GEOMETRY AS EVIDENCE, NOT EXCLUSION: SUPPORT-AWARE CORRECTION FOR SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial executors may assess only some candidate answers, yet multiplying their scores with a full-question prediction excludes every unassessed candidate at any positive exponent. This conflates preference within the assessed set with probability allocation across its boundary. At exponent one, 88–92% of new errors across three natural-language 3D grounding settings occur when the initially correct parent winner is excluded. Geometry-Specific Correction (GSC) is a training-free rule expressing geometry as bounded evidence relative to a same-support relation-null reference. It preserves parent support and exactly recovers the parent under null evidence. A controlled comparison isolates allocation: a smoothed product and unclipped GSC have identical conditional predictions on each side of the boundary, yet the reference allocation improves natural-language 3D accuracy by 1.88–2.22 points. With one correction strength selected on SR3D and transferred unchanged, GSC improves all 11 parent–benchmark settings across five benchmarks, with positive paired intervals and gains up to 7.37 points in 3D grounding and 6.48 in RGB reasoning. Bounded-product controls attain comparable 3D accuracy, while GSC additionally specifies exact relation-null behavior. On ViewSpatial, omitting specified frame operations removes 60–80% of the gain and swapping entity roles reverses it. Support allocation is therefore a consequential design choice for partial spatial evidence.

1 INTRODUCTION

A geometric routine and a full-question model answer different questions. For *the white pillow on the chair next to a blue pillow*, proximity alone does not resolve color, category, or the intended anchor. An executor can measure a relation while the parent evaluates the complete expression. How should that partial evidence change the answer?

A geometric distribution encodes both preference *within* assessed candidates and a boundary separating them from unassessed alternatives. Multiplication conflates these decisions: every candidate outside the support receives zero probability at any positive exponent. In four 3D development and expansion sweeps, exponent $1/16$ already incurs 74–86% as many new errors as exponent two; GSC incurs at most one (Appendix D.2).

Geometry-Specific Correction (GSC) uses neutral evidence outside the assessed set and a relation-null reference within it. The reference retains support while removing relational preference. It specifies both how assessed candidates are reranked and how they compete with unassessed alternatives, with finite influence and exact parent recovery under null evidence.

Reference-relative scoring is established in calibration and evidence aggregation (Zhao et al., 2021; Holtzman et al., 2021; Deisenroth & Ng, 2015). Holding conditional predictions on both sides of the support boundary fixed, changing only probability allocation improves natural-language 3D accuracy by 1.88–2.22 points and reverses on structured SR3D, where the parent is rarely right on executed questions with no correct proposal in S (9% versus 20–28%).

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

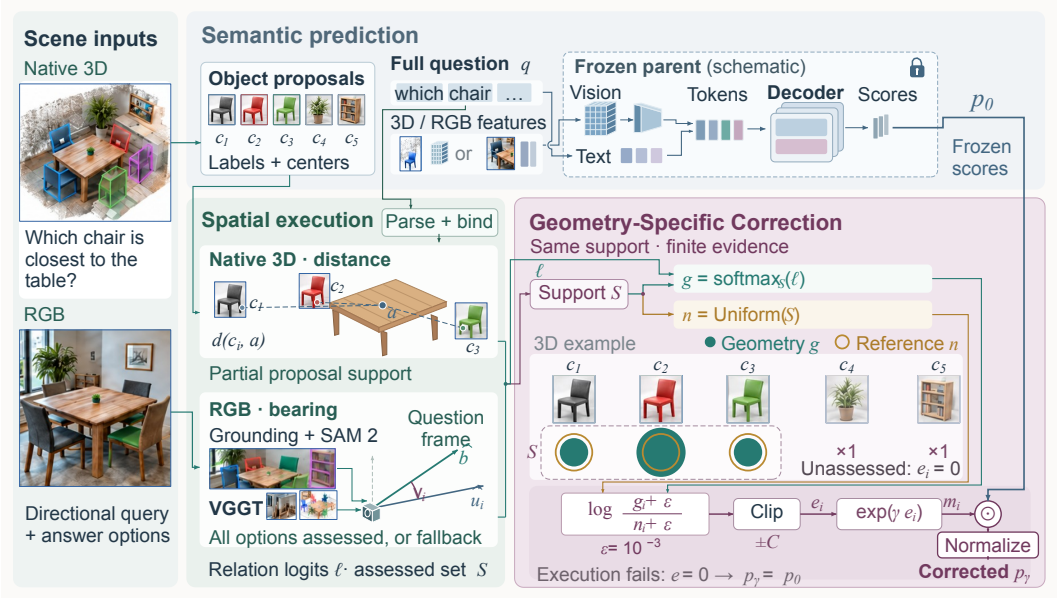

Figure 1: **Geometry-Specific Correction (GSC).** The frozen parent evaluates the full question. Geometry supplies relation logits on an assessed support; a reference on that same support converts them into clipped relative evidence. Unassessed candidates receive unit multipliers, followed by reweighting and normalization. Native 3D permits partial proposal support; RGB assesses every legal option or falls back. Disk areas encode geometry and equal outlines its reference on a common scale; scene depictions and the parent cutaway are illustrative.

Our contributions are:

- A weak-product exclusion result connected to full-manifest error counts and small-strength behavior on fixed development and expansion subsets.
- A same-support evidence rule and a controlled comparison in which only boundary allocation changes, moving natural-language accuracy by 1.88–2.22 points and reversing on SR3D.
- A frozen-score evaluation with 11 positive paired effects, repairs that neither the parent nor geometry makes alone, and ViewSpatial role-swap interventions that reverse the gains.

2 When Partial Geometry Becomes Hard Exclusion

Let $\mathcal{C}$ be the candidate set and p_0 the semantic distribution produced by a frozen parent on the full question. A geometric executor returns a normalized distribution g with positive scores on its assessment support $S \subseteq \mathcal{C}$ and zero elsewhere. In native 3D, predicted categories and parsed entities determine S, and only executable relations contribute to g. Consequently, $i \notin S$ means that the executor did not assess candidate i under this program; it need not mean that i is an impossible answer. The problem is interpreting an assessment zero as evidence of impossibility, rather than multiplication of valid likelihoods.

The weak-product limit already excludes candidates. Consider $q_\eta(i) \propto p_0(i) g_i^\eta$ for $\eta > 0$. If $p_0(S) > 0$, then

$$\lim_{\eta \downarrow 0} q_\eta(i) = \frac{p_0(i)\mathbf{1}[i \in S]}{p_0(S)} = p_0(i \mid S). \tag{1}$$

Inside S, $g_i^\eta \to 1$; outside, it remains zero at every positive strength. Decreasing the product strength weakens relational preferences but never withdraws this hard exclusion. Even uniform geometry on S conditions the parent on that set. Figure 2 isolates the distinction between assessment and answer eligibility.

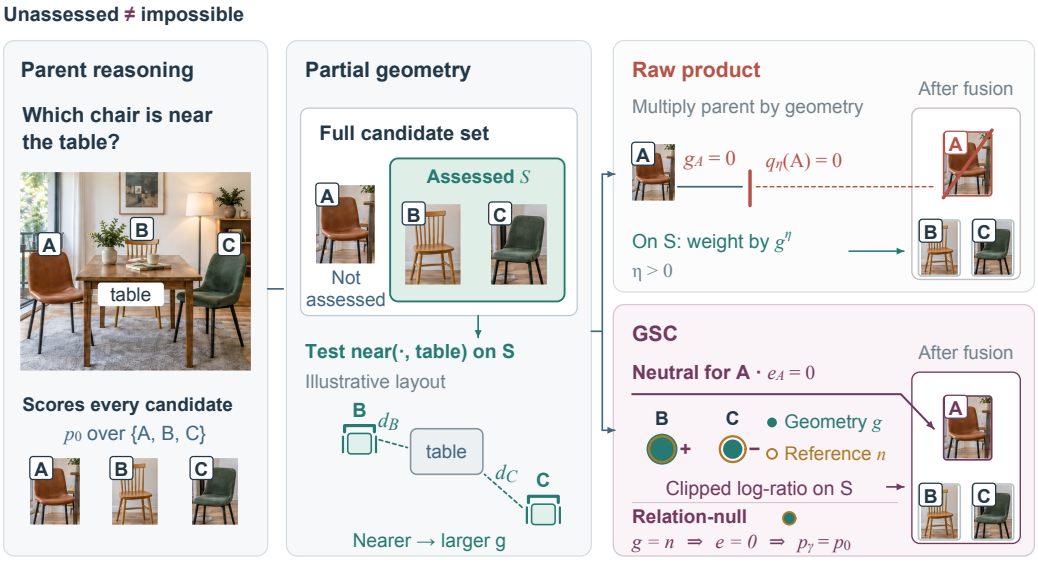

Figure 2: **Unassessed is not impossible.** Both rules share the full-candidate parent p_0 and geometry g, zero-extended outside S. Here A is unassessed ($g_A = 0$): raw product excludes it, while GSC assigns $e_A = 0$. The correspondence line tracks candidates, not scores. Photographs track object identity; the near-test layout and disk areas are illustrative, with filled disks for g and equal gold rings for the same-support reference n. Outputs show candidate eligibility; the separate relation-null strip shows $g = n \Rightarrow p_\gamma = p_0$ for the full distribution.

Smoothing removes literal zeros but leaves the relative allocation to be determined. The interface in Figure 1 treats partial geometry as evidence without redefining the answer space; the analysis below examines both hard exclusion and finite alternatives.

3 Geometry-Specific Correction

3.1 Evidence relative to assessment support

For a successfully executed question, let ℓ_i be the summed relation score for $i \in S$. Define geometry and a relation-null reference on the *same* support:

$$g_i = \frac{\mathbf{1}[i \in S]\exp(\ell_i)}{\sum_{j \in S}\exp(\ell_j)}, \qquad n_i = \frac{\mathbf{1}[i \in S]}{|S|}. \tag{2}$$

The reference holds assessment support fixed while suppressing relational preference. GSC applies

$$e_i = \mathrm{clip}\left(\log \frac{g_i + \epsilon}{n_i + \epsilon}, -C, C\right), \qquad p_\gamma(i) = \frac{p_0(i)\exp(\gamma e_i)}{\sum_j p_0(j)\exp(\gamma e_j)}. \tag{3}$$

We use $\epsilon = 10^{-3}$, $C = 2$, and $\gamma = 2$. Failed execution sets $e = 0$ and returns p_0. The correction has no learned parameters; its log ratio is a defined contrast against the executor's null preference, not an assumed calibrated likelihood ratio.

3.2 Separating ranking from support allocation

If $g = n$, then $p_\gamma = p_0$. Outside S, $g_i = n_i = 0$ gives $e_i = 0$, a unit multiplier; normalization can still change an individual absolute probability. The condition $g = n$ removes relational preference, not the possibility that assessment support itself is informative.

Write $\alpha = p_0(S)$, with $0 < \alpha < 1$, and $M_\gamma = \mathbb{E}_{i \sim p_0(\cdot|S)}[\exp(\gamma e_i)]$. Then

$$\frac{p_\gamma(S)}{1 - p_\gamma(S)} = \frac{\alpha}{1 - \alpha} M_\gamma. \tag{4}$$

Individual multipliers change the conditional prediction inside S; their parent-weighted average controls allocation across its boundary. A common evidence offset b on S leaves conditional predictions unchanged but multiplies support odds by $\exp(\gamma b)$.

This yields a directly testable comparison. Let $q_{\mathrm{sm}} \propto p_0(g+\epsilon)^\gamma$ be the smoothed product and $q_{\mathrm{ref}} \propto p_0[(g+\epsilon)/(n+\epsilon)]^\gamma$ be unclipped GSC.

Corollary 1 (Boundary-only comparison)**.** *For the uniform reference in Eq. 2, with $s = |S|$, the conditional distributions agree wherever defined:*

$$q_{\mathrm{ref}}(\cdot \mid T) = q_{\mathrm{sm}}(\cdot \mid T), \quad T \in \{S, S^c\},$$
$$\frac{q_{\mathrm{ref}}(S)}{1 - q_{\mathrm{ref}}(S)} = \left(\frac{\epsilon}{1/s + \epsilon}\right)^\gamma \frac{q_{\mathrm{sm}}(S)}{1 - q_{\mathrm{sm}}(S)}. \tag{5}$$

The odds identity requires $0 < p_0(S) < 1$.

Thus any changed prediction in this matched pair crosses the assessment boundary. Section 5.2 measures its accuracy consequences; Appendix A gives the proof and boundary cases.

3.3 Decision margins and bounded influence

For candidates with positive parent probability,

$$\log \frac{p_\gamma(i)}{p_\gamma(j)} = \log \frac{p_0(i)}{p_0(j)} + \gamma(e_i - e_j), \qquad \left|\Delta \log \frac{p(i)}{p(j)}\right| \le 2\gamma C. \tag{6}$$

Candidate i overtakes parent winner a when $\gamma(e_i - e_a) > \log[p_0(a)/p_0(i)]$, with equality governed by tie breaking. Odds among unassessed candidates are unchanged. The general bound is eight nats at our constants; it limits influence without guaranteeing that every update is beneficial.

3.4 Native 3D and RGB instantiations

Native 3D. The parent scores object identifiers in a fixed Mask3D proposal pool (Schult et al., 2023). Predicted labels bind candidates and anchors; box centers determine proximity, distance ranking, vertical ordering, between, and room-middle scores. Successfully bound relation logits are summed on the target-class support. Partial relation execution is allowed while the parent reads the entire expression. Non-ranking relations can use all proposals when no target class is found. Anchor hypotheses are scored individually and then averaged (Appendix B).

RGB. Grounding DINO and SAM 2 locate entities; VGGT supplies cameras and world points (Liu et al., 2023; Ravi et al., 2024; Wang et al., 2025). Typed programs compare option directions with a measured bearing in the question's camera, simulated-person, or compass frame. For angular violation $v_i = \arccos(\mathrm{clip}(\langle \hat{b}, \hat{u}_i \rangle, -1, 1))$, we set $\ell_i = -v_i/0.7$. Every legal option must be executable, otherwise the question returns p_0. Successful RGB execution uses $S = \mathcal{C}$. Native 3D therefore tests incomplete assessment; RGB tests transfer of the same frozen rule under complete option support.

4 Experiments

4.1 Frozen protocol and comparisons

Benchmarks and parents. Native 3D uses full validation manifests of SR3D ($N = 17{,}726$), NR3D (7,899), and ScanRefer (9,508) (Achlioptas et al., 2020; Chen et al., 2020). NR3D follows the recorded `correct_guess=True` filter. Correctness requires selected-proposal IoU ≥ 0.5; unreachable targets and unexecuted questions remain in the denominator. These are detected-proposal results, not ground-truth-instance classification scores. The QuatRoPE parent, adapted on 1,600 SR3D training expressions, covers all three datasets; the independently trained Chat-Scene parent covers SR3D/NR3D (Zhou et al., 2026; Huang et al., 2024). Both are frozen before correction.

RGB uses official ViewSpatial-Bench (5,712 questions) and MMSI-Bench (1,000), their permitted images, and InternVL2.5-8B, Qwen3-VL-8B, and Qwen2.5-VL-7B (Li et al., 2025; Yang et al.,

Table 1: **Full native 3D grounding.** Detected-proposal accuracy at IoU 0.5 (%); Δ [95% CI] compares GSC with P_0 in points. Execution coverage is 73.65%/33.93%/28.06% on SR3D/NR3D/ScanRefer. Best per row is bold; shading identifies GSC.

Parent	Benchmark	P_0	GEO	MIX	Prod_2	BP	GSC	Δ vs. P_0 [95% CI]
Structured expressions								
QuatRoPE	SR3D	35.70	39.83	43.15	**45.06**	43.23	43.07	+7.37 [5.84, 8.94]
Chat-Scene	SR3D	35.26	39.74	41.58	**44.56**	42.37	42.28	+7.02 [5.72, 8.46]
Natural expressions								
QuatRoPE	NR3D	33.76	29.97	33.31	33.04	**35.32**	35.14	+1.38 [0.86, 1.94]
Chat-Scene	NR3D	34.05	30.57	34.03	33.57	35.76	**35.79**	+1.73 [1.16, 2.32]
QuatRoPE	ScanRefer	49.86	44.55	48.81	48.42	50.32	**50.33**	+0.46 [0.17, 0.76]

2026a; Chen et al., 2024; Bai et al., 2025a;b). P_0 uses first-token option-letter scoring; *Free* generates an answer deterministically. The final interface retains every legal option; a cached-score audit and an explicit option-parser repair are documented in Appendix B.

Same-evidence controls. Fusion arms share p_0, candidates, geometry, parser, and fallback. GEO chooses the geometric maximum on execution and the parent otherwise. MIX uses $(p_0 + g)/2$ with the extended executor. Prod_η uses $p_0 g^\eta$; Prod_2 matches GSC's *exponent*, not its intervention magnitude. Bounded product (BP) uses $p_0 \exp[\gamma \,\text{clip}(\log(g+\epsilon), -C, C)]$. *Hint* supplies the same RGB angular violations as text and adds a model re-scoring call. It was run on two backbones; the Qwen2.5-VL condition was not run. These readout interfaces have different inference budgets.

Selection and uncertainty. We selected γ once from $\{0.5, 1, 2\}$ on 400 SR3D development questions; RGB temperature was calibrated synthetically. Paired 95% intervals use 20,000 question-weighted cluster-bootstrap draws: physical scenes for 3D and recorded image groups for RGB. Development exposure, scoring failures, and stricter grouping checks are documented in Appendix C. Excluding known development rows leaves QuatRoPE gains of +7.30/+1.38/+0.47 points on SR3D/NR3D/ScanRefer. Removing 40 inspected MMSI questions leaves +2.50 to +3.96 points across backbones. All corresponding paired intervals remain positive; these checks concern recorded exposure. Merging ViewSpatial rescans reduces the bootstrap from 1,338 to 1,190 clusters; the smallest lower interval bound across backbones remains +1.85 points. Deltas use unrounded counts, so they need not equal differences between displayed accuracies.

4.2 Native 3D grounding

GSC improves all five full-manifest settings (Table 1): +7.37/+7.02 points on QuatRoPE/Chat-Scene SR3D, +1.38/+1.73 on NR3D, and +0.46 on ScanRefer, with positive paired intervals. On executed questions, the corresponding gains are +10.01/+9.54, +4.07/+5.11, and +1.65 points; full-denominator results remain primary (Appendix D.2).

Against the exponent-matched product on natural expressions, GSC gains +2.10 points on QuatRoPE NR3D (95% CI [1.48, 2.77]), +1.90 on ScanRefer ([1.39, 2.43]), and +2.22 on Chat-Scene NR3D ([1.54, 2.93]). Prod_2 instead has the highest SR3D accuracy. This reversal motivates the boundary-allocation comparison and operating curves in Section 5.

4.3 RGB tests transfer under complete option support

GSC improves all six official RGB settings and free generation from the same backbones, with positive paired intervals for both comparisons (Table 2; Appendix D). ViewSpatial gains over P_0 are +2.50 to +6.48 points; MMSI gains are +2.80 to +4.90. On executed ViewSpatial questions they are +9.26 to +23.95 points, with exact fallback elsewhere.

GEO has higher point accuracy than GSC in all six RGB settings. When the executable relation nearly resolves the option-level question, direct execution can justify a stronger departure from the parent. Across the broader 11-setting evaluation, however, GEO and MIX each degrade three natural-language 3D settings, while GSC and BP improve every setting (Appendix D.2).

Table 2: **Official RGB spatial reasoning.** Accuracy (%); Δ [95% CI] is GSC minus P_0 in points. Geometry executes on 1,545/5,712 ViewSpatial and 88/1,000 MMSI questions. †: Hint adds a model call; a dash denotes an unrun condition. Best per row is bold; GSC is shaded.

Benchmark	Backbone	Free	P_0	GEO	BP	Hint†	GSC	Δ vs. P_0 [95% CI]
ViewSpatial	InternVL2.5	41.49	41.07	**48.02**	47.48	44.07	47.55	+6.48 [5.36, 7.57]
	Qwen3-VL	40.41	40.90	**45.85**	43.01	45.26	43.40	+2.50 [1.89, 3.10]
	Qwen2.5-VL	36.80	36.96	**43.98**	41.77	—	42.38	+5.43 [4.42, 6.42]
MMSI	InternVL2.5	28.30	27.60	**31.50**	31.00	28.10	31.10	+3.50 [2.30, 4.80]
	Qwen3-VL	31.10	31.70	**35.90**	33.90	34.90	34.50	+2.80 [1.80, 3.90]
	Qwen2.5-VL	27.70	27.60	**32.60**	32.40	—	32.50	+4.90 [3.60, 6.30]

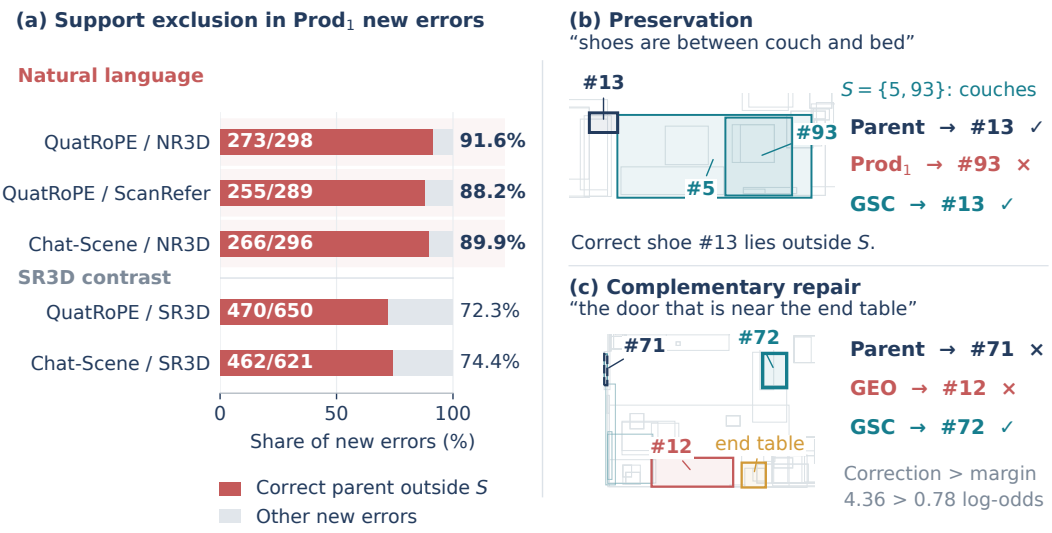

Figure 3: **Preservation and complementary repair.** Left: Prod_1 new errors on the full 3D manifests, partitioned by whether the initially correct parent winner lies outside S; labels give counts over all new errors. Right: cropped top views of actual predicted proposals; the shoes layout is rigidly rotated. The shoes case preserves correct #13 outside a misbound couch support. The door case selects correct #72 although both parent and GEO are wrong; Prod_1 also repairs this example.

Hint's higher Qwen3-VL ViewSpatial accuracy shows the value of presenting evidence in the prompt; this interface intervenes at a different stage from decision-level correction.

5 Understanding the Correction

5.1 Support exclusion characterizes product regressions

In 88–92% of Prod_1 new errors on natural-language grounding, the initially correct parent winner lies outside S (Figure 3). These counts diagnose ordinary multiplication; Appendix D.2 also reports exponent-two counts.

Correct-answer availability exposes the tradeoff. On 1,172 executed QuatRoPE NR3D questions with no correct proposal in S, Prod_2 is necessarily wrong; GSC retains 19.54% accuracy, close to the parent's 19.97%. When a correct proposal is assessed, Prod_2 instead exceeds GSC, 62.53% versus 58.36% on 1,508 questions. Their full-manifest difference decomposes as $+2.90 - 0.80 = +2.10$ points across these two strata (Table 18). These are annotation-defined diagnostics, not test-time selection rules.

Table 3: **Isolating support allocation.** Full QuatRoPE-adapted accuracy (%), exponent two. The paired † rows have identical conditional distributions on S and S^c; $\Delta^\dagger$ is unclipped GSC minus Smooth_2 in points, with paired 95% intervals. Null invariant means $g = n \Rightarrow p_\gamma = p_0$; clip denotes explicit threshold C.

Rule	SR3D	NR3D	ScanRefer	Null invariant	Explicit clip
Prod_2	45.06	33.04	48.42	No	No
Neutral outside	35.98	33.35	48.94	No	Yes
BP	43.23	35.32	50.32	No	Yes
$\text{Smooth}_2^\dagger$	45.01	33.05	48.45	No	No
Unclipped GSC†	43.07	35.11	50.34	Yes	No
GSC	43.07	35.14	50.33	Yes	Yes
$\Delta^\dagger$ [95% CI]	-1.94 [-2.96, -0.78]	+2.05 [1.43, 2.72]	+1.88 [1.37, 2.41]	—	—

The shoes case preserves a correct answer outside a misbound support. The door case shows complementary repair: geometric evidence supplies a 4.36-nat advantage to correct #72, exceeding the parent's 0.78-nat margin for #71, although neither parent nor GEO selects #72 alone.

5.2 Support allocation changes decisions at fixed conditional predictions

Smooth_2 and unclipped GSC differ only in support allocation (Corollary 1). This raises NR3D accuracy from 33.05 to 35.11 and ScanRefer from 48.45 to 50.34; exact paired effects appear in Table 3. Chat-Scene NR3D gains +2.22 [1.56, 2.92] points. All changed predictions cross the boundary. At median support sizes three or four, Smooth_2 gives S roughly 6.3×10^4–1.1×10^5 times the odds of unclipped GSC: a large support preference survives smoothing.

Support coverage alone does not explain the SR3D reversal. A correct proposal lies in S for about 70% of executed SR3D and ScanRefer questions alike. When none is assessed, the parent is still correct on 19.6–28.1% of natural-language questions, versus 8.8–9.3% on SR3D; on questions whose assessed support contains a correct proposal, Prod_2 exceeds GSC by 7.6–8.4 points on SR3D versus 2.0–4.2 on natural-language grounding (Table 18). The tradeoff depends on recoverable parent knowledge outside S as well as useful commitment inside it.

At our constants, BP cannot decrease support odds; GSC can move them in either direction through M_γ (Appendix A). After an irrelevant global score shift, BP is a one-sided evidence rule relative to a fixed floor; GSC uses signed evidence relative to a normalized relation-null distribution on S. All five 3D GSC-minus-BP intervals include zero, and three of six RGB intervals favor GSC. At $\gamma = 4$ on the four development and expansion sweeps, BP makes 1.24–2.25 times as many new errors as GSC, with repair counts differing by at most five (Table 19). Removing GSC clipping changes the 3D accuracies only slightly; the controls separate finite boundary penalties from the reference's exact null semantics.

Complete-support interpretation. When $S = \mathcal{C}$, unclipped GSC equals the normalized smoothed product $p_0(g + \epsilon)^\gamma$: the constant reference cancels. With clipping, reference centering sets the lower saturation threshold ($g_i \approx 0.033$ for GSC versus 0.134 for BP in four-option tasks). RGB thus tests transfer and grounding dependence of the frozen correction; the controlled native 3D comparison isolates boundary allocation.

5.3 Correction strength exposes repair–harm tradeoffs

Let R count repaired parent errors and H count newly incorrect answers. Figure 4 plots $100(R - H)/N$ against $100H/N$, reusing frozen inputs.

Equal exponents do not match decision changes. On SR3D expansion, Prod_2 gains +10.58 points with 85 new errors, versus GSC's +9.33 with 34. At $\gamma = 4$, GSC reaches +11.08 with 62 new errors, exceeding the best sampled raw-product gain with fewer errors.

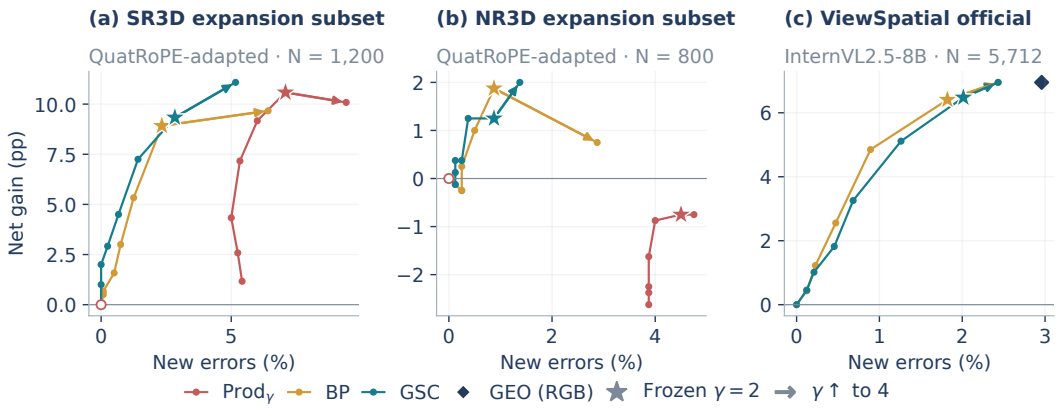

Figure 4: **GSC improves the sampled raw-product tradeoff.** In the 3D subsets, each positive-strength raw point has a sampled GSC point with higher net gain and fewer new errors. Panels use SR3D expansion ($N = 1{,}200$), NR3D expansion (800), and official ViewSpatial with InternVL2.5 (5,712). Stars mark frozen $\gamma = 2$; arrows indicate increasing strength. Raw zero strength is separate from its positive trajectory; GEO is an RGB reference point.

Every positive raw point in all four 3D development and expansion sweeps has a better sampled GSC point (finite grids, possibly different strengths; Appendix D.2). The reported configuration remains $\gamma = 2$.

Under complete RGB support, GEO can profitably intervene more strongly. On InternVL2.5 ViewSpatial it repairs 566 answers and harms 169, versus GSC's 485/115. On InternVL2.5 MMSI, GEO repairs 41 versus 37 with the same two new errors. Across all 11 settings, GSC's maximum observed new-error rate is 2.28%, versus 11.55% for GEO.

5.4 Frame and role-swap interventions establish evidence dependence

Omitting the specified frame operations removes 60–80% of the ViewSpatial gain. Swapping entity roles *reverses* it on all three backbones, leaving accuracy 8.18, 2.50, and 6.04 points below the parent. These interventions show that the gains depend materially on frame alignment and correct entity bindings; they concern external evidence, not the parent's internal representation.

Scope. GSC depends on proposal recall, parsing, geometry, the parent scoring interface, and correction strength. Incorrect evidence can still overwhelm a correct semantic margin: in the white-pillow case, a 1.63-nat update exceeds the parent's 1.19-nat margin (Appendix E). Cached-input profiles measure 1.9 ms per SR3D correction and 76 μs for RGB fusion arithmetic; perception and parent inference precede these stages (Appendix D.7).

6 Related Work

Spatial evidence acquisition and representation. Chat-Scene and QuatRoPE build object-level representations for 3D language grounding (Huang et al., 2024; Zhou et al., 2026), while ZSVG3D executes spatial programs over scene objects (Yuan et al., 2024). MSSR acquires and prunes information toward a minimally sufficient set (Guo et al., 2026); Geometrically-Constrained Agent formalizes a task's reference frame and objective to constrain subsequent tool calls (Chen et al., 2026). Perception Programs (P^2) express tool outputs as compact language-native summaries (Janjua et al., 2026), and Geo3R presents measurements through structured geometric cards (Wang et al., 2026). Their prompt-level use of evidence is complementary in scope to GSC's decision-level update of a fixed candidate distribution.

GCA reports 47.6% on MMSI with Qwen3-VL-Thinking and iterative tool/code execution (Chen et al., 2026). This is a higher-accuracy operating point with repeated model calls.

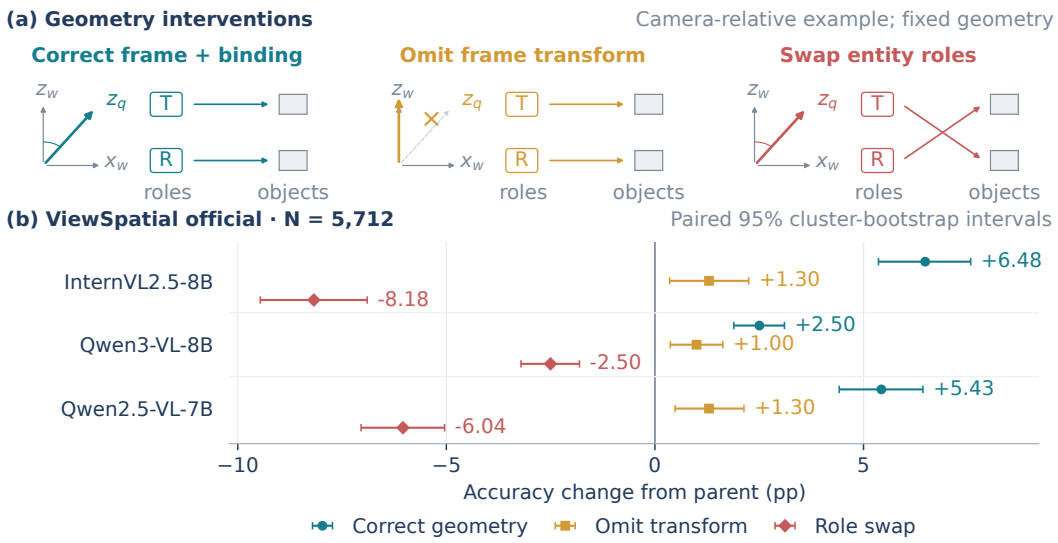

Figure 5: **Gains depend on grounded spatial evidence.** Top: a camera-relative example with correct frame transformation, omitted frame operations, and swapped target (T)/reference (R) roles. Bottom: gains over fixed P_0 on official ViewSpatial ($N = 5{,}712$), with paired cluster-bootstrap 95% intervals conditional on that deterministic swap. Appendix D.5 specifies the swapped roles; geometry is re-executed with the original execution mask.

GSC studies a deterministic update of already-computed candidate scores. Combining the two is an untested possibility.

Test-time correction. GRASP uses geometric counterfactuals to rectify intermediate trajectories and output distributions (Yan et al., 2026); Trace, Verify, and Correct revises reasoning after identifying contradicted spatial evidence (Yang et al., 2026b). DExperts and visual contrastive decoding contrast expert or input-conditioned distributions (Liu et al., 2021; Leng et al., 2024). RED derives multiplicative decoding through KL-constrained optimization (Yamaguchi et al., 2026). GSC instead studies the authority of an external partial assessment after parent scores and measurements are fixed. We account for regressions through negative flips (Yan et al., 2021).

Reference-relative evidence. Content-free calibration estimates answer bias from uninformative inputs (Zhao et al., 2021); domain-conditional PMI reweights answers by their task-conditioned prior likelihood (Holtzman et al., 2021). Prior-corrected Bayesian committee machines provide a precedent for reference denominators and recovery of prior behavior (Deisenroth & Ng, 2015). Probabilistic belief revision distinguishes likelihood evidence from Jeffrey-style posterior constraints (Chan & Darwiche, 2005). These connections place exponential tilting and reference correction within established methodology. Our contribution concerns partial geometric assessment: a same-support relation-null reference, neutral evidence outside that support, and controlled comparisons showing that allocation across the assessment boundary changes accuracy even when within-support conditional predictions agree.

7 Conclusion

Combining partial assessment with a full-question prediction requires deciding both how to rank assessed candidates and how to allocate probability across the boundary. GSC makes those decisions explicit through a same-support reference and finite evidence. Controlled comparisons isolate the consequences of allocation; cross-setting improvements, complementary repairs, and grounding interventions establish its practical relevance. The support-offset identity suggests reliability-aware allocation as a direction for future work.

REPRODUCIBILITY STATEMENT

The anonymous code repository provides browsable source; the supplementary archive `anonymous_replay_bundle.zip` contains the complete frozen inputs and standalone CPU implementation of fusion and paired statistics. It reconstructs the native 3D geometry from saved predicted boxes, labels, and question text, and rebuilds RGB geometric distributions from frozen angular violations. The replay regenerates all three main tables, reproduces the support-allocation analysis, and verifies predictions before and after an explicit cached-score option-interface repair. Perception and parent-model inference precede this replay boundary. Appendix F gives the tested commands and output paths; Appendices B–C specify the implementation and statistical units.

STATEMENT ON THE USE OF AI

AI tools assisted with **prose refinement**, **literature surveying and prior-art discovery**, and **codebase refinement and script debugging**. Verification included algebraic checks against the implementation, deterministic replay of frozen predictions, paired statistical recomputation, source checks for cited literature, independent compilation, and visual inspection. These checks and the replay boundary are documented in the supplement. The authors are responsible for the experimental record, claims, citations, and final submission.

REFERENCES

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision*, 2020. URL https://referit3d.github.io/.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Hei Chan and Adnan Darwiche. On the Revision of Probabilistic Beliefs Using Uncertain Evidence. *Artificial Intelligence*, 163(1):67–90, 2005. doi: 10.1016/j.artint.2004.09.005. URL https://www.sciencedirect.com/science/article/pii/S0004370204001559.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D Object Localization in RGB-D Scans using Natural Language. In *European Conference on Computer Vision*, 2020. URL https://arxiv.org/abs/1912.08830.

Zeren Chen, Xiaoya Lu, Zhijie Zheng, Pengrui Li, Lehan He, Yijin Zhou, Jing Shao, Bohan Zhuang, and Lu Sheng. Geometrically-Constrained Agent for Spatial Reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38689–38699, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_Geometrically-Constrained_Agent_for_Spatial_Reasoning_CVPR_2026_paper.html.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, et al. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*, 2024. URL https://arxiv.org/abs/2412.05271.

Marc Deisenroth and Jun Wei Ng. Distributed Gaussian Processes. In *International Conference on Machine Learning*, pp. 1481–1490, 2015. URL https://proceedings.mlr.press/v37/deisenroth15.html.

Yejie Guo, Yunzhong Hou, Wufei Ma, Meng Tang, and Ming-Hsuan Yang. Pursuing Minimal Sufficiency in Spatial Reasoning. In *International Conference on Learning Representations*,

2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c4ff64d68ba491b9048f00d25690d363-Abstract-Conference.html.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. In *Conference on Empirical Methods in Natural Language Processing*, pp. 7038–7051, 2021. doi: 10.18653/v1/2021.emnlp-main.564. URL https://aclanthology.org/2021.emnlp-main.564/.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, and Zhou Zhao. Chat-Scene: Bridging 3D Scene and Large Language Models with Object Identifiers. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2312.08168.

Muhammad Kamran Janjua, Hugo Silva, Di Niu, and Bahador Rashidi. Don't Show Pixels, Show Cues: Unlocking Visual Tool Reasoning in Language Models via Perception Programs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Janjua_Dont_Show_Pixels_Show_Cues_Unlocking_Visual_Tool_Reasoning_in_CVPR_2026_paper.html.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. ViewSpatial-Bench: Evaluating Multi-perspective Spatial Localization in Vision-Language Models. *arXiv preprint arXiv:2505.21500*, 2025. URL https://arxiv.org/abs/2505.21500.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts. In *Annual Meeting of the Association for Computational Linguistics*, 2021. URL https://arxiv.org/abs/2105.03023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv preprint arXiv:2303.05499*, 2023. URL https://arxiv.org/abs/2303.05499.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, et al. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In *IEEE International Conference on Robotics and Automation*, 2023. URL https://arxiv.org/abs/2210.03105.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.11651.

Mingyu Wang, Weilin Jin, Wenbo Li, Haoyang Huang, Tong Jia, and Ying Li. Geo3R: Mitigating Spatial Reasoning Hallucination in Multimodal Large Language Models. *arXiv preprint arXiv:2607.21085*, 2026. URL https://arxiv.org/abs/2607.21085.

Shin'ya Yamaguchi, Kosuke Nishida, and Daiki Chijiwa. Rationale-Enhanced Decoding for Multi-modal Chain-of-Thought. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19241–19252, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yamaguchi_Rationale-Enhanced_Decoding_for_Multi-modal_Chain-of-Thought_CVPR_2026_paper.html.

Jiadong Yan, Ke Zhang, Chenyang Zhao, Shoushan Li, and Xizhao Luo. GRASP: Awakening latent spatial reasoning in LVLMs via training-free geometric rectification. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63744.

Sijie Yan, Yuanjun Xiong, Kaustav Kundu, Shuo Yang, Siqi Deng, Meng Wang, Wei Xia, and Stefano Soatto. Positive-Congruent Training: Towards Regression-Free Model Updates. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2011.09161.

Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, Dahua Lin, Tai Wang, and Jiangmiao Pang. MMSI-Bench: A Benchmark for Multi-Image Spatial Intelligence. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/fe5d4bd3e2af823701b5d8f70a3c7602-Abstract-Conference.html.

Yang Yang, Jiawei Chen, Tairan Chen, and Zhaoxia Yin. Trace, Verify, and Correct: A Training-Free Framework for Spatial Reasoning in Multimodal LLMs. *arXiv preprint arXiv:2608.04759*, 2026b. URL https://arxiv.org/abs/2608.04759.

Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. Visual Programming for Zero-shot Open-Vocabulary 3D Visual Grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.15383.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate Before Use: Improving Few-shot Performance of Language Models. In *International Conference on Machine Learning*, pp. 12697–12706, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

Shengli Zhou, Minghang Zheng, Feng Zheng, and Yang Liu. Scalable Object Relation Encoding for Better 3D Spatial Reasoning in Large Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhou_Scalable_Object_Relation_Encoding_for_Better_3D_Spatial_Reasoning_in_CVPR_2026_paper.html.

A PROPERTIES OF THE CORRECTION

Proposition 1 (Neutrality and support)**.** *For $\epsilon > 0$ and finite $C, \gamma \geq 0$, Eq. 3 defines a distribution with the same support as p_0. If $g = n$, it equals p_0. For any two candidates outside S with positive parent probability, their relative odds are unchanged.*

Proof. Each multiplier $\exp(\gamma e_i)$ is finite and strictly positive. Hence a candidate has zero corrected probability exactly when its parent probability is zero. If $g = n$, every log ratio is zero and the normalization is one. Outside S, $g_i = n_i = 0$, so $e_i = 0$; the shared normalization cancels in the odds of two such candidates. □

Neutral evidence does not keep each absolute probability fixed. Mass can move between supported and unsupported candidates through normalization. Likewise, preserving the odds among unsupported candidates does not prevent a supported candidate from overtaking them.

Proposition 2 (Probability allocation across assessment support)**.** *Suppose $e_i = 0$ for $i \notin S$ and $0 < \alpha = p_0(S) < 1$. Then Eq. 4 holds, and within S,*

$$p_\gamma(i \mid S) = \frac{p_0(i \mid S)\exp(\gamma e_i)}{M_\gamma}.$$

Proof. The contribution to the normalizer from S is αM_γ, and the contribution from its complement is $1-\alpha$. Hence $p_\gamma(S) = \alpha M_\gamma/(\alpha M_\gamma + 1 - \alpha)$. Taking odds gives Eq. 4; conditioning the individual probabilities on S gives the second identity. If $\alpha = 0$ or 1, finite positive multipliers preserve that total mass. The odds expression is used only for $0 < \alpha < 1$. □

The sign of the change is determined by $M_\gamma - 1$, not by support size alone. Neutral evidence outside S preserves relative odds within the complement, not necessarily its total probability. If $g = n$, all multipliers equal one, including inside S.

Proof of Corollary 1. Let $s = |S| > 0$, $a = (1/s + \epsilon)^\gamma$, and $b = \epsilon^\gamma$. On S, the unnormalized reference weights equal the smoothed-product weights divided by the common constant a. On S^c, the two weights are $p_0(i)$ and $b\,p_0(i)$, respectively. Conditioning cancels these constants separately on each side, proving the conditional identities. If $U = \sum_{i\in S} p_0(i)(g_i + \epsilon)^\gamma$ and $V = p_0(S^c)$, their support odds are $U/(aV)$ and $U/(bV)$, whose ratio is b/a. This gives Eq. 5. If $p_0(S) = 0$ or 1, both rules preserve that total mass and agree globally on the side with positive mass. The result is for unclipped updates, equal γ, ϵ, and a uniform same-support reference; clipping can also change within-support predictions. With a shared deterministic tie rule, any different winners must lie on opposite sides of the boundary.

A common evidence offset controls support allocation. Let $e'_i = e_i + b$ on S and $e'_i = 0$ outside, where e_i is the final evidence after clipping. For $i, j \in S$, $e'_i - e'_j = e_i - e_j$, so their odds are unchanged. Meanwhile $M'_\gamma = \exp(\gamma b) M_\gamma$; Eq. 4 therefore multiplies the support's odds by $\exp(\gamma b)$. This identity concerns an offset to final evidence; moving pre-clipping scores can additionally change which candidates saturate.

Proposition 3 (Bounded-product support preference)**.** *Let $0 < \epsilon \leq e^{-C}$, $C \geq 0$, $\gamma \geq 0$, and $0 < p_0(S) < 1$. With $b_i = \mathrm{clip}(\log(g_i + \epsilon), -C, C)$, bounded product cannot decrease the odds of S against its complement.*

Proof. Outside S, $b_i = -C$; inside S, $b_i \geq -C$. Its support-odds multiplier is therefore

$$\mathbb{E}_{i\sim p_0(\cdot|S)} \exp\{\gamma(b_i + C)\} \geq 1.$$

The inequality is strict when $\gamma > 0$ and positive parent mass is assigned to at least one inside candidate with $b_i > -C$. Otherwise equality is possible. In particular, a uniform $g = n$ can still increase support odds under BP; uniformity alone does not imply a strict increase in every configuration. □

A reparameterization of bounded product. Adding C to every clipped BP log score changes only the common normalizer. Thus BP is equivalently generated by

$$\widetilde{b}_i = \mathrm{clip}\left(\log \frac{g_i + \epsilon}{e^{-C}}, 0, 2C\right).$$

For $\epsilon \le e^{-C}$, $\widetilde{b}_i = 0$ outside S and is nonnegative inside. This is a fixed-reference, one-sided evidence rule; it is not the same as GSC's normalized same-support reference and symmetric clipping. An effective reference value would satisfy $n_\star + \epsilon = e^{-C}$, giving a nominal size $1/(e^{-C} - \epsilon) \approx 7.44$ at the reported constants. Empirical support medians are three or four, so similarity to this nominal size is not an established explanation of BP's observed accuracy. Under four-option complete support, BP's lower clipping threshold is $g \approx 0.1343$, compared with 0.0330 for GSC. Matching or similar accuracies does not make their null semantics identical.

Proposition 4 (Variational interpretation and odds)**.** *For $\gamma > 0$, p_γ uniquely maximizes*

$$\mathcal{J}(p) = \sum_i p(i)e_i - \frac{1}{\gamma} D_{\mathrm{KL}}(p \| p_0) \tag{7}$$

over distributions supported on $\mathrm{supp}(p_0)$*. Its pairwise log odds satisfy Eq. 6.*

Proof. Let $Z = \sum_i p_0(i) \exp(\gamma e_i)$. Substitution gives $\mathcal{J}(p) = \gamma^{-1}\{\log Z - D_{\mathrm{KL}}(p \| p_\gamma)\}$. The maximum is attained uniquely at $p = p_\gamma$. Dividing two corrected probabilities gives Eq. 6, and $|e_i - e_j| \le 2C$ gives the bound. □

This is the standard KL-regularized exponential tilt, not a new optimization theorem. With the reported constants, the worst-case pairwise shift is eight nats; the bound should not be interpreted as a small-error guarantee. The actual shift can be much smaller.

Weak and strong evidence limits. Equation 1 follows because $g_i^\eta \to 1$ for $i \in S$ and remains zero outside S for every $\eta > 0$. Smoothed products and GSC instead converge to p_0 as their strength tends to zero. At the other extreme, let $E = \arg\max_{i: p_0(i) > 0} e_i$. Then

$$\lim_{\gamma \to \infty} p_\gamma(i) = \frac{p_0(i)\mathbf{1}[i \in E]}{\sum_{j \in E} p_0(j)}. \tag{8}$$

Clipping can produce ties in E even when g has a unique maximizer. Thus the strong-correction limit is not generally the GEO decision. The replay finds multiple maximum clipped-evidence candidates in 1,592 executed SR3D questions and 778 ScanRefer questions under the QuatRoPE adaptation.

Uniform reference and clipping. When $S = \mathcal{C}$ and clipping is removed, the denominator $n_i + \epsilon$ is constant. Therefore $p_\gamma(i) \propto p_0(i)(g_i + \epsilon)^\gamma$. With clipping, shifting the origin before clipping can change which candidates saturate. This explains why RGB reference and non-reference controls can differ despite uniform n; it does not by itself explain which will be more accurate.

A constructed null-evidence example. Let $p_0 = (0.55, 0.30, 0.15)$, $S = \{2, 3\}$, and $g = n = (0, 0.5, 0.5)$. Geometry supplies no within-support preference. GSC returns p_0 exactly. The raw product returns $(0, 2/3, 1/3)$; BP with the reported constants returns approximately $(0.082, 0.612, 0.306)$. Both change the winner because support itself becomes evidence. This example illustrates the reference's neutrality property. The empirical comparison in Table 9 determines its practical importance on the tested distributions.

Coordinate changes. Distance-based predicates are invariant to joint translations and rotations of all positions. Vertical predicates require the vertical axis to transform with the scene. RGB angles are invariant when bearings and their frame basis co-transform; applying a transformation to only one side changes the evaluated claim. Fixed metric temperatures in the 3D executor do not give scale invariance: multiplying coordinates while holding temperature fixed can change fusion, even if a distance ranking is unchanged.

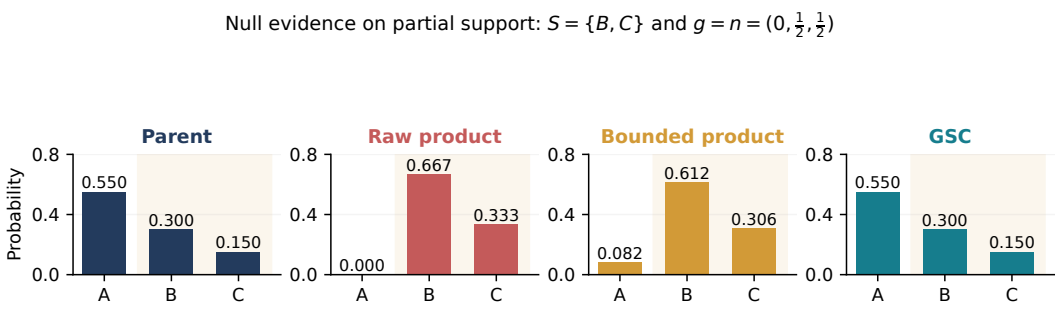

Figure 6: Constructed null-evidence example, not a benchmark result. Geometry is uniform on the shaded support $S = \{B, C\}$. Both raw product and BP change the winner even though $g = n$; GSC preserves the parent exactly. BP uses $\epsilon = 10^{-3}$, $C = 2$, and $\gamma = 2$.

B Implementation Details

B.1 3D executor and semantic interface

Mask3D supplies up to 100 proposals per scene. The semantic branch scores object identifiers; its candidate ordering is aligned with the cached proposal assignment before fusion. All geometry methods in a matched comparison use the same proposal pool, semantic scores, and assignment. The QuatRoPE-adapted parent was updated on 1,600 SR3D training expressions from 489 physical scenes for 250 optimizer steps, before correction was studied. Only rank-16 LoRA parameters were updated (scale 16), using AdamW with learning rate 10^{-5}, zero weight decay, micro-batch one, accumulation eight, 10 warmup steps, gradient norm bound one, and seed zero. The parent retains the original geometric encoding and uses `knn=0`. Chat-Scene uses the independently trained checkpoint `ckpt_01_3446`. Correction itself has no learned parameters or gradient updates.

Candidate likelihood. Both native parents teacher-force the three-token answer $y_k = (\texttt{<OBJ}k\texttt{>}, \texttt{.}, \texttt{</s>})$, using identifier tokens `<OBJ000>` through `<OBJ099>`. They sum autoregressive token log probabilities, without length normalization:

$$s_k = \sum_{t=1}^{3} \log P(y_{k,t} \mid \text{prefix}, y_{k,<t}), \qquad p_0(k) = \frac{\exp(s_k - \max_j s_j)}{\sum_j \exp(s_j - \max_\ell s_\ell)}.$$

Display index k maps to the proposal through a fixed seed-zero assignment shared across parents; references in the question use that same assignment. The prefix combines the official conversational system prompt, *The conversation centers around an indoor scene. Object information: [*, all object embeddings, *]. USER:*, the complete saved grounding prompt, and *ASSISTANT:*. The manifest's prompt is used verbatim, for example *Share the ID of the object that best fits the description "<description>".*, rather than a bare description. The supplement's `data/protocol.json` provides the full system string, all nine prompt templates, exact spacing/tokenization rules, and source-code hashes. Object encoders remain parent-specific.

The parser chooses the earliest word-boundary match to a predicted class in the question, preferring the longer class name on a positional tie. An anchor class is the nearest matching class within an eight-word window after the relation phrase. All proposals of the matched anchor class remain hypotheses. Supported relation logits use the following functions, with proposal centers x_i, anchor

sets A, B, and global vertical coordinate z:

$$d(i, A) = \frac{1}{|A|} \sum_{a \in A} \|x_i - x_a\|_2, \tag{9}$$

$$\ell_i^{\text{closest}} = -d(i, A)/\tau_r, \qquad \ell_i^{\text{farthest}} = d(i, A)/\tau_r, \tag{10}$$

$$\ell_i^{\text{near}} = -d(i, A)/\tau_{\text{near}}, \qquad \ell_i^{\text{far}} = d(i, A)/\tau_{\text{near}}, \tag{11}$$

$$\ell_i^{\text{above/below}} = \pm \frac{1}{|A|\tau_z} \sum_{a \in A} (z_i - z_a), \tag{12}$$

$$\ell_i^{\text{between}} = -\frac{1}{|A||B|\tau_r} \sum_{a \in A, b \in B} \left\| x_i - \frac{x_a + x_b}{2} \right\|_2, \tag{13}$$

$$\ell_i^{\text{middle}} = -\|x_i - \bar{x}\|_2/\tau_{\text{near}}. \tag{14}$$

Here $\bar{x}$ is the mean center of the scene proposals. Between is a midpoint-distance proxy rather than a complete logical test of betweenness; above and below are signed-height proxies rather than support or contact detectors. Closest, farthest, and between use $\tau_r = 0.5$; near, far, and middle use $\tau_{\text{near}} = 0.75$; vertical predicates use $\tau_z = 0.5$, with coordinates in meters. These are fixed executor choices. Distances are evaluated per anchor hypothesis and then averaged; averaging coordinates first would implement a different function.

If a target class is identified, the support contains its proposals. If not, non-ranking predicates may execute on all proposals; closest and farthest are skipped without a target class. Logits of successfully bound predicates are summed. At least one executable relation is sufficient to obtain g, even if other phrases remain unsupported. A question with no executable relation returns exactly the parent distribution. The rule-based parser and the geometric proxies are potential sources of binding and interpretation errors.

B.2 RGB PERCEPTION AND EXECUTION

The geometry pipeline uses Grounding DINO Swin-T with box and text thresholds 0.35 and 0.25, SAM 2.1 Hiera-Large masks, and VGGT-1B camera/depth/world-point estimates. These models process the permitted benchmark images; their outputs are cached and reused across VLM backbones and fusion arms. Frame conventions are part of the executor: camera directions use right/up/forward axes, and a scene-simulation query of the form *stand at A facing B* defines its forward axis through the anchored A-to-B direction with horizontal projection. Diagonal option directions are normalized before angular comparison.

The implemented families are angular direction judgments. Missing entity points, degenerate direction vectors, unresolved frames, and options requiring unsupported orientation evidence trigger fallback. For multiple choice, *every option* must be executable; the code never scores only a favorable subset of options. In contrast, native 3D supports partial relation execution within an expression. These are different execution contracts.

Synthetic temperature selection. The fixed RGB temperature uses a synthetic angular-scale rule without benchmark labels. With NumPy `default_rng(20260907)`, generate 100,000 scenes for each $(\sigma, N) \in \{0.1, 0.2\} \times \{20, 100\}$ in lexicographic order. Draw separation $d \sim U(0.5, 3)$ and depth scale $z \sim U(1, 4)$; perturb reference zero and target $(0, 0, d)$ independently by $\mathcal{N}(0, \sigma^2 z^2 I/N)$. This approximates aggregate location error directly, rather than simulating N points and taking their median. In each cell, measure the angle of the perturbed target-minus-reference vector against $(0, 0, 1)$ and its 95th percentile a_{95}. Set $c = \max a_{95}$ and $w = \min(\pi/2 - a_{95})$, then choose the nearest grid value to $\sqrt{cw}$ from $\{0.05, 0.10, \ldots, 1.50\}$. The values $c = 0.397627$, $w = 1.173169$, and $\sqrt{cw} = 0.682996$ radians select $\tau = 0.70$. This geometric-mean rule balances a worst-cell correct-angle bound and an orthogonal-direction margin proxy; it does not optimize RGB accuracy or likelihood.

B.3 Option probabilities, generation, and failures

The option-scoring prompt appends *Answer with the option's letter from the given choices directly.* to the original question and legal images. An assistant continuation starts with `Answer:`. The scorer requests one token, temperature zero, and the top 20 log probabilities at that position. Tokens are stripped and uppercased. For each option letter, the first matching token variant in the returned list is retained; the available letter scores are exponentiated and normalized. Missing letters receive zero. This is a truncated-token scoring interface, not a sum over all tokenizations. An audit of all 20,136 official question–backbone records finds no missing legal-option scores or numerical zeros in the cached scorer distributions.

If no letter is present, a uniform option fallback is used. Request errors in the recorded four-choice protocols also use a uniform distribution and remain in the denominator. The final official ViewSpatial snapshots have no such failures for any of the three models. On official MMSI, InternVL2.5 has three API failures and the two Qwen models have none. These failures remain in the final tables.

The free-generation baseline uses deterministic decoding with a 1,024-token limit. Answer parsing prioritizes an explicit answer tag, then the last answer declaration, then a standalone option letter. All six final official free-generation files have parseable outputs on every question. Numerical hints expose the same option-level angular violations as text; this is our matched-evidence prompting control, not a reproduction of the complete perception-program system of Janjua et al. (2026).

Option-parser audit and repair. The historical fusion loader split option strings only at commas. On 51 of the 1,000 MMSI questions, period or semicolon separators caused it to retain only A, although the original model-score cache contained all four legal options. This omitted 153 option entries per backbone and the target on 39 questions. All 51 questions were non-executed fallbacks; none belonged to the first-50 tool-agent profile. We restore the complete cached scores for *all* affected questions, independently of target correctness, and apply the same corrected fallback to every fusion arm. Hint already retained the complete cached fallback scores, so its distributions and accuracies remain unchanged. Free-generation outputs are unchanged. No model is queried and no geometric execution is added.

These corrected scores define the primary results throughout the paper. MMSI parent/GSC accuracies become 27.60/31.10 for InternVL2.5, 31.70/34.50 for Qwen3-VL, and 27.60/32.50 for Qwen2.5-VL, compared with pre-repair values 27.10/30.60, 31.00/33.80, and 27.10/32.00. The exact pre-repair reference is `expected_legacy.json` (SHA-256 prefix `f3ec9653161e634d`); the replay README gives its full hash and input-record hashes. Because the repair affects only shared fallback decisions, all GSC-minus-parent effects, repair/harm counts, and paired intervals are unchanged. GSC-minus-Free comparisons are recomputed. The anonymous bundle preserves the historical records and reference metrics, checks them first, and then checks the explicitly repaired primary results.

B.4 Replaying controls without model inference

The 3D analysis loads the frozen semantic-score cache, predicted validation attributes, saved candidate assignments, and the existing executor. It reconstructs the IoU of each predicted proposal against the target box. Replayed GSC candidate IDs and correctness agree with every saved prediction on the five full manifests. GEO, BP, and the neutral-outside/uncentered-inside control are then computed from exactly those scores and measurements. Their full-manifest results were derived during manuscript preparation; no hyperparameter was selected on these replay results.

The neutral-outside control is $e_i = \mathbf{1}[i \in S]\,\mathrm{clip}(\log(g_i + \epsilon), -C, C)$. Its poor performance is specific to this evidence origin. BP instead evaluates the same clipped log score outside S as well. Calling either one simply *without normalization* would obscure an important difference.

C Evaluation protocol and development exposure

Native 3D manifests. SR3D contains validation-scene questions with scorable target annotations. NR3D additionally follows the recorded `correct_guess=True` filter, yielding 7,899 questions from 8,584 validation rows before that filter. The proposal evaluation does not condition on target

recall, anchor availability, or parsing success. At IoU 0.5, 3,304 SR3D targets and 1,783 NR3D targets are unreachable by the proposal pool; these questions remain in the denominator and cannot receive a correct prediction. The complete ScanRefer manifest has 9,508 questions.

Accuracy is the fraction whose selected proposal box has IoU at least 0.5 with the annotated target box. It is not merely exact agreement with the index of the highest-IoU proposal. Both parent and correction arms are evaluated using the same geometry and target annotations. This proposal-based adapted setting differs from an evaluation in which ground-truth object instances are candidates.

Development and diagnostic exposure. The SR3D development set has 400 questions and was used for selecting correction strength. The full SR3D manifest contains it. Earlier NR3D and ScanRefer development subsets contain 600 and 400 questions, respectively. Removing the known development rows leaves 17,326 SR3D, 7,299 NR3D, and 9,108 ScanRefer questions (Table 4). Parent development and method design also involved earlier subset inspection; an exclusion cannot erase this history.

RGB angular temperature was calibrated synthetically, and γ was not tuned on RGB answers. However, implementation diagnostics inspected MMSI examples to resolve coordinate conventions. A traceable union of 40 diagnostic question IDs can be reconstructed from the recorded inspected groups. Table 5 removes these IDs, leaving 960 questions. The positive effects persist. This is a sensitivity check against the known inspected examples, not proof of a completely unexposed development process.

Table 4: GSC after excluding known 3D development rows. Changes are against the corresponding parent on the same remaining questions.

Stack	Benchmark	N	GSC	Δ [95% CI]
Quat.	SR3D	17326	42.67	+7.30 [5.75, 8.88]
Quat.	NR3D	7299	35.64	+1.38 [0.84, 1.95]
Quat.	ScanRefer	9108	50.29	+0.47 [0.17, 0.78]
Chat-Scene	SR3D	17326	42.02	+6.94 [5.65, 8.41]
Chat-Scene	NR3D	7299	35.96	+1.73 [1.13, 2.33]

Bootstrap units. For question k, let $b_k, m_k \in \{0, 1\}$ denote parent and method correctness. The paired effect is $100N^{-1}\sum_k(m_k - b_k)$. Each bootstrap replicate samples clusters with replacement and divides the summed correctness difference by the summed number of sampled questions. We use 20,000 draws and seed 20260907, reporting percentile intervals.

Native 3D uses 123 physical scenes for SR3D, 130 for NR3D, and 141 for ScanRefer. ViewSpatial's original grouping has 1,338 clusters: 279 scan identifiers and 1,059 other image groups, with 93 singleton clusters overall. The other-image groups are not all singletons. Merging rescans of the same physical scene reduces this to 1,190 clusters and preserves positive intervals. MMSI has 1,000 image groups that are singletons under the recorded grouping; its cluster bootstrap therefore equals question-level resampling. Repeated evaluations of a dataset under different parents are not independent new datasets, and overlapping validation subsets are not added into a larger test sample.

Table 5: Sensitivity to more conservative ViewSpatial grouping and exclusion of known MMSI diagnostics. All effects compare GSC with P_0.

Sensitivity	Backbone	N	Δ [95% CI]
ViewSpatial: merge rescans	InternVL2.5-8B	5712	+6.48 [5.13, 7.76]
ViewSpatial: merge rescans	Qwen3-VL-8B	5712	+2.50 [1.85, 3.14]
ViewSpatial: merge rescans	Qwen2.5-VL-7B	5712	+5.43 [4.25, 6.56]
MMSI: omit 40 diagnostics	InternVL2.5-8B	960	+2.71 [1.67, 3.85]
MMSI: omit 40 diagnostics	Qwen3-VL-8B	960	+2.50 [1.56, 3.54]
MMSI: omit 40 diagnostics	Qwen2.5-VL-7B	960	+3.96 [2.71, 5.31]

Repair and harm denominators. Define $R = \sum_k \mathbf{1}[b_k = 0, m_k = 1]$ and $H = \sum_k \mathbf{1}[b_k = 1, m_k = 0]$. Then $\Delta\text{Acc} = 100(R - H)/N$. Figure 8 plots $100R/N$ and $100H/N$. A conditional negative-flip rate would instead divide H by the number of initially correct questions. A relative reduction in new errors between methods is $(H_{\text{baseline}} - H_{\text{method}})/H_{\text{baseline}}$. These quantities have different denominators. We do not use a lower H alone as evidence of a superior method: the identity operation has $R = H = 0$.

D COMPLETE RESULTS AND CONTROLS

Table 6: **Complete official RGB interface comparison, including MIX.** Accuracy (%); best per row is bold and GSC is shaded. Hint requires an additional model call and was not run for Qwen2.5-VL.

Benchmark	Backbone	Free	P_0	GEO	MIX	BP	Hint†	GSC
ViewSpatial	InternVL2.5	41.49	41.07	**48.02**	46.88	47.48	44.07	47.55
	Qwen3-VL	40.41	40.90	**45.85**	42.31	43.01	45.26	43.40
	Qwen2.5-VL	36.80	36.96	**43.98**	40.91	41.77	—	42.38
MMSI	InternVL2.5	28.30	27.60	**31.50**	30.40	31.00	28.10	31.10
	Qwen3-VL	31.10	31.70	**35.90**	33.60	33.90	34.90	34.50
	Qwen2.5-VL	27.70	27.60	**32.60**	32.00	32.40	—	32.50

Table 7: Paired official RGB comparisons with deterministic free generation. All six GSC-minus-Free intervals are positive. Free and option scoring are alternative readouts, not identical compute budgets.

Benchmark	Backbone	Free	GSC	GSC − Free [95% CI]
ViewSpatial	InternVL2.5	41.49	47.55	+6.06 [4.88, 7.21]
ViewSpatial	Qwen3-VL	40.41	43.40	+2.99 [2.20, 3.78]
ViewSpatial	Qwen2.5-VL	36.80	42.38	+5.58 [4.49, 6.66]
MMSI	InternVL2.5	28.30	31.10	+2.80 [0.90, 4.70]
MMSI	Qwen3-VL	31.10	34.50	+3.40 [1.40, 5.40]
MMSI	Qwen2.5-VL	27.70	32.50	+4.80 [3.00, 6.70]

Table 8: Official RGB paired effects and correctness transitions.

Benchmark	Backbone	N	R	H	Δ [95% CI]
ViewSpatial	InternVL2.5-8B	5712	485	115	+6.48 [5.36, 7.57]
ViewSpatial	Qwen3-VL-8B	5712	212	69	+2.50 [1.89, 3.10]
ViewSpatial	Qwen2.5-VL-7B	5712	404	94	+5.43 [4.42, 6.42]
MMSI	InternVL2.5-8B	1000	37	2	+3.50 [2.30, 4.80]
MMSI	Qwen3-VL-8B	1000	29	1	+2.80 [1.80, 3.90]
MMSI	Qwen2.5-VL-7B	1000	51	2	+4.90 [3.60, 6.30]

The stricter-parser mixture predates the extended executor and is useful as a historical control, rather than a same-evidence ablation. Its accuracies are 43.22/34.56/49.84 on SR3D/NR3D/ScanRefer for the QuatRoPE adaptation and 42.18/34.80 on SR3D/NR3D for Chat-Scene. The main tables use extended-parser MIX to maintain the matched comparison. Same-executor GEO reaches 39.83/29.97/44.55 with the QuatRoPE adaptation and 39.74/30.57 with Chat-Scene. Thus, geometry-only decisions improve structured SR3D but lose accuracy on the natural-language NR3D and ScanRefer evaluations.

D.1 EXPONENT-MATCHED AND CLIPPING CONTROLS

We replay three additional controls with strength two: the raw product $p_0 g^2$; the smoothed, unclipped product $p_0(g + \epsilon)^2$; and unclipped GSC, $p_0[(g + \epsilon)/(n + \epsilon)]^2$. All are normalized over the same candidate set and use the same fallback when no relation executes. No setting is selected on these results.

Table 9: Same-executor 3D controls recomputed on all full manifests. BP is competitive with GSC; every GSC-minus-BP interval includes zero.

Stack	Benchmark	P_0	GEO	BP	GSC − BP [95% CI]
Quat.	SR3D	35.70	39.83	43.23	-0.16 [-0.88, 0.62]
Quat.	NR3D	33.76	29.97	35.32	-0.18 [-0.51, 0.16]
Quat.	ScanRefer	49.86	44.55	50.32	+0.01 [-0.29, 0.30]
Chat-Scene	SR3D	35.26	39.74	42.37	-0.08 [-0.63, 0.60]
Chat-Scene	NR3D	34.05	30.57	35.76	+0.03 [-0.28, 0.31]

Table 10: Full native 3D repair–harm counts for raw product and GSC.

Stack	Benchmark	R (Prod)	H (Prod)	R (GSC)	H (GSC)
Quat.	SR3D	1979	650	1711	404
Quat.	NR3D	228	298	179	70
Quat.	ScanRefer	155	289	113	69
Chat-Scene	SR3D	1884	621	1600	355
Chat-Scene	NR3D	239	296	201	64

Table 11: Support-related failures. Outside S: initially correct parent winners outside the geometric support; Kept: those still correct under GSC. The last columns partition the product's new errors by whether its parent winner was outside S.

Stack	Benchmark	Outside S	Kept	Prod harm	Outside harm	Share
Quat.	SR3D	1015	961	650	470	72.3%
Quat.	NR3D	422	401	298	273	91.6%
Quat.	ScanRefer	515	504	289	255	88.2%
Chat-Scene	SR3D	1036	995	621	462	74.4%
Chat-Scene	NR3D	389	378	296	266	89.9%

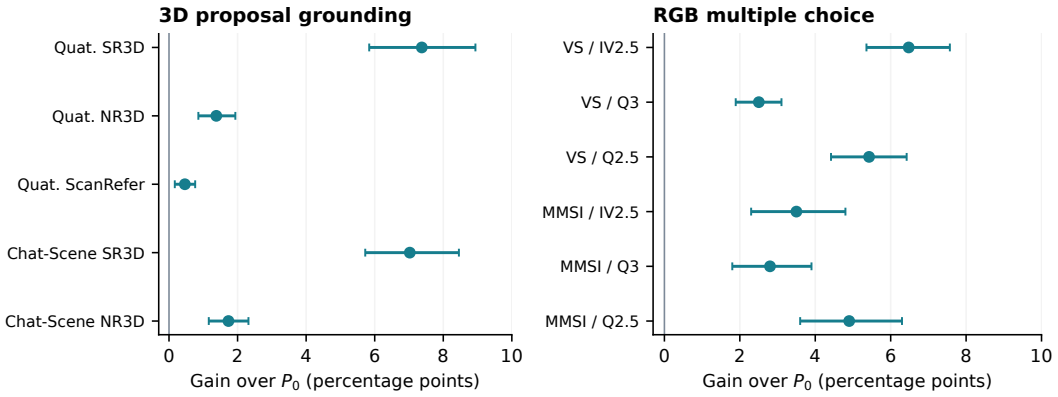

Figure 7: All 11 main GSC effects against their corresponding frozen parent, with paired cluster-bootstrap 95% intervals. The settings share datasets and are not treated as independent replications.

Table 12: Component controls on all five full 3D manifests, including neutral-outside for both parents. Smooth_2 adds $\epsilon = 10^{-3}$ to geometry but has no reference or clipping. The small difference between GSC and its unclipped version prevents attributing the gains to clipping alone.

Parent	Benchmark	Prod_2	Smooth_2	Unclipped	BP	Neutral outside	GSC
QuatRoPE	SR3D	45.06	45.01	43.07	43.23	35.98	43.07
QuatRoPE	NR3D	33.04	33.05	35.11	35.32	33.35	35.14
QuatRoPE	ScanRefer	48.42	48.45	50.34	50.32	48.94	50.33
Chat-Scene	SR3D	44.56	44.52	42.39	42.37	36.47	42.28
Chat-Scene	NR3D	33.57	33.61	35.83	35.76	34.07	35.79

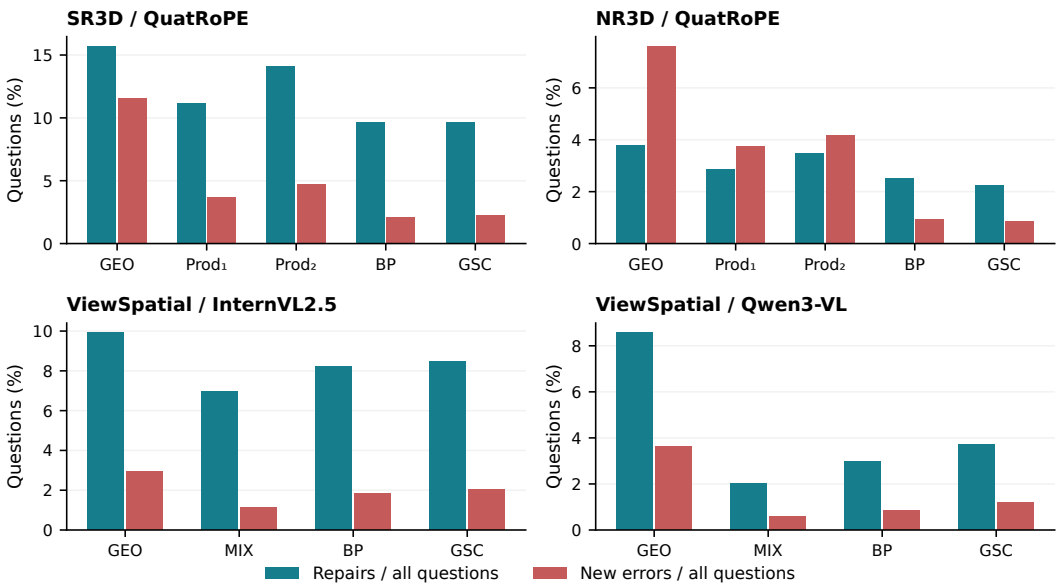

Figure 8: Repair and new-error counts divided by the full question denominator. This supplements the net-gain operating curves with the individual components at the frozen strength.

The raw exponent-two product exceeds GSC on SR3D by 1.99 points with the QuatRoPE adaptation (GSC-minus-product interval [-3.03, -0.81]) and 2.27 points with Chat-Scene ([-3.05, -1.33]). GSC instead exceeds that product on NR3D by 2.10 points ([1.48, 2.77]) and 2.22 points ([1.54, 2.93]), respectively; its ScanRefer advantage is 1.90 points ([1.39, 2.43]). Smoothing alone barely changes the raw exponent-two results.

GSC's reference places candidates outside S at zero evidence even without clipping. Removing clipping changes its accuracy by 0.00, -0.04, and +0.01 points on the three QuatRoPE-adaptation benchmarks and +0.11/+0.04 points on Chat-Scene SR3D/NR3D. For BP, clipping raises the absolute-score floor outside S from $\log \epsilon$ to $-C$. The empirical distinction is therefore the treatment of incomplete support through a suitable evidence origin or floor, rather than a universal accuracy gain from clipping as an isolated component.

D.2 Boundary allocation and interface audits

All quantities in this subsection are computed from frozen per-question caches with the same paired bootstrap as the main tables. The portable analysis command reproduces these tables, the conditional identities, and the finite-grid checks. No model is trained or queried for these analyses.

Table 13: **Boundary-only comparison on full 3D manifests.** Unclipped GSC minus Smooth_2; paired intervals and deltas use unrounded correctness counts. The two distributions agree conditionally on each side of S. Every changed prediction crosses that boundary.

Parent	Benchmark	N	Smooth_2	Unclipped	Δ [95% CI]
QuatRoPE	SR3D	17,726	45.01	43.07	-1.94 [-2.96, -0.78]
Chat-Scene	SR3D	17,726	44.52	42.39	-2.13 [-2.88, -1.21]
QuatRoPE	NR3D	7,899	33.05	35.11	+2.05 [1.43, 2.72]
Chat-Scene	NR3D	7,899	33.61	35.83	+2.22 [1.56, 2.92]
QuatRoPE	ScanRefer	9,508	48.45	50.34	+1.88 [1.37, 2.41]

The maximum observed discrepancy between the conditional distributions is below 8×10^{-16}; no changed prediction remains on the same side of S. The natural-language gains therefore isolate support allocation. On structured SR3D the change is detrimental, consistently with the value of

support selection for those expressions. This is an exact contrast between the stated controls, not an attribution of every difference between clipped GSC and other methods to allocation alone.

Table 14: **Where correction executes.** GSC versus its parent on executed questions. Coverage uses the full manifest denominator; gains and paired 95% intervals use only executed questions. Point gains equal full-denominator gains divided by coverage, but intervals are recomputed on the executed clusters.

Parent	Benchmark	Executed / N	Δ_{exec} [95% CI]	R	H
QuatRoPE	SR3D	13,055/17,726	+10.01 [8.11, 11.84]	1,711	404
Chat-Scene	SR3D	13,055/17,726	+9.54 [7.87, 11.21]	1,600	355
QuatRoPE	NR3D	2,680/7,899	+4.07 [2.59, 5.61]	179	70
Chat-Scene	NR3D	2,680/7,899	+5.11 [3.47, 6.66]	201	64
QuatRoPE	ScanRefer	2,668/9,508	+1.65 [0.60, 2.74]	113	69
InternVL2.5	ViewSpatial	1,545/5,712	+23.95 [20.59, 27.22]	485	115
Qwen3-VL	ViewSpatial	1,545/5,712	+9.26 [7.22, 11.29]	212	69
Qwen2.5-VL	ViewSpatial	1,545/5,712	+20.06 [17.03, 23.06]	404	94
InternVL2.5	MMSI	88/1,000	+39.77 [28.41, 51.14]	37	2
Qwen3-VL	MMSI	88/1,000	+31.82 [21.59, 42.05]	29	1
Qwen2.5-VL	MMSI	88/1,000	+55.68 [44.32, 67.05]	51	2

Table 15: **Observed cross-setting stability.** Number of evaluated settings, settings with negative point gain, minimum gain over the corresponding parent, and maximum new-error rate. The extrema describe this evaluation matrix and are not out-of-distribution guarantees.

Rule	Settings	$\Delta < 0$	Min. Δ	Max. harm count	Max. $100H/N$
GEO	11	3	-5.31	2,048/17,726	11.55
MIX	11	3	-1.05	736/17,726	4.15
Prod_2	5	3	-1.44	838/17,726	4.73
Neutral outside	5	2	-0.93	900/17,726	5.08
BP	11	0	+0.45	377/17,726	2.13
GSC	11	0	+0.46	404/17,726	2.28

Table 16: **Official RGB GSC-minus-BP paired comparisons.** Three intervals are strictly positive; the others include zero. These results are not statistical equivalence tests.

Parent	Benchmark	N	GSC−BP [95% CI]	R	H
InternVL2.5	ViewSpatial	5,712	+0.07 [-0.10, 0.24]	15	11
Qwen3-VL	ViewSpatial	5,712	+0.39 [0.10, 0.67]	42	20
Qwen2.5-VL	ViewSpatial	5,712	+0.61 [0.32, 0.91]	51	16
InternVL2.5	MMSI	1,000	+0.10 [0.00, 0.30]	1	0
Qwen3-VL	MMSI	1,000	+0.60 [0.20, 1.10]	6	0
Qwen2.5-VL	MMSI	1,000	+0.10 [-0.20, 0.50]	2	1

Across the four 3D development and expansion grids, each positive-exponent raw point has at least one sampled GSC point with strictly lower H and higher $(R-H)/N$. The strengths of the compared points may differ; the common zero-strength origin is excluded. This finite-grid statement neither asserts same-exponent dominance nor interpolates between points. At exponent $1/16$, the raw product already incurs 74–86% of its exponent-two new-error count, whereas GSC incurs at most one new error. These observations are consistent with persistent support exclusion but do not measure a constant error floor as strength tends continuously to zero: which remaining candidate wins can still change, and multiple proposals can count as correct.

The relation-null condition does not declare support statistically uninformative. BP's one-sided reference and GSC's signed reference can both recover useful corrections, while structured SR3D favors stronger support commitment. All reported main results retain the preselected $\gamma = 2$.

Table 17: **Exclusion diagnosis at both product exponents.** Numerators count new errors whose initially correct parent winner lies outside S; denominators count all new errors. An alternative proposal can satisfy the IoU criterion, so these are decision diagnostics rather than a one-to-one causal labeling.

Parent	Benchmark	Prod_1: out/H	%	Prod_2: out/H	%
QuatRoPE	SR3D	470/650	72.31	486/838	58.00
Chat-Scene	SR3D	462/621	74.40	484/796	60.80
QuatRoPE	NR3D	273/298	91.61	278/331	83.99
Chat-Scene	NR3D	266/296	89.86	269/327	82.26
QuatRoPE	ScanRefer	255/289	88.24	262/320	81.88

Table 18: **Correct-proposal availability in the assessed support.** Executed questions split by whether S contains any proposal with IoU ≥ 0.5. When no correct proposal lies in S, raw products are necessarily wrong; the checked normalizers are positive throughout. Accuracy is conditional on each annotation-defined split; the absent stratum can include targets unreachable in the entire proposal pool.

Parent	Benchmark	Correct in S	N_{exec}	P_0	Prod_2	GSC
QuatRoPE	SR3D	Present	9,183	47.29	69.07	61.44
		Absent	3,872	8.81	0.00	9.01
Chat-Scene	SR3D	Present	9,183	46.41	68.28	59.87
		Absent	3,872	9.30	0.00	9.53
QuatRoPE	NR3D	Present	1,508	50.80	62.53	58.36
		Absent	1,172	19.97	0.00	19.54
Chat-Scene	NR3D	Present	1,508	49.40	62.14	58.62
		Absent	1,172	19.62	0.00	19.45
QuatRoPE	ScanRefer	Present	1,873	61.61	66.20	64.18
		Absent	795	28.05	0.00	27.55

Table 19: **Positive-strength behavior on development and expansion subsets.** The $1/16$ columns compare new-error counts against the preselected exponent two. The $\gamma = 4$ columns give BP/GSC new-error (H_4) and repair (R_4) counts. Dev/expansion labels preserve the original subset roles.

Subset	N	Raw harms $H_{1/16}/H_2$	%	GSC harms $H_{1/16}$	H_4 BP/GSC	R_4 BP/GSC	Dominated raw points
SR3D dev.	400	29/39	74.36	1	42/27	61/66	7/7
SR3D expansion	1,200	65/85	76.47	0	77/62	193/195	7/7
NR3D dev.	600	18/22	81.82	0	18/8	11/11	7/7
NR3D expansion	800	31/36	86.11	1	23/11	29/27	7/7

A development-selected semantic-margin gate. A gate applies Prod_1 only when the parent's top-two log-score margin is below t, otherwise retaining the parent prediction. We select t from the 5th–95th percentiles of SR3D development margins in 5-percentile steps, maximizing accuracy on those 400 questions with ties resolved toward smaller thresholds, then freeze $t = 4.032814$ nats. It admits 90.80% of full SR3D and 94.19% of full NR3D questions and reaches 43.26% and 33.13% accuracy. Paired differences from GSC are $+0.19$ $[-0.80, 1.07]$ and -2.01 $[-2.68, -1.40]$ points, respectively. Admission need not entail executable geometry or a changed answer. These results concern this particular gate, not confidence gating generally.

D.3 Alternative ViewSpatial image construction

The `code_input` condition uses an alternative implementation-specific image construction for the same 5,712 ViewSpatial questions. It is not an additional dataset and is not pooled with the official-image results. Its execution coverage is 1,297/5,712. The completed semantic-score files include 728, 39, and 228 API failures for InternVL2.5, Qwen3-VL, and Qwen2.5-VL, respectively. These receive uniform fallback and remain in all paired denominators. This condition is reported as a robustness observation, not used to establish the primary result.

Table 20: Supplemental ViewSpatial code-input condition. API failures refer to the semantic-score call and are included in the denominator.

Backbone	P_0	GEO	MIX	BP	GSC	API failures
InternVL2.5-8B	40.27	46.41	45.17	45.89	45.97	728
Qwen3-VL-8B	41.46	45.50	42.47	43.12	43.45	39
Qwen2.5-VL-7B	37.22	43.33	40.55	41.35	41.86	228

D.4 Evidence strength and geometric sensitivity

Figure 9 uses the recorded offline sweep $\gamma \in \{0, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4\}$ on fixed RGB scores and geometry. It shows operating points after selection of the reported $\gamma = 2$; the sweep is descriptive, not a new procedure for selecting the best test-set value. Increasing strength generally permits additional repairs and additional errors. The curves need not be monotone in net accuracy or dominate each other. At sufficiently strong correction, clipping can leave multiple tied evidence maxima.

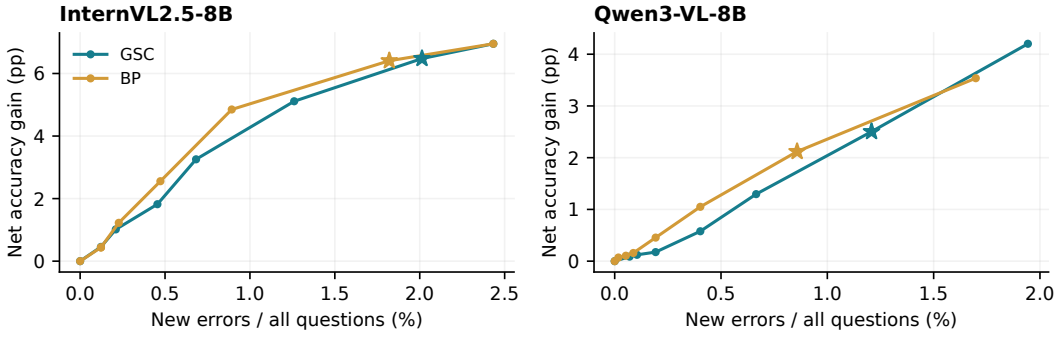

Figure 9: ViewSpatial strength sweeps for GSC and BP. Stars indicate the frozen $\gamma = 2$ point. Each trajectory varies strength without changing perception or the parent.

The recorded 3D geometry perturbations add random-direction center offsets at specified fractions of scene diameter or randomly replace predicted category labels. A 5%-diameter center perturbation leaves a +9.17-point GSC gain on the 1,200-question SR3D expansion subset. A 25% label-flip perturbation leaves +4.50 points there, while the NR3D effect becomes weak. These subset observations are not full-validation robustness estimates. Box-size perturbations are not informative about the relation executor because the implemented relation scores consume centers, not sizes. Likewise, unchanged distance rankings under a unit rescaling do not imply unchanged fused decisions when metric temperatures remain fixed.

D.5 FRAME AND NUMERICAL-HINT CONTROLS

Table 21: ViewSpatial absolute accuracies under frame and binding interventions.

Backbone	Correct frame	No transform	Role swap
InternVL2.5-8B	47.55	42.37	32.90
Qwen3-VL-8B	43.40	41.89	38.39
Qwen2.5-VL-7B	42.38	38.25	30.92

Operational intervention definitions. In Figure 5, `binding_shuffle` is one *deterministic within-question role swap*. Camera-relative queries exchange target and reference point-cloud bindings; scene-simulation queries exchange origin and target while retaining `forward_to`. The program recomputes the derived frame and option angular violations. There is no across-question or option-score permutation, no random seed, and no averaging over realizations. There is no rejection of unchanged outcomes: swapped role slots differ, but numerical geometry or predictions can coincide.

The original execution mask gates every intervention: original fallbacks remain fallbacks, and executed questions are reevaluated. Invalid modified geometry could trigger fallback in code, but it does not do so in these runs. All three official ViewSpatial comparisons retain 1,545 executed and 4,167 fallback questions; InternVL2.5 MMSI retains 88 and 912. Parent distributions, full denominators, and the frame-conversion *procedure* remain fixed, while the recomputed numerical frame can change. Paired cluster-bootstrap intervals use 20,000 draws and seed 20260907, conditional on this one intervention; they do not estimate variability over random swaps.

For MMSI, camera–camera questions exchange source/destination cameras, including the source basis. Camera–object questions advance an integer photo index by one modulo the frame count; the `last` sentinel is unchanged. Compass questions read the anchor from the reference role and the anchor-reference from the anchor role, retaining target–reference bearing roles. These rules allow unchanged cases and do not enforce a derangement.

The `no_transform` condition omits specified frame operations. ViewSpatial camera-relative queries compare the option vector directly with the world-space bearing; scene-simulation queries omit horizontal projection of forward and target-bearing vectors but retain simulated-basis construction. MMSI branches omit camera-basis conversion, stated matrices, horizontal projection, or declared compass-anchor rotation where applicable. The full contract is in `data/protocol.json`. For InternVL2.5 on MMSI, correct geometry reaches 31.10%, omitted transforms reach 27.60%, and swapped roles reach 27.80%, against a 27.60% parent. The omitted-transform arm repairs 18 examples and harms 18; equality in net accuracy does not mean that its outputs equal P_0.

For the RGB numerical-hint arm, per-option angular violations are converted to degrees with one decimal place and appended to the question. The model is re-scored using the same first-token option interface as p_0; non-executable questions and failed hint requests retain their original p_0. This differs from the earlier native-3D hint experiment, which used free generation. The paired GSC-minus-hint intervals are [2.65, 4.31] and [1.90, 4.30] points for InternVL2.5 on ViewSpatial and MMSI, respectively. They are [-2.55, -1.17] and [-1.40, 0.60] for Qwen3-VL. The signs demonstrate backbone dependence. Numerical hints were evaluated for these two models; we do not impute results for Qwen2.5-VL.

D.6 LOCAL TOOL-AGENT PROFILE

We ran a local MSSR adaptation (Guo et al., 2026) with Qwen3-VL on the first 50 questions of each frozen official manifest, using one or five iterations. This is a nonrandom profile with backend and adapter effects, not a full benchmark reproduction. With unanswered questions counted wrong, MMSI accuracies are 48% (one iteration) and 40% (five), compared with 32% for both P_0 and GSC on these 50 questions. On ViewSpatial, they are 36% and 28%, compared with 36% for P_0 and 48% for GSC. Answered counts are 47/37 on MMSI and 48/33 on ViewSpatial for one/five iterations. The profile shows complementary evidence-acquisition behavior and substantial execution sensitivity; it does not establish overall superiority over MSSR.

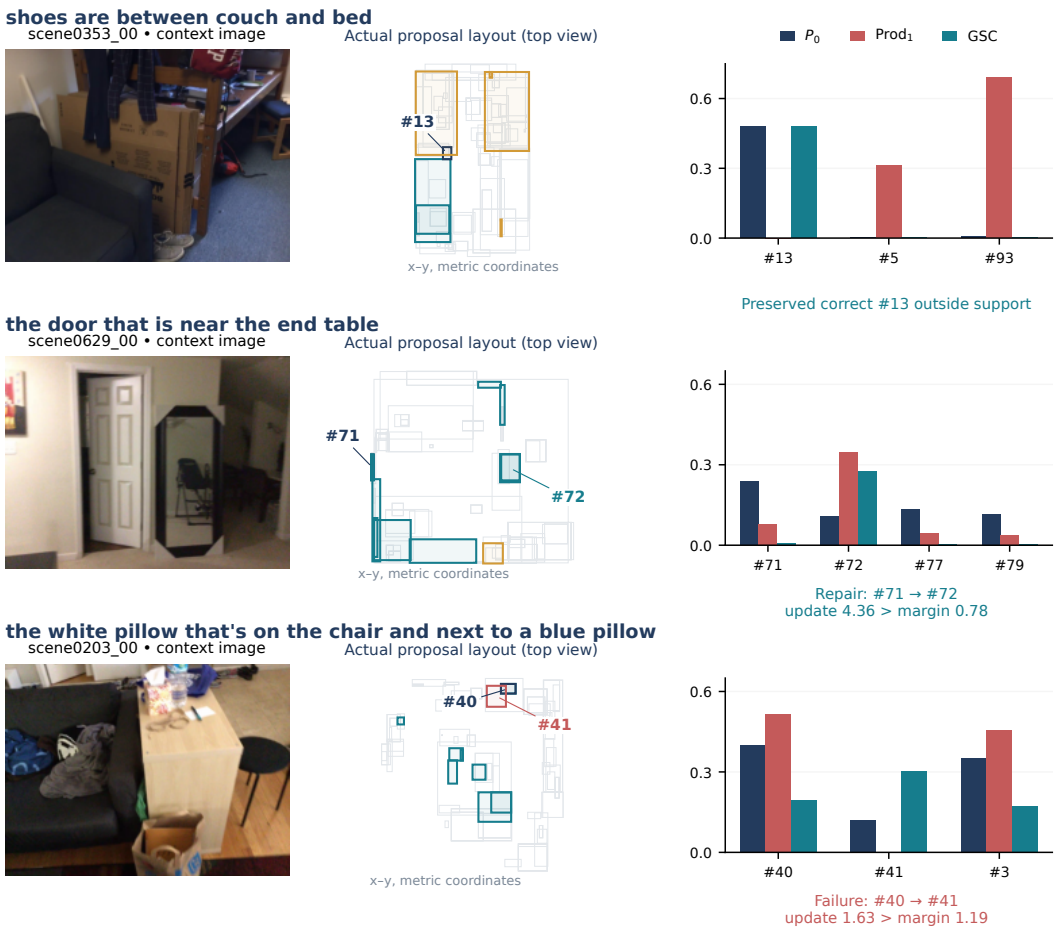

Figure 10: **Three actual decisions from frozen caches.** Left: original photographs from the corresponding ScanNet scans, used only as visual context. Middle: top views of the measured predicted proposal boxes, with original proposal IDs. Right: cached parent, raw-product, and GSC probabilities for selected candidates; other candidates remain in the normalization. No synthetic scene is paired with empirical predictions. The photographs are not registered proposal overlays and are not additional model inputs in the native 3D runs.

D.7 Compute accounting

The native 3D CPU profile uses 400 SR3D development questions. Cached-input loading, parsing, geometry, and fusion take 0.756 seconds in total, about 1.9 ms per question; a 600-question NR3D profile averages 1.27 ms. The RGB arithmetic profile times the fusion routine on 400 executed ViewSpatial records using a uniform parent placeholder. It includes geometry normalization, reference construction, both GSC and BP updates, and Python/NumPy overhead, averaging 76 μs per question. Cached-world geometric execution and I/O, measured separately, average 0.466 s per question. Perception-cache construction is amortized over reused inputs, and semantic-score throughput differs by backbone. These are different stages and workloads, not end-to-end latencies or cross-modality speed comparisons; they do not establish a headline speedup over the tool-agent profile.

E Real-scene decisions and failures

Preservation under category-binding failure. NR3D sample `nr3d-000217` asks for shoes between a couch and a bed. The predicted class is singular *shoe*, while the question says *shoes*; the parser's word-boundary rule instead selects *couch* as the earliest matching category. Its support is therefore $S = \{5, 93\}$, excluding correct proposal 13. The parent gives proposal 13 probability

0.477899. GSC leaves its evidence at zero and retains it with probability 0.479235; Prod_1 assigns it zero and selects proposal 93. This is a concrete parser-induced support mismatch, not a demonstration that the executor resolved the complete query. The example belongs to an earlier inspected NR3D development subset and is used illustratively; aggregate claims use the full manifests and the exclusion checks.

Repair through complementary geometry. For SR3D sample `sr3d-000046`, the parent chooses door 71. The executed near-end-table relation favors correct door 72 sufficiently to overcome the semantic margin: $2(e_{72} - e_{71}) = 4.362$ versus $\log[p_0(71)/p_0(72)] = 0.781$. GSC raises proposal 72 from 0.108809 to 0.2761 and selects it, although the GEO maximum itself is incorrect. Prod_1 also repairs this example. It illustrates complementary use of the two sources, rather than a correction unique to GSC.

Failure when partial evidence suppresses the correct answer. NR3D sample `nr3d-001313` asks for a white pillow on a chair next to a blue pillow. The executor binds *next to pillow* and averages over nine predicted pillow anchors; it does not resolve the white/blue distinction or the complete support relation. Correct proposal 40 receives negative evidence, while incorrect proposal 41 (predicted as clothes) is outside S and receives zero evidence. The resulting 1.635 log-odds shift exceeds the 1.188 semantic margin. GSC chooses 41; the raw product retains a correct pillow. The same decision equation therefore explains a repair, a preservation, and a regression.

F Replay boundary and interpretation

Download the supplementary `anonymous_replay_bundle.zip`, extract it, and enter its `anonymous_replay` directory. With Python 3.10 or newer, run:

```
python -m pip install -r requirements.txt
python replay.py --rebuild-geometry --output output
python mechanism_analysis.py --replay-output output \
    --output output/analysis
python option_interface_audit.py \
    --output output/analysis/rgb_options_standalone.json
```

The first replay writes the three primary table fragments under `output/tables`, along with the full RGB and component comparisons. The analysis writes `output/analysis/audit_results.json`, boundary-change records, and seven diagnostic tables including `support_allocation.tex`. The final command independently checks the option-interface audit. The package README maps each output to the paper; assertions verify saved predictions, old and repaired reference metrics, and the allocation identities.

The bundle reconstructs all five native 3D GSC prediction sequences from frozen candidate scores, predicted scene attributes, question text, and candidate assignments. It rebuilds RGB distributions from saved angular violations and all five fusion arms for every official question–backbone pair, including recorded Free decisions and Hint scores. Tables 1–3, paired intervals, and support-error counts require neither the source experiment directories nor a model or GPU.

Perception and semantic inference precede this boundary; their models, prompts, thresholds, and fallback contracts are specified above. The bundle excludes model weights and full benchmark imagery and distinguishes reexecuted 3D geometry from frozen RGB angles.