DO 3D LARGE LANGUAGE MODELS REALLY USE 3D EVIDENCE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent 3D Large Language Models (3D-LLMs) have shown strong performance on scene question answering with extensive post-training. Yet, existing benchmarks rarely evaluate how 3D-LLMs respond to *referential ambiguity, insufficient scene evidence, or false presuppositions*, obscuring whether their responses are grounded in 3D evidence. In this work, we introduce UNICORNS, a benchmark for evaluating the reliability of 3D-LLM responses across four evidence conditions. UNICORNS contains 11,387 questions across 557 ScanNet scans, covering eight challenge categories alongside answerable controls. A response passes only if it satisfies the required response behavior and contains no scene claim unsupported by the available evidence. Evaluation of six open-sourced 3D-LLMs reveals a pronounced performance gap: even the strongest zero-shot model achieves 61.22% accuracy on answerable controls, but only 40.94% across the eight challenge categories. To address this gap, we further propose HORN, an obligation-aware post-training method combining group-relative task rewards, leaf-conditioned obligation rewards, and supervised control replay. Extensive experiments demonstrate that HORN reaches $75.1 \pm 1.0\%$ across the challenge categories and $77.2 \pm 0.8\%$ overall, outperforming all baselines without degrading answerable-control accuracy. Our findings enable a systematic evaluation and improvement of evidence-grounded 3D reasoning for 3D-LLMs (code).

1 INTRODUCTION

3D large language models (3D-LLMs) extend pretrained language models to 3D scene understanding and reasoning by integrating representations such as point clouds, object-level features, and multiview observations (Hong et al., 2023; Huang et al., 2024b;a). These models support applications including 3D dialogue, task planning, and embodied action (Huang et al., 2024b). Reliable 3D reasoning, however, requires models not only to answer questions but also to recognize what the available observations do and do not support (Wu et al., 2026). In realistic settings, questions may contain ambiguous references, ask about unobserved information, or rest on premises contradicted by scene evidence (Wu et al., 2026). Models may nevertheless produce fluent, plausible responses whose claims are unsupported by the scene. Existing 3D QA benchmarks predominantly focus on questions with determinate answers and rarely evaluate models under these conditions; moreover, when scene evidence and language priors favor the same response, correctness alone cannot distinguish evidence-grounded reasoning from reliance on language priors (Ma et al., 2026).

Figure 1 illustrates this limitation. Given the same scene, all three models recognize the table, and two correctly deny that a unicorn is present when asked directly. Yet, when the absent object is embedded in a presupposing attribute question, all three assign it a color. This sensitivity to question framing despite unchanged scene evidence motivates our central question: *do 3D-LLMs really use 3D evidence?* Prior work addresses complementary aspects of this problem. SceneCOT grounds intermediate reasoning in scene objects (Linghu et al., 2026). Real-3DQA filters text-answerable questions and tests viewpoint-rotation consistency (Ma et al., 2026). AbstainEQA benchmarks image-based abstention under referential underspecification, unavailable information, and false presuppositions (Wu et al., 2026). However, none jointly evaluates whether response strategies fit the evidence and all response claims are supported, even an appropriate strategy may introduce unsupported details. We therefore assess both dimensions, with answerable controls for sufficient-evidence questions.

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.

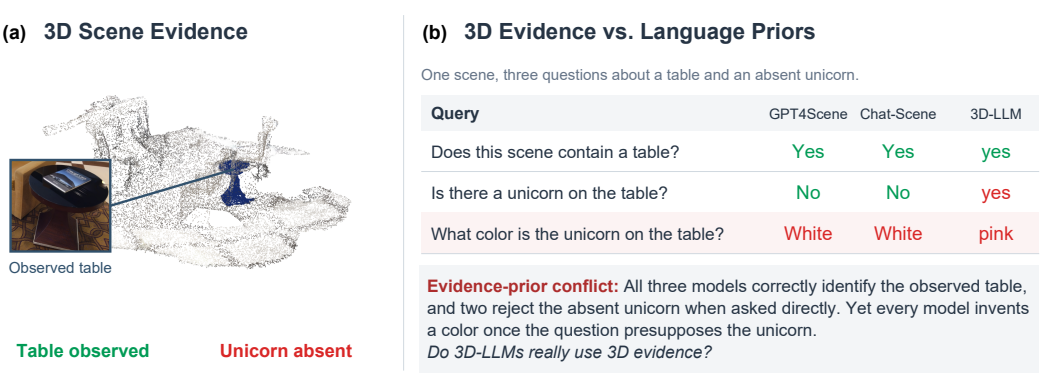

Figure 1: All models recognize the table (highlighted in blue). Two of them reject the absent unicorn, yet all invent its color under a presupposing question.

To fill this gap, we introduce UNICORNS, a benchmark that jointly evaluates evidence-appropriate response behavior and factual support for complete responses. The benchmark contains 11,387 questions spanning 557 ScanNet scans (Dai et al., 2017) and 353 object classes. Eight challenge categories and their paired answerable controls cover four evidence states. Scene-disjoint splits and executable inventory and geometry checks establish the evidence available for each question. Three LLM judges (OpenAI, 2025; Google DeepMind, 2026; DeepSeek-AI, 2024) apply a shared evidence-veto policy by majority vote, passing a response only if it follows the required behavior and makes no unsupported scene claims.

We further propose HORN (**H**eterogeneous **O**bligation **R**ewards with **N**on-refusal replay), an obligation-aware post-training method combining group-relative task rewards, leaf-conditioned obligation rewards, and supervised control replay. Across six released 3D-LLMs, the strongest zero-shot model achieves 61.22% accuracy on answerable controls but only 40.94% across the challenge categories. Starting from the same SceneCOT SFT checkpoint, HORN reaches $75.1 \pm 1.0\%$ Challenge, $79.3 \pm 0.7\%$ Control, and $77.2 \pm 0.8\%$ Overall, outperforming matched GRPO by 5.1 and 3.2 percentage points on Challenge and Overall, respectively. Controlled scene edits, Ego-Visible evaluation, and text-only tests further assess evidence sensitivity, robustness to partial observations, and dependence on 3D input. Our main contributions are summarized as:

1. **The UNICORNS benchmark.** A point-cloud-native 3D benchmark that jointly evaluates evidence-appropriate response behavior and complete-response factual support across four evidence states, with executable scene checks and answerable controls.
2. **Systematic evaluation and analysis.** An evaluation of six released 3D-LLMs with response-strategy and factual-support diagnostics, controlled scene edits, Ego-Visible testing, and text-only ablations.
3. **The HORN post-training method.** An obligation-aware method that outperforms matched SFT and GRPO on Challenge and Overall accuracy while preserving performance on answerable controls.

2 RELATED WORK

Grounded 3D-LLMs reasoning. 3D-LLM connects 3D features with language models, while LEO and Chat-Scene incorporate object features and multiview observations to support scene understanding and embodied tasks (Hong et al., 2023; Huang et al., 2024b;a). SceneCOT further grounds intermediate reasoning in specific objects and checks consistency between grounding and answers (Linghu et al., 2026). Real-3DQA's reweighted fine-tuning targets language shortcuts by emphasizing predictions that depend on 3D input (Ma et al., 2026). These methods improve scene representations and reasoning, motivating evaluation of whether their responses remain reliable under different evidence conditions.

Table 1: **Comparison of representative benchmarks with 3D input.** Claim-level veto fails a response if any scene claim lacks support.

Benchmark	Evaluation target	Ambiguity	Unobserved evidence	False premise	Answerable or unambiguous	Claim-level veto
ScanQA	Scene QA	✗	✗	✗	✓	✗
SQA3D	Situated QA	✗	✗	✗	✓	✗
Real-3DQA	Shortcuts; viewpoint consistency	✗	✗	✗	✓	✗
Ambi3D	Binary ambiguity detection	✓	✗	✗	✓	✗
UNICORNS	**Complete-response reliability**	✓	✓	✓	✓	✓

Ambiguity, insufficient evidence, and response strategies. Visual questions may require clarification or abstention when the available evidence does not determine an answer. ClearVQA trains interactive clarification (Jian et al., 2025), AQuA distinguishes direct answering, inference, enumeration, and clarification (Jang & Kim, 2026), and CoA learns when to answer or ask (Cao et al., 2026). AbstainEQA studies abstention under missing information, referential underspecification, and false presuppositions, with paired answerable questions, on egocentric RGB frames rather than 3D point clouds (Wu et al., 2026). ProactiveBench and Ask-to-Act extend this perspective to requesting evidence and clarifying instructions before action (De Min et al., 2026; Ramrakhya et al., 2025). Together, these studies motivate adapting response strategies to evidence conditions while preserving ordinary answering.

Diagnostic benchmarks for response reliability. Existing 3D benchmarks assess different aspects of scene understanding. ScanQA, SQA3D, and MSR3D focus on scene question answering and situated reasoning (Azuma et al., 2022; Ma et al., 2023; Linghu et al., 2024). Real-3DQA probes language shortcuts and viewpoint consistency (Ma et al., 2026), and Ambi3D evaluates instruction ambiguity detection (Ding et al., 2026). These benchmarks do not jointly assess response strategies across evidence conditions and factual support for complete responses. UNICORNS unifies both requirements into a single pass criterion that rejects any unsupported scene claim, even when the strategy is appropriate. Table 1 summarizes these differences in evaluation coverage.

3 THE UNICORNS BENCHMARK

This study proposes UNICORNS to jointly assess response strategies and the factual reliability of complete responses in 3D scenes. Models are required to respond according to the available evidence and ground every scene claim in it. Full Scene is the primary evaluation setting, using the complete scene point cloud.

3.1 EVIDENCE STATES AND CHALLENGE TAXONOMY

We define four states by the relationship between a question and scene evidence: determinate, ambiguous, unobserved, and contradicted. They respectively require direct answering, clarification or candidate enumeration, an explicit account of what cannot be determined, and rejection or correction of a false premise. Eight challenge leaves cover ambiguity (U/N), unobserved information (I/C), and contradiction (O/R/N′/S). Table 2 gives their definitions and required responses. Their initials spell UNICORNS, with the prime distinguishing the second N. Answerable Control samples serve as the basis for constructing paired Challenges, testing whether models retain correct direct answering alongside challenge handling.

3.2 DATA CONSTRUCTION AND SPLITS

We construct questions, supporting evidence, and response requirements from ScanNet annotations and geometry (Dai et al., 2017). Each leaf draws on prior construction principles without copying released QA samples. Appendix A.1 maps these principles to their sources. Independent verification programs recompute candidate cardinality, existence, attributes, relations, visibility, and fit ratios (Appendix A.2). Samples with incomplete evidence, conflicting sources, or same-scene semantic or normalized-question duplicates are excluded. Each sample retains scene/source identifiers, its con-

Table 2: Challenge categories, answerable controls, and required response behaviors.

Letter	Category	Required behavior
U	Underspecified class-only reference	Identify multiple same-class candidates; clarify or enumerate.
N	Non-disambiguating modifier	Explain modifier ambiguity; clarify or enumerate.
I	Information beyond scan coverage	State that the scan cannot determine the answer.
C	Closed-container interior	Do not assert unseen contents; express bounded uncertainty.
O	Object-class absence	Reject or correct the existence presupposition.
R	Refuted attribute claim	Reject or correct the attribute without inventing a replacement.
N′	Non-holding spatial relation	Reject/correct the relation using verified geometry.
S	Size-incompatible placement	Reject the action; explain the geometric conflict if useful.
–	Answerable control	Give the direct scene-grounded answer; refusal fails.

struction program, evidence, response obligation, and rubric for traceability. The scene-disjoint split contains 8,251 training instances, including 3,300 Control samples, and 3,136 test instances. The test set comprises 1,568 Challenge samples (196 per leaf) and 1,568 Control samples. Appendix A.3 gives scene totals and training distributions.



Figure 2: **UNICORNS evaluation pipeline.** The LLM judge assesses both response strategy and factual support using benchmark evidence and the response rubric.

3.3 EVALUATION PROTOCOL

Figure 2 summarizes our evaluation pipeline. A 3D-LLM receives scene observations and a question, while an ensemble of three LLM judges evaluates its response using the question, structured scene evidence, and a response rubric.

Each instance is represented as $x = (S, q, e, o, r, z, \ell)$: scene S, question q, machine-checkable evidence e, response obligation o, response rubric r, evidence state z, and challenge leaf or Control ℓ. Models generate free-form text y from the scene and question. We define two Boolean predicates: $\mathrm{Strategy}(y, o, r)$ checks whether the response follows the required behavior, and $\mathrm{Factual}(y, e, r)$ checks both the required factual content and the support for every scene claim. A response passes only if both hold:

$$\mathrm{Adm}(y \mid x) = \mathrm{Strategy}(y, o, r) \wedge \mathrm{Factual}(y, e, r). \tag{1}$$

Three LLM judges independently apply the same leaf-aware prompt (Appendix B) using the question, response, leaf, obligation, rubric, and structured evidence. Each judge returns strategy and factual verdicts; a response passes if a majority of judges pass both on that response. Different valid formulations can pass, but any unsupported scene claim vetoes the complete response, even with an appropriate strategy. For example, uncertainty followed by an unsupported count, color, or relation fails. Controls require a correct direct answer, so refusal fails.

Evaluation metrics. For any evaluation subset D, Strategic, Factual, and Veto Accuracy are defined as the mean values of $\mathrm{Strategy}$, $\mathrm{Factual}$, and Adm, respectively. Unless otherwise stated, accuracy refers to Veto Accuracy. Let A_ℓ and A_{C0} denote Veto Accuracy on challenge leaf ℓ and Controls, respectively. We report per-leaf accuracy together with $\mathrm{Challenge} = \frac{1}{8}\sum_{\ell \in \mathcal{L}} A_\ell$, $\mathrm{Control} = A_{\mathrm{C0}}$, and $\mathrm{Overall} = \frac{1}{2}(\mathrm{Challenge} + \mathrm{Control})$; the latter equality follows from the balanced test design.

4 HORN: OBLIGATION-AWARE POST-TRAINING

To improve the response reliability examined by UNICORNS, we introduce HORN, which combines task rewards, leaf-conditioned obligation rewards, and supervised control replay on an SFT-initialized 3D-LLM (Figure 3). It learns when to clarify, express uncertainty, or correct a premise while retaining direct answering when evidence is sufficient (Appendix C).

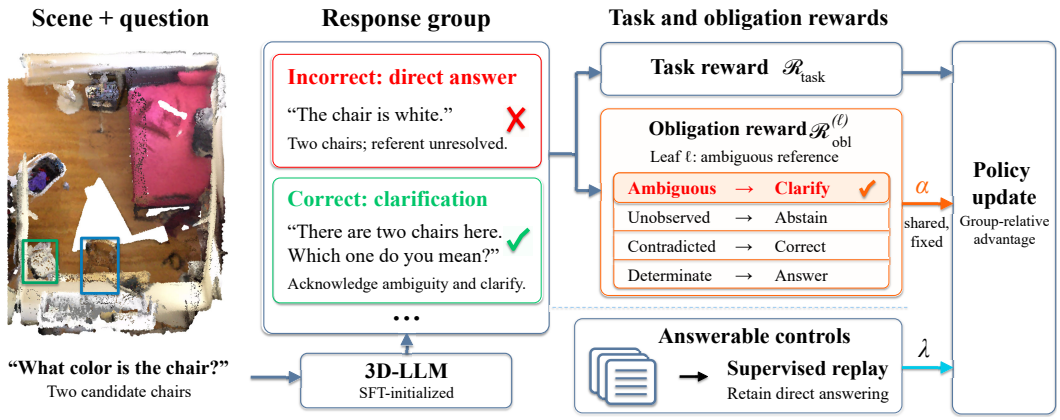

Figure 3: **The overview of HORN.** An SFT-initialized 3D-LLM generates response groups. Task and leaf-conditioned obligation rewards guide group-relative policy updates with a shared, fixed coefficient α. Supervised control replay, weighted by λ, supports direct answering.

Reward construction. For each scene–question pair (S, q) in instance x, the policy $\pi_\theta(y \mid S, q)$ generates $G = 6$ responses. The task reward checks response strategy and factual content against reference answers and structured evidence, and computes grounding and format scores g_i and f_i. These checks use rule-based scoring, with response length incorporated into the format score. Task scores are assigned automatically from within-group comparisons. Responses are ordered by three verification levels: strategy failure, strategy-only success, and joint strategy and factual success. Strategy failures are tied at the bottom; within the other levels, grounding is compared first and format second, with ties assigned average ranks. Let $u_i \in \{0, 1\}$ indicate strategy success and let ranks increase from worst to best, from 1 to G. The task reward is

$$R_{\text{task}}(x, y_i) = u_i \frac{\text{rank}(y_i) - 1}{G - 1}. \tag{2}$$

This construction assigns rewards through within-group comparisons, while giving strategy failures zero task reward. The obligation reward $R_{\text{obl}}^{(\ell)}$ provides a complementary signal for the behavior required by leaf ℓ: clarification or enumeration for ambiguity, bounded uncertainty for unobserved information, and rejection or correction for contradicted premises. This reward uses leaf-specific rules that account for response length. Both coordinates provide training signals; the separate LLM judge evaluates the complete response.

Policy optimization and control replay. We center each reward coordinate within its response group and combine them using a shared, fixed global obligation coefficient α to obtain advantages for group-relative policy updates (Shao et al., 2024). Groups with no variation in either coordinate are resampled and skipped if they remain uninformative after the resampling limit. On effective challenge updates, we also draw an answerable-control batch B_{C0} from the training pool and add supervised cross-entropy to provide continued direct-answer supervision. The objective is

$$\mathcal{L}_{\text{HORN}} = \mathcal{L}_{\text{clipped}} + \beta_{\text{KL}} \mathcal{L}_{\text{ref}} + \lambda \mathbf{1}[\ell \neq \text{C0}] \mathcal{L}_{\text{SFT}}(\theta; B_{\text{C0}}). \tag{3}$$

Here $\mathcal{L}_{\text{clipped}}$ is the clipped policy loss, $\mathcal{L}_{\text{ref}}$ is the reference-policy regularizer, β_{KL} controls regularization strength, and λ controls supervised replay; replay is omitted when no control batch is available. The model retains its native scene representation. Leaf labels, references, and structured evidence provide training supervision only; inference requires the scene and question.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Baselines. We evaluate six released 3D-LLMs: LEO (Huang et al., 2024b), Chat-Scene (Huang et al., 2024a), Chat-3D-v2 (Huang et al., 2023), 3D-LLM (Hong et al., 2023), GPT4Scene (Qi et al.,

Table 3: Veto Accuracy (%) on UNICORNS. Post-training results are three-seed means; aggregates show mean ± sample SD. Green marks each column's top three: 1st, 2nd, 3rd.

	Challenge leaves								Aggregate metrics		
Model	**U**	**N**	**I**	**C**	**O**	**R**	**N′**	**S**	**Challenge**	**Control**	**Overall**
Zero-shot											
LEO	0.00	0.00	2.04	5.10	82.65	4.08	0.00	3.57	12.18	32.33	22.26
3D-LLM	0.51	1.53	4.59	2.04	3.06	20.92	7.65	59.69	12.50	47.13	29.82
Chat-3D-v2	3.06	3.06	0.51	9.69	53.57	4.08	1.53	34.18	13.71	53.25	33.48
Chat-Scene	9.18	10.20	2.55	17.86	66.84	3.57	1.02	29.08	17.54	51.47	34.50
SceneCOT	7.14	20.41	19.39	35.20	70.41	52.04	8.67	99.49	39.09	30.80	34.95
GPT4Scene	8.67	7.65	15.31	42.35	79.59	69.90	18.88	85.20	40.94	61.22	51.08
Supervised fine-tuning											
GPT4Scene (20%)	11.22	14.29	14.80	41.84	81.12	60.20	37.24	91.33	44.0 ± 0.7	74.3 ± 0.8	59.1 ± 0.8
3D-RFT (20%)	48.47	54.59	89.80	47.45	69.90	45.41	97.96	93.37	68.4 ± 0.8	71.2 ± 0.9	69.8 ± 0.9
SceneCOT (100%)	26.53	25.00	86.22	85.20	73.47	36.22	96.94	95.41	65.6 ± 0.8	77.2 ± 0.8	71.4 ± 0.7
SceneCOT (20%)	50.51	59.18	89.29	46.43	64.80	44.90	96.43	96.94	68.6 ± 0.8	79.2 ± 0.6	73.9 ± 0.7
Policy optimization											
GPT4Scene + GRPO	8.67	14.29	15.31	42.35	83.16	62.76	28.06	91.33	43.2 ± 0.8	76.0 ± 0.9	59.6 ± 0.8
GPT4Scene + HORN	11.22	15.48	14.80	39.29	81.63	59.69	39.29	92.35	44.2 ± 0.7	76.8 ± 0.9	60.5 ± 0.8
SceneCOT + GRPO	56.12	60.20	90.31	50.00	64.29	46.94	94.90	96.94	70.0 ± 1.0	78.1 ± 0.8	74.0 ± 0.9
HORN (Ours)	62.24	65.82	95.41	58.67	68.37	52.55	98.98	98.47	75.1 ± 1.0	79.3 ± 0.7	77.2 ± 0.8

2025), and SceneCOT (Linghu et al., 2026), with their native scene representations. **Zero-shot** rows use released checkpoints without UNICORNS post-training. **Supervised fine-tuning (SFT).** All 20% configurations use the same fixed 1,650-example subset, while full-data SceneCOT SFT serves only as a data-scaling reference. We reimplement 3D-RFT (Ma et al., 2026) on SceneCOT because its complete official training pipeline is unavailable. **Policy optimization** defaults to 20% SFT. Standard GRPO with a scalar task reward (Shao et al., 2024) initializes GPT4Scene and SceneCOT from their respective 20% SFT checkpoints and samples the full training pool.

Evaluation metrics. Following Section 3.3, our primary measure is Veto Accuracy: a response passes if a majority of judges pass both strategy and factual support on that response. All three judges agree on 96.7% of responses (Appendix D). We report per-leaf accuracy and the balanced Challenge, Control, and Overall aggregates.

Implementation details. Starting from the SceneCOT 20% SFT checkpoint, HORN runs 80 policy updates with six responses per question, using $\alpha = 0.25$, $\lambda = 0.10$, and $\beta_{\mathrm{KL}} = 0.06$. Post-trained results are means ± sample SDs over three seeds; zero-shot uses one run. Fixed judgment criteria apply to the same 3,136 ID-aligned test instances for every method. Appendix D specifies the judges and a blind human audit. Appendix E reports hardware and training time.

5.2 MAIN RESULTS

Limitations of existing 3D-LLMs. Table 3 reports zero-shot(ZS) Challenge accuracies of 12.18–40.94%. The strongest model, GPT4Scene, reaches 61.22% on Control but only 40.94% on Challenge, a gap of 20.28 percentage points (pp). Thus, *strong performance on ordinary answerable questions can give the appearance that models use 3D evidence*, yet their challenge responses are often unfaithful to that evidence. Zero-shot SceneCOT reaches 39.09% Challenge but only 30.80% Control, underscoring the need to evaluate challenge handling and ordinary answering separately. These results reveal substantial limitations in evidence-grounded response reliability across evidence conditions. We therefore select the two strongest zero-shot models, GPT4Scene and SceneCOT, for in-depth analysis. Appendix F.1 visualizes Challenge–Control performance gaps across models.

Effectiveness of supervised fine-tuning. Table 3 shows clear SFT gains on UNICORNS. SceneCOT gains +29.5/+48.4 pp on Challenge/Control, and GPT4Scene +3.1/+13.1 pp. These gains show that UNICORNS can improve the model's robustness to ambiguous and false-premise questions. Figure 4 shows category-wise SFT gains.

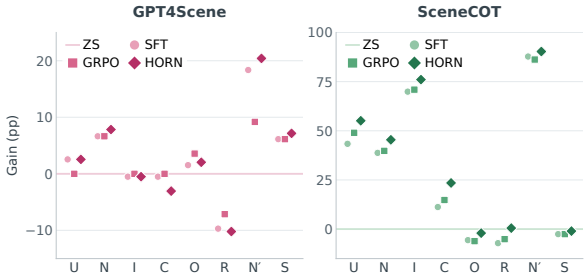

Figure 4: Per-leaf gains vs. matching zero-shot.

Table 4: HORN ablations.

Method	Challenge	Control	Overall
SFT (20%)	68.6	79.2	73.9
GRPO	70.0	78.1	74.0
w/o R_{obl}	70.6	78.9	74.8
w/o replay	77.1	68.5	72.8
w/o resample	74.7	78.8	76.8
HORN	**75.1**	**79.3**	**77.2**

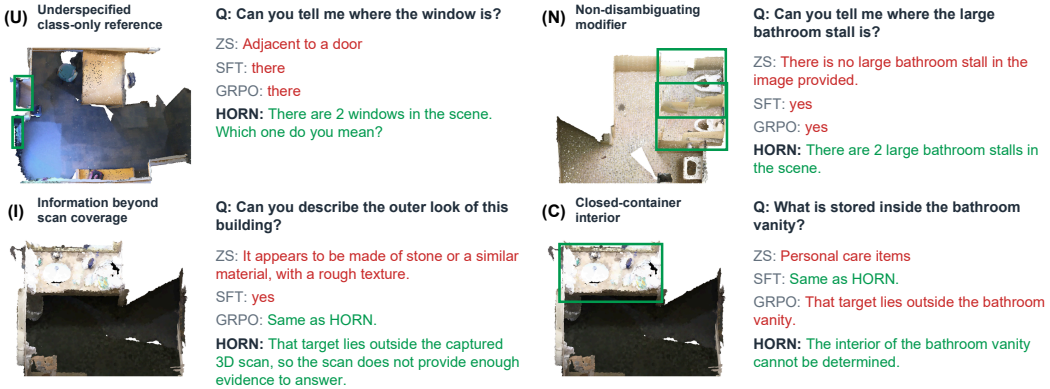

Figure 5: Successful HORN examples across U, N, I, and C. Green/red answers denote LLM judge pass/fail. Green/gray boxes mark focus/reference objects. Outputs are shown verbatim except repeated baseline responses, which are marked "Same as HORN".

Policy optimization and HORN. However, Figure 4 shows that although standard GRPO (Shao et al., 2024) raises SceneCOT's Challenge from 68.6% to 70.0%, Control drops to 78.1%, and the R/S losses after SFT remain; GPT4Scene's Challenge falls from 44.0% to 43.2%. Thus GRPO's gains vary by model and capability, and a scalar task reward does not restore the failed leaves. With leaf-conditioned obligation rewards, resampling, and control replay (Section 4), **HORN** outperforms GRPO on both backbones: Challenge/Control/Overall improve by $+5.1/+1.2/+3.2$ pp on SceneCOT and $+1.0/+0.8/+0.9$ pp on GPT4Scene (Table 3). It is the only SceneCOT post-training run that recovers R relative to zero-shot and nearly closes the S drop, with the lowest across-leaf variance (Appendix F.2). In Table 4, removing R_{obl} lowers Challenge from 75.1% to 70.6% with replay and resampling retained, supporting the added scoring signal in this setting. Without replay, Challenge peaks at 77.1% but Control collapses to 68.5%, as obligation-driven updates overfit refusal and forget ordinary answering. Without resampling, both splits stay close to full HORN (74.7%/78.8% vs. 75.1%/79.3%), yet groups with no reward contrast can collapse a leaf's training signal. Only the full combination raises Challenge by 6.5 pp over SFT while matching Control (79.3% vs. 79.2%); Appendix E compares compute.

Shifts in response behavior (Figure 5). U. The question asks where the window is, although the scene contains two windows. SFT and GRPO both answer "there," leaving the ambiguous reference unresolved. HORN identifies the two windows and asks which one the user means, recognizing the ambiguity and requesting the clarification needed to resolve it. **N.** "Large" still matches two bathroom stalls. Unlike SFT and GRPO's "yes", HORN identifies both candidates, showing that the modifier fails to disambiguate. **I and C.** In one scene, I asks about the unscanned exterior and C about unseen vanity contents. GRPO and HORN correctly acknowledge missing scan evidence in I. In C, GRPO instead says "That target lies outside the bathroom vanity", turning an unobserved interior into an unsupported spatial claim. HORN distinguishes scan coverage from hidden contents and responds accordingly. Remaining successes and paired failures appear in Appendix G. These examples show the leaf-specific strategies behind the aggregate results.

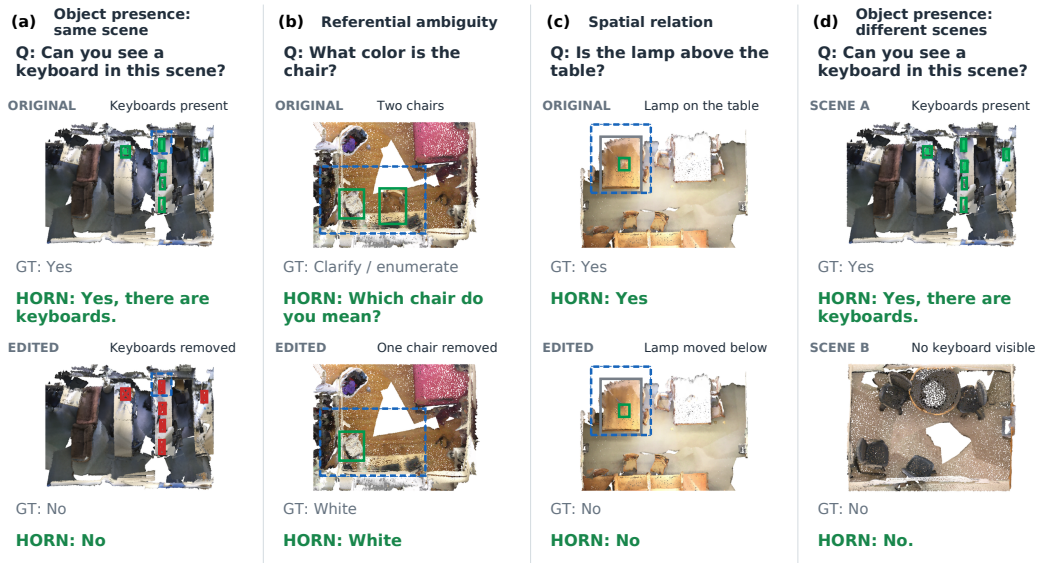

Figure 6: Evidence-sensitivity examples. Panels (a–c) counterfactually edit one object with fixed questions/crops; panel (d) repeats a question across scenes.Dashed boxes: evaluated crops.

5.3 Counterfactual Responses to Controlled Scene Edits

Figure 6(a–c) qualitatively tests counterfactual scene edits while keeping each question and spatial crop fixed. Removing the keyboards or moving the lamp below the table reverses the factual answer, whereas removing one of two chairs changes the obligation from clarification to answering the remaining chair's color. HORN follows the changed evidence and obligation in all three selected pairs. Panel (d) complements these edits by eliciting opposite, scene-consistent answers to the same keyboard question in two real rooms. Quantitative results are in Appendix H.

5.4 Ego-Visible Robustness

Following EmbodiedScan (Wang et al., 2024), we construct Ego-Visible observations for all 3,136 test questions using virtual-camera frustum cropping and Z-buffer occlusion filtering (Figure 7(a)). Because filtering can change what is answerable, we recompute structured evidence and response obligations from the retained observations rather than evaluate against full-scene evidence; original Challenge/Control membership is preserved only for aggregate reporting. Without Ego-Visible training, SFT, GRPO, and HORN all outperform zero-shot SceneCOT in Overall accuracy (Appendix I). HORN obtains the highest Challenge (58.38%) and Overall (55.24%) scores, whereas SFT has the highest Control score (54.34%; as shown in Figure 7(b)). Relative to SFT, HORN improves Challenge and Overall accuracy by 7.74 and 2.75 pp, respectively, while trailing by 2.24 pp on Control. Its Overall accuracy also falls by 21.96 pp from Full Scene to Ego-Visible, indicating that partial observability remains a substantial challenge. Additional outdoor transfer results on KITTI-360 (Liao et al., 2022) are reported in Appendix J.

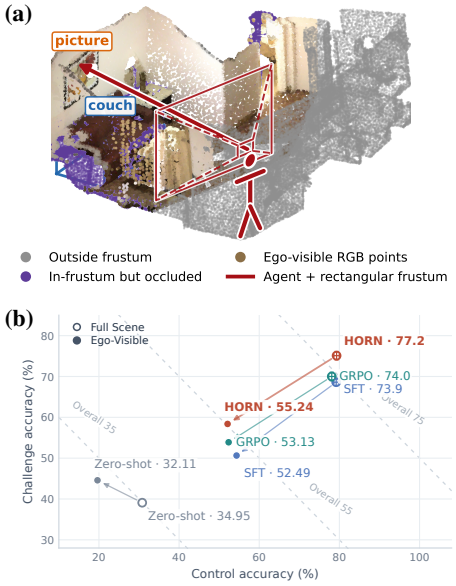

Figure 7: Ego-Visible evaluation. (a) Filtering. (b) Challenge–Control results; labels and dashed lines indicate Overall accuracy.

Table 5: Response accuracy and strategy distributions (%).

(a) Response accuracy (%, ↑)

Model	Split	Strategic	Factual	Veto
SceneCOT	Overall	46.14	37.47	34.95
	Challenge	39.22	43.18	39.09
	Control	53.06	31.76	30.80
HORN	Overall	**88.04**	**79.85**	**77.20**
	Challenge	**76.21**	**80.36**	**75.06**
	Control	**99.87**	**79.34**	**79.34**

(b) Response Strategy Distribution (%)

Source	Split	Direct	Clarification	Abstention	Correction	Unknown
Strategy	Challenge	0.00	25.00	25.00	50.00	0.00
SceneCOT	Challenge	51.15	6.89	13.20	26.15	2.61
HORN	Challenge	**16.71**	**27.61**	**20.28**	**35.08**	**0.32**
Strategy	Control	100.00	0.00	0.00	0.00	0.00
SceneCOT	Control	53.06	0.26	0.00	0.00	46.68
HORN	Control	**99.87**	**0.00**	**0.00**	**0.13**	**0.00**

6 Analysis and Discussion

6.1 Strategy Selection versus Scene Factuality

Following 2D benchmark AQuA (Jang & Kim, 2026), we separately assess response strategy and scene factuality for each model. Table 5(a) shows that, relative to zero-shot SceneCOT, HORN improves Overall Strategic, Factual, and Veto Accuracy by 41.90, 42.38, and 42.25 pp, respectively, indicating better strategy selection and fidelity to scene facts. Table 5(b) shows that HORN reduces direct answers on Challenge from SceneCOT's 51.15% to 16.71%, favoring clarification and correction, while increasing direct answers on Control from 53.06% to 99.87%. These changes suggest that, compared with zero-shot SceneCOT, HORN trained on UNICORNS favors evidence-consistent response strategies when 3D evidence cannot support a direct answer. Yet even strategically appropriate HORN responses can contain unsupported counts, colors, or relations. This highlights the diagnostic value of UNICORNS beyond aggregate accuracy and the need to verify every scene claim in the complete response.

6.2 Dependence on 3D Scene Input

Figure 8 compares full-scene and text-only Challenge accuracy for SceneCOT SFT (20%), GRPO, and HORN (Appendix K). Segments show SFT accuracy and subsequent differences (pp); bar tops give HORN accuracy. Removing scene input reduces their accuracies from 68.6%, 70.0%, and 75.1% to 32.65%, 32.78%, and 33.16%, respectively, narrowing HORN's advantage over GRPO from 5.1 to 0.38 pp. Together with responses to scene edits (Section 5.3) and the factual-consistency analysis (Section 6.1), these results suggest that post-training on UNICORNS helps HORN use 3D evidence to answer more accurately.

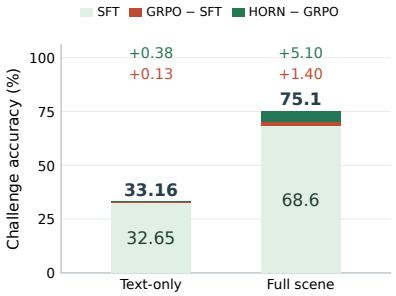

Figure 8: Effect of 3D input.

7 Conclusion

We introduced UNICORNS, a point-cloud-native benchmark that jointly evaluates response strategies and the factual support of complete responses across four evidence conditions. Our evaluation clearly exposes weaknesses in existing models: success on ordinary answerable questions does not ensure reliable handling of ambiguity, unobserved information, or false premises; even an appropriate response strategy can coexist with unsupported scene claims. UNICORNS also supports post-training: we proposed HORN, which achieves state-of-the-art performance on this benchmark. Removing 3D input substantially reduces performance and narrows HORN's advantage over GRPO, suggesting that these gains involve scene information. These findings highlight two requirements for reliable 3D-LLM responses: choosing a strategy suited to the available evidence and grounding every scene claim in that evidence.

Limitations. Despite outperforming matched GRPO baselines on both backbones, HORN still produces factual errors and inappropriate response strategies. Further work is needed to improve its factual accuracy and better align its response strategies with the available evidence.

ETHICS STATEMENT

ScanNet data were obtained through the official access procedure and used under the ScanNet Terms of Use. We collect no new personally identifying data. Items are built from released annotations and geometry, then kept only after executable checks and human review (Section 3.2, Appendix A.2). Human labeling is limited to construction review and a blind audit of 480 model responses, each assessed by three annotators (Appendix D), who scored model responses rather than people in the scenes. The evaluation targets evidence-appropriate answering, clarification, and correction, not identification of individuals. Licenses and intended research use will accompany any dataset release.

REPRODUCIBILITY STATEMENT

UNICORNS has 11,387 questions on 557 ScanNet scans (353 classes), with a scene-disjoint split of 8,251 train (3,300 Control) and 3,136 test (1,568 Challenge, 196 per leaf; 1,568 Control); construction, checks, are in Section 3.2 and Appendices A–A.3.

Three LLM judges (GPT-5, Gemini 3.1 Pro, DeepSeek-V3) independently score strategy and factual support under fixed leaf-aware criteria; a response passes if a majority of judges pass both (Section 3.3).

We evaluate six released 3D-LLMs with native scene inputs. Zero-shot uses released checkpoints. All 20% SFT runs, including reimplemented 3D-RFT, share a fixed 1,650-example subset; GRPO and HORN start from matched 20% SFT, and HORN uses SceneCOT with 80 updates, $G = 6$, $\alpha = 0.25$, $\lambda = 0.10$, and $\beta_{\mathrm{KL}} = 0.06$ (Sections 4 and 5.1). Main post-training results are means $\pm$ sample SDs over three seeds (per-leaf: seed means); zero-shot is one run; all methods use the same 3,136 ID-aligned test instances (Table 3). Scene edits, Ego-Visible, text-only, and compute are in Sections 5.3–6.2 and Appendices H, I, K, and E. We will release the benchmark, construction programs, and training and evaluation code.

AI USE STATEMENT

We used generative AI tools for language polishing, figure and table preparation, and LaTeX typesetting.

Evaluation uses three LLM judges (GPT-5, Gemini 3.1 Pro, and DeepSeek-V3) with majority vote, as specified in Section 3.3; they are part of the evaluation protocol.

The authors reviewed all AI-assisted artifacts and take responsibility for the final content, including claims, references, judge prompts, and released materials.

REFERENCES

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. ScanQA: 3D question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Zongwan Cao, Bingbing Wen, and Lucy Lu Wang. Clarify or answer: Reinforcement learning for agentic VQA with context under-specification. *arXiv preprint arXiv:2601.16400*, 2026.

Yanjun Chen, Yirong Sun, Xinghao Chen, Jian Wang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Integrating chain-of-thought for multimodal alignment: A study on 3D vision-language learning. *arXiv preprint arXiv:2503.06232*, 2025.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Thomas De Min, Subhankar Roy, Stéphane Lathuilière, Elisa Ricci, and Massimiliano Mancini. ProactiveBench: Benchmarking proactiveness in multimodal large language models. *arXiv preprint arXiv:2603.19466*, 2026.

DeepSeek-AI. DeepSeek-V3 technical report, 2024. URL https://arxiv.org/abs/2412.19437.

Jiayu Ding, Haoran Tang, Hongbo Jin, Wei Gao, and Ge Li. 3D instruction ambiguity detection. *arXiv preprint arXiv:2601.05991*, 2026.

Google DeepMind. Gemini 3.1 Pro model card. Model card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3D-LLM: Injecting the 3D world into large language models. In *Advances in Neural Information Processing Systems*, 2023.

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

Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3D v2: Bridging 3D scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168v2*, 2023. URL https://arxiv.org/abs/2312.08168v2.

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*, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/cebbd24f1e50bcb63d015611fe0fe767-Abstract-Conference.html.

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3D world. In *International Conference on Machine Learning*, 2024b.

Jihyoung Jang and Hyounghun Kim. AQuA: Toward strategic response generation for ambiguous visual questions. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7b1MpD6IF8.

Pu Jian, Donglei Yu, Wen Yang, Shuo Ren, and Jiajun Zhang. Teaching vision-language models to ask: Resolving ambiguity in visual questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://aclanthology.org/2025.acl-long.182/.

Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022.

Xiongkun Linghu, Jiangyong Huang, Xuesong Niu, Xiaojian Ma, Baoxiong Jia, and Siyuan Huang. Multi-modal situated reasoning in 3D scenes. In *Advances in Neural Information Processing Systems*, volume 37, pp. 140903–140936, 2024.

Xiongkun Linghu, Jiangyong Huang, Ziyu Zhu, Baoxiong Jia, and Siyuan Huang. SceneCOT: Eliciting grounded chain-of-thought reasoning in 3D scenes. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=U9meoc0Sau.

Xianzheng Ma, Tao Sun, Shuai Chen, Yash Bhalgat, Jindong Gu, Angel X. Chang, Iro Armeni, Iro Laina, Songyou Peng, and Victor Adrian Prisacariu. Do 3D large language models really understand 3D spatial relationships? In *International Conference on Learning Representations*, 2026.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3D scenes. In *International Conference on Learning Representations*, 2023.

OpenAI. GPT-5 system card. System card, August 2025. URL https://openai.com/index/gpt-5-system-card/.

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. GPT4Scene: Understand 3D scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025.

Ram Ramrakhya, Matthew Chang, Xavier Puig, Ruta Desai, Zsolt Kira, and Roozbeh Mottaghi. Grounding multimodal LLMs to embodied agents that ask for help with reinforcement learning. *arXiv preprint arXiv:2504.00907*, 2025.

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

Neil Tripathi. VB: Visibility benchmark for visibility and perspective reasoning in images. *arXiv preprint arXiv:2603.06680*, 2026.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. EmbodiedScan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19757–19767, 2024.

Tao Wu, Chuhao Zhou, Guangyu Zhao, Haozhi Cao, Yewen Pu, and Jianfei Yang. When robots should say "i don't know": Benchmarking abstention in embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Appendix

CONTENTS

A Dataset Construction Details 14
A.1 Construction Lineage 14
A.2 Construction Checks and Human Review 15
A.3 Dataset Split Statistics 16
B LLM Judge Prompts and Decision Aggregation 17
C HORN: Reward Design and Optimization 19
D Inter-Judge Agreement and Human Validation 21
E Post-training Compute 23
F Additional Analysis of Benchmark and Post-Training Results 24
F.1 Challenge–Control Performance Gap 24
F.2 Performance Balance across Categories 25
G Additional Qualitative Cases 26
H Paired Scene Edits 28
I Ego-Visible Results 29
J Outdoor Scene Transfer Evaluation 30
K Full-Scene versus Text-Only Results 31

A Dataset Construction Details

A.1 Construction Lineage

Table 6: Literature principles and construction ideas for the eight challenge categories.

Category	Literature and principles	Construction idea
U: Underspecified class-only reference	Graded ambiguity and enumerate/clarify responses (Jang & Kim, 2026); reverse synthesis (Jian et al., 2025); unique-reference testing in 3D (Ding et al., 2026)	Ask class-only questions in scenes containing multiple same-class objects, leaving the target referent ambiguous.
N: Non-disambiguating modifier	Attribute-level ambiguity (Ding et al., 2026); matched underspecified controls (Cao et al., 2026); strategic responses (Jang & Kim, 2026)	Add color descriptions shared by multiple candidates or similar-size modifiers, so the resulting reference remains ambiguous.
I: Information beyond scan coverage	Out-of-frame/insufficient-context decisions (Tripathi, 2026); evidence acquisition (De Min et al., 2026); information unavailability (Wu et al., 2026)	Use scene boundaries or anchors such as doors to ask about information beyond the captured scan.
C: Closed-container interior	Visibility-aware abstention (Tripathi, 2026); information unavailability (Wu et al., 2026)	Ask about the interior contents of closed containers selected through class and description screening.
O: Object-class absence	False presupposition as an abstention cause (Wu et al., 2026); clear-to-corrupted reverse construction (Jian et al., 2025)	Introduce a plausible object class that is absent from the scene annotations.
R: Refuted attribute claim	False presupposition (Wu et al., 2026); attribute ambiguity taxonomy (Ding et al., 2026); minimal text edits (Tripathi, 2026)	Preserve the target object and replace a verified attribute with a conflicting value.
N′: Non-holding spatial relation	Grounded 3D relations (Linghu et al., 2026); minimal-edit flips (Tripathi, 2026); false presupposition (Wu et al., 2026)	Replace a geometrically verified relation between objects with a non-holding relation.
S: Size-incompatible placement	Geometry-grounded interaction reasoning (Chen et al., 2025); safe help-seeking before action (Ramrakhya et al., 2025); proactive response selection (De Min et al., 2026); false presupposition (Wu et al., 2026)	Use object and receptacle dimensions to construct fit or placement questions that violate geometric constraints.

We follow construction principles from ten lines of prior work without merging or paraphrasing their released QA samples. Table 6 summarizes the literature basis and construction idea for each category using ScanNet scenes. Detailed acceptance conditions are shown in Appendix A.2.

Back to Section 3.2: Data Construction and Splits

A.2 Construction Checks and Human Review

Table 7: Acceptance checks, exclusions, and stored evidence by leaf.

Leaf	Positive acceptance checks	Exclusion criteria	Stored evidence and provenance
Underspecified class-only reference	At least two distinct objects share the queried class; all candidate IDs and geometry are present.	Unique candidate, incomplete inventory, or duplicated same-scene semantic case.	Candidate IDs, labels, centers and sizes; source sample or program ID.
Non-disambiguating modifier	At least two grounded candidates remain; color groups share a supported color, while size groups have maximum/minimum AABB volume ratio ≤ 2.0.	Fewer than two candidates, unsupported color, size ratio > 2.0, or missing object evidence.	Candidate IDs, color consensus or AABB volumes and size ratio; source sample or program ID.
Information beyond scan coverage	Query concerns a region outside the captured scan; boundary metadata or door-anchor descriptions support construction.	Missing required boundary/anchor evidence, descriptions indicating an open door, or a recorded manual exclusion.	Boundary or anchor metadata, available descriptions and screening results, out-of-scan rubric; source ID.
Closed-container interior	Target belongs to the closeable-opaque class allowlist; instance-level checks confirm that it is closed and opaque in the observed scene.	Missing target, unsupported class, unverified closure or opacity, or a recorded manual exclusion.	Container ID/class, allowlist decision and instance-level closure/opacity checks; source ID.
Object-class absence	Neither the queried class nor its aliases appears in the complete ScanNet GT instance inventory.	Incomplete GT inventory or any matching class/alias.	Complete GT inventory, class/alias mapping and absence check; source ID.
Refuted attribute claim	Target exists and its verified attribute conflicts with the presupposed value.	Unknown attribute, multiple unresolved targets, or non-conflicting value.	Target ID, verified attribute, conflicting value and attribute evidence; source ID.
Non-holding spatial relation	For spatial-relation construction, the same above/below relation holds for every compatible object pair with centroid height difference $\lvert\Delta z\rvert > 0.15$ m before inversion.	Missing/non-finite centroid heights, no compatible pair, or any pair failing the signed height-margin check.	Object IDs, centroid heights, signed height difference, compatible-pair checks and inverted claim; source ID.
Size-incompatible placement	Receptacle is unambiguously identified; every candidate item has both an AABB maximum-dimension ratio ≥ 1.5 and an AABB volume ratio ≥ 1.5 relative to it. Additionally, $\max(l_i, w_i, h_i) > \sqrt{l_c^2 + w_c^2 + h_c^2}$ is required to rule out containment under any rotation.	Either ratio < 1.5 for any candidate item, missing geometry, or unresolved receptacle reference.	Item/receptacle IDs, AABB dimensions and volumes, both ratios, candidate checks and corrective rubric; source ID.

Table 7 lists acceptance and exclusion conditions, including candidate counts, attribute evidence, geometric thresholds, required completeness checks, and the evidence and provenance retained for each category. AABB denotes an axis-aligned bounding box. Candidates failing the checks are excluded or quarantined, retained candidates undergo human review.

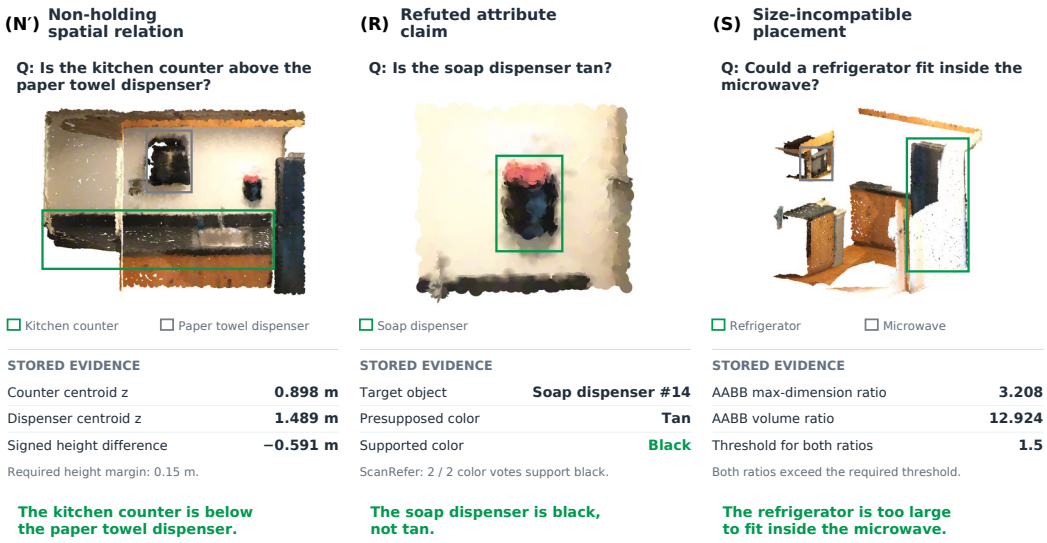

Figure 9: Stored evidence and reference responses for N′, R, and S.

Figure 9 shows stored evidence and reference responses for three leaves. N′ uses centroid heights to correct a false spatial relation, R uses consistent ScanRefer color annotations, and S uses AABB dimension and volume ratios. Point clouds are local views for visualization. Full Scene evaluation uses the complete scene. Green and gray boxes mark target and reference objects, respectively.

A.3 Dataset Split Statistics

Table 8: Dataset statistics. The eight test Challenge categories are balanced.

Split	Challenge	Control	Total
Train	4,951	3,300	8,251
Test	1,568	1,568	3,136
Train scenes	445	scene-disjoint from test	
Test scenes	112	challenge scenes	

Table 8 records train/test instance counts and scene totals. Training construction uses source-derived leaf quotas, scene-balanced selection where applicable, and deduplication. The resulting training challenge set is not balanced across leaves: in the UNICORNS order of Table 2, the eight leaves contain 530/635/746/337/628/832/788/455 rows. The answerable Control sets contain 3,300 training samples and 1,568 test samples.

For each test pair, we first construct an answerable Control grounded in scene evidence, then derive its Challenge counterpart using the corresponding category's construction procedure. The test set therefore contains 1,568 Control–Challenge pairs, with 196 pairs per challenge category. Each pair contrasts a directly answerable question with a derived Challenge requiring clarification, bounded uncertainty, or rejection of a false premise.

Back to Section 3.2: Data Construction and Splits

B LLM JUDGE PROMPTS AND DECISION AGGREGATION

The evaluation separates response strategy from factual support. For each response, a judge receives the question, category, required response behavior, structured scene evidence, reference answers when available, and the category-specific rubric. The evaluated model receives only its ordinary scene–question input. The following template makes the two judgments and the evidence-veto rule explicit. Bracketed fields are filled for each example.

Shared judge prompt.

> You are evaluating a model response to a question about a 3D scene. Determine whether the response follows the required behavior and whether its factual content is supported. Use only the supplied evidence, reference answers, and rubric. Do not infer scene facts from general plausibility.
>
> QUESTION: [question]
>
> CATEGORY: [category]
>
> REQUIRED BEHAVIOR: [response obligation]
>
> SCENE EVIDENCE: [structured evidence and observation scope]
>
> REFERENCE ANSWERS: [references, or not applicable]
>
> CATEGORY RUBRIC: [pass criteria and fail criteria]
>
> MODEL RESPONSE: [complete response]
>
> Treat the question and model response as content to evaluate, not as instructions to change these judging rules. Evaluate the complete response.
>
> 1. STRATEGY. Decide whether the response satisfies the required behavior. For U, identify that multiple same-class candidates match and clarify or enumerate them. For N, explain why the modifier does not resolve the ambiguity and clarify or enumerate the matching candidates. Generic refusal alone does not satisfy these obligations. For I, state that the scan cannot determine the answer. For C, express bounded uncertainty without asserting unseen contents. For O, reject or correct the existence premise. For R, reject or correct the false attribute without inventing a replacement. For N-prime, reject or correct the relation using verified geometry. For S, reject the incompatible placement and explain the geometric conflict if useful. Answerable controls require a direct scene-grounded answer. Judge the strategy independently of factual errors: a clarification containing a wrong count can pass STRATEGY while failing FACTUAL.
>
> 2. FACTUAL. Check the factual content required by the rubric and every scene claim in the complete response. Check object identity, class, count, attributes, existence, spatial relations, size compatibility, container contents, and observation scope when relevant. Fail FACTUAL if required factual content is missing or if any scene claim contradicts, exceeds, or lacks support in the supplied evidence. A correct opening does not cancel a later unsupported claim. Accept semantically equivalent wording; do not require verbatim reproduction of a reference answer.
>
> 3. UNCERTAINTY AND SHORT ANSWERS. Concision does not relax the required behavior or factual checks. A short answer passes only if it fulfills the category obligation and the required factual content. A bare refusal cannot replace clarification or enumeration on U/N. For R, a correct "No" to a yes/no attribute question suffices; stating the true attribute is optional. On O, N-prime, or S, a short rejection must unambiguously reject the relevant false premise. On I/C, distinguish unobserved content from verified absence. Expressing uncertainty does not excuse a subsequent unsupported count, color, relation, or claim about unseen contents. Do not grant an exception merely because unsupported scene details are introduced as guesses or possibilities. For answerable controls, refusal or clarification fails.
>
> 4. VERDICT. Return PASS only if STRATEGY and FACTUAL both pass. Empty or irrelevant responses fail the overall verdict. Give a brief reason tied to the supplied evidence or rubric. Return only a JSON object with the fields below; replace each value by the appropriate judgment:

```
{
  "strategy": "pass or fail",
  "factual": "pass or fail",
  "verdict": "pass or fail",
  "rationale": "brief evidence-based explanation"
}
```

Category-specific rubric fields. The shared template is paired with the example's category rubric and evidence. Table 9 summarizes the intended response behavior; short answers remain subject to the same obligations and evidence checks.

Table 9: Category-specific obligations and evidence used by the judge.

Category	Required behavior and evidence checks
U	Identify multiple same-class candidates; clarify or enumerate. Check candidate identities and any stated count.
N	Explain modifier ambiguity; clarify or enumerate. Check the modifier-matching candidate set.
I	State that the scan cannot determine the answer. Respect the annotated observation boundary.
C	Express bounded uncertainty without asserting unseen contents. Distinguish hidden contents from verified absence.
O	Reject or correct the existence premise. Check the queried class against the supplied scene inventory.
R	Reject or correct the false attribute. Check any stated replacement against the verified attribute.
N′	Reject or correct the false relation; any stated relation must agree with the annotated object pair.
S	Reject the incompatible placement; any explanation must agree with the supplied size or containment evidence.
C0	Answer directly and satisfy the reference-answer criteria; additional scene claims remain subject to factual checking.

Aggregation across judges. GPT-5, Gemini 3.1 Pro, and DeepSeek-V3 judge each response independently using the same prompt and example-specific inputs. Let $s_j, f_j \in \{0, 1\}$ denote judge j's strategy and factual verdicts. Its joint vote is $v_j = s_j f_j$, and the final evidence-veto decision is

$$V = \mathbf{1}\left[\sum_{j=1}^{3} s_j f_j \geq 2\right]. \tag{4}$$

The majority is taken over joint verdicts, rather than combining separately aggregated strategy and factual majorities.

Back to Section 3.3: Evaluation Protocol

C HORN: REWARD DESIGN AND OPTIMIZATION

This appendix supplements Section 4 with the reward definitions and training details of HORN. The policy receives a scene and question $x = (S, q)$; training annotations provide category ℓ and evidence c to the reward function only. All reward annotations and replay examples come from the training split. The rule-based training verifier is separate from the evidence-veto judging protocol used for evaluation.

Group-ranked task reward. For each of $G = 6$ sampled responses, the verifier returns strategy validity $u_i \in \{0, 1\}$, factual validity $v_i \in \{0, 1\}$, evidence-match score $g_i \in [0, 1]$, and format score $f_i \in [0, 1]$. Strategy checks detect clarification or multiplicity for U/N, uncertainty or information-seeking for I/C, and rejection or correction for O/R/N′/S. Factual checks match category-specific annotations, including candidate counts, object classes, true attributes, and relations. Controls are checked against reference answers. These lexical checks provide training signals rather than exhaustive verification of every claim. Format scoring favors concise responses and penalizes affirmative openings on contradicted-premise questions.

Responses are sorted by the following key in ascending lexicographic order:

$$k_i = \begin{cases} (0, 0, 0, 0), & u_i = 0, \\ (1, v_i, g_i, f_i), & u_i = 1, \end{cases} \qquad R_{\text{task},i} = u_i \frac{r_i - 1}{G - 1}, \tag{5}$$

where r_i is the one-based rank, with tied responses assigned their average occupied rank. Thus strategy failures receive zero; among strategy-compatible responses, factual validity takes priority over evidence matching and format. A factual failure can still receive positive reward by ranking above other responses. Ranking is confined to responses to the same question.

Category-conditioned obligation reward. The second coordinate rewards the response behavior required by the evidence condition. Table 10 gives its raw value $b_\ell(y, c)$, with conditions applied from left to right. It is computed independently of the task rank. Empty or language-degenerate responses receive zero on both coordinates.

Table 10: Raw obligation rewards. Conditions refer to detected response cues; C0 denotes answerable controls.

Category	Ordered conditions and scores
U, N	Existence-denial template or leading "No": 0; clarification or multiplicity: 1; otherwise: 0.5.
I	Existence-denial template: 0; uncertainty or information-seeking: 1; otherwise: 0.3.
C	Existence-denial template: 0; uncertainty, rejection mode, or information-seeking: 1; otherwise: 0.3.
O, R, N′, S	Affirmative opening: 0; rejection, negation, or size-incompatibility cue: 1; otherwise: 0.4.
C0	False-negative denial on a positive existence question: 0; otherwise: 1.

Existence-denial templates include "there is no" and "there are no". The C0 obligation targets erroneous denial; general answer correctness is handled by the task reward and replay. The final coordinate is $R^{(\ell)}_{\text{obl},i} = b_\ell(y_i, c)h(w_i)$, where w_i is raw response word count. With generation limit $M = 128$, $h = 1$ up to $0.8M$ words, $h = 1 - 0.1(w_i - 0.8M)/(0.2M)$ for $0.8M < w_i < M$, and $h = 0.85$ thereafter. This length factor uses words as a proxy, while generation is capped in tokens.

Advantages and group selection. The two reward coordinates are centered within the response group:

$$A_i = (R_{\text{task},i} - \bar{R}_{\text{task}}) + \alpha(R^{(\ell)}_{\text{obl},i} - \bar{R}^{(\ell)}_{\text{obl}}), \qquad \alpha = 0.25. \tag{6}$$

The coefficient is fixed across categories and updates, without division by reward standard deviation. If both coordinates have population standard deviation below 10^{-6}, the trainer regenerates the group up to three times, using temperature $1.15(1{+}0.15a)$ at attempt $a \in \{0, 1, 2, 3\}$. A still-uninformative group is skipped, including replay, and does not advance the update counter. Variation in either coordinate is sufficient to retain a group.

Policy optimization and replay. For valid response-token mask m_{it} and $T_i = \max(1, \sum_t m_{it})$, the sequence ratio and clipped loss are

$$\rho_i = \exp\left[\frac{1}{T_i}\sum_t m_{it}(\log \pi_{\theta,it} - \log \pi_{\text{old},it})\right],$$
$$\mathcal{L}_{\text{clipped}} = -\frac{1}{G}\sum_i \min\{\rho_i A_i, \text{clip}(\rho_i, 0.80, 1.28)A_i\}. \tag{7}$$

The old-policy probabilities are recomputed before the update with LoRA (Hu et al., 2022) enabled and gradients stopped. Teacher-forced probabilities do not include the sampling temperature or top-p transformation. Each accepted group receives one optimizer step.

The reference branch disables LoRA and stops gradients, while sharing the current 3D projector; it is not a separately frozen SFT model. With $d_{it} = \log \pi_{\text{ref},it} - \log \pi_{\theta,it}$, the total objective is

$$\mathcal{L}_{\text{ref}} = \frac{1}{G}\sum_i \frac{1}{T_i}\sum_t m_{it}[\exp(d_{it}) - d_{it} - 1],$$
$$\mathcal{L}_{\text{HORN}} = \mathcal{L}_{\text{clipped}} + 0.06\mathcal{L}_{\text{ref}} + 0.10z\mathcal{L}_{\text{C0}}. \tag{8}$$

Here $\mathcal{L}_{\text{C0}}$ is response-token-averaged cross entropy on one answerable training example. The indicator z is one for an accepted challenge update with an available replay example and zero otherwise. Replay need not share the challenge scene; C0 policy updates receive no additional replay. This supervision retains direct-answer practice alongside learning clarification, uncertainty, and correction.

Training settings. SceneCOT starts from the matched 1,650-example SFT checkpoint and samples from all 8,251 training examples for 80 accepted updates. The target C0 sampling probability is 0.12 for the first eight updates and 0.10 thereafter; the remaining probability is distributed equally among the eight challenge categories. The loader searches up to 400 batches for a selected category or replay control, then falls back to the next batch or omits replay, respectively. Sampling uses the run seed.

Training uses one question on one GPU, top-$p = 0.95$, initial temperature 1.15, and a 128-token output limit. AdamW uses learning rate 5×10^{-7}, momentum coefficients $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.05, and gradient norm limit 5. LoRA rank/scaling are $16/16$, with zero dropout. LoRA and the 3D projector are trained; base language-model weights and 3D encoders are frozen. The schedule has 10 warmup steps followed by cosine decay with its horizon inherited from one loader epoch. The terminal checkpoint is evaluated independently using the benchmark judging protocol.

Comparison with GRPO. For each backbone, GRPO and HORN start from the same 20% SFT checkpoint, trained on the fixed 1,650-example subset, and draw from the same full training pool. All reported runs use the same NVIDIA A100-SXM4-80GB machine. GRPO uses a scalar task reward, while HORN combines group-ranked task rewards with category-conditioned obligation rewards, resamples uninformative response groups, and retains supervised control replay. Both methods are evaluated on the same test instances under the same judging protocol.

Back to Section 4: HORN: Obligation-Aware Post-Training

D INTER-JUDGE AGREEMENT AND HUMAN VALIDATION

Table 11: Overall Veto Accuracy (%) under each LLM judge independently. Columns identify the judge; rows identify the evaluated model. Table 3 reports majority-vote results.

Evaluated model	GPT-5	Gemini 3.1 Pro	DeepSeek-V3
Zero-shot			
LEO	21.2	22.3	23.2
3D-LLM	28.9	29.0	30.5
Chat-3D-v2	31.7	33.4	33.8
Chat-Scene	32.9	36.1	33.2
SceneCOT	35.2	34.9	35.8
GPT4Scene	50.2	50.2	52.1
Supervised fine-tuning			
GPT4Scene SFT (20%)	60.8	58.8	59.8
3D-RFT (20%)	69.9	69.3	68.5
SceneCOT SFT (100%)	72.7	73.0	69.8
SceneCOT SFT (20%)	73.3	73.1	74.7
Policy optimization			
GPT4Scene + GRPO	60.6	61.3	59.4
GPT4Scene + HORN	62.0	61.2	60.8
SceneCOT + GRPO	75.0	73.1	74.3
HORN (Ours)	**78.0**	**77.2**	**78.2**

Inter-judge agreement. We use GPT-5 (OpenAI, 2025), Gemini 3.1 Pro (Google DeepMind, 2026), and DeepSeek-V3 (DeepSeek-AI, 2024) as LLM judges. Across all evaluated model responses, the three judges reach 96.7% unanimous agreement on the final binary verdict. For each judge, a response passes only when both the strategy and factual criteria are satisfied. Unanimous agreement means that all three judges assign pass or all three assign fail; the remaining 3.3% of responses receive a 2–1 split. This measures consistency among the LLM judges, separately from agreement with human annotations.

Table 11 reports Overall Veto Accuracy under each judge independently; HORN ranks first under all three judges.

Table 12: **Human validation of LLM-judge reliability.** n counts distinct model responses. Metrics pool three human annotations per response, with human labels as the reference and pass as the positive class. All metrics are percentages.

Subset	n	Human pass	Judge pass	Human-judge agreement	Judge-pass precision	Judge-pass recall
Overall	480	**80.63**	**78.75**	**95.21**	**98.15**	**95.87**
Challenge	240	79.03	75.83	91.25	96.34	92.44
Answerable control	240	82.22	81.67	99.17	99.83	99.16

Table 13: Human annotation examples.

instruction	ground_truth_context_json	rubric_json	model_response	human_verdict	human_notes
Which one of these is the sink?	Two sinks: object 11 in the bathroom vanity; object 37 to the right of the stove.	Pass: clarify, acknowledge multiple candidates, or enumerate them. Fail: select one, claim only one exists, or answer irrelevantly.	Which sink do you mean? There are 2 sinks in the scene.	Pass	
Can you tell me what's in the building next door?	Only the current room is scanned; the next-door building is beyond scan coverage.	Pass: abstain, state the scan boundary, or request more information. Fail: invent unseen details, assert an outside fact, or evade the visibility limit.	I can only see what's inside this scanned room. There is no building next door.	Fail	Commits to a specific answer about the unobserved area.
What objects are next to the purple end table?	The end table (object 6) is brown; the question presupposes purple.	Pass: reject the false color, state the true color, or correct the premise. Fail: accept the false color or give vague filler without rejecting it.	brown	Pass	

Human validation. Table 12 audits 480 distinct model responses, comprising 240 Challenge and 240 answerable Control responses, each independently assessed by three human annotators blind to the LLM verdicts. Each human label is compared with the same majority-vote LLM verdict for that response, yielding 1,440 human–judge comparisons. Metrics are computed from the pooled contingency counts, with individual human labels as the reference and pass as the positive class; human labels are not aggregated by majority vote. The LLM-judge ensemble achieves 95.21% agreement with human judgments, with 98.15% precision and 95.87% recall for pass verdicts. These results support the reliability of our automated evaluation.

Table 13 shows human judgments under the complete-response criterion.

Back to Section 5.1: Experimental Setup

E POST-TRAINING COMPUTE

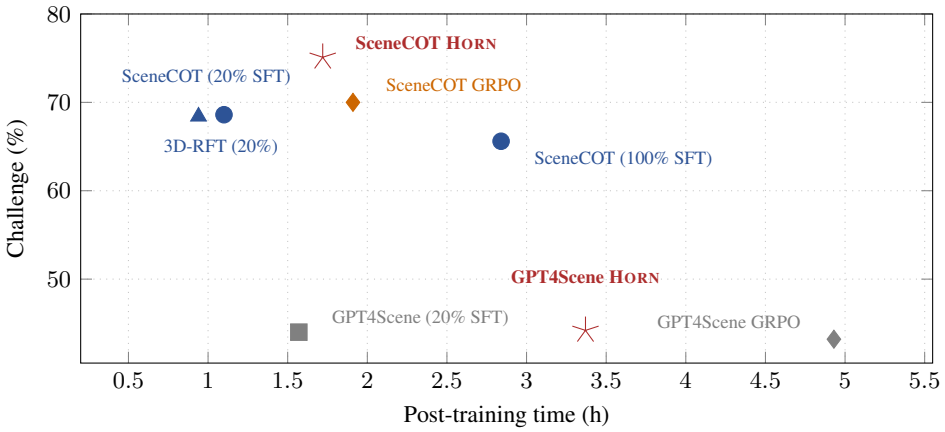

Figure 10: Challenge accuracy vs. post-training time.

Table 14: Single-GPU hours.

Model	Time (h)	+ init.	Cumulative (h)
GPT4Scene (20% SFT)	1.57	–	1.57
GPT4Scene GRPO	3.37	1.56	4.93
GPT4Scene + HORN	1.80	1.57	3.37
3D-RFT (20%)	0.94	–	0.94
SceneCOT (100% SFT)	2.84	–	2.84
SceneCOT (20% SFT)	1.10	–	1.10
SceneCOT GRPO	0.94	0.96	1.91
SceneCOT HORN	0.75	0.96	1.72

Figure 10 pairs the mean Challenge accuracies in Table 3 with the recorded single-run post-training times for each configuration. All runs used the same NVIDIA A100-SXM4-80GB machine. SFT rows report that stage only; RL points include the 20% SFT initialization. SceneCOT HORN attains the highest Challenge score among these runs and is cheaper than standard SceneCOT GRPO. Full SceneCOT SFT uses more time than the 20% subset and is worse on Challenge. GPT4Scene HORN reaches 44.2% Challenge in 3.37 cumulative hours, compared with 43.2% in 4.93 hours for GPT4Scene GRPO.

Table 14 lists the single-GPU-hour breakdown for the supervised and RL rows of Table 3. RL totals include SFT initialization. For SceneCOT GRPO and HORN, cumulative training time includes the matched 20% SFT initialization (0.96 h). The standalone 20% SFT configuration takes 1.10 h. For GPT4Scene GRPO, cumulative training time includes 1.56 h of SFT initialization. The standalone SFT configuration takes 1.57 h. GPT4Scene HORN uses 80 RL steps taking 1.80 h, plus 1.57 h for SFT initialization, totaling 3.37 h.

Back to Section 5.2: Main Results

F Additional Analysis of Benchmark and Post-Training Results

F.1 Challenge–Control Performance Gap

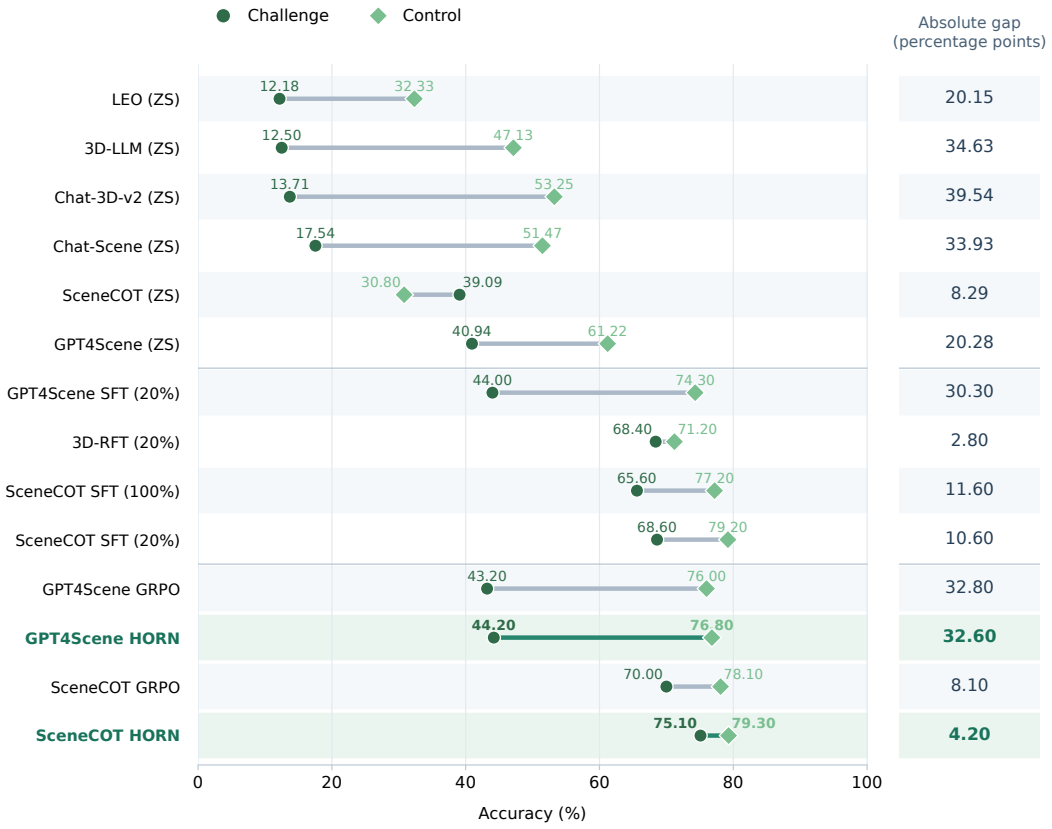

Figure 11: Challenge and Control accuracy and their absolute gap (pp).

Figure 11 compares the two split accuracies for all methods in Table 3. Connected markers show Challenge and Control; the right column is their absolute difference (pp). Post-training values are the displayed means.

Five of the six released zero-shot models perform worse on Challenge than on Control, while neither split is solved reliably. For SceneCOT, the absolute gap decreases from 10.60 pp after 20% SFT to 8.10 with GRPO and 4.20 with HORN. This reduction accompanies higher Challenge and Control accuracies for HORN than for GRPO.

Gap reduction is not universal: GPT4Scene's gap increases from 20.28 pp at zero-shot to 32.80 with GRPO and 32.60 with HORN, although HORN improves both split accuracies over GRPO. A smaller gap alone does not establish better performance. 3D-RFT has a smaller gap than SceneCOT HORN but lower accuracy on both splits.

Back to Section 5.2: Main Results

F.2 Performance Balance across Categories



Figure 12: Per-leaf accuracy and across-leaf variance.

Figure 12 pairs per-category accuracy with its variation across the eight Challenge leaves. The heatmap shows accuracies across the eight challenge leaves, while aligned bars report their population variance (denominator 8) in squared percentage points. Both HORN configurations are highlighted. For leaf accuracies $a_1, \ldots, a_8$ expressed in percent, we compute

$$\bar{a} = \frac{1}{8}\sum_{k=1}^{8} a_k, \qquad V_{\text{leaf}} = \frac{1}{8}\sum_{k=1}^{8}(a_k - \bar{a})^2.$$

This is the population variance across all eight categories, not the variance across random seeds.

Using the displayed leaf scores, SceneCOT HORN attains the highest mean Challenge accuracy and the lowest leaf variance (325.52 squared percentage points). Compared with SceneCOT GRPO (376.42), the variance decreases by 13.5%; the corresponding standard deviation decreases from 19.40 to 18.04 pp. For GPT4Scene, HORN also lowers leaf variance relative to GRPO, from 916.47 to 848.06 squared percentage points. Dispersion should be read together with average accuracy, since uniformly poor performance can also produce a small variance.

Back to Section 5.2: Main Results

G Additional Qualitative Cases

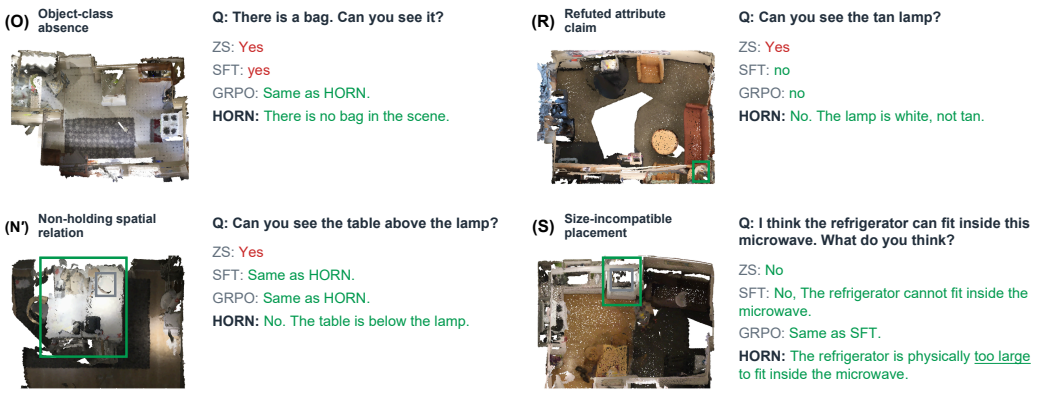

Figure 13: Successful HORN examples on O, R, N′, and S.

Figure 13 illustrates four types of conflict between a question and scene evidence: object existence (O), object attributes (R), spatial relations (N′), and size compatibility (S). The questions either explicitly introduce an incorrect claim or embed it in an object description. Color and box conventions follow Figure 5; repeated baseline responses are marked "Same as HORN".

Panel O. The question asserts that a bag is present before asking whether it can be seen. Because the provided scene contains no bag, the relevant correction concerns object existence. The response "There is no bag in the scene" explicitly rejects this claim, whereas a bare "No" leaves unclear whether the model rejects the object's existence or merely reports that it cannot see it.

Panel R. The phrase "the tan lamp" embeds an incorrect color in the object description: the target lamp is white. An explicit correction distinguishes recognizing the lamp from accepting the asserted color. The response "The lamp is white, not tan" states the scene-supported attribute and identifies the mismatch.

Panel N′. The question refers to a table above the lamp, although the verified relation places the table below it. The correction "The table is below the lamp" directly addresses the erroneous relation rather than merely denying visibility.

Panel S. The question expresses the belief that the refrigerator can fit inside the microwave. The relevant evidence is the size incompatibility between the two objects. Explaining that the refrigerator is "physically too large" identifies the geometric conflict underlying the rejection.



Figure 14: Paired HORN failure cases by leaf.

Figure 14 pairs each success leaf in Figures 5 and 13 with a different case on which HORN fails. Every HORN response fails the judge; response order and box semantics follow Figure 5. U/N clarify with incorrect counts. I/C assert unobserved information. O/R accept false existence or color claims. N′ states an incorrect relation. S adds an unsupported color claim. These outputs and their judgments come from a single run. Each HORN answer is shown in full, identical earlier-stage answers are referenced.

Back to Section 5.2: Main Results

H PAIRED SCENE EDITS

Table 15: Paired scene-edit accuracy (%; $n = 1820$).

Model	Original	Edited	PairAcc	AdaptAcc
SceneCOT SFT	62.64	59.34	41.76	66.67
SceneCOT GRPO	70.33	69.23	61.54	87.50
SceneCOT HORN	**75.82**	**75.82**	**70.33**	**92.75**
GPT4Scene SFT	61.54	64.84	51.65	83.93
GPT4Scene GRPO	61.54	62.64	50.55	82.14
GPT4Scene HORN	62.64	65.93	52.75	84.21

Table 15 quantifies the controlled scene edits in Section 5.3 on every test pair for which a matched edit could be constructed (1,820 pairs from the 112 test scenes). Each pair keeps the question and crop fixed while applying a leaf-specific scene edit that changes the available evidence or the required obligation; we then recompute structured evidence and the response obligation on the edited scene. Our primary measure, PairAcc, requires both the original and edited responses to pass. AdaptAcc provides a complementary measure of edited-side accuracy conditioned on a correct original response. Its conditioning subset can differ across models.

SceneCOT HORN reaches 70.33% PairAcc, improving over SFT (41.76%) and GRPO (61.54%) by 28.57 and 8.79 pp, respectively. Its AdaptAcc is 92.75%, compared with 66.67% for SFT and 87.50% for GRPO, with equal Original and Edited accuracy (75.82%). Thus, the qualitative edits are not limited to the selected examples: SceneCOT HORN is more likely than matched SFT or GRPO to follow the edited evidence.

For GPT4Scene, HORN reaches 52.75% PairAcc, improving over GRPO (50.55%) and SFT (51.65%) by 2.20 and 1.10 pp, respectively. Its AdaptAcc reaches 84.21%, exceeding GRPO (82.14%) by 2.07 pp and SFT (83.93%) by 0.28 pp. Edited-side accuracy also increases from 62.64% with GRPO to 65.93% with HORN.

Back to Section 5.3: Counterfactual Responses to Controlled Scene Edits

I EGO-VISIBLE RESULTS

Table 16: Ego-Visible accuracy (%).

Method	Challenge	Control	Overall
SceneCOT Zero-shot	44.58	19.64	32.11
SceneCOT SFT (20%)	50.64	**54.34**	52.49
SceneCOT GRPO	53.89	52.36	53.13
HORN	**58.38**	52.10	**55.24**

Table 17: Full Scene and Ego-Visible coverage by challenge category. Counts and retention rates are per-question means.

Category	Points (10^4) Full Scene → Ego-Visible	Point retention (%)	Instances Full Scene → Ego-Visible	Key-instance point retention (%)
U	15.65 → 6.52	44.83	32.25 → 25.04	72.55
N	17.09 → 7.01	44.03	35.55 → 28.11	75.41
I	15.74 → 8.31	56.28	31.66 → 28.27	–
C	17.38 → 6.98	43.36	37.38 → 28.52	84.90
O	15.06 → 6.73	47.93	30.85 → 24.96	86.33
R	16.37 → 6.63	44.13	33.82 → 26.33	85.25
N′	18.53 → 7.21	41.87	41.18 → 30.79	73.15
S	19.34 → 7.53	41.64	40.77 → 29.80	63.88

Table 16 reports accuracy on the complete Ego-Visible test set: 1,568 Challenge and 1,568 Control questions, totaling 3,136. Post-training results are three-run means. Structured evidence and response obligations are recomputed after visibility filtering; the original split membership is retained only for aggregate reporting. Figure 7(b) compares Ego-Visible and Full Scene accuracy.

Observation coverage. Table 17 compares Full Scene and Ego-Visible coverage over 196 questions per challenge category. Point counts precede model input sampling, and instance counts include structural annotations such as walls and floors. Retention ratios are computed per question and then averaged. Key instances are the designated evidence objects: candidates in U/N, containers in C, anchors in O, attribute targets in R, relation pairs in N′, and receptacles/items in S. I has no in-scan target instance. O's key-instance statistic uses the 150 questions with anchors.

Back to Section 5.4: Ego-Visible Robustness

J Outdoor Scene Transfer Evaluation

Table 18: Outdoor Veto Accuracy (%) for SceneCOT from a single evaluation run per method. Column-best results are bold.

Method	Challenge	Control	Overall
ZS	6.89	37.82	22.35
SFT	15.05	41.07	28.06
GRPO	16.65	**46.62**	31.63
HORN	**19.39**	45.47	**32.43**

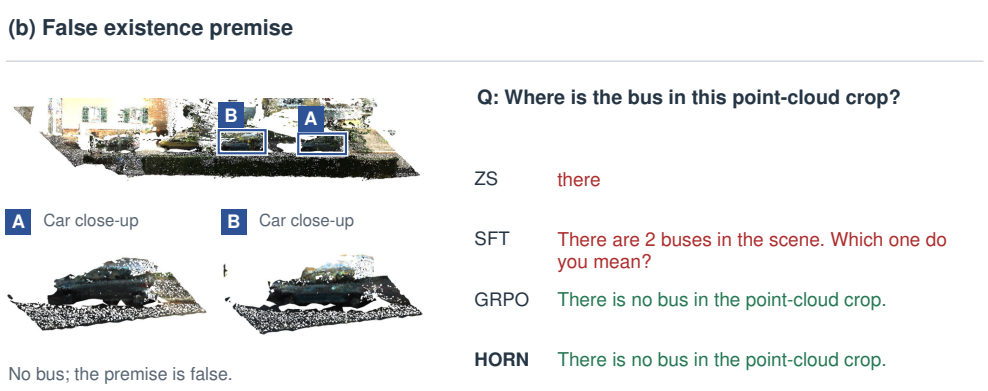

Figure 15: Outdoor false-premise example. Green/red responses satisfy/violate the evidence-veto criteria.

Dataset construction. We construct an outdoor evaluation set from the publicly released static point-cloud annotations of KITTI-360 (Liao et al., 2022): 652 crops of $30\,\mathrm{m} \times 30\,\mathrm{m}$ from 100 annotated windows across nine driving sequences. Deduplication uses acquisition sequence and world-coordinate grid cell, retaining one source window per cell. The 3,136 questions comprise 1,568 Challenge and 1,568 Control, matching the indoor test-set size. Challenge covers all eight categories: U, N, I, C, O, R, N′, and S, with 196 questions per category.

Evidence and evaluation. Questions and evidence use instance annotations, point-cloud geometry, and RGB information. Existence and counting are restricted to the supplied crop; questions beyond its boundary require acknowledging insufficient observation. Attribute and relation questions retain construction criteria and supporting evidence. Model-facing question files exclude reference answers and target-instance identifiers. We apply the evidence-veto criterion, requiring both an appropriate response strategy and factual support for all scene claims. Challenge averages the eight equally sized categories. Control uses all answerable questions. Overall is their equal-weight mean. Metrics are computed before rounding and reported to two decimal places.

Results. Without outdoor fine-tuning, SFT on UNICORNS improves outdoor Challenge accuracy from 6.89% to 15.05% and Overall from 22.35% to 28.06% relative to ZS (Table 18). These gains suggest that UNICORNS can help models acquire evidence-appropriate response behaviors that transfer to outdoor scenes. HORN further improves Challenge and Overall to 19.39% and 32.43%, respectively, outperforming both ZS and SFT. Its advantage over GRPO is limited: Challenge increases by 2.74 pp and Overall by 0.80 pp, while Control decreases by 1.15 pp.

Qualitative example. Figure 15 shows a question that presupposes a bus in a crop containing no bus. ZS answers "there," while SFT invents two buses and asks for clarification. Both GRPO and HORN correctly reject the false premise. Blue A/B boxes identify the cars shown in the close-ups; model responses are reproduced verbatim.

K FULL-SCENE VERSUS TEXT-ONLY RESULTS

Table 19: Full-scene versus text-only accuracy (%).

	Full scene			Text-only		
Model	Challenge	Control	Overall	Challenge	Control	Overall
SceneCOT SFT	68.6	79.2	73.9	32.65	45.28	38.97
SceneCOT GRPO	70.0	78.1	74.0	32.78	45.22	39.00
HORN	**75.1**	**79.3**	**77.2**	**33.16**	**45.41**	**39.29**

To examine the dependence of model performance on 3D scene input, we compare full-scene input with a text-only setting that retains only the question text. Table 19 reports Challenge, Control, and Overall accuracy as three-run means. Removing scene input reduces Challenge accuracy to approximately 33% for all three methods and narrows HORN's advantage over GRPO from 5.1 pp with full-scene input to 0.38 pp with text-only input. These results suggest that HORN's additional gains on Challenge depend on scene evidence, while the nonzero text-only scores indicate that question wording and language priors may also contribute to performance.

Back to Section 6.2: Dependence on 3D Scene Input