, so it is not a reproduction of SCOUT's complete algorithm. Reference-prefix and Gaze-only distillation separately test prefix source and response coverage before identical GF-RL.

C.4 Supplied states, frozen strata and uncertainty

On 3DSR-I100, identical-state diagnostics freeze checkpoint, image, question, Scan bindings and decoding. At a complete Gaze boundary, compare replay, independent oracle attributes, raw expert attributes, and bounded relation-preserving/reversing substitutions. Preserve nonedited fields, identities, frame, scale and validity. Use the same eligible questions and serialized states for plain, masked and committed students under both committed-produced and mask-produced anchors; analyze anchors separately. Perturbations, tie tolerance and exclusions are fixed in advance. These conditional continuations assess supplied-state use alongside full-protocol native answering.

Freeze the D1000 SFT relation strata once, retaining correct/incorrect student and expert states, agreement, both conflict directions, unsupported operators and unclassified cases. Paired comparisons partition the same canonical questions into both-right, repair, regression and both-wrong outcomes. Net diagnostic QA gain equals repairs minus regressions divided by that cohort. Common-valid numerical intersections retain their own counts, with separate numerical-error and coverage denominators. These strata locate answer changes within D1000.

Reference-noise controls perturb only training-reference centers at shared prespecified levels, using the same corrupted cache for teacher/checker and across methods, while keeping evaluation labels clean. Principal noise comparisons use full-protocol native QA on the shared Qwen2.5-VL-7B-Instruct backbone; attribute and joint readouts retain D1000. Compare the retention effect at each shared noise level. Settings are not retuned on test outcomes.

For full-protocol QA, aggregate option permutations within each scoring form before pairing, retaining original and flipped forms as separate scoring units; D1000 uses canonical-question predictions. Resample source-image clusters jointly across variants, carrying every question with identical multiplicity; use physical scenes only with an audited mapping. Freeze the resampling configuration and primary contrast family. Intervals are conditional on the evaluated checkpoints and use saved question-level predictions.

Table 10: Additional student geometry, arithmetic and coverage readouts. The selective mask uses the same committed-priority teacher.

	Before GF-RL		Arith.↑		Coverage	
Teacher policy	Attr.↑	Joint↑	Before	After	Before	After
Plain conditioning	65.7	54.5	90.6	95.7	960/1000	973/1000
Expert priority	65.2	53.6	89.2	95.0	959/1000	972/1000
Conflict-selective mask	71.2	59.3	93.9	97.1	968/1000	982/1000
CS-OPSD	70.0	60.5	94.8	99.2	965/1000	990/1000

Before/after refer to the policy-specific OPSD stage and subsequent GF-RL. Table 5 reports QA at both stages and post-RL attributes/joint correctness. Attribute and joint rates here use all 1,000 canonical 3DSR-D1000 questions. Invalid attributes count as failures. Numeric coverage counts valid predictions, while arithmetic is conditional on an executable operation. These seed-29 outcomes use shared raw references, SFT and RL recipe. Attributes use independent annotations.

D Additional Results and Diagnostics

D.1 Student readouts and training-route controls

The additional readouts distinguish geometric and arithmetic changes from output eligibility. Table 15 records paired answer changes on D1000.

Table 11: Raw-reference training schedules.

After common SFT	QA_F	$Attr._D$	$Joint_D$	$Center_D$	$Dim._D$
Longer CS-OPSD	63.4	71.5	61.2	0.1222	0.1804
Longer GF-RL	63.6	75.6	64.1	0.1053	0.1552
Continued SFT then GF-RL	63.9	74.4	62.8	0.1109	0.1688
Geometry CE then GF-RL	64.0	76.3	64.9	0.1034	0.1494
GF-RL then CS-OPSD	63.7	74.8	63.8	0.1072	0.1610
CS-OPSD then GF-RL	69.2	85.8	78.4	0.0810	0.1260

F: full 3DSRBench-real QA. D: D1000 attribute/joint rates, with numerical errors on the same 940-case valid intersection. Routes share SFT and raw references. Longer single-stage controls are distinct from B/C. The full endpoint reuses Table 5.

These endpoints compare full-benchmark answering across the stated training schedules, with D1000 readouts showing their associated attribute quality.

Table 12: Category detail for training-route alternatives.

Variant	Overall	Height	Loc.	Orient.	Multi
Longer CS-OPSD	63.4	56.2	79.6	53.1	56.3
Longer GF-RL	63.6	56.4	80.3	52.4	56.7
Geometry CE then GF-RL	64.0	56.8	80.5	53.0	57.1
GF-RL then CS-OPSD	63.7	56.2	79.9	52.9	57.1

Overall uses all 5,157 official processed forms; Height/Location/Orientation/Multi-Object use 690/1,709/1,036/1,722 forms. Circular option permutations are aggregated within each form. Forms are scoring units, not independent questions or scenes. E and the refined-reference variant use the main-table endpoints. Conditional-state and independent-attribute diagnostics retain their separate populations.

D.2 Teacher behavior and paired student diagnostics

Against masking, committed priority adds 1/20 correct answers when the student state is right and the expert is wrong, and 3/30 when the student state is wrong and the expert is right. These two subsets retain their separate denominators. The two conflict strata contribute 4 of the 39 net additional correct D1000 answers over masking; unsupported and unclassified cases contribute 30 and 3, respectively. The both-wrong stratum has one fewer correct answer than rescaled masking. This partial depth audit locates changes without attributing the full-benchmark gain to conflicts.

Table 15 locates repairs and regressions on the same D1000 canonical questions. These paired diagnostics remain separate from the full-protocol QA comparison.

Table 13: Teacher behavior on the balanced 3DSR-T400 prefix bank.

Student / expert	Teacher policy	Factual	Arith.	Trace	Invalid
Correct / agrees	Plain conditioning	96	93	96	2
Correct / agrees	Expert priority	97	95	97	1
Correct / agrees	CS-OPSD	97	96	97	1
Correct / conflicts	Plain conditioning	67	63	67	3
Correct / conflicts	Expert priority	24	22	24	3
Correct / conflicts	CS-OPSD	95	94	95	2
Incorrect / agrees	Plain conditioning	6	88	90	4
Incorrect / agrees	Expert priority	3	92	94	3
Incorrect / agrees	CS-OPSD	2	95	96	2
Incorrect / conflicts	Plain conditioning	58	35	38	4
Incorrect / conflicts	Expert priority	88	7	9	3
Incorrect / conflicts	CS-OPSD	3	94	95	2

Every rate uses all 100 prefixes in its cell. Malformed outputs fail. 3DSR-T400 balances 400 prefixes from five SFT samples per original camera-depth question, with 100 per stratum. Questions may repeat. Teacher contexts reuse each prefix. The bank measures behavior conditional on the four balanced prefix strata. In the incorrect/conflicting cell, committed priority has 3% factual correctness and 95% trace consistency. Expert priority has 88% factual correctness.

Table 14: Frozen-SFT strata: final answers and paired mask-to-CS-OPSD changes.

A. Final correct answers / stratum size

Teacher policy	S^+E^-	S^-E^+	S^+E^+	S^-E^-	Other
Plain conditioning	17/20	20/30	65/70	15/20	683/860
Expert priority	15/20	24/30	65/70	16/20	679/860
Conflict-selective mask	18/20	23/30	66/70	16/20	702/860
CS-OPSD	19/20	26/30	67/70	17/20	735/860

B. Mask-to-CS-OPSD transitions on the same strata

Stratum	N	Both right	Wins	Losses	Both wrong	Net
Correct / agree	70	64	3	2	1	1
Correct / conflict	20	18	1	0	1	1
Incorrect / agree	20	15	2	1	2	1
Incorrect / conflict	30	22	4	1	3	3
Unsupported / unclassified	860	686	49	16	109	33

D1000 uses 1,000 canonical questions. Of 175 camera-depth questions, 140 are classifiable and 35 unclassified; the other 825 have unsupported operators. The 50 conflict cases are 35.7% of the 140 classifiable cases. $S^{\pm}E^{\pm}$ denotes initial student/expert relation correctness. Other combines the 35 and 825 cases. Panel B totals are 805/59/20/116 (both right/repairs/regressions/both wrong), matching Table 15.

Table 15: Answer repairs and regressions on 3DSR-D1000 canonical questions.

Comparison	Repairs	Regress.	Both right	Both wrong	Net pp
B: after CS-OPSD → D: answer-only RL	48	23	772	157	+2.5
B: after CS-OPSD → E: GF-RL	94	25	770	111	+6.9
D: answer-only RL → E: GF-RL	63	19	801	117	+4.4
Plain + GF-RL → E: GF-RL	92	28	772	108	+6.4
Mask + GF-RL → E: GF-RL	59	20	805	116	+3.9
G: answer RL + CE → E: GF-RL	53	22	811	114	+3.1

Each row partitions the 1,000 3DSR-D1000 questions. B/D/E have 795/820/864 correct answers, respectively. The ledger supplies one compatible joint B/D/E allocation. Repairs are wrong-to-right and regressions right-to-wrong. Their difference equals the QA count gain. The paired counts locate answer repairs and regressions on the same questions.

D.3 Using supplied states: native prefixes and shared anchors

Supplied-state comparisons separate attribute quality from how each frozen checkpoint uses those attributes. Table 16 retains native-prefix and reciprocal shared-anchor results as distinct conditional tests.

Table 16: Conditional state use under native prefixes and two shared anchors.

A. Shared committed-E anchor

Student	Attributes supplied	State correct	QA	Agreement
Plain conditioning	Unedited replay	90	83	87
Plain conditioning	Oracle attributes	100	90	90
Plain conditioning	Raw expert	92	88	92
Plain conditioning	Relation preserved	90	82	86
Plain conditioning	Relation reversed	10	35	69
Conflict-selective mask	Unedited replay	90	85	91
Conflict-selective mask	Oracle attributes	100	94	94
Conflict-selective mask	Raw expert	92	91	95
Conflict-selective mask	Relation preserved	90	84	90
Conflict-selective mask	Relation reversed	10	24	82
CS-OPSD	Unedited replay	90	92	96
CS-OPSD	Oracle attributes	100	98	98
CS-OPSD	Raw expert	92	92	96
CS-OPSD	Relation preserved	90	90	98
CS-OPSD	Relation reversed	10	15	91

B. Plain student on its own native prefixes

Plain conditioning	Unedited replay	87	81	88
Plain conditioning	Oracle attributes	100	90	90
Plain conditioning	Raw expert	92	87	93
Plain conditioning	Relation preserved	87	80	87
Plain conditioning	Relation reversed	13	35	70

C. Shared mask-produced anchor: QA only

Student	Oracle	Preserved	Reversed
Plain conditioning	90	81	34
Conflict-selective mask	95	84	25
CS-OPSD	98	89	16

Percentages use the same 100 eligible 3DSR-I100 questions, distinct from the initial 140-case SFT audit. A and C supply identical serialized states to all students within each condition. The committed student's native-prefix outcomes equal its five rows in A. B retains the distinct plain-native outcomes. State correctness scores the symbolic answer from supplied attributes. Agreement compares it with the generated answer. Oracle/preserved/reversed state correctness is 100/90/10% for A and 100/89/11% for C. Eligibility requires complete Gaze, supported independent depth labels and valid edited states. Exclusions and invalid continuations remain part of the reporting protocol. No teacher or answer text is supplied. These conditional continuations are separate from native RGB-only benchmark results.

The oracle committed–mask gaps are +4/+3 points under the committed/mask anchors. The differences have the same direction under both anchors. Reversed states retain a factual-answer deficit.

Native-prefix attribute substitutions. On 3DSR-I100, oracle attributes give committed priority 98% correct continuations versus 90% for plain conditioning. Committed replay-to-oracle counts contain 7 repairs and 1 regression. The ledger retains compatible paired allocations for preserved and reversed states.

Failure under reversed relations. Reversed relations give committed-priority QA 15%. Following an inaccurate state can hurt factual answers even when internal agreement is high.

D.4 Supervision controls and reward sensitivity

Reward-weight sensitivity. With answer weight one, attribute weights 0.5/1.0/2.0 give full-benchmark QA 68.7/69.2/68.6%, D1000 attribute accuracy 84.8/85.8/85.1% and joint accuracy 77.5/78.4/77.6%. Weight one reuses E; the weights are fixed before evaluation.

Table 17: Supervision-amount and instruction-wording controls after identical GF-RL.

OPSD variant	QA_F	$Attr._D$	$Joint_D$
Conflict mask	66.0	81.2	71.0
Random continuation mask	65.8	80.5	71.2
Rescaled conflict mask	66.7	81.1	72.0
Committed paraphrase 1	68.8	85.1	77.7
Committed paraphrase 2	69.0	85.4	77.9
CS-OPSD	69.2	85.8	78.4

F: full-benchmark QA; D: D1000 attribute/joint correctness. Random masking matches removed continuation-token counts. Rescaled masking changes retained-loss normalization. Paraphrases preserve the priority rule.

D.5 Closest alternatives and paired answer changes

Conflict-selective and rescaled masking share the committed-state teacher and subsequent GF-RL. Ordinary masking keeps the response-length denominator; rescaled masking compensates for removed supervision with a validation-fixed cap. Together with answer RL and attribute CE, these controls test continuation retention and numerical feedback. Full-protocol paired counts measure net answer changes; D1000 supplies separate attribute/joint and frozen-stratum diagnostics.

Table 18: Four closest comparisons with a shared full-method endpoint.

Comparator	ΔQA_F	$Wins/losses_F$	$\Delta Attr._D$	$\Delta Joint_D$
Conflict mask + GF-RL	+3.2	264/99	+4.6	+7.4
Rescaled mask + GF-RL	+2.5	236/107	+4.7	+6.4
D: answer RL	+3.8	289/93	+13.8	+14.9
G: answer RL + attribute CE	+2.6	233/99	+3.8	+5.9

F: full 3DSRBench-real, with E QA 69.2%. D: D1000, with E attribute/joint accuracy 85.8/78.4%. Differences are E minus the comparator. Wins/losses count discordant forms within the same 5,157-form evaluation after circular aggregation; forms sharing an image are dependent. Teacher controls share SFT and GF-RL. D/E/G share B. Attribute and joint metrics assess centers, dimensions and orientations.

The full-benchmark QA margins are 3.8/2.6 points over answer RL/attribute CE, and 3.2/2.5 points over selective/rescaled masking.

D.6 Answer supervision and normalization–credit controls

These additional controls compare two design choices while retaining the shared endpoints.

Table 19: Restoring Think supervision while masking conflict-following Answer KL.

OPSD supervision	QA_F	QA_D	$Attr._D$	$Joint_D$	ΔQA_F
Think+Answer mask	66.0	82.5	81.2	71.0	+3.2
Rescaled Think+Answer mask	66.7	83.4	81.1	72.0	+2.5
Answer-only mask	68.8	86.0	85.2	77.7	+0.4
Rescaled Answer-only mask	68.9	86.1	85.4	77.9	+0.3
Full CS-OPSD	69.2	86.4	85.8	78.4	–

F: full 3DSRBench-real; D: D1000. ΔQA is E minus the row. Answer-only arms retain Scan/Gaze/Think KL, use the same conflict flag and share GF-RL. Ordinary masks use denominator T; rescaled masks use capped T/R. Comparator rows and E reuse their shared endpoints.

The Answer-only masks recover 2.8/2.2 full-QA points relative to the ordinary/rescaled full-continuation masks. Full CS-OPSD adds 0.4/0.3 points over the Answer-only arms. Restoring Think supervision closes most of the gap.

Table 20: Normalization population crossed with temporal credit.

Normalization	Credit	QA_F	QA_D	$Attr._D$	$Joint_D$
Type-pooled	Segment	65.6	83.2	79.8	70.4
Type-pooled	Boundary	67.8	85.0	83.0	75.0
Object-field	Segment	66.2	83.1	80.9	71.6
Object-field	Boundary	69.2	86.4	85.8	78.4

F: full-benchmark QA; D: D1000. All arms share B, individual raw rewards, weights and answer normalization. Type pooling changes the statistics population; individual events remain. Object-field rows reuse segment credit and E. Field–type gains are +0.6/+1.4 QA points under segment/boundary credit; boundary–segment gains are +2.2/+3.0 under type/field normalization, with +0.8 interaction from unrounded counts.