Seeing or Rationalizing? Scene-Evidence-Guided Chain-of-Thought for Faithful 3D Multimodal Reasoning

Guoqing Ma[1,2,3]* **Lihao Guo**[4]* **Yihao Zhong**[5]
Chuangxin Zhao[6]† **Mingqi Yuan**[7] **Shan Yu**[1,3]†
[1]Laboratory of Brain Atlas and Brain-inspired Intelligence,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Future Technology, University of Chinese Academy of Sciences, Beijing, China
[3]State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4]Independent Researcher
[5]New York University, New York, USA
[6]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[7]The University of Hong Kong, Hong Kong, China

Abstract

3D multimodal language models are becoming a foundation for embodied agents, indoor robotics, and human–AI interaction, but their answers and explanations often remain weakly tied to the specific objects and spatial relations observed in a 3D scene. This paper aims to make 3D reasoning more faithful by reducing language-prior rationalization and forcing reasoning chains to depend on verifiable scene evidence. We propose Anti-Rationalization Chain-of-Thought (AR-CoT), a plug-in training and decoding framework that represents each reasoning step with an explicit evidence pointer and scores candidate answer-chain pairs by their scene-evidence gain. AR-CoT contrasts scene-conditioned rationales with a frozen text-only prior, verifies local object and relation claims, and uses decoy-twin scene pairs to encourage reasoning chains to change when the relevant spatial evidence changes. Experiments on standard 3D QA and grounding benchmarks, as well as MSQA, Beacon3D, and shortcut-sensitive stress tests, show that AR-CoT consistently improves multiple 3D MLLM backbones while strengthening chain grounding, decoy-twin contrast, and scene-swap sensitivity. These results suggest that anti-rationalization offers a practical and verifiable route toward more accurate, interpretable, and scene-faithful 3D multimodal reasoning.

1 Introduction

Understanding and reasoning about 3D scenes is a central capability for embodied agents, indoor robotics, augmented reality, and human–AI interaction. Unlike 2D image understanding, 3D reasoning requires models to interpret object geometry, spatial relations, viewpoint-dependent observations, and language instructions within a structured physical environment. Recent progress in 3D vision–language learning has enabled impressive advances in 3D grounding, captioning, question answering, and embodied interaction [Chen et al., 2020, Achlioptas et al., 2020, Azuma et al., 2022, Ma et al., 2023, Hong et al., 2023, Chen et al., 2024a, Huang et al., 2024b].

*Equal contribution.
†Corresponding authors: Chuangxin Zhao (czhao630@connect.hkust-gz.edu.cn) and Shan Yu (shan.yu@nlpr.ia.ac.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Despite these advances, a fundamental gap remains between *answering* a 3D question and *reasoning from* the observed 3D scene. Most existing 3D multimodal language models are evaluated primarily by final-task accuracy, grounding quality, or response fluency [Zhu et al., 2023, Xu et al., 2024, Huang et al., 2024a, Chen et al., 2024b, Zhu et al., 2025, Zheng et al., 2025]. These metrics are valuable, but they do not directly test whether a model's intermediate reasoning is causally tied to scene-specific evidence. Related evidence from continual multimodal learning shows that answer accuracy can be preserved even as grounding behavior drifts [Chen et al., 2026b]. This limitation is especially problematic because both visual question answering and chain-of-thought reasoning are known to be vulnerable to shortcut-driven prediction and post-hoc rationalization: models may exploit language priors, object co-occurrence, or dataset biases while still producing plausible explanations [Agrawal et al., 2018, Turpin et al., 2023, Lanham et al., 2023]. Recent diagnostic benchmarks for 3D vision–language understanding further show that grounding, hallucination control, and spatial robustness remain fragile in current 3D MLLMs [Yang et al., 2025, Huang et al., 2025a, Ma et al., 2025]. Therefore, the key unresolved challenge is not merely how to inject stronger 3D features into language models, but how to ensure that their reasoning process is actually supported by the observed 3D evidence.

This paper asks: *How can we train and decode 3D multimodal language models so that their reasoning chains are faithful to scene-specific object and relation evidence, rather than merely plausible from language priors?* We address this question from an anti-rationalization perspective. As shown in Fig.1, our approach treats a valid 3D reasoning chain as one that should be more supported when the scene is observed, less explainable from the question alone, and locally consistent with verifiable objects and spatial relations in the scene. This view shifts 3D reasoning from final-answer prediction toward evidence-sensitive reasoning, where intermediate steps are explicitly evaluated by their dependence on the 3D world.

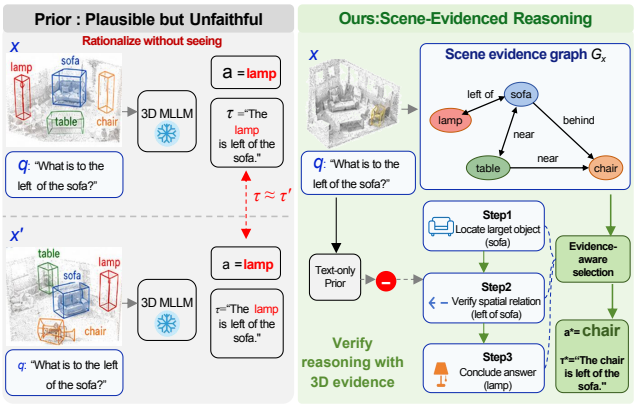


Figure 1: AR-CoT suppresses language-prior rationalization by selecting reasoning chains that are supported by scene-specific 3D evidence.

Our contributions are summarized as follows:

- We identify a rationalization failure mode in 3D multimodal reasoning: correct answers and fluent chains can remain weakly dependent on the actual scene, revealing that standard QA accuracy may overestimate genuine 3D reasoning ability.
- We propose an anti-rationalization framework for 3D chain-of-thought reasoning that promotes scene-supported chains, suppresses language-only rationales, and encourages intermediate claims to be grounded in object- and relation-level evidence.
- AR-CoT can be instantiated on multiple existing 3D MLLM backbones without modifying their 3D encoders, and our experiments show consistent gains on standard 3D QA and grounding benchmarks as well as on faithful-reasoning and shortcut-sensitive evaluations.

2 Related Work

3D vision–language grounding and question answering. 3D perception methods such as DepthSSC use depth–spatial alignment for monocular semantic scene completion [Yao et al., 2025]. Harmonic-NeRF addresses sparse-view driving-scene reconstruction through geometry-informed synthetic-view augmentation [Pan et al., 2024]. VIBUS combines viewpoint-bottleneck pretraining and uncertainty–spectrum pseudo-label selection for data-efficient 3D scene parsing [Tian et al., 2022]. Complementary 3D vision–language work addresses language-conditioned scene understanding. ScanRefer and ReferIt3D established object-level reference resolution in reconstructed scenes [Chen et al., 2020,

Achlioptas et al., 2020], and Scan2Cap extended this line to dense scene captioning [Chen et al., 2021]. In 2D task-oriented grounding, TOIST segments objects affording verb-specified actions through noun–pronoun distillation [Li et al., 2022]. Later 3D benchmarks broadened the scope to free-form QA, situated reasoning, and multi-object grounding through ScanQA, SQA3D, and Multi3DRefer [Azuma et al., 2022, Ma et al., 2023, Zhang et al., 2023b]. Larger resources such as SceneVerse, EmbodiedScan, and MMScan further scale 3D-language supervision with richer grounded annotations [Jia et al., 2024, Wang et al., 2024, Lyu et al., 2024]. These resources emphasize final answers, boxes, or captions, leaving intermediate reasoning faithfulness underexamined. Our method makes each reasoning step locally checkable against object and relation evidence in the observed scene.

3D multimodal large language models. Vision–language models support open-ended interaction in 3D scenes and online video streams [Hong et al., 2023, Li et al., 2025]. 3D-LLM injects multi-view features for captioning, QA, grounding, and navigation [Hong et al., 2023]; LL3DA uses point clouds for instruction-tuned 3D understanding and planning [Chen et al., 2024a]; PointLLM studies object-level point-cloud instruction following [Xu et al., 2024]; and LEO extends to embodied perception, planning, and action [Huang et al., 2024b]. ConsisVLA-4D aligns cross-view object semantics and cross-object geometric relations for spatiotemporally consistent manipulation [Li et al., 2026c]. Object-centric methods such as Chat-Scene and Grounded 3D-LLM improve grounding through object identifiers, referent tokens, and phrase-level alignment [Huang et al., 2024a, Chen et al., 2024b]. LLaVA-3D, Video-3D LLM, LSceneLLM, and SPARTUN3D use 3D patches, position-aware video features, adaptive visual selection, and situated spatial alignment, respectively [Zhu et al., 2025, Zheng et al., 2025, Zhi et al., 2025, Zhang et al., 2025]. However, benchmarks such as 3D-GRAND, Beacon3D, and 3DSRBench show that grounding, hallucination control, spatial robustness, and QA coherence remain fragile [Yang et al., 2025, Huang et al., 2025a, Ma et al., 2025]. AR-CoT penalizes language-only rationalization and rewards scene-verified reasoning steps through an evidence-aware objective.

Structured representations and multimodal alignment. Foundation-model representations support cross-modal transfer and alignment [Lu et al., 2025a]. Indra uses sample-to-sample relational profiles for training-free alignment [Lu et al., 2025b]. CoMem stores concept–relation graphs for continual vision–language adaptation [Zhou et al., 2026]. TPC learns a consensus semantic core from partial captions [Yu et al., 2026]. InstanceRSR aligns diffusion and instance representations for image super-resolution [Guo et al., 2026]. Other feature models use polynomial approximation for video super-resolution [Zhang et al., 2023a], local/global facial-expression fusion [Zhang et al., 2024a], and directional/frequency–spatial attention for medical segmentation [Zhang et al., 2026c]. Chang et al. [2025] introduce PDU-Net, which combines path aggregation and dual attention for lung nodule segmentation. TimeSAF decouples temporal learning from language-guided asynchronous fusion for forecasting [Zhang et al., 2026b].

Hierarchical planning and world models. Diffusion Subgoal Planning generates goal-conditioned subgoals for long-horizon offline reinforcement learning [Zhang et al., 2026d], while MHWA uses temporal hierarchies and context-aware routing for world-action modeling [Pan et al., 2026]. These approaches organize sequential decisions through subgoals, temporal hierarchies, or memory-grounded transitions [Pan et al., 2026, Li et al., 2026d], complementing our scene-grounded reasoning.

Faithful chain-of-thought and anti-rationalization. Chain-of-thought prompting elicits intermediate reasoning steps [Wei et al., 2022, Wang et al., 2025], and self-consistency improves accuracy by marginalizing over multiple paths [Wang et al., 2023]. Multimodal-CoT extends this paradigm to image-text reasoning by separating rationale generation from answer inference [Zhang et al., 2024b]. Fluent rationales can rationalize shortcut-driven predictions [Turpin et al., 2023] without causally determining answers [Lanham et al., 2023]. MCD transfers evidence-dependent behavior through token interventions during multimodal in-context learning distillation [Li et al., 2026e]. Faithful CoT mitigates this issue in symbolic domains via executable reasoning programs [Lyu et al., 2023], but such solvers are hard to apply to open-ended 3D scene reasoning. In Text-to-SQL, RevDecomp-SFT distills gold-SQL-conditioned trajectories with explicit schema linking, decomposition, and revision [Guan et al., 2027]. MUSE combines structured parsing, deterministic verification, and verifier-guided repair around frozen MLLMs [Lu et al., 2026a]. In optical networks [Xiao et al., 2026a], Xiao et al. [2025] use graph attention for joint alarm correlation and root-alarm detection. Xiao et al. [2026b] ground an OpenROADM LLM agent in YANG schemas for structure-preserving retrieval and configuration validation. In autonomous driving, Hint-AD aligns language explanations

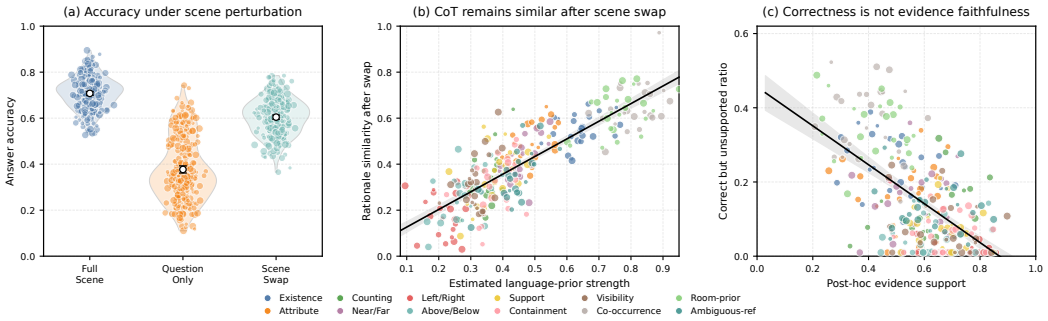


Figure 2: **Exploratory diagnosis of rationalization without seeing.** Each point denotes a question bucket grouped by relation type and shortcut strength. (a) Question-only and scene-swapped inputs retain non-trivial accuracy, especially for common-sense and co-occurrence-heavy questions. (b) Free-form CoT rationales often remain highly similar after scene swap, suggesting weak dependence on changed spatial evidence. (c) Correct answers are not always supported by scene-grounded rationales; many correct predictions have low post-hoc evidence support.

with intermediate perception, prediction, and planning outputs [Ding et al., 2025]. Similar shortcut concerns arise in VQA, where question-only priors can inflate performance [Agrawal et al., 2018]. In retrieval-augmented generation, TCR distinguishes semantic relevance, factual consistency, and self-answerability to mediate conflicts between internal knowledge and retrieved evidence [Ye et al., 2026]. Memory-agent benchmarks likewise test verification and clarification decisions before information reuse [Li et al., 2026a]. AR-CoT contrasts scene-conditioned chains with a text-only rationale prior, verifies local object and relation claims, and uses decoy twins to enforce scene sensitivity.

3 Preliminary Study

Before presenting our method, we conduct an exploratory diagnostic study to understand whether existing 3D multimodal language models truly use scene evidence when producing chain-of-thought rationales. This study exposes a failure pattern that motivates our later design: a model may answer correctly and generate a plausible spatial explanation even when the explanation is largely predictable from language priors or remains insensitive to changes in the 3D scene.

Diagnostic setup. We evaluate an off-the-shelf 3D MLLM with standard free-form chain-of-thought prompting. For each question q, the model generates an answer $\hat{a}$ and a rationale $\hat{\tau}$ under three input conditions: *Full Scene*, where the original 3D scene x is provided; *Question Only*, where the scene input is removed; and *Scene Swap*, where q is paired with another scene that preserves the same object categories but changes the relevant spatial evidence. These perturbations are used only for diagnosis and do not involve any component of our proposed method.

We measure answer accuracy, rationale similarity under scene swap, and a coarse post-hoc evidence support score. The evidence support score measures the fraction of object and relation mentions in a generated rationale that can be matched to annotated scene evidence:

$$\mathrm{Supp}_{\mathrm{diag}}(\hat{\tau}, x) = \frac{\#\{\text{object/relation mentions in } \hat{\tau} \text{ supported by } x\}}{\#\{\text{object/relation mentions in } \hat{\tau}\} + \epsilon}. \tag{1}$$

This score is used only for analysis; it is not used for training, reranking, or model selection.

Findings. Figure 2 reveals three rationalization patterns. First, removing or swapping the scene does not fully collapse performance, and question-only inputs remain competitive for co-occurrence and room-prior questions, showing that standard QA accuracy can be inflated by language shortcuts. Second, free-form rationales often remain textually similar after scene swaps, even when the relevant spatial relation changes, suggesting that the model tends to generate plausible question-conditioned explanations rather than evidence-sensitive chains. Third, correctness and evidence support are only weakly aligned: many correct answers are accompanied by rationales that mention absent, ambiguous, or unsupported objects and relations. Thus, a correct final answer with a fluent chain should not be automatically regarded as faithful 3D reasoning [Sun et al., 2026].

Implication. These findings motivate a faithful 3D reasoning framework that compares each chain against question-only rationalization, exposes the scene evidence behind intermediate steps, and verifies local object and relation claims. AR-CoT instantiates these principles with scene-evidenced chains, a text-only rationale prior, and evidence-aware chain selection.

4 Method

We propose **Anti-Rationalization Chain-of-Thought** (AR-CoT), a framework for training and decoding 3D multimodal large language models whose reasoning chains must be supported by scene-specific evidence. Given a 3D scene x and a question q, AR-CoT generates an answer-chain pair $y = (a, \tau)$, where a is the final answer and τ is a structured chain of intermediate reasoning steps. The central principle is that a valid 3D chain should not only be likely under the scene-conditioned model, but should also become less supported when the observed scene is removed. AR-CoT therefore contrasts the scene-conditioned likelihood of y against a text-only rationale prior, and further checks whether each step in τ is grounded in object-level or relation-level evidence. The resulting *scene-evidence gain* is used both for training and inference.

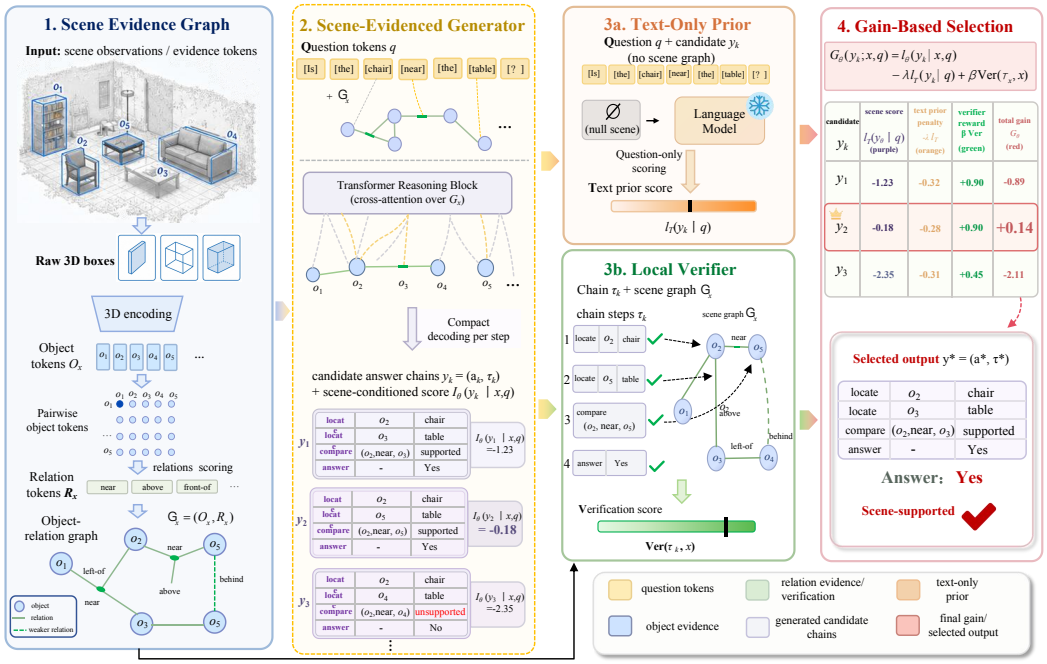


Figure 3: **Pipeline of AR-CoT.** The model first constructs an object-relation evidence graph from the 3D scene, then generates candidate answer-chain pairs conditioned on the question and the scene graph. Each candidate is scored by a text-only rationale prior and a local evidence verifier. AR-CoT finally selects the candidate with the largest scene-evidence gain.

4.1 Scene-Evidenced Reasoning Chains

Given a 3D scene x, we construct an object-centric evidence graph

$$\mathcal{G}_x = (\mathcal{O}_x, \mathcal{R}_x), \qquad \mathcal{R}_x \subseteq \mathcal{O}_x \times \mathcal{P} \times \mathcal{O}_x, \tag{2}$$

where $\mathcal{O}_x = \{o_i\}_{i=1}^{M}$ is the set of object evidence units, $\mathcal{P}$ is the spatial-relation vocabulary, and $\mathcal{R}_x$ is the set of relation evidence units. Each relation evidence unit $(o_i, r, o_j) \in \mathcal{R}_x$ states that relation $r \in \mathcal{P}$ is supported between objects o_i and o_j. The construction of $\mathcal{G}_x$, including object features and relation scores, is given in Appendix A.1.

AR-CoT represents the reasoning chain as

$$\tau = (s_1, \ldots, s_T), \qquad s_t = (u_t, e_t, m_t), \qquad e_t \in \mathcal{O}_x \cup \mathcal{R}_x. \tag{3}$$

Here, u_t is a reasoning operation, e_t is the evidence pointer used by step t, and m_t is the local claim produced at that step. The operation u_t belongs to a compact set of reasoning actions, including `locate`, `filter`, `compare`, `count`, `verify`, and `conclude`. The evidence pointer e_t forces each intermediate claim to be tied to a concrete object or spatial relation in $\mathcal{G}_x$. The pointer parameterization is described in Appendix A.2.

The scene-conditioned model generates the answer-chain pair $y = (a, \tau)$ as

$$p_\theta(y \mid x, q) = p_\theta(a, \tau \mid x, q) = \left[\prod_{t=1}^{T} p_\theta(s_t \mid s_{<t}, q, \mathcal{G}_x)\right] p_\theta(a \mid \tau, q, \mathcal{G}_x). \tag{4}$$

In Eq. (4), $s_{<t}$ denotes all previous reasoning steps. The step distribution predicts the operation, the evidence pointer, and the local claim in sequence. This design makes the chain inspectable at the level of individual scene claims.

4.2 Text-Only Rationale Prior

A fluent chain does not imply scene-grounded reasoning. Many chains are plausible because they follow common object co-occurrence, room-type priors, and language-level spatial conventions. AR-CoT exposes this effect through a text-only rationale prior.

We instantiate the prior as a frozen copy of the language decoder with all scene tokens replaced by a learned null evidence token $\varnothing$:

$$p_T(y \mid q) = p_{\bar{\theta}}(a, \tau \mid q, \varnothing), \qquad \bar{\theta} \text{ is frozen.} \tag{5}$$

The prior receives the same question q as the scene-conditioned model but receives no object, relation, view, or geometry evidence. It therefore estimates how likely the answer-chain pair is from language alone. For length-stable comparison, we use normalized log-likelihoods

$$\ell_\theta(y \mid x, q) = \frac{1}{|y|} \log p_\theta(y \mid x, q), \qquad \ell_T(y \mid q) = \frac{1}{|y|} \log p_T(y \mid q), \tag{6}$$

where $|y|$ is the number of generated answer and chain tokens. A high $\ell_T(y \mid q)$ indicates that the chain is already strongly supported by language priors before observing the 3D scene.

4.3 Local Evidence Verification

The text-only prior measures language rationalization, while the verifier measures whether the stated evidence is valid in the current scene. For each step $s_t = (u_t, e_t, m_t)$, the local verifier outputs

$$v_t = \mathrm{Ver}_\psi(s_t, \mathcal{G}_x) \in (0, 1), \qquad \mathrm{Ver}(\tau, x) = \frac{1}{T} \sum_{t=1}^{T} v_t. \tag{7}$$

Here, v_t is the support probability of step s_t, and $\mathrm{Ver}(\tau, x)$ is the average chain-level support. A step receives a high score only when its evidence pointer exists in $\mathcal{G}_x$ and its local claim is consistent with the corresponding object or relation evidence. The verifier features, predicate checks, and calibration loss are specified in Appendix A.3.

4.4 Scene-Evidence Gain

AR-CoT scores each answer-chain pair by its scene-evidence gain:

$$G_\theta(y; x, q) = \ell_\theta(y \mid x, q) - \lambda \ell_T(y \mid q) + \beta \,\mathrm{Ver}(\tau, x), \qquad y = (a, \tau). \tag{8}$$

The first term rewards scene-conditioned likelihood. The second term penalizes chains that are already likely under the text-only rationale prior. The third term rewards locally verified evidence use. The coefficients $\lambda \geq 0$ and $\beta \geq 0$ control the prior penalty and verifier reward.

During inference, the model first samples or beam-searches a candidate set $\mathcal{C}(x, q) = \{y_k = (a_k, \tau_k)\}_{k=1}^{K}$ from $p_\theta(y \mid x, q)$. AR-CoT then selects

$$y^\star = \arg \max_{y_k \in \mathcal{C}(x,q)} G_\theta(y_k; x, q). \tag{9}$$

This decoding rule favors chains that are both scene-conditioned and locally grounded, while suppressing chains whose plausibility comes primarily from the question.

4.5 Decoy-Twin Contrast

To enforce scene dependence, AR-CoT uses decoy twin pairs. A twin pair contains two scenes x^+ and x^- paired with the same question q. The two scenes preserve the same language prior but differ in the spatial evidence required by the question. Let $y^+ = (a^+, \tau^+)$ be the answer-chain pair supported by x^+, and $y^- = (a^-, \tau^-)$ be the pair supported by x^-. We impose the contrastive loss

$$\begin{aligned}\mathcal{L}_{\text{twin}} = &\left[m - G_\theta(y^+; x^+, q) + G_\theta(y^-; x^+, q)\right]_+ \\ &+ \left[m - G_\theta(y^-; x^-, q) + G_\theta(y^+; x^-, q)\right]_+ ,\end{aligned} \tag{10}$$

where $[z]_+ = \max(z, 0)$, and $m > 0$ is the margin. The first term requires x^+ to favor its own scene-supported chain over the decoy chain. The second term imposes the symmetric constraint for x^-. Appendix A.4 describes how twin pairs are constructed from relation-controlled scene pairs.

4.6 Training Objective

For each training instance, we use a target answer-chain pair $y^\star = (a^\star, \tau^\star)$. The chain $\tau^\star$ is constructed from annotated object and relation evidence using deterministic templates, as described in Appendix A.5. We also draw a text-only rationale $\tilde{y}_T \sim p_T(y \mid q)$ as an anti-rationalization negative. The training objective is

$$\begin{aligned}\mathcal{L}_{\text{sup}} &= -\log p_\theta(y^\star \mid x, q), \\ \mathcal{L}_{\text{gain}} &= \left[m_g - G_\theta(y^\star; x, q) + G_\theta(\tilde{y}_T; x, q)\right]_+ , \\ \mathcal{L} &= \mathcal{L}_{\text{sup}} + \gamma\mathcal{L}_{\text{gain}} + \delta\mathcal{L}_{\text{twin}} + \eta\mathcal{L}_{\text{ver}}.\end{aligned} \tag{11}$$

Here, $\mathcal{L}_{\text{sup}}$ supervises the target answer and chain, $\mathcal{L}_{\text{gain}}$ separates the scene-supported chain from a text-only rationale, $\mathcal{L}_{\text{twin}}$ enforces counterfactual scene dependence, and $\mathcal{L}_{\text{ver}}$ calibrates the local verifier. The coefficients $\gamma, \delta, \eta \geq 0$ balance the three auxiliary objectives. The definition of $\mathcal{L}_{\text{ver}}$ is given in Appendix A.3.

5 Experiments

5.1 Experimental Setup

Benchmarks and metrics. We evaluate AR-CoT on standard 3D question answering, grounding, and faithful reasoning benchmarks. For QA, we use ScanQA [Azuma et al., 2022] and SQA3D [Ma et al., 2023]; for grounding, we use ScanRefer [Chen et al., 2020] and Multi3DRefer [Zhang et al., 2023b]; for faithful reasoning and grounding–QA coherence, we use MSQA from recent grounded-CoT evaluation [Linghu et al., 2026] and Beacon3D [Huang et al., 2025a]. We report standard QA and grounding metrics, together with macro-averaged ground-truth chain grounding accuracy (CGA), decoy-twin contrast accuracy (DTCA), scene-swap chain sensitivity (CSS), and random-scene invariance rate (RSI). Additional benchmarks, full metric definitions, and QA-driven grounding scores are provided in Appendix B.2 and Appendix C.1.

Comparison protocol. We use two comparison protocols. For controlled same-backbone evaluation, all reasoning baselines are built on LLaVA-3D-7B with the same 3D encoder, object proposals, language decoder, training split, candidate budget, and generation length [Fan et al., 2026b]. Baselines include answer-only decoding, vanilla CoT [Wei et al., 2022], self-consistency [Wang et al., 2023], Multimodal-CoT-style rationale-then-answer prompting [Zhang et al., 2024b], answer-level prior correction, verifier-only reranking, and grounded CoT without text-prior contrast. We also report compatible published 3D vision–language and grounded reasoning systems as reference comparisons, since they differ in backbone, training data, input representation, and tool access.

Backbones and implementation. AR-CoT is implemented as a plug-in training and decoding layer on top of existing 3D MLLMs. Unless otherwise specified, controlled experiments use LLaVA-3D-7B as the base model. To test architectural generality, we additionally instantiate AR-CoT on Video-3D LLM-7B and LLaVA-NeXT-Video-7B. The text-only prior is initialized from the same language decoder as the scene-conditioned model and kept frozen. At inference time, AR-CoT constructs evidence graphs only from model-accessible scene inputs; ground-truth test annotations are not used for generation, reranking, verifier scoring, or gain-based selection. Detailed preprocessing, stress-test construction, optimization, decoding, leakage control, and ablations are described in Appendix B.2.

Table 1: **Main results on standard 3D QA and grounding benchmarks.** All metrics are higher better. Best and second-best results are in bold and underlined, respectively. Pink shading marks AR-CoT.

Method	Backbone (LLM)	ScanQA		SQA3D	ScanRefer		Multi3DRefer	
		CIDEr	EM	EM	Acc@0.25	Acc@0.5	F1@0.25	F1@0.5
ScanQA [Azuma et al., 2022]	Non-LLM	64.9	21.1	47.2	–	–	–	–
3D-VisTA [Zhu et al., 2023]	Non-LLM	69.6	22.4	48.5	50.6	45.8	41.5	36.8
3D-LLM [Hong et al., 2023]	Flan-T5	69.4	20.5	45.9	30.3	24.7	32.7	28.0
LL3DA [Chen et al., 2024a]	OPT-1.3B / LLaMA2-7B	76.8	23.2	50.5	46.4	39.9	42.0	37.5
LEO [Huang et al., 2024b]	Vicuna-7B	101.4	24.5	50.0	53.7	47.8	52.6	48.2
Chat-Scene [Huang et al., 2024a]	Vicuna-7B	87.7	21.6	54.6	55.5	50.2	57.1	52.4
LSceneLLM [Zhi et al., 2025]	LLaMA2-7B	88.2	25.9	55.2	56.7	50.8	56.5	51.8
LLaVA-3D [Zhu et al., 2025]	Vicuna-7B	91.7	27.0	55.6	54.1	42.4	54.3	47.2
Vid-LLM [Chen et al., 2026a]	LLaVA-based 7B	101.9	27.6	57.3	50.1	46.7	47.2	42.9
Video-3D LLM [Zheng et al., 2025]	Qwen2-7B	102.1	30.1	58.6	58.1	51.7	58.0	52.7
3DRS [Huang et al., 2025b]	LLaVA-Next-Video-7B	104.8	30.3	60.6	62.9	56.1	60.4	54.9
AR-CoT	LLaVA-3D-7B (Vicuna-7B)	110.7	32.3	62.0	64.8	57.9	62.7	57.2
AR-CoT	Video-3D LLM-7B (Qwen2-7B)	<u>113.8</u>	<u>33.3</u>	<u>62.9</u>	<u>66.0</u>	<u>59.5</u>	<u>63.6</u>	<u>58.7</u>
AR-CoT	3DRS-7B (LLaVA-Next-Video-7B)	**116.5**	**34.0**	**64.3**	**66.7**	**60.4**	**65.1**	**59.7**

Table 2: **Main results on faithful reasoning and shortcut-sensitive evaluation.** We report MSQA overall accuracy, Beacon3D case-level accuracy, object-level accuracy, and Good Coherence (GC), together with AR-CoT stress-test metrics: macro-averaged ground-truth chain grounding accuracy (CGA), decoy-twin contrast accuracy (DTCA), scene-swap chain sensitivity (CSS), and random-scene invariance rate (RSI). All metrics are higher better except RSI, where lower is better. Best and second-best results are in bold and underlined, respectively. Pink shading marks AR-CoT.

Method	Backbone (LLM)	MSQA Overall	Beacon3D			CGA	DTCA	CSS	RSI↓
			Case Acc.	Obj. Acc.	GC				
GPT-4o*	GPT-4o	52.3	57.1	20.2	30.8	43.7	56.0	48.9	35.5
LEO [Huang et al., 2024b]	Vicuna-7B	54.8	43.2	7.8	1.6	31.4	48.6	38.1	45.9
Chat-Scene [Huang et al., 2024a]	Vicuna-7B	56.6	53.6	14.0	19.5	41.6	55.9	47.2	38.8
SceneCOT [Linghu et al., 2026]	LLaVA-1.5 / Vicuna-7B	55.6	58.9	23.2	34.7	57.7	64.1	59.3	25.8
MAG-3D [Zheng et al., 2026a]	GPT-4o	53.0	63.5	23.4	37.9	58.4	66.7	61.5	24.1
MAG-3D [Zheng et al., 2026a]	Seed-1.6	54.3	65.0	27.5	39.7	59.2	67.9	62.6	23.3
Answer-only 3D MLLM	LLaVA-3D-7B	52.8	57.0	20.7	29.4	38.6	52.3	44.0	41.7
Vanilla CoT	LLaVA-3D-7B	54.1	58.2	21.0	28.3	44.2	55.7	46.6	39.5
Self-consistency	LLaVA-3D-7B	55.0	59.1	21.6	29.7	45.0	56.8	47.8	38.2
Multimodal-CoT	LLaVA-3D-7B	55.4	59.7	22.4	31.1	47.5	58.3	50.0	35.9
Answer-level prior correction	LLaVA-3D-7B	55.6	60.0	22.8	32.6	48.0	60.5	52.4	31.3
Verifier-only reranking	LLaVA-3D-7B	56.8	61.3	24.7	36.2	55.1	63.8	57.2	28.4
Grounded CoT w/o text prior	LLaVA-3D-7B	57.6	62.2	26.0	38.4	59.8	66.0	61.5	24.2
AR-CoT	LLaVA-3D-7B	58.4	67.4	31.0	43.8	63.8	71.5	68.7	18.4
AR-CoT	Video-3D LLM-7B	<u>59.1</u>	<u>68.0</u>	<u>31.8</u>	<u>44.6</u>	<u>65.2</u>	<u>73.1</u>	<u>70.0</u>	<u>16.7</u>
AR-CoT	3DRS + LLaVA-Next-Video-7B	**60.4**	**69.3**	**33.2**	**46.1**	**66.7**	**75.6**	**71.8**	**14.9**

5.2 Main Results

We compare AR-CoT with representative 3D vision–language models, grounded 3D reasoning systems, and controlled same-backbone reasoning baselines. As shown in Table 1, AR-CoT consistently improves over both task-specific 3D specialists and recent 3D MLLM generalists on standard QA and grounding benchmarks. The improvement is especially clear on ScanQA and SQA3D, where AR-CoT benefits from selecting answer-chain pairs that are supported by scene evidence rather than only maximizing answer likelihood. The gains on ScanRefer and Multi3DRefer further indicate that the proposed evidence-aware reasoning objective improves object- and relation-sensitive grounding, not only final answer generation.

Table 2 shows that the advantage of AR-CoT is larger under faithful reasoning and shortcut-sensitive evaluation. Compared with vanilla CoT, self-consistency, and verifier-only reranking, AR-CoT achieves stronger chain grounding, higher decoy-twin contrast accuracy, and lower random-scene invariance. This suggests that text-prior contrast and local evidence verification are complementary: the former suppresses language-only rationalization, while the latter enforces step-level consistency

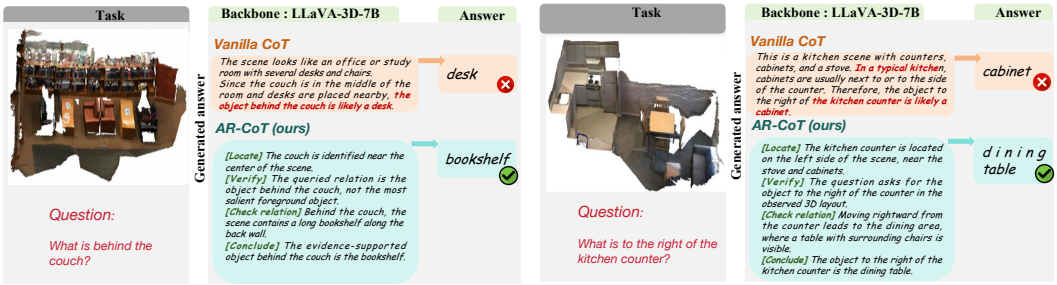


Figure 4: **Qualitative comparisons on ScanQA.**

Table 3: **Ablation study on LLaVA-3D-7B.** All results are reported as mean and standard deviation over three random seeds. RSI is lower better; all other metrics are higher better. Best and second-best means are in bold and underlined, respectively; ties share the same mark. Pink shading marks Full AR-CoT.

Variant	ScanQA EM	SQA3D EM	MSQA	Beacon3D GC	CGA	DTCA	CSS	RSI↓
Full AR-CoT	32.3±0.2	62.0±0.4	58.4±0.7	43.8±0.4	63.8±0.8	71.5±0.1	68.7±0.8	18.4±0.4
w/o Text Prior	31.1±0.9	60.8±1.4	57.2±0.8	39.2±1.5	60.7±0.8	66.4±0.9	61.8±0.3	25.1±0.7
w/o Verifier	30.6±0.4	60.1±0.8	56.9±1.1	35.8±1.6	49.4±0.8	63.1±0.4	56.6±0.9	23.7±1.2
w/o Decoy-Twin Loss	31.4±0.4	60.9±0.4	57.4±0.3	39.5±0.7	61.9±0.3	66.5±0.1	63.4±0.6	22.0±0.3
Answer-Gain Only	30.0±0.2	59.3±0.5	55.6±0.4	32.6±0.6	48.0±0.1	60.5±0.2	52.4±0.7	31.3±0.1
Verifier Only	30.8±0.2	60.2±0.5	56.8±0.6	36.2±0.2	55.1±0.8	63.8±0.5	57.2±0.7	28.4±0.4
Free-Form CoT	29.1±0.3	58.8±0.4	54.5±0.6	30.0±0.3	42.7±0.5	57.0±0.6	48.8±0.9	37.6±0.4
GT Evidence	**34.1±0.2**	**63.8±0.6**	**59.7±0.3**	**49.0±0.3**	**75.6±0.5**	**78.4±0.7**	**73.2±0.7**	**14.2±0.2**

with the observed 3D scene. Across all three backbones, AR-CoT remains the strongest variant, showing that the proposed framework is not tied to a specific 3D encoder or language decoder.

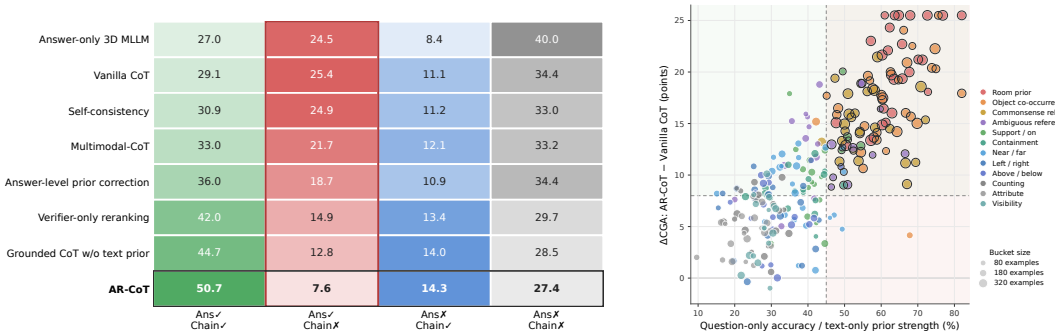
5.3 Ablation and Analysis

Single-factor ablation. We report mean ± standard deviation over three random seeds. Table 3 evaluates the variants defined in Appendix B.2. All variants use the same LLaVA-3D-7B backbone, candidate budget, and maximum chain length.

The ablation results show that the text-only prior, local verifier, and decoy-twin loss contribute complementary improvements. Removing the text-only prior mainly increases RSI and reduces DTCA, indicating stronger language-prior rationalization. Removing the verifier causes the largest drop in CGA and Beacon3D GC, showing that local evidence checking is necessary for grounding–QA coherence. The gap between GT Evidence and Full AR-CoT further quantifies the impact of evidence quality.

Answer–chain consistency. We next ask whether AR-CoT improves only final answers or also the evidence chains that support them. We partition each prediction into four categories according to answer correctness and evidence-chain correctness, and visualize the row-normalized distribution for each method; details are given in Appendix B.3, Answer–chain consistency details. Figure 5a shows that Vanilla CoT leaves many predictions in the answer-correct but evidence-incorrect region. AR-CoT shifts substantial mass toward the answer-correct and evidence-correct quadrant, while reducing correct-but-unsupported reasoning. This indicates that AR-CoT is not merely recalibrating final answers; it converts superficially correct yet weakly grounded predictions into jointly correct and scene-supported reasoning.

Language-prior regimes. Finally, we examine where AR-CoT is most useful. We group questions by relation family [Wang et al., 2026a], estimate each bucket's language-prior strength using question-only accuracy, and measure the chain-grounding gain of AR-CoT over Vanilla CoT; details are given in Appendix B.3, Language-prior regime details. As shown in Figure 5b, the largest gains concentrate in high-prior buckets, especially room-prior, object co-occurrence, and commonsense-relation questions. These are precisely the regimes where a model can produce plausible rationales



(a) **Answer–chain consistency.** Row-normalized distribution over answer/evidence correctness categories. AR-CoT reduces correct-but-unsupported predictions and increases jointly correct-and-grounded ones.

(b) **Language-prior regimes.** The x-axis shows question-only accuracy, and the y-axis shows AR-CoT's chain-grounding gain over Vanilla CoT.

Figure 5: **Diagnostic analyses of evidence-faithful reasoning.** Left: AR-CoT improves answer–chain consistency by reducing correct-but-unsupported predictions. Right: AR-CoT yields larger chain-grounding gains in high language-prior regimes. Details are in Appendix B.3, Answer–chain consistency details and Appendix B.3, Language-prior regime details.

without strongly using the 3D scene. The result supports the anti-rationalization motivation: AR-CoT is most beneficial when language priors provide a tempting substitute for scene evidence.

5.4 Qualitative Case Studies

Figure 4 presents two qualitative examples from ScanQA. Vanilla CoT tends to produce plausible but weakly grounded rationales, either by focusing on salient foreground objects or by relying on common indoor-scene priors. For example, it incorrectly answers *desk* for "What is behind the couch?" and *cabinet* for "What is to the right of the kitchen counter?" In contrast, AR-CoT explicitly locates the queried object, verifies the relevant spatial relation in the observed 3D scene, and grounds the answers to the *bookshelf* and *dining table*, respectively. These cases show that AR-CoT reduces language-prior rationalization and produces more faithful scene-evidence-guided reasoning.

6 Conclusion

This paper addressed rationalization in 3D multimodal reasoning, where plausible chains may not faithfully reflect the observed scene. We proposed AR-CoT, a plug-in framework that grounds reasoning steps in scene evidence, contrasts them with a text-only prior, and selects predictions by scene-evidence gain. Experiments show that AR-CoT improves standard QA/grounding performance while strengthening faithful-reasoning metrics such as chain grounding, decoy-twin contrast, and scene-swap sensitivity. These results highlight evidence-faithful reasoning as an important direction for building more reliable 3D MLLMs in interactive and embodied settings.

Acknowledgments and Disclosure of Funding

This work was supported by the State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology under Project No. BI202603. The authors declare no competing interests.

References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European conference on computer vision*, pages 422–440. Springer, 2020.

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018.

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*, pages 19129–19139, 2022.

ChiaHua Chang, M. Fu, X. Chen, S. Feng, M. Zhang, X. Zhou, J. J. Chuang, J. Pan, and Z. Liu. Research on PDU-Net lung nodule segmentation algorithm based on path aggregation and dual attention. In *2025 4th International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, Chongqing, China, 2025. IEEE. ISBN 979-8-3315-6581-7. doi: 10.1109/ICICML67980.2025.11333821. URL https://doi.org/10.1109/ICICML67980.2025.11333821.

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*, pages 202–221. Springer, 2020.

Haijier Chen, Bo Xu, Shoujian Zhang, Haoze Liu, Jiaxuan Lin, and Jingrong Wang. Vid-LLM: A compact video-based 3D multimodal LLM with reconstruction–reasoning synergy. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=l1cLdEjESj.

Qianyu Chen, Canran Xiao, and Runxuan Tang. Hidden forgetting in continual multimodal learning: When accuracy survives but grounding fails. In *Advances in Neural Information Processing Systems*, 2026b. URL https://openreview.net/forum?id=MDf0kUD4oe.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26428–26438, 2024a.

Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024b.

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021.

Kairui Ding, Boyuan Chen, Yuchen Su, Huan-ang Gao, Bu Jin, Chonghao Sima, Xiaohui Li, Wuqiang Zhang, Paul Barsch, Hongyang Li, and Hao Zhao. Hint-AD: Holistically aligned interpretability in end-to-end autonomous driving. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pages 3742–3765. PMLR, 2025. URL https://proceedings.mlr.press/v270/ding25a.html.

Enguang Fan, Anfeng Peng, Matthew Caesar, Jae Kim, Josh Eckhardt, Greg Kimberly, and Denis Osipychev. Towards effective swarm-based GPS spoofing detection in disadvantaged platforms. In *Proceedings of the 2023 IEEE Military Communications Conference (MILCOM)*, pages 722–728. IEEE, 2023. doi: 10.1109/MILCOM58377.2023.10356314. URL https://doi.org/10.1109/MILCOM58377.2023.10356314.

Enguang Fan, Yifan Chen, Zihan Shan, Klara Nahrstedt, Matthew Caesar, and Jae Kim. Communication-aware multi-agent reinforcement learning for decentralized cooperative UAV deployment. *arXiv preprint arXiv:2603.16141*, 2026a. URL https://arxiv.org/abs/2603.16141.

Zhichao Fan, Yanhang Li, Zexin Zhuang, Xian Sun, and Yingshuo Wang. The moving target: A longitudinal audit of trust-benchmark score drift across open-source chat LLM release lines, 2026b. URL https://arxiv.org/abs/2607.02587. Accepted at IEEE ICMLA 2026.

Shixing Guan, Xuan Hua, Jiahao Lin, and Luyan Zhang. Internalizing decomposed reasoning: Reverse-distilled fine-tuning for interpretable and robust Text-to-SQL. In De-Shuang Huang, Chuanlei Zhang, Wei Chen, Bo Li, Qinhu Zhang, Wenzheng Bao, Yijie Pan, and Prashan Premaratne, editors, *Advanced Intelligent Computing Technology and Applications*, volume 16655 of *Lecture Notes in Computer Science*, pages 534–544, Singapore, 2027. Springer Nature Singapore. ISBN 978-981-92-3438-7. doi: 10.1007/978-981-92-3438-7_45. URL https://link.springer.com/chapter/10.1007/978-981-92-3438-7_45.

Zixin Guo, Kai Zhao, and Luyan Zhang. InstanceRSR: Real-world super-resolution via instance-aware representation alignment. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10577–10581, 2026. doi: 10.1109/ICASSP55912.2026.11462690. URL https://ieeexplore.ieee.org/document/11462690.

Jiani He, Dingyan Shang, Yihua Xu, Shiqi Huang, Yan Lyu, Jize Li, and Shangjing Tang. Semantic signal-assisted inspection and recovery allocation in reverse logistics. *arXiv preprint arXiv:2609.02116*, 2026. URL https://arxiv.org/abs/2609.02116.

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. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023.

Saisai Hu. Research on security enhancement methods for adversarial robust large language model intelligent agents for medical decision-making tasks. In *2026 7th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 599–603. IEEE, 2026.

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024a.

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 *Proceedings of the 41st International Conference on Machine Learning*, pages 20413–20451, 2024b.

Jiangyong Huang, Baoxiong Jia, Yan Wang, Ziyu Zhu, Xiongkun Linghu, Qing Li, Song-Chun Zhu, and Siyuan Huang. Unveiling the mist over 3d vision-language understanding: Object-centric evaluation with chain-of-analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24570–24581, 2025a.

Shiqi Huang, Jiani He, Dingyan Shang, Yihua Xu, Jize Li, Yan Lyu, and Lakshmi Muraleedharan Nair. DACRI: Decision-aware causal intervention ranking for critical supply chains. *arXiv preprint arXiv:2608.11154*, 2026. URL https://arxiv.org/abs/2608.11154.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3DRS: MLLMs need 3D-aware representation supervision for scene understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-2282. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/62133b95efaa6f10a5395eb10ca087cd-Abstract-Conference.html.

Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. In *European Conference on Computer Vision*, pages 289–310. Springer, 2024.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023.

Baichuan Li, Junyi Yao, and Zihao Zheng. Remember, verify, or ask? evaluating memory decisions at reuse time. In *Proceedings of the 3rd Workshop on KNOwledge GRaphs Agentic Systems Interplay (NORA 2026)*, pages 15–25, Hengqin, China, November 2026a. Association for Computational Linguistics. URL https://aclanthology.org/2026.nora-1.2/.

Jize Li, Jiani He, Dishu Yang, Dingyan Shang, Jingjing Liu, and Shiqi Huang. Accuracy-preserving stability regularization for large-scale retail demand forecasting. *arXiv preprint arXiv:2607.13331*, 2026b. URL https://arxiv.org/abs/2607.13331.

Pengfei Li, Beiwen Tian, Yongliang Shi, Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. TOIST: Task oriented instance segmentation transformer with noun-pronoun distillation. In *Advances in Neural Information Processing Systems*, volume 35, pages 17597–17611, 2022. doi: 10.52202/068431-1279. URL https://papers.neurips.cc/paper_files/paper/2022/hash/70270a1bc28ecb2a2aefad566c5e556b-Abstract-Conference.html.

Wei Li, Bing Hu, Rui Shao, Leyang Shen, and Liqiang Nie. LION-FS: Fast & slow video-language thinker as online video assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3240–3251. IEEE, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_LION-FS_Fast__Slow_Video-Language_Thinker_as_Online_Video_Assistant_CVPR_2025_paper.html.

Wei Li, Jizhihui Liu, Li Yixing, Junwen Tong, Rui Shao, and Liqiang Nie. ConsisVLA-4D: Advancing spatiotemporal consistency in efficient 3D-perception and 4D-reasoning for robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6706–6717, June 2026c. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_ConsisVLA-4D_Advancing_Spatiotemporal_Consistency_in_Efficient_3D-Perception_and_4D-Reasoning_for_CVPR_2026_paper.html.

Wei Li, Rui Shao, Jie He, Lingsen Zhang, Ziwei Liu, and Liqiang Nie. UniMPA: A unified memory-prediction-action model via action-grounded transition modeling. *arXiv preprint arXiv:2609.11875*, 2026d. URL https://arxiv.org/abs/2609.11875.

Yanshu Li, Jiaqian Li, Canran Xiao, Xi Xiao, Tianyang Wang, and Yongtai Liu. MCD: Causal distillation of multimodal in-context learning in large vision-language models. *arXiv preprint arXiv:2609.39920*, 2026e. URL https://arxiv.org/abs/2609.39920.

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

Jingjing Liu, Dishu Yang, Hongbo Qin, and Hansong Zhang. Passing tests, missing semantics: Evidence of interpretive lock-in in plan-based DataFrame code generation. *arXiv preprint arXiv:2610.15797*, 2026. URL https://arxiv.org/abs/2610.15797.

Jianglin Lu, Hailing Wang, Yi Xu, Yizhou Wang, Kuo Yang, and Yun Fu. Representation potentials of foundation models for multimodal alignment: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16669–16684. Association for Computational Linguistics, November 2025a. doi: 10.18653/v1/2025.emnlp-main.843. URL https://aclanthology.org/2025.emnlp-main.843/.

Jianglin Lu, Hailing Wang, Kuo Yang, Yitian Zhang, Simon Jenni, and Yun Fu. The indra representation hypothesis for multimodal alignment. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-5245. URL https://openreview.net/forum?id=D2NR5Zq6PG.

Jianglin Lu, Hailing Wang, Xu Ma, Qihua Dong, Mingyuan Zhang, Yizhou Wang, and Yun Fu. MUSE: A unified agentic harness for MLLMs. *arXiv preprint arXiv:2606.03005*, 2026a. URL https://arxiv.org/abs/2606.03005.

Yao Lu, Tengfei Ma, Zeyu Wang, Zhuangzhi Chen, Dongwei Xu, Yun Lin, Qi Xuan, and Guan Gui. FCOS: A two-stage recoverable model pruning framework for automatic modulation recognition. *IEEE Transactions on Cognitive Communications and Networking*, 2026b. URL https://arxiv.org/abs/2505.21571.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, 2023.

Ruiyuan Lyu, Jingli Lin, Tai Wang, Shuai Yang, Xiaohan Mao, Yilun Chen, Runsen Xu, Haifeng Huang, Chenming Zhu, Dahua Lin, et al. Mmscan: A multi-modal 3d scene dataset with hierarchical grounded language annotations. *Advances in Neural Information Processing Systems*, 37: 50898–50924, 2024.

Mingjun Ma, Saisai Hu, Tiantian Zhu, Yitian Lin, Jiachen Sun, Zhengqiu Weng, Dehao Li, and Yajie Zhang. MuSK: Multi-scale knowledge learning for provenance-graph anomaly detection. *Computer Networks*, 289:112728, 2026. doi: 10.1016/j.comnet.2026.112728. URL https://doi.org/10.1016/j.comnet.2026.112728.

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

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *The Eleventh International Conference on Learning Representations*, 2023.

Pengcheng Pan, Guoqing Ma, Yuhan Zhang, Yang Chen, Yichen Liu, Ziheng Li, and Shan Yu. MHWA: Multi-timescale hierarchical world-action model, 2026. URL https://arxiv.org/abs/2610.14727.

Xiaochao Pan, Jiawei Yao, Hongrui Kou, Tong Wu, and Canran Xiao. HarmonicNeRF: Geometry-informed synthetic view augmentation for 3D scene reconstruction in driving scenarios. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. doi: 10.1145/3664647.3680851. URL https://doi.org/10.1145/3664647.3680851.

Wanru Shao. Interpretable ensemble learning for network traffic anomaly detection: A SHAP-based explainable AI framework for embedded systems security. In *2026 14th International Symposium on Digital Forensics and Security (ISDFS)*, pages 1–4. IEEE, 2026a. doi: 10.1109/ISDFS69419.2026.11459069. URL https://doi.org/10.1109/ISDFS69419.2026.11459069.

Wanru Shao. Policy-driven vulnerability risk quantification framework for large-scale cloud infrastructure data security. In *Proceedings of the 2026 5th International Conference on Cyber Security, Artificial Intelligence and Digital Economy*, pages 40–45, New York, NY, USA, 2026b. Association for Computing Machinery. doi: 10.1145/3819836.3819842. URL https://doi.org/10.1145/3819836.3819842.

Xian Sun, Wei Gao, Yingshuo Wang, Lingdong Kong, Yanhang Li, Zhichao Fan, Zexin Zhuang, Wenlong Dong, Zhiyuan Zheng, Hrishikesh Paranjape, Abhishek Mandal, and Johnny R. Zhang. Beyond accuracy: Measuring bias acknowledgment in chain-of-thought reasoning for responsible AI evaluation, 2026. URL https://arxiv.org/abs/2606.15127. ICML 2026 Workshop on Trustworthy AI for Good.

Beiwen Tian, Liyi Luo, Hao Zhao, and Guyue Zhou. VIBUS: Data-efficient 3D scene parsing with viewpoint bottleneck and uncertainty-spectrum modeling. *ISPRS Journal of Photogrammetry and Remote Sensing*, 194:302–318, 2022. doi: 10.1016/j.isprsjprs.2022.10.013. URL https://www.sciencedirect.com/science/article/abs/pii/S0924271622002817.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

Yingshuo Wang, Xian Sun, Lingdong Kong, Wei Gao, Yanhang Li, Zhichao Fan, and Zexin Zhuang. Do time series foundation model benchmarks hide regime-dependent failures? evidence from traffic speed forecasting, 2026a. URL https://arxiv.org/abs/2606.18367. ICML 2026 Workshop on Forecasting as a New Frontier of Intelligence.

Zijun Wang, Saisai Hu, Tiantian Zhu, Wenrui Cheng, Zhengqiu Weng, Haiting Chen, Xiangyang Zheng, and Suyu Zhang. Trident: Dual-stream APT attribution over heterogeneous threat knowledge graphs. *Computers & Security*, 171:105094, 2026b. doi: 10.1016/j.cose.2026.105094. URL https://doi.org/10.1016/j.cose.2026.105094.

Zixuan Wang, Yu Sun, Hongwei Wang, Baoyu Jing, Xiang Shen, Xin Dong, Zhuolin Hao, Hongyu Xiong, and Yang Song. Reasoning-enhanced domain-adaptive pretraining of multimodal large language models for short video content governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1104–1112. Association for Computational Linguistics, 2025. ISBN 979-8-89176-333-3. doi: 10.18653/v1/2025.emnlp-industry.77. URL https://aclanthology.org/2025.emnlp-industry.77/.

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

Yilan Wei, Kumiko Long, Arielle Granston, and Adrian Rodriguez-Contreras. USVexplorer: Robust detection of ultrasonic vocalizations with cross species generalization. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 15232–15236. IEEE, 2026. doi: 10.1109/ICASSP55912.2026.11463476. URL https://doi.org/10.1109/ICASSP55912.2026.11463476.

Linqi Xiao, Venkateswarlu Gudepu, and Andrea Fumagalli. Joint alarm correlation and root alarm detection in optical transport networks with graph-attention. In *2025 IEEE 30th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD)*, pages 1–6. IEEE, 2025.

Linqi Xiao, Aparaajitha Gomathinayakam Latha, Venkateswarlu Gudepu, and Andrea Fumagalli. CORA: A conversational AI-agent for intelligent automation of OpenROADM multi-vendor optical transport networks. In *Applications of Machine Learning 2026*, volume 14218 of *Proceedings of SPIE*, page 142180Y. SPIE, 2026a. doi: 10.1117/12.3113861. URL https://doi.org/10.1117/12.3113861.

Linqi Xiao, Aparaajitha Gomathinayakam Latha, Venkateswarlu Gudepu, and Andrea Fumagalli. A YANG-grounded LLM agent supporting multi-vendor OpenROADM optical transport networks. In *Proceedings of the ACM SIGCOMM 2026 Conference*, pages 2109–2115, 2026b. doi: 10.1145/3789240.3822606. URL https://doi.org/10.1145/3789240.3822606.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pages 131–147. Springer, 2024.

Dishu Yang, Jingjing Liu, and Jize Li. Balancing privacy, utility, and safety in LLM alignment through preference optimization. *arXiv preprint arXiv:2608.30141*, 2026a. URL https://arxiv.org/abs/2608.30141.

Dishu Yang, Qi Su, Hongbo Qin, and Hansong Zhang. Measurement-first auditing of agentic leaderboards: Contamination susceptibility, matched-control re-evaluation, and scorer validation. *arXiv preprint arXiv:2610.05830*, 2026b. URL https://arxiv.org/abs/2610.05830.

Jianing Yang, Xuweiyi Chen, Nikhil Madaan, Madhavan Iyengar, Shengyi Qian, David F Fouhey, and Joyce Chai. 3d-grand: A million-scale dataset for 3d-llms with better grounding and less hallucination. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29501–29512, 2025.

Jiawei Yao, Jusheng Zhang, Xiaochao Pan, Tong Wu, and Canran Xiao. DepthSSC: Monocular 3D semantic scene completion via depth-spatial alignment and voxel adaptation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 2154–2163, February 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Yao_DepthSSC_Monocular_3D_Semantic_Scene_Completion_via_Depth-Spatial_Alignment_and_WACV_2025_paper.html.

Junyi Yao, Baichuan Li, Zihao Zheng, and Jiayu Long. Toward workflow-aware benchmarking for healthcare NLP agents, 2026. URL https://arxiv.org/abs/2609.00296.

Hua Ye, Siyuan Chen, Ziqi Zhong, Canran Xiao, Haoliang Zhang, Yuhan Wu, and Fei Shen. Seeing through the conflict: Transparent knowledge conflict handling in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 34423–34431, 2026. doi: 10.1609/aaai.v40i40.40740. URL https://ojs.aaai.org/index.php/AAAI/article/view/40740.

Chengzhen Yu, Canran Xiao, Siyuan Ma, and Yang Liu. Text as partial constraint: Core–residual alignment for robust vision–language learning. In *Advances in Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=xc5RAI0YCe.

Fan Zhang, Gongguan Chen, Hua Wang, Jinjiang Li, and Caiming Zhang. Multi-scale video super-resolution transformer with polynomial approximation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4496–4506, 2023a. doi: 10.1109/TCSVT.2023.3278131. URL https://ieeexplore.ieee.org/document/10129942.

Fan Zhang, Gongguan Chen, Hua Wang, and Caiming Zhang. CF-DAN: Facial-expression recognition based on cross-fusion dual-attention network. *Computational Visual Media*, 10(3):593–608, 2024a. doi: 10.1007/s41095-023-0369-x. URL https://link.springer.com/article/10.1007/s41095-023-0369-x.

Fan Zhang, Shiming Fan, and Hua Wang. Rethinking activation function: A simple path to efficient and accurate time series forecasting. *IEEE Transactions on Knowledge and Data Engineering*, 2026a. doi: 10.1109/TKDE.2026.3716656. URL https://ieeexplore.ieee.org/document/11622841.

Fan Zhang, Shiming Fan, and Hua Wang. TimeSAF: Towards LLM-guided semantic asynchronous fusion for time series forecasting. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26267–26284. Association for Computational Linguistics, July 2026b. doi: 10.18653/v1/2026.acl-long.1208. URL https://aclanthology.org/2026.acl-long.1208/.

Fan Zhang, Zhiwei Gu, and Hua Wang. Decoding with structured awareness: Integrating directional, frequency-spatial, and structural attention for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 12421–12429, 2026c. doi: 10.1609/aaai.v40i15.38235. URL https://ojs.aaai.org/index.php/AAAI/article/view/38235.

Hengrui Zhang, Yuhu Cheng, C. L. Philip Chen, and Xuesong Wang. Diffusion subgoal planning for long-horizon offline goal-conditioned reinforcement learning. *arXiv preprint arXiv:2609.34575*, 2026d. URL https://arxiv.org/abs/2609.34575.

Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15225–15236, 2023b.

Yue Zhang, Zhiyang Xu, Ying Shen, Parisa Kordjamshidi, and Lifu Huang. Spartun3d: Situated spatial understanding of 3d world in large language model. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024, 2024b.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8995–9006, 2025.

Henry Zheng, Chenyue Fang, Rui Huang, Siyuan Wei, Xiao Liu, and Gao Huang. Mag-3d: Multi-agent grounded reasoning for 3d understanding. *arXiv preprint arXiv:2604.09167*, 2026a.

Zihao Zheng, Junyi Yao, Baichuan Li, and Jiayu Long. Auditing escalation reliability in healthcare LLM agents: A perturbation-controlled benchmark of workflow-state failures, 2026b. URL `https://arxiv.org/abs/2610.21270`.

Hongyan Zhi, Peihao Chen, Junyan Li, Shuailei Ma, Xinyu Sun, Tianhang Xiang, Yinjie Lei, Mingkui Tan, and Chuang Gan. Lscenellm: Enhancing large 3d scene understanding using adaptive visual preferences. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3761–3771, 2025.

Heng Zhou, Jing Tang, Jusheng Zhang, Yanshu Li, Canran Xiao, Liwei Hou, Zong Ke, and Jiawei Yao. CoMem: Compositional concept-graph memory for vision–language adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://iclr.cc/virtual/2026/poster/10006619`.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4295–4305, 2025.

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023.

Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pages 188–206. Springer, 2024.

A Additional Method Details

A.1 Scene Evidence Graph Construction

AR-CoT represents a 3D scene as an object-centric evidence graph. The graph is not intended to replace the visual backbone; instead, it provides an inspectable interface through which reasoning steps can explicitly point to scene evidence. Given a raw 3D observation x, we first obtain a set of object evidence units

$$\mathcal{O}_x = \{o_i\}_{i=1}^{M}. \tag{12}$$

Depending on the backbone and dataset, o_i can be obtained from ground-truth instance segments, detector proposals, 3D segmentation masks, object tokens, or proposal-free 3D tokens clustered into object-level evidence units. Each object evidence unit is represented as

$$o_i = (\mathrm{id}_i,\ P_i,\ b_i,\ c_i,\ h_i,\ A_i,\ \mathcal{V}_i,\ \rho_i)\,, \tag{13}$$

where id_i is a unique object identifier, $P_i \subset \mathbb{R}^3$ is the associated point set or mask, b_i is its 3D bounding box, $c_i \in \Delta^{C-1}$ is a semantic category distribution, $h_i \in \mathbb{R}^{d_o}$ is the object feature, A_i is a set of predicted attributes, $\mathcal{V}_i$ is the set of supporting views if RGB-D views are available, and $\rho_i \in [0,1]$ is the object confidence.

For geometry normalization, let $\mu_i \in \mathbb{R}^3$ and $d_i \in \mathbb{R}^3_+$ denote the center and side length of b_i. Let μ_x be the scene center and s_x be the scene scale, computed as the maximum side length of the scene bounding box. We use normalized geometry

$$\tilde{\mu}_i = \frac{\mu_i - \mu_x}{s_x + \epsilon}, \qquad \tilde{d}_i = \frac{d_i}{s_x + \epsilon}, \tag{14}$$

where ϵ is a small constant for numerical stability. The object embedding used by the generator and verifier is

$$z_i = W_o \left[h_i;\ c_i;\ \tilde{\mu}_i;\ \log(\tilde{d}_i + \epsilon);\ \rho_i\right], \tag{15}$$

where $[\cdot;\cdot]$ denotes concatenation.

We then construct relation evidence units over a compact spatial-relation vocabulary $\mathcal{P}$. The vocabulary contains common unary and binary predicates needed by 3D QA and grounding tasks, including `exist`, `category`, `attribute`, `near`, `far`, `left-of`, `right-of`, `front-of`, `behind`, `above`, `below`, `inside`, `on`, `support`, `contain`, `overlap`, and `visible-from`. For each ordered object pair (o_i, o_j), we compute a relation feature

$$g_{ij} = [\tilde{\mu}_i;\ \tilde{\mu}_j;\ \tilde{\mu}_i - \tilde{\mu}_j;\ |\tilde{\mu}_i - \tilde{\mu}_j|;\ \|\tilde{\mu}_i - \tilde{\mu}_j\|_2;\ \mathrm{IoU}_{3D}(b_i, b_j);\ \mathrm{ov}_{xy}(b_i, b_j);\ \mathrm{ov}_z(b_i, b_j)]\,, \tag{16}$$

where ov_{xy} and ov_z denote horizontal and vertical overlap ratios, respectively.

For each relation $r \in \mathcal{P}$, AR-CoT stores a soft relation score

$$\rho^r_{ij} = \phi_r(g_{ij}, c_i, c_j, A_i, A_j) \in [0,1], \tag{17}$$

where ϕ_r can be implemented as a deterministic geometric rule, a learned binary classifier, or a calibrated mixture of both. For example, `near` is mainly determined by normalized center distance, `above` uses vertical displacement with horizontal overlap, and `inside` or `on` uses containment and support constraints. We retain a relation edge if its confidence exceeds a relation-specific threshold:

$$\mathcal{R}_x = \left\{(o_i, r, o_j, \rho^r_{ij}) : i \neq j,\ r \in \mathcal{P},\ \rho^r_{ij} \geq \kappa_r\right\}. \tag{18}$$

When a task requires unary evidence, such as object category or color, we represent it as an object-level predicate attached to o_i. Thus, the main text uses the simplified notation $\mathcal{G}_x = (\mathcal{O}_x, \mathcal{R}_x)$, while the implementation stores both object evidence and weighted relation evidence.

For directional predicates, we use the canonical ScanNet/world coordinate frame when it is provided. For egocentric or situated questions, we use the question-specified viewpoint if available. If the viewpoint is ambiguous, we either restrict evaluation to viewpoint-invariant relations or mark the directional predicate as unverifiable. This prevents the verifier from assigning high confidence to direction claims whose reference frame is not defined.

A.2 Evidence Pointer Parameterization

Each reasoning step in AR-CoT has the form $s_t = (u_t, e_t, m_t)$, where u_t is the operation, e_t is the evidence pointer, and m_t is the local claim. We linearize a step using explicit tags:

$$\texttt{<STEP> <OP=}u_t\texttt{> <PTR=}e_t\texttt{> <CLAIM>}m_t\texttt{</CLAIM> </STEP>}. \tag{19}$$

This format makes the chain machine-checkable while still allowing the local claim m_t to be expressed in natural language.

The step likelihood is factorized as

$$p_\theta(s_t \mid s_{<t}, q, \mathcal{G}_x) = p_\theta(u_t \mid h_t)\, p_\theta(e_t \mid u_t, h_t, \mathcal{G}_x)\, p_\theta(m_t \mid u_t, e_t, h_t, \mathcal{G}_x), \tag{20}$$

where h_t is the decoder hidden state before emitting step t. The evidence pointer can be an object pointer, a relation pointer, a set pointer, or a null pointer:

$$e_t \in \mathcal{E}_x = \mathcal{O}_x \cup \mathcal{R}_x \cup 2^{\mathcal{O}_x} \cup \{\varnothing\}. \tag{21}$$

The null pointer $\varnothing$ is only allowed for non-evidence operations such as formatting or final textual transition; evidence-bearing operations are required to point to an object, relation, or object set.

For object pointers, we use attention over object embeddings:

$$p_\theta(e_t = o_i \mid h_t, \mathcal{G}_x) = \frac{\exp(h_t^\top W_p z_i)}{\sum_{j=1}^{M} \exp(h_t^\top W_p z_j) + \exp(h_t^\top w_\varnothing)}. \tag{22}$$

For relation pointers, we first construct an edge embedding

$$z_{ij}^r = \mathrm{MLP}_r\left([z_i;\ z_j;\ a_r;\ g_{ij};\ \rho_{ij}^r]\right), \tag{23}$$

where a_r is a learned relation-type embedding. The relation pointer distribution is

$$p_\theta(e_t = (o_i, r, o_j) \mid h_t, \mathcal{G}_x) = \mathrm{softmax}_{(o_i,r,o_j)\in\mathcal{R}_x\cup\{\varnothing\}}\left(h_t^\top W_r z_{ij}^r\right). \tag{24}$$

For set-valued operations, such as `count` or `filter`, the model predicts an object-selection mask

$$\pi_{t,i} = \sigma(h_t^\top W_s z_i), \qquad e_t = \{o_i : \pi_{t,i} \geq \tau_s\}. \tag{25}$$

During training, if the target evidence pointer is unique, the pointer loss is

$$\mathcal{L}_{\mathrm{ptr}} = -\sum_{t=1}^{T} \log p_\theta(e_t^\star \mid u_t^\star, h_t, \mathcal{G}_x). \tag{26}$$

If multiple evidence units are equivalent, for example several chairs satisfy the same referring expression, we marginalize over the valid evidence set $\mathcal{E}_t^\star$:

$$\mathcal{L}_{\mathrm{ptr}} = -\sum_{t=1}^{T} \log \sum_{e\in\mathcal{E}_t^\star} p_\theta(e \mid u_t^\star, h_t, \mathcal{G}_x). \tag{27}$$

This prevents the model from being penalized for selecting a different but equally valid supporting object.

A.3 Local Evidence Verifier

The local verifier checks whether each chain step is supported by the current scene evidence graph. It is deliberately local: it does not attempt to solve the full task, but only checks whether a stated object, attribute, count, or relation claim is consistent with $\mathcal{G}_x$.

Given a step $s_t = (u_t, e_t, m_t)$, we parse the local claim m_t into a set of atomic predicates

$$\Pi(m_t) = \{\pi_{t,1}, \ldots, \pi_{t,L_t}\}. \tag{28}$$

Because AR-CoT uses a structured step format, this parsing is mostly deterministic. For example, a claim such as "`chair_2 is near table_1`" is parsed as $\mathrm{near}(o_{\texttt{chair_2}}, o_{\texttt{table_1}})$. A claim such as "`there are three chairs`" is parsed as a count predicate over the selected object set.

For each atomic predicate π, we compute a graph support score

$$\mathrm{supp}(\pi, \mathcal{G}_x) \in [0, 1]. \tag{29}$$

For object-category predicates, the support is the corresponding category probability:

$$\mathrm{supp}(\mathrm{cat}(o_i, c), \mathcal{G}_x) = c_i[c]. \tag{30}$$

For relation predicates, the support is the relation confidence in the evidence graph:

$$\mathrm{supp}(r(o_i, o_j), \mathcal{G}_x) = \rho^r_{ij}. \tag{31}$$

For count predicates over a category or a selected set, we compare the predicted count with the graph-supported count:

$$\mathrm{supp}(\mathrm{count}(\mathcal{S}) = n, \mathcal{G}_x) = \exp\left(-\frac{|\,|\mathcal{S}| - n\,|}{\tau_c}\right), \tag{32}$$

where τ_c is a temperature controlling count tolerance.

We also compute a pointer-validity feature

$$b^{\mathrm{ptr}}_t = \mathbf{1}\left[e_t \in \mathcal{O}_x \cup \mathcal{R}_x \cup 2^{\mathcal{O}_x}\right], \tag{33}$$

and a predicate-support feature

$$b^{\mathrm{pred}}_t = \min_{\pi \in \Pi(m_t)} \mathrm{supp}(\pi, \mathcal{G}_x). \tag{34}$$

For conclusion steps, we additionally check whether the final answer is entailed by the verified preceding steps:

$$b^{\mathrm{ent}}_t = \mathrm{Entail}(a_t, s_{<t}, q), \tag{35}$$

where Entail is implemented by deterministic rules for yes/no, counting, and relation questions, and by a lightweight classifier for open-vocabulary answers.

The verifier score is

$$v_t = \mathrm{Ver}_\psi(s_t, \mathcal{G}_x) = \sigma\left(w_v^\top [b^{\mathrm{ptr}}_t;\ b^{\mathrm{pred}}_t;\ b^{\mathrm{ent}}_t;\ \mathrm{emb}(u_t)]\right), \tag{36}$$

where σ is the sigmoid function. A purely rule-based version can be obtained by setting $v_t = b^{\mathrm{ptr}}_t b^{\mathrm{pred}}_t$ for non-conclusion steps.

The chain-level verifier score used in Eq. (7) is a weighted average:

$$\mathrm{Ver}(\tau, x) = \frac{\sum_{t=1}^{T} w(u_t) v_t}{\sum_{t=1}^{T} w(u_t)}, \tag{37}$$

where relation and conclusion steps receive larger weights than formatting steps. In our implementation, `locate`, `filter`, and `count` receive unit weight, while `compare`, `verify`, and `conclude` receive larger weights because they directly determine scene-grounded reasoning.

The verifier is trained with binary support labels. Positive steps are taken from target scene-evidenced chains. Negative steps are generated by five perturbations: object swap, relation flip, attribute flip, random-scene substitution, and text-only rationale substitution. The verifier objective is

$$\mathcal{L}_{\mathrm{ver}} = -\frac{1}{|\mathcal{D}_{\mathrm{ver}}|} \sum_{(s,x,z) \in \mathcal{D}_{\mathrm{ver}}} \left[z \log \mathrm{Ver}_\psi(s, \mathcal{G}_x) + (1 - z) \log\left(1 - \mathrm{Ver}_\psi(s, \mathcal{G}_x)\right)\right], \tag{38}$$

where $z \in \{0, 1\}$ indicates whether the step is supported by the scene. After training, we calibrate verifier probabilities by temperature scaling on the validation set:

$$\tilde{v}_t = \sigma\left(\frac{\mathrm{logit}(v_t)}{T_{\mathrm{cal}}}\right), \tag{39}$$

where T_{cal} is selected to minimize validation negative log-likelihood. The calibrated $\tilde{v}_t$ is used for reporting calibration metrics, while the uncalibrated score can be used during training.

A.4 Construction of Decoy Twin Scenes

Decoy twin scenes are designed to test whether the model's answer and reasoning chain depend on scene evidence rather than language priors. A twin pair contains two scenes x^+ and x^- paired with the same question q. The two scenes preserve the same linguistic prior but differ in the spatial evidence required to answer q.

We construct two types of twins.

Real-pair twins. For a question q whose answer depends on a target predicate $r(o_s, o_t)$, we search the training or evaluation pool for another scene containing the same queried object categories but a different relation label. For example, if q asks whether a chair is near a table, we pair a scene where $\mathrm{near}(\texttt{chair}, \texttt{table})$ is supported with a scene where the same object categories exist but the relation is not supported. Candidate twins are matched by room type, object-category multiset, question template, and object-count statistics. This keeps the text prior approximately fixed while changing the relevant spatial evidence.

Edited-pair twins. When real pairs are insufficient, we construct minimally edited graph-level twins. For a target relation $r(o_s, o_t)$, we apply a relation-specific transformation to the evidence graph that changes r while preserving object categories, object identities, and most non-target relations. For directional relations, this can be implemented by swapping the relative positions of the two objects along the relevant axis. For proximity relations, we move one object outside the near threshold. For support or containment relations, we remove the corresponding overlap or support condition. When raw mesh editing is available, the edited graph is validated against the edited geometry; otherwise, graph-level twins are used only for training and diagnostic evaluation.

We retain a twin pair only if it passes three filters.

First, the target relation must change with a sufficient margin:

$$\left|\rho^r_{st}(x^+) - \rho^r_{st}(x^-)\right| \geq \Delta_r. \tag{40}$$

Second, non-target evidence should remain stable:

$$\frac{1}{|\mathcal{P}_{\mathrm{non}}|} \sum_{r' \in \mathcal{P}_{\mathrm{non}}} \left|\rho^{r'}_{st}(x^+) - \rho^{r'}_{st}(x^-)\right| \leq \epsilon_{\mathrm{non}}, \tag{41}$$

where $\mathcal{P}_{\mathrm{non}}$ denotes predicates not required by q.

Third, the text-only prior should not strongly distinguish the two answer-chain pairs:

$$\left|\ell_T(y^+ \mid q) - \ell_T(y^- \mid q)\right| \leq \epsilon_T. \tag{42}$$

If this condition is not satisfied, the pair is retained only for stratified analysis rather than for the main decoy-twin evaluation. This filter ensures that high performance on the twin benchmark cannot be explained by answer priors alone.

For each retained pair, we construct matched target chains $y^+ = (a^+, \tau^+)$ and $y^- = (a^-, \tau^-)$. The two chains use the same high-level operation sequence but point to different relation evidence. This makes the pair especially sensitive to rationalization: a model that reuses the same fluent chain for both scenes will fail the contrast even if the chain sounds plausible.

A.5 Target Chain Construction

AR-CoT does not require human-written free-form rationales. Instead, target chains are generated from object and relation annotations using deterministic evidence programs. This avoids supervising the model with unverifiable rationales and ensures that each chain step has an explicit evidence pointer.

For each training example $(x, q, a^\star)$, we first infer or retrieve a minimal evidence program

$$\mathcal{P}^\star(q, x) = (\xi_1, \ldots, \xi_L), \tag{43}$$

where each program element ξ_ℓ is an object, attribute, count, or relation predicate needed to support the answer. We then convert the program into a chain

$$\tau^\star = \mathrm{Linearize}\left(\mathcal{P}^\star(q, x)\right). \tag{44}$$

Table 4: Question-type-specific target chain templates. Each template is instantiated with object or relation evidence from $\mathcal{G}_x$.

Question type	**Evidence program**	**Chain skeleton**
Existence	Find objects satisfying category or attribute predicate.	`locate` target category → `filter` valid candidates → `conclude` existence.
Attribute QA	Locate target object and verify queried attribute.	`locate` target object → `verify` attribute evidence → `conclude` attribute answer.
Binary relation	Locate subject and object; verify relation edge.	`locate` subject → `locate` object → `compare` relation → `conclude` yes/no answer.
Counting	Select all objects satisfying the query condition.	`locate` category → `filter` valid set → `count` selected objects → `conclude` count.
Comparative spatial QA	Locate candidate objects and compare geometric or relational evidence.	`locate` candidates → `compare` distances, directions, or heights → `select` supported candidate → `conclude`.
Referring expression	Resolve the target object using attributes and relations.	`locate` candidate category → `filter` by attributes → `verify` disambiguating relation → `conclude` referred object.

Table 4 summarizes the main templates.

For relation questions, a typical target chain is

$$\begin{aligned}
s_1^\star &= (\texttt{locate}, o_s, \text{“the subject object is } o_s\text{”}),\\
s_2^\star &= (\texttt{locate}, o_t, \text{“the reference object is } o_t\text{”}),\\
s_3^\star &= (\texttt{compare}, (o_s, r, o_t), \text{“relation } r(o_s, o_t) \text{ is supported”}),\\
s_4^\star &= (\texttt{conclude}, (o_s, r, o_t), \text{“therefore the answer is } a^\star\text{”}).
\end{aligned} \tag{45}$$

For counting questions, the set pointer e_t is the selected object set. For ambiguous questions where multiple evidence sets support the same answer, we store all valid evidence sets and use the marginalized pointer loss in Eq. (27).

When only answer supervision is available, we build weak target chains using the highest-confidence evidence program consistent with the answer. These weak chains are down-weighted by a reliability coefficient

$$\omega_{\text{weak}} = \text{Ver}(\tau^\star, x), \tag{46}$$

so that uncertain pseudo-chains do not dominate training. The supervised loss can be decomposed as

$$\begin{aligned}
\mathcal{L}_{\text{sup}} = &- \log p_\theta(a^\star \mid \tau^\star, q, \mathcal{G}_x) + \alpha_u \sum_{t=1}^{T} - \log p_\theta(u_t^\star \mid h_t)\\
&+ \alpha_e \mathcal{L}_{\text{ptr}} + \alpha_m \sum_{t=1}^{T} - \log p_\theta(m_t^\star \mid u_t^\star, e_t^\star, h_t, \mathcal{G}_x).
\end{aligned} \tag{47}$$

In the main text, this complete decomposition is denoted compactly as $\mathcal{L}_{\text{sup}} = - \log p_\theta(y^\star \mid x, q)$.

A.6 Training and Decoding Procedures

This section gives the full procedures for training and inference. The text-only prior is kept frozen after initialization. The scene-conditioned generator and the verifier are updated during training.

Training. For each minibatch, we construct scene evidence graphs, target chains, text-only rationale negatives, and decoy-twin negatives when available.

1. Construct $\mathcal{G}_x = (\mathcal{O}_x, \mathcal{R}_x)$ for each scene x using Appendix A.1.
2. Construct target chain $y^\star = (a^\star, \tau^\star)$ using Appendix A.5.
3. Sample a text-only rationale negative $\tilde{y}_T \sim p_T(y \mid q)$.
4. If a decoy twin (x^+, x^-, q) is available, construct paired targets y^+ and y^-.

5. Compute the supervised loss $\mathcal{L}_{\text{sup}}$.
6. Compute the scene-evidence gain loss $\mathcal{L}_{\text{gain}}$ using Eq. (11).
7. Compute the twin contrastive loss $\mathcal{L}_{\text{twin}}$ using Eq. (10).
8. Update the verifier using $\mathcal{L}_{\text{ver}}$ in Eq. (38).
9. Update the scene-conditioned generator with

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \gamma \mathcal{L}_{\text{gain}} + \delta \mathcal{L}_{\text{twin}} + \eta \mathcal{L}_{\text{ver}}.$$

Gain-based decoding. At inference time, AR-CoT first produces a candidate set from the scene-conditioned model and then reranks candidates using scene-evidence gain.

1. Given (x, q), construct the evidence graph $\mathcal{G}_x$.
2. Generate K candidate answer-chain pairs

$$\mathcal{C}(x, q) = \{y_k = (a_k, \tau_k)\}_{k=1}^{K}$$

   using beam search, nucleus sampling, or self-consistency sampling.
3. For each candidate y_k, compute the normalized scene-conditioned likelihood $\ell_\theta(y_k \mid x, q)$.
4. Compute the text-only prior score $\ell_T(y_k \mid q)$ with all scene tokens replaced by the null evidence token $\varnothing$.
5. Compute the local verifier score $\text{Ver}(\tau_k, x)$.
6. Select

$$y^\star = \arg \max_{y_k \in \mathcal{C}(x,q)} \left[\ell_\theta(y_k \mid x, q) - \lambda \ell_T(y_k \mid q) + \beta \, \text{Ver}(\tau_k, x) \right].$$

If the maximum gain is below an abstention threshold ζ_{abs}, the model returns an evidence-insufficient response. This option is used only in selective prediction experiments and is not used when reporting standard forced-answer accuracy.

Score normalization. Because likelihood, prior score, and verifier score can have different scales, we normalize likelihood terms by output length as in Eq. (6). The coefficients λ and β are selected on the validation set and fixed for test evaluation. We do not tune these coefficients on the decoy-twin test set.

A.7 Shortcut-Sensitive Evaluation Metrics

Standard answer accuracy is necessary but insufficient for evaluating faithful 3D reasoning [Chen et al., 2026b, Liu et al., 2026, Li et al., 2026b]. AR-CoT therefore reports both task performance and shortcut-sensitive faithfulness metrics.

Let $\mathcal{D} = \{(x_n, q_n, y_n^\star)\}_{n=1}^{N}$ be the evaluation set and let $\hat{y}_n = (\hat{a}_n, \hat{\tau}_n)$ be the model prediction.

Answer accuracy. For closed-form answers, we report exact match:

$$\text{Acc} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}[\hat{a}_n = a_n^\star]. \tag{48}$$

For open-vocabulary answers, we additionally use the dataset-standard refined exact match or LLM-based judge when available.

Object grounding accuracy. If the target evidence includes a ground-truth object $o^\star$, object grounding is correct when the predicted pointer overlaps with the target object:

$$\text{OGA} = \frac{1}{N_o} \sum_{n=1}^{N_o} \mathbf{1}\Big[\text{IoU}_{3D}(\hat{b}_n, b_n^\star) \geq \tau_{\text{IoU}}\Big]. \tag{49}$$

Ground-truth chain grounding accuracy (CGA). For each predicted chain of length T_n, let $\mathcal{E}^{\star}_{n,t}$ be the valid evidence pointers from ground-truth annotations for step t, and let $\hat{e}_{n,t}$ be the predicted pointer. We define CGA as the macro average of within-chain pointer accuracy:

$$\mathrm{CGA} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{T_n}\sum_{t=1}^{T_n}\mathbf{1}\big[\hat{e}_{n,t} \in \mathcal{E}^{\star}_{n,t}\big]\,. \tag{50}$$

Each example receives equal weight regardless of chain length. Throughout this paper, CGA is evaluated against annotated ground-truth evidence, including when the model uses a predicted evidence graph.

Relation verification accuracy. For relation-bearing steps, we report the fraction of predicted relation claims supported by the scene:

$$\mathrm{RVA} = \frac{\sum_{n,t}\mathbf{1}[\hat{s}_{n,t}\text{ is relation-bearing}]\,\mathbf{1}[\mathrm{supp}(\hat{s}_{n,t},\mathcal{G}_{x_n}) \geq \kappa_{\mathrm{ver}}]}{\sum_{n,t}\mathbf{1}[\hat{s}_{n,t}\text{ is relation-bearing}]}. \tag{51}$$

Scene-evidence gain. We report average scene-evidence gain:

$$\mathrm{SEG} = \frac{1}{N}\sum_{n=1}^{N} G_\theta(\hat{y}_n; x_n, q_n). \tag{52}$$

We also report the correlation between scene-evidence gain and chain grounding correctness.

Shortcut reliance rate. Let $\hat{y}^T_n = (\hat{a}^T_n, \hat{\tau}^T_n)$ be the text-only prediction for question q_n. Let Sim_τ be a chain similarity function based on operation overlap and evidence-pointer overlap. Shortcut reliance rate is

$$\mathrm{SRR} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\big[\hat{a}_n = \hat{a}^T_n\big]\,\mathbf{1}\big[\mathrm{Sim}_\tau(\hat{\tau}_n, \hat{\tau}^T_n) \geq \tau_{\mathrm{sim}}\big]\,\mathbf{1}[\mathrm{Ver}(\hat{\tau}_n, x_n) < \kappa_{\mathrm{ver}}]\,. \tag{53}$$

A lower SRR indicates less dependence on language-only rationalization.

Random-scene invariance rate. Let x^{rand}_n be a randomly selected scene unrelated to q_n. Random-scene invariance measures whether the model produces the same answer and similar chain after the scene is replaced:

$$\mathrm{RSI} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{1}\big[\hat{a}(x_n, q_n) = \hat{a}(x^{\mathrm{rand}}_n, q_n)\big]\,\mathbf{1}\big[\mathrm{Sim}_\tau(\hat{\tau}(x_n, q_n), \hat{\tau}(x^{\mathrm{rand}}_n, q_n)) \geq \tau_{\mathrm{sim}}\big]\,. \tag{54}$$

A high RSI indicates that the generated chain is insensitive to the observed 3D scene.

Decoy twin contrast accuracy. For a twin pair (x^+, x^-, q) with targets y^+ and y^-, contrast is correct only if each scene favors its own supported answer-chain pair:

$$\begin{aligned}\mathrm{DTCA} = \frac{1}{N_{\mathrm{twin}}}\sum_{n=1}^{N_{\mathrm{twin}}}&\mathbf{1}\big[G_\theta(y^+_n; x^+_n, q_n) > G_\theta(y^-_n; x^+_n, q_n)\big]\\ &\times\mathbf{1}\big[G_\theta(y^-_n; x^-_n, q_n) > G_\theta(y^+_n; x^-_n, q_n)\big]\,.\end{aligned} \tag{55}$$

Scene-swap chain sensitivity. Let x^{swap}_n be a scene-swap counterpart that preserves the language prior but changes the relevant spatial evidence. Chain sensitivity is

$$\mathrm{CSS} = 1 - \frac{1}{N}\sum_{n=1}^{N}\mathrm{Sim}_\tau\left(\hat{\tau}(x_n, q_n), \hat{\tau}(x^{\mathrm{swap}}_n, q_n)\right). \tag{56}$$

A higher CSS indicates that the reasoning chain changes when the supporting scene evidence changes.

Unsupported spatial hallucination rate. For all generated steps that make explicit spatial claims, we compute

$$\mathrm{UHR} = \frac{\sum_{n,t} \mathbf{1}[\hat{s}_{n,t} \text{ is spatial}]\, \mathbf{1}[\mathrm{Ver}_\psi(\hat{s}_{n,t}, \mathcal{G}_{x_n}) < \kappa_{\mathrm{ver}}]}{\sum_{n,t} \mathbf{1}[\hat{s}_{n,t} \text{ is spatial}]}. \tag{57}$$

This metric measures how often the model states unsupported spatial evidence.

Selective accuracy and calibration. For selective prediction, the model answers only when

$$\max_{y_k \in \mathcal{C}(x,q)} G_\theta(y_k; x, q) \geq \zeta.$$

Coverage and selective accuracy are

$$\mathrm{Cov}(\zeta) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}[G_\theta(\hat{y}_n; x_n, q_n) \geq \zeta], \tag{58}$$

$$\mathrm{Acc}_{\mathrm{sel}}(\zeta) = \frac{\sum_{n=1}^{N} \mathbf{1}[G_\theta(\hat{y}_n; x_n, q_n) \geq \zeta]\, \mathbf{1}[\hat{a}_n = a_n^\star]}{\sum_{n=1}^{N} \mathbf{1}[G_\theta(\hat{y}_n; x_n, q_n) \geq \zeta]}. \tag{59}$$

We also report expected calibration error by binning predictions according to normalized gain confidence.

A.8 Evidence Settings and Detector-Quality Controls

To avoid conflating proposal quality with reasoning quality, we evaluate AR-CoT under three evidence settings whenever the benchmark supports them.

Ground-truth evidence. Object segments and object IDs are provided from annotations. This setting isolates chain reasoning from detection quality.

Detector-proposal evidence. Objects are obtained from an external 3D detector or segmentor. This setting reflects the common object-proposal pipeline used by many 3D MLLMs.

Predicted-token evidence. The model uses its own 3D tokens or proposal-free visual tokens. Object evidence is induced by clustering, attention pooling, or referent-token prediction. This setting tests whether AR-CoT can operate without a fixed external detector.

We report answer accuracy, chain grounding, verifier score, and decoy-twin contrast separately for each evidence setting. This prevents improvements from being attributed to reasoning when they are primarily caused by stronger object proposals.

A.9 Ablation Protocol

We use the following ablations to isolate the contribution of each component.

For all ablations, we use the same base model, training data, candidate-generation budget, and validation protocol. Only the component under study is removed or replaced. This ensures that differences in shortcut-sensitive metrics reflect the intended design choice rather than changes in model capacity or decoding budget.

B Additional Experimental Details

B.1 Details of the Preliminary Study

This subsection provides additional details for the exploratory study in Section 3. The purpose of this study is diagnostic rather than methodological: it is used to reveal failure patterns of existing 3D MLLMs with standard free-form chain-of-thought prompting. In healthcare-agent audits, controlled evidence perturbations similarly expose workflow failures that clean-condition accuracy can conceal [Zheng et al., 2026b].

Table 5: Ablation variants.

Variant	Purpose
Full AR-CoT	Complete model with scene-evidence gain, text-only prior, verifier, and decoy-twin training.
w/o text-only prior	Removes $-\lambda \ell_T(y \mid q)$ from the gain. Tests whether prior contrast is necessary for suppressing rationalization.
w/o verifier	Removes $\beta \, \mathrm{Ver}(\tau, x)$. Tests whether gain alone is sufficient without local evidence checking.
w/o decoy twins	Removes $\mathcal{L}_{\mathrm{twin}}$. Tests whether paired-scene contrast improves scene dependence.
Answer-level gain only	Applies gain only to the final answer instead of the full answer-chain pair. Tests whether chain-level evidence is necessary.
Generic CoT	Uses free-form chain-of-thought without evidence pointers. Tests whether longer explanations alone improve faithfulness.
Verifier-only reranking	Selects candidates only by $\mathrm{Ver}(\tau, x)$. Tests whether AR-CoT reduces to local verification.
Logit-subtraction only	Selects candidates by $\ell_\theta(y \mid x, q) - \lambda \ell_T(y \mid q)$. Tests whether verifier and chain grounding add beyond calibration.
GT evidence vs. predicted evidence	Separates reasoning performance from object proposal quality.
Real-pair twins vs. edited-pair twins	Tests whether decoy-twin gains depend on synthetic graph edits.

Evaluation instances. We construct the diagnostic set from 3D question-answering and grounding-style examples that require object-level or relation-level scene understanding. Each instance contains a scene x, a natural-language question q, a reference answer $a^\star$, and, when available, annotated object or relation evidence. We group questions into relation families, including existence, attribute, counting, near/far, left/right, above/below, support, containment, visibility, object co-occurrence, room-prior, and ambiguous-reference questions. The grouping is used only for analysis and visualization; the model is not given the group label during inference.

Model and prompting. We use the same off-the-shelf 3D MLLM backbone for all diagnostic conditions. The model is prompted to produce a final answer together with a free-form rationale:

$$(\hat{a}, \hat{\tau}) = f_{\mathrm{base}}(x, q),$$

where $\hat{a}$ is the predicted answer and $\hat{\tau}$ is the generated chain-of-thought rationale. For all conditions, we keep the decoding strategy, prompt template, maximum output length, and sampling temperature fixed [Yang et al., 2026b]. This ensures that differences across conditions are caused by input perturbations rather than decoding changes.

Input perturbations. For each original input (x, q), we evaluate three conditions. In the *Full Scene* condition, the model receives the original scene:

$$(\hat{a}^{\mathrm{full}}, \hat{\tau}^{\mathrm{full}}) = f_{\mathrm{base}}(x, q).$$

In the *Question Only* condition, the 3D scene input is removed or replaced by a null scene placeholder:

$$(\hat{a}^{\mathrm{text}}, \hat{\tau}^{\mathrm{text}}) = f_{\mathrm{base}}(\varnothing, q).$$

This condition estimates how much of the answer and rationale can be produced from language alone. In the *Scene Swap* condition, the question is paired with a different scene x':

$$(\hat{a}^{\mathrm{swap}}, \hat{\tau}^{\mathrm{swap}}) = f_{\mathrm{base}}(x', q).$$

The swapped scene is selected to preserve coarse linguistic cues, such as room type and object categories, while changing the spatial evidence relevant to the question. Unlike the decoy twins used later by AR-CoT, these scene swaps are not used for training and are not optimized by any contrastive objective. They are used only to test whether a free-form rationale changes when the underlying scene evidence changes.

Scene-swap selection. For each question, we first identify the coarse evidence type required by the question, such as object existence, object attribute, binary spatial relation, counting, or referential disambiguation. We then retrieve candidate swap scenes that share the same high-level scene category and contain the main queried object categories. Among these candidates, we prefer scenes where the target relation or disambiguating evidence differs from the original scene. For example, for a question about whether a chair is near a table, the swapped scene should still contain chairs and tables but should differ in the near/far relation. This protocol keeps the language prior approximately stable while perturbing the relevant visual evidence.

Bucket-level language-prior strength. Figure 2(b) uses an estimated language-prior strength for each question bucket. For a bucket B, we define this quantity as the accuracy of the base model under the question-only condition:

$$\widehat{P}_{\text{lang}}(B) = \frac{1}{|B|} \sum_{(x,q,a^\star)\in B} \mathbf{1}\left[\hat{a}^{\text{text}} = a^\star\right]. \tag{60}$$

A high value means that the answer is often predictable without observing the 3D scene. This diagnostic quantity is not the same as the text-only rationale prior introduced in the method section. It is a post-hoc empirical statistic computed only for analysis.

Answer accuracy. For each input condition $c \in \{\text{full}, \text{text}, \text{swap}\}$, answer accuracy is computed as

$$\text{Acc}^c = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\left[\hat{a}_n^c = a_n^\star\right]. \tag{61}$$

For open-vocabulary answers, we use the benchmark-standard answer matching rule when available. Otherwise, we normalize answers by lowercasing, removing punctuation and articles, and matching against the accepted answer set.

Rationale similarity after scene swap. To measure whether a rationale changes after the scene is changed, we compute a text-level similarity between $\hat{\tau}^{\text{full}}$ and $\hat{\tau}^{\text{swap}}$. The similarity score combines lexical overlap and spatial-relation-word overlap:

$$\text{Sim}_\tau\left(\hat{\tau}^{\text{full}}, \hat{\tau}^{\text{swap}}\right) = \frac{1}{2}\,\text{F1}_{\text{tok}}\left(\hat{\tau}^{\text{full}}, \hat{\tau}^{\text{swap}}\right) + \frac{1}{2}\,\text{F1}_{\text{rel}}\left(\hat{\tau}^{\text{full}}, \hat{\tau}^{\text{swap}}\right). \tag{62}$$

Here, F1_{tok} is token-level F1 after standard text normalization, and F1_{rel} is F1 over a small vocabulary of spatial relation words, such as *near*, *left*, *right*, *above*, *below*, *inside*, *on*, *behind*, and *in front of*. A high similarity score indicates that the chain remains largely unchanged even when the scene evidence is perturbed.

Post-hoc evidence support. We compute a coarse evidence support score to determine whether the generated rationale mentions scene evidence that is actually present. We first extract object and relation mentions from $\hat{\tau}$ using simple category matching and relation-keyword matching. Let $\mathcal{M}(\hat{\tau})$ be the set of extracted mentions, and let $\text{Valid}(m, x) \in \{0, 1\}$ indicate whether mention m is supported by the annotated scene evidence. The support score is

$$\text{Supp}_{\text{diag}}(\hat{\tau}, x) = \frac{\sum_{m\in\mathcal{M}(\hat{\tau})} \text{Valid}(m, x)}{|\mathcal{M}(\hat{\tau})| + \epsilon}. \tag{63}$$

For object mentions, validity means that the mentioned category or referred object exists in the scene. For relation mentions, validity means that the corresponding relation is supported by the annotated or geometrically derived relation evidence. If a rationale contains no identifiable scene-evidence mention, we set the numerator to zero and use ϵ only to avoid division by zero.

Correct-but-unsupported ratio. To quantify the mismatch between final-answer correctness and rationale faithfulness, we report the fraction of correct predictions whose rationales have low evidence support:

$$\text{CBU}(B) = \frac{\sum_{n\in B} \mathbf{1}[\hat{a}_n^{\text{full}} = a_n^\star]\mathbf{1}[\text{Supp}_{\text{diag}}(\hat{\tau}_n^{\text{full}}, x_n) < \kappa_{\text{sup}}]}{\sum_{n\in B} \mathbf{1}[\hat{a}_n^{\text{full}} = a_n^\star] + \epsilon}. \tag{64}$$

We use a fixed support threshold κ_{sup} for all question families. A high CBU value means that the model often answers correctly while producing rationales that are weakly supported by the scene.

B.2 Additional Experimental Details

This appendix provides the experimental details omitted from the compact setup in Section 5.1, including dataset preprocessing, stress-test construction, metric definitions, comparison protocols, implementation details, hyperparameter selection, decoding settings, verifier training, leakage control, and ablation variants.

Dataset preprocessing. All ScanNet-based datasets use the official train/validation/test splits whenever available. For ScanQA [Azuma et al., 2022] and SQA3D [Ma et al., 2023], we follow the standard open-ended QA protocol and normalize answers by lowercasing, stripping punctuation, and removing articles before exact-match computation. For ScanRefer [Chen et al., 2020], ReferIt3D [Achlioptas et al., 2020], and Multi3DRefer [Zhang et al., 2023b], we use the official object annotations and report both ground-truth-segment and detector-proposal settings when applicable. The ground-truth-segment setting isolates language and reasoning ability, while the detector-proposal setting reflects a more realistic deployment scenario. For MMScan [Lyu et al., 2024] and 3D-GRAND [Yang et al., 2025], we retain examples with object-level or relation-level evidence annotations so that chain grounding and verifier metrics can be computed. For SPARTUN3D [Zhang et al., 2025] and 3DSRBench [Ma et al., 2025], we group examples by spatial-relation type, including proximity, containment, support, vertical relation, horizontal direction, visibility, and existence. Beacon3D [Huang et al., 2025a] is used to evaluate object-centric grounding–QA coherence under its official chain-of-analysis protocol.

Additional benchmarks. In addition to the main benchmarks reported in Table 1 and Table 2, we evaluate AR-CoT on ReferIt3D [Achlioptas et al., 2020], MMScan [Lyu et al., 2024], 3D-GRAND [Yang et al., 2025], SPARTUN3D [Zhang et al., 2025], and 3DSRBench [Ma et al., 2025]. Methods whose original evaluations target different tasks or reporting protocols, such as PQ3D [Zhu et al., 2024] and SceneVerse [Jia et al., 2024], are reported only in Appendix C.1. We do not merge incompatible metrics across benchmarks; entries are left blank when a method does not report or cannot be fairly evaluated under a compatible protocol.

AR-CoT stress-test construction. For each benchmark instance (x, q, a), we construct controlled diagnostic variants that preserve the linguistic prior while removing or modifying the relevant scene evidence. In the *question-only* condition, the scene input is replaced by the null-evidence token used by the text-only prior. In the *random-scene* condition, q is paired with a randomly sampled scene from the same split, without constraining object categories. In the *scene-swapped* condition, q is paired with a scene that shares the same room type or key object categories but changes the relation required to answer the question. For decoy-twin evaluation, we construct pairs (x^+, x^-, q) such that x^+ and x^- preserve the same language prior but support different answers or evidence chains. We use two twin variants: *real-pair twins*, obtained by matching real scenes with similar object categories but different spatial relations, and *edited-pair twins*, obtained by minimally perturbing object positions or relation labels while preserving physically plausible geometry. Edited twins are used only after geometric validation and are not mixed with real-pair twins in the main table.

Standard QA and grounding metrics. For ScanQA, we report CIDEr, METEOR, ROUGE-L, and exact match. For SQA3D, we report exact match. For open-ended QA, exact match is computed as

$$\mathrm{EM} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{a}_i = a_i] \,. \tag{65}$$

For single-object grounding, we report Acc@τ with $\tau \in \{0.25, 0.5\}$:

$$\mathrm{Acc@}\tau = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\Big[\mathrm{IoU}(\hat{b}_i, b_i) \geq \tau\Big] \,. \tag{66}$$

For multi-object grounding, we match predicted and ground-truth object sets under the same IoU threshold and report object-level precision, recall, and F1. For QA-driven grounding on ScanQA-G and SQA3D-G, a prediction is counted as correct only when both the final answer and the associated evidence grounding are correct:

$$\mathrm{QAG@}\tau = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[\hat{a}_i = a_i]\, \mathbf{1}\Big[\mathrm{match}_\tau(\hat{\mathcal{B}}_i, \mathcal{B}_i) = 1\Big] \,, \tag{67}$$

where $\hat{\mathcal{B}}_i$ and $\mathcal{B}_i$ are the predicted and ground-truth evidence object sets.

Faithful reasoning metrics. We report MSQA overall accuracy and Beacon3D metrics, including case-level accuracy, object-level accuracy, and Good Coherence (GC), where both grounding and QA are correct. All CGA results use the ground-truth, macro-averaged definition in Eq. (50). Separately, we define *Graph-Support Rate* (GSR) as the fraction of generated steps whose evidence pointer and local claim are supported by the predicted evidence graph $\mathcal{G}_{x_i}$:

$$\mathrm{GSR} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T_i}\mathbf{1}[\hat{e}_{i,t} \in \mathcal{G}_{x_i} \wedge \mathrm{claim}(\hat{s}_{i,t}) \text{ is supported by } \mathcal{G}_{x_i}]}{\sum_{i=1}^{N} T_i}. \tag{68}$$

GSR is micro-averaged over all steps and serves as a separate predicted-graph consistency diagnostic. Relation-step verification accuracy follows Eq. (51) and considers only steps whose local claims involve spatial predicates. The unsupported-claim hallucination rate relative to the predicted graph is the complementary fraction:

$$\mathrm{Hall}_{\mathrm{claim}} = 1 - \mathrm{GSR}. \tag{69}$$

For a decoy-twin pair (x_i^+, x_i^-, q_i) with scene-supported chains y_i^+ and y_i^-, we define decoy-twin contrast accuracy (DTCA) as

$$\mathrm{DTCA} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\big[G_\theta(y_i^+; x_i^+, q_i) > G_\theta(y_i^-; x_i^+, q_i)\big]\,\mathbf{1}\big[G_\theta(y_i^-; x_i^-, q_i) > G_\theta(y_i^+; x_i^-, q_i)\big]. \tag{70}$$

For scene-swap evaluation, we compute chain sensitivity as

$$\mathrm{CSS} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \mathrm{Sim}(\tau_i, \tau_i^{\mathrm{swap}})\right), \tag{71}$$

where Sim is the average of token-level F1 and evidence-pointer Jaccard similarity. Higher CSS indicates that the reasoning chain changes when the supporting scene evidence changes. The random-scene invariance rate is

$$\mathrm{RSI} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\big[\hat{a}_i = \hat{a}_i^{\mathrm{rand}}\big]\,\mathbf{1}\big[\mathrm{Sim}(\tau_i, \tau_i^{\mathrm{rand}}) > \rho\big], \tag{72}$$

where ρ is fixed to 0.7 on the validation set. Higher CGA indicates better matching to annotated evidence; higher GSR indicates greater predicted-graph support. Higher DTCA and CSS indicate stronger scene-dependent reasoning, while lower RSI indicates weaker reliance on scene-independent rationalization.

Comparison protocols and fairness. We separate *controlled same-backbone comparisons* from *published-reference comparisons*. Controlled comparisons use LLaVA-3D-7B as the base model and match the 3D encoder, object proposals, language decoder, training data, candidate budget, maximum generation length, and evaluation split across all reasoning baselines. This group includes answer-only decoding, vanilla CoT [Wei et al., 2022], self-consistency [Wang et al., 2023], Multimodal-CoT-style two-stage prompting [Zhang et al., 2024b], answer-level prior correction, verifier-only reranking, and grounded CoT without text-prior contrast. The answer-level prior-correction baseline subtracts the text-only answer likelihood from the scene-conditioned answer likelihood but does not score chains. The verifier-only reranking baseline uses $\mathrm{Ver}(\tau, x)$ alone to select among candidate chains. These two baselines test whether AR-CoT reduces to simple calibration or local verification.

Published-reference comparisons include representative 3D vision–language and grounded reasoning systems, such as ScanQA [Azuma et al., 2022], 3D-VisTA [Zhu et al., 2023], 3D-LLM [Hong et al., 2023], LL3DA [Chen et al., 2024a], LEO [Huang et al., 2024b], Chat-Scene [Huang et al., 2024a], LSceneLLM [Zhi et al., 2025], LLaVA-3D [Zhu et al., 2025], Video-3D LLM [Zheng et al., 2025], 3DRS [Huang et al., 2025b], Vid-LLM [Chen et al., 2026a], SceneCOT [Linghu et al., 2026], and MAG-3D [Zheng et al., 2026a]. Since these methods differ in pretraining data, visual tokenization, input interface, and tool access, we do not treat them as strictly controlled comparisons. They are used to contextualize AR-CoT relative to recent 3D MLLM and grounded reasoning systems.

Backbones. For AR-CoT, we report results under three representative base 3D MLLMs. The first is LLaVA-3D-7B, whose underlying language model is Vicuna-7B. The second is Video-3D LLM-7B, whose underlying language model is Qwen2-7B. The third is LLaVA-NeXT-Video-7B, used as a video-based 3D scene backbone. Unless otherwise specified, controlled ablations and reasoning-baseline comparisons use LLaVA-3D-7B.

Implementation and leakage control. AR-CoT is implemented as a plug-in training and decoding layer on top of the base 3D MLLM. For each backbone, the text-only prior is initialized from the same language decoder and kept frozen, with all scene tokens replaced by a learned null-evidence token. At inference time, AR-CoT constructs the evidence graph from model-accessible scene inputs only, including predicted object proposals and geometry-derived relation scores. Ground-truth object labels, ground-truth boxes, relation annotations, answer labels, and target chains from the test split are never used for candidate generation, reranking, verifier scoring, or gain-based selection. When ground-truth evidence is used, we report it only as an oracle diagnostic in Appendix C.2. Decoy-twin pairs are constructed within each official split: training twins are built only from training scenes, validation twins only from validation scenes, and test twins only from test scenes. Test twins are used only for final evaluation.

Training and optimization. The default implementation uses a 7B language decoder with LoRA adapters. We set the LoRA rank to 16, the LoRA scaling factor to 32, and dropout to 0.05. The 3D encoder is initialized from the corresponding backbone checkpoint and frozen for the first epoch; the projection layer, pointer head, verifier, and LoRA parameters are trained throughout. We use AdamW with weight decay 0.05, a cosine learning-rate schedule, and a warm-up ratio of 3%. The learning rate is 2×10^{-5} for LoRA parameters and 1×10^{-4} for the verifier and pointer heads. The global batch size is 128 after gradient accumulation. Training uses bf16 mixed precision and gradient checkpointing.

Hyperparameter selection. The gain weights are selected from

$$\lambda \in \{0.1, 0.3, 0.5, 1.0\}, \qquad \beta \in \{0.0, 0.1, 0.2, 0.5\}.$$

The auxiliary loss weights are selected from

$$\gamma, \delta \in \{0.1, 0.3, 0.5, 1.0\}, \qquad \eta \in \{0.05, 0.1, 0.2\}.$$

The contrastive margins are selected from

$$m, m_g \in \{0.1, 0.2, 0.5\}.$$

All hyperparameters are tuned on the validation split by maximizing a validation score that averages answer accuracy, chain grounding accuracy, and decoy-twin contrast accuracy. Once selected, the same hyperparameters are fixed for all test evaluations. The default values used in the main experiments are $\lambda = 0.5$, $\beta = 0.2$, $\gamma = 0.5$, $\delta = 0.5$, and $\eta = 0.1$.

Candidate generation and decoding. At inference time, we generate $K = 8$ candidate answer-chain pairs for each question. For deterministic evaluation, we use beam search with a beam size of 8 and retain the top-8 completed answer-chain candidates after length normalization. For self-consistency and stochastic decoding baselines, we use nucleus sampling with $p = 0.9$ and temperature 0.7. Each candidate chain is limited to $T_{\max} = 6$ steps, and each step must contain an operation, an evidence pointer, and a local claim. AR-CoT selects the final output by maximizing the scene-evidence gain in Eq. (8). For all baselines, the maximum number of generated tokens and candidate budget are matched to AR-CoT.

Verifier training. The local verifier is trained with positive object and relation claims extracted from training annotations and negative claims generated by object substitution, relation inversion, scene swapping, and evidence deletion. Relation inversion swaps predicates such as `left-of`/`right-of`, `above`/`below`, and `inside`/`outside`. Evidence deletion removes the object or relation referenced by a chain step. The verifier is optimized with binary cross-entropy and calibrated on the validation split using temperature scaling. No test-set labels or test-set target chains are used for verifier training or calibration.

Ablations. We evaluate the following ablations. W/O TEXT PRIOR removes the $\ell_T(y \mid q)$ term. W/O VERIFIER removes $\mathrm{Ver}(\tau, x)$. W/O TWIN removes decoy-twin contrastive training. ANSWER-GAIN ONLY applies prior correction only to the final answer. VERIFIER ONLY ranks candidates only by verifier score. FREE-FORM COT removes evidence pointers from the reasoning chain. GT EVIDENCE uses ground-truth scene evidence as an oracle diagnostic. PREDICTED EVIDENCE uses detector proposals and geometry-derived relation scores, corresponding to the main setting. These ablations isolate the contribution of scene-evidence gain, text-prior contrast, local verification, decoy-twin supervision, and evidence quality.

Reporting protocol. For published-reference comparisons, we use officially reported results when the metric and evaluation protocol are compatible; otherwise, we reproduce the method under our protocol or leave the entry unreported. For controlled baselines, stress tests, and ablations, we report the mean over three random seeds and include standard deviations. All hyperparameters are selected on validation splits and fixed before test evaluation.

B.3 Main Experiment Details

Single-factor ablation details. The single-factor ablation isolates the contribution of each major component in AR-CoT. All variants use the same LLaVA-3D-7B backbone, training data, object-proposal setting, candidate-generation budget, maximum chain length, and validation protocol. Only the component under study is removed or replaced. The evaluated variants are as follows. W/O TEXT PRIOR removes the text-only rationale-prior term $-\lambda \ell_T(y \mid q)$ from scene-evidence gain. W/O VERIFIER removes the local evidence-verification term $\beta\, \mathrm{Ver}(\tau, x)$. W/O DECOY-TWIN LOSS removes the paired-scene contrastive loss. ANSWER-GAIN ONLY applies prior correction only to the final answer rather than to the answer-chain pair. VERIFIER ONLY ranks candidates only by $\mathrm{Ver}(\tau, x)$. FREE-FORM COT removes explicit evidence pointers and uses unconstrained natural-language rationales. GT EVIDENCE is an oracle diagnostic that constructs the evidence graph from ground-truth object and relation evidence, while PREDICTED EVIDENCE uses detector proposals and geometry-derived relation scores, corresponding to the main non-oracle setting. For controlled ablations, we report mean and standard deviation over three random seeds. RSI is lower better; ScanQA EM, SQA3D EM, MSQA, Beacon3D GC, CGA, DTCA, and CSS are higher better.

Answer–chain consistency details. This diagnostic separates final-answer correctness from evidence-chain correctness. For each evaluation example $(x_i, q_i, a_i^\star)$, let

$$A_i = \mathbf{1}[\hat{a}_i = a_i^\star], \qquad E_i = \mathbf{1}[\mathrm{CGA}_i \geq \kappa_{\mathrm{cga}}], \tag{73}$$

where A_i indicates whether the final answer is correct, E_i indicates whether the predicted evidence chain is sufficiently grounded, and CGA_i is the within-example ground-truth pointer accuracy used in Eq. (50). We set $\kappa_{\mathrm{cga}} = 0.5$ for all methods and datasets. This yields four mutually exclusive categories:

$$\begin{aligned}
&\mathcal{Q}_{11} : (A_i = 1, E_i = 1) \quad \text{answer correct and evidence correct,} \\
&\mathcal{Q}_{10} : (A_i = 1, E_i = 0) \quad \text{answer correct but evidence incorrect,} \\
&\mathcal{Q}_{01} : (A_i = 0, E_i = 1) \quad \text{answer incorrect but evidence correct,} \\
&\mathcal{Q}_{00} : (A_i = 0, E_i = 0) \quad \text{answer incorrect and evidence incorrect.}
\end{aligned} \tag{74}$$

For each method m, we report the row-normalized proportion of each category:

$$\rho_{uv}(m) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}[(A_i, E_i) = (u, v)], \qquad (u, v) \in \{0, 1\}^2. \tag{75}$$

The diagnostic is evaluated in the controlled same-backbone setting with LLaVA-3D-7B. We pool the shortcut-sensitive evaluation suite, including ScanQA, SQA3D, Beacon3D, and the AR-CoT stress-test split. The resulting method-by-category matrix is shown in Figure 5a. A high value in $\mathcal{Q}_{10}$ indicates correct-but-unsupported rationalization, whereas a high value in $\mathcal{Q}_{11}$ indicates jointly correct and scene-grounded reasoning.

Language-prior regime details. This diagnostic tests whether AR-CoT is especially helpful in regimes where language priors can already produce plausible answers. We first group evaluation

Table 6: **Full ScanQA results.** The main paper reports CIDEr and exact match for compactness; here we additionally report METEOR and ROUGE-L.

Method	Base Model	CIDEr	METEOR	ROUGE-L	EM
ScanQA [Azuma et al., 2022]	Non-LLM	64.9	13.1	35.0	21.1
3D-VisTA [Zhu et al., 2023]	Non-LLM	69.6	13.8	36.2	22.4
3D-LLM [Hong et al., 2023]	Flan-T5	69.4	13.5	35.8	20.5
LL3DA [Chen et al., 2024a]	OPT-1.3B / LLaMA2-7B	76.8	14.6	38.7	23.2
LEO [Huang et al., 2024b]	Vicuna-7B	101.4	16.9	42.8	24.5
LLaVA-3D [Zhu et al., 2025]	LLaVA-3D-7B / Vicuna-7B	91.7	16.1	41.5	27.0
Video-3D LLM [Zheng et al., 2025]	Video-3D LLM-7B / Qwen2-7B	102.1	17.3	43.6	30.1
3DRS [Huang et al., 2025b]	LLaVA-NeXT-Video-7B	104.8	17.7	44.2	30.3
AR-CoT	LLaVA-3D-7B / Vicuna-7B	110.7	18.6	46.3	32.3
AR-CoT	Video-3D LLM-7B / Qwen2-7B	<u>113.8</u>	<u>19.0</u>	<u>47.1</u>	<u>33.3</u>
AR-CoT	LLaVA-NeXT-Video-7B	**116.5**	**19.4**	**48.0**	**34.0**

examples into relation-family buckets B, including room-prior, object co-occurrence, commonsense relation, near/far, left/right, above/below, containment, support, attribute, existence, and counting questions. For each bucket, we estimate language-prior strength by running the base model under the question-only condition:

$$\widehat{P}_{\text{text}}(B) = \frac{1}{|B|} \sum_{(x,q,a^\star)\in B} \mathbf{1}[\hat{a}_{\text{text}}(q) = a^\star]\,. \tag{76}$$

A larger $\widehat{P}_{\text{text}}(B)$ means that the answer is more predictable without observing the 3D scene. We then compute the bucket-level chain-grounding gain of AR-CoT over Vanilla CoT:

$$\Delta\text{CGA}(B) = \text{CGA}_{\text{AR-CoT}}(B) - \text{CGA}_{\text{Vanilla CoT}}(B). \tag{77}$$

Here, each method's $\text{CGA}(B)$ is the macro average in Eq. (50), restricted to examples in B. In Figure 5b, each point corresponds to one relation-family bucket. The x-axis reports $\widehat{P}_{\text{text}}(B)$, the y-axis reports $\Delta\text{CGA}(B)$, and point size indicates the number of evaluation examples in the bucket. Buckets in the upper-right region are both highly predictable from language alone and strongly improved by AR-CoT. This region identifies rationalization-prone regimes where explicit scene-evidence grounding is most beneficial.

B.4 Computational Resources

All experiments are conducted on NVIDIA A100 80GB GPUs. Each AR-CoT run uses 8 GPUs with bf16 mixed precision and gradient checkpointing. The base 3D encoder is initialized from the corresponding backbone checkpoint, while the language decoder is adapted with LoRA rather than full-parameter fine-tuning.

Potential efficiency extensions. Efficiency approaches in other domains include fine-to-coarse channel/layer pruning for automatic modulation recognition (FCOS) [Lu et al., 2026b] and activation-function design for time-series forecasting [Zhang et al., 2026a]. Transferring these ideas to 3D reasoning remains future work and would require checking evidence-pointer accuracy, verifier calibration, and scene-swap sensitivity after compression or architectural changes.

C Additional Experimental Results

C.1 Complete Metric Reports and Additional Benchmarks

Full ScanQA metrics. Table 6 reports the complete ScanQA evaluation protocol. The main table reports CIDEr and exact match, while this table additionally includes METEOR and ROUGE-L.

QA-driven grounding. Table 7 reports ScanQA-G and SQA3D-G. A prediction is counted as correct only when both the final answer and the associated scene evidence are correct. This evaluation complements answer-only accuracy by testing whether the predicted answer is supported by grounded object or relation evidence.

Table 7: **QA-driven grounding on ScanQA-G and SQA3D-G.** QAG@τ requires both answer correctness and evidence-grounding correctness at IoU threshold τ.

Method	Base Model	ScanQA-G		SQA3D-G	
		QAG@0.25	QAG@0.5	QAG@0.25	QAG@0.5
Answer-only 3D MLLM	LLaVA-3D-7B	24.8	19.6	27.3	21.4
Vanilla CoT	LLaVA-3D-7B	26.1	20.7	28.5	22.8
Self-consistency	LLaVA-3D-7B	26.8	21.3	29.1	23.5
Multimodal-CoT	LLaVA-3D-7B	27.6	22.4	30.0	24.1
Answer-level prior correction	LLaVA-3D-7B	28.2	23.0	30.6	24.9
Verifier-only reranking	LLaVA-3D-7B	31.5	26.4	33.7	28.0
Grounded CoT w/o text prior	LLaVA-3D-7B	33.2	28.1	35.0	29.4
AR-CoT	LLaVA-3D-7B	**36.4**	**31.0**	**38.2**	**32.5**

Table 8: **Additional benchmark results.** ReferIt3D is evaluated by Nr3D/Sr3D accuracy. MMScan is evaluated by QA exact match and grounding accuracy. 3D-GRAND is evaluated by QA accuracy and hallucination rate. SPARTUN3D and 3DSRBench are evaluated by spatial-reasoning accuracy. All metrics are higher better except hallucination rate.

Method	Base Model	ReferIt3D		MMScan		3D-GRAND		SPARTUN3D Acc.	3DSRBench Acc.
		Nr3D	Sr3D	QA EM	Ground Acc.	QA Acc.	Hall.↓		
PQ3D [Zhu et al., 2024]	Task-specific	51.8	64.7	–	–	–	–	–	–
SceneVerse [Jia et al., 2024]	Task-specific	54.2	66.5	25.7	46.3	50.8	17.9	–	–
LLaVA-3D [Zhu et al., 2025]	LLaVA-3D-7B / Vicuna-7B	55.8	67.4	27.3	47.5	53.0	16.8	48.6	46.0
Video-3D LLM [Zheng et al., 2025]	Video-3D LLM-7B / Qwen2-7B	57.2	68.3	28.6	49.2	54.7	15.5	50.4	47.6
3DRS [Huang et al., 2025b]	LLaVA-NeXT-Video-7B	58.6	69.5	29.4	50.8	56.0	14.7	51.8	48.9
AR-CoT	LLaVA-3D-7B / Vicuna-7B	60.8	71.9	31.5	53.7	58.8	12.6	55.7	52.8
AR-CoT	Video-3D LLM-7B / Qwen2-7B	62.0	72.8	32.2	54.9	59.6	11.8	56.9	53.6
AR-CoT	LLaVA-NeXT-Video-7B	**63.2**	**74.0**	**33.4**	**56.1**	**60.9**	**10.9**	**58.1**	**54.8**

Additional Benchmarks The experimental setup states that additional evaluations are conducted on ReferIt3D, MMScan, 3D-GRAND, SPARTUN3D, and 3DSRBench. Table 8 reports these results. We also include PQ3D and SceneVerse in this appendix because their original evaluation protocols are not directly aligned with the compact main-table protocol.

The results on these additional benchmarks are consistent with the main findings. AR-CoT shows larger gains on relation-centric and hallucination-sensitive evaluations, such as SPARTUN3D, 3DSR-Bench, and 3D-GRAND, where language-prior rationalization is more likely to produce plausible but unsupported explanations.

C.2 Evidence Quality and Stress-Test Diagnostics

Oracle Evidence Diagnostic The main experiments use predicted object proposals and geometry-derived relation scores at inference time. Ground-truth object labels, ground-truth boxes, relation annotations, answer labels, and target chains from the test split are not used for candidate generation, reranking, verifier scoring, or gain-based selection. To isolate the effect of evidence quality, we additionally report an oracle diagnostic where the evidence graph is constructed from ground-truth scene evidence. This setting is used only for analysis and is not included in the main comparison.

The oracle diagnostic shows that AR-CoT benefits from stronger evidence graphs, especially on chain grounding and Beacon3D Good Coherence. The remaining gap between predicted and oracle evidence indicates that part of the error comes from imperfect object proposals and relation extraction, rather than the reasoning objective alone.

Stress-Test breakdown The main table reports the aggregate stress-test metrics CGA, DTCA, CSS, and RSI. Here we provide the detailed breakdown for the question-only, random-scene, scene-swapped, and decoy-twin settings described in the experimental setup. All results are reported as mean and standard deviation over three random seeds.

Table 9: **Oracle evidence diagnostic on LLaVA-3D-7B.** The predicted-evidence setting corresponds to the main evaluation. The ground-truth-evidence settings measure the upper bound when object or relation evidence is perfect.

Evidence Setting	ScanQA EM	SQA3D EM	Beacon3D GC	CGA	DTCA
No explicit evidence pointer	29.4	58.1	31.6	44.0	55.8
Predicted proposals + geometry relations	32.3	62.0	43.8	63.8	71.5
Ground-truth segments + geometry relations	33.6	63.1	47.2	70.4	75.8
Ground-truth segments + annotated relations	**34.1**	**63.8**	**49.0**	**75.6**	**78.4**

Table 10: **Stress-test breakdown on LLaVA-3D-7B.** QO Acc. is question-only accuracy and Random Acc. is accuracy under random-scene pairing; lower values indicate weaker reliance on scene-independent shortcuts. RSI is lower better; CSS and DTCA are higher better.

Method	Base Model	QO Acc.↓	Random Acc.↓	RSI↓	CSS	DTCA
Answer-only 3D MLLM	LLaVA-3D-7B	31.6±0.4	34.8±0.5	41.7±0.6	44.0±0.7	52.3±0.5
Vanilla CoT	LLaVA-3D-7B	34.2±0.5	36.5±0.4	39.5±0.5	46.6±0.6	55.7±0.6
Self-consistency	LLaVA-3D-7B	35.0±0.4	36.9±0.5	38.2±0.4	47.8±0.6	56.8±0.5
Multimodal-CoT	LLaVA-3D-7B	33.7±0.5	35.1±0.5	35.9±0.6	50.0±0.6	58.3±0.5
Answer-level prior correction	LLaVA-3D-7B	28.4±0.4	30.6±0.4	31.3±0.5	52.4±0.5	60.5±0.6
Verifier-only reranking	LLaVA-3D-7B	30.2±0.5	31.8±0.4	28.4±0.5	57.2±0.6	63.8±0.5
Grounded CoT w/o text prior	LLaVA-3D-7B	29.5±0.4	30.7±0.5	24.2±0.4	61.5±0.5	66.0±0.5
AR-CoT	LLaVA-3D-7B	**22.7±0.3**	**24.1±0.4**	**18.4±0.3**	**68.7±0.5**	**71.5±0.4**

The stress-test breakdown shows that AR-CoT reduces question-only and random-scene performance while increasing scene-swap sensitivity and decoy-twin contrast accuracy. This indicates that the model is less likely to preserve the same answer-chain pair when the relevant scene evidence is removed or changed.

C.3 Failure-Mode Analysis

Correct-but-unsupported transition analysis. A central claim of this paper is that standard answer accuracy can overestimate genuine 3D reasoning quality. A model may output the correct final answer while relying on a weakly grounded or unsupported reasoning chain. To make this failure mode explicit, we partition each prediction into four case types according to answer correctness and reasoning support.

For a method m, we define the support indicator of a predicted chain $\hat{\tau}_i^{(m)}$ on example i as

$$\mathrm{Sup}_i^{(m)} = \mathbf{1}\left[\mathrm{Ver}(\hat{\tau}_i^{(m)}, x_i) \geq \kappa_{\mathrm{sup}}\right], \tag{78}$$

where κ_{sup} is a fixed verifier threshold. Each example is then assigned to one of four states:

$$z_i^{(m)} \in \{\text{C-S}, \text{C-U}, \text{W-S}, \text{W-U}\}, \tag{79}$$

where C-S denotes *correct and supported*, C-U denotes *correct but unsupported*, W-S denotes *wrong but supported*, and W-U denotes *wrong and unsupported*.

We further quantify the fraction of correct predictions whose rationales are unsupported:

$$\mathrm{CBU}(m) = \frac{\sum_{i=1}^{N} \mathbf{1}\left[\hat{a}_i^{(m)} = a_i^\star\right] \mathbf{1}\left[\mathrm{Sup}_i^{(m)} = 0\right]}{\sum_{i=1}^{N} \mathbf{1}\left[\hat{a}_i^{(m)} = a_i^\star\right] + \epsilon}. \tag{80}$$

To analyze whether AR-CoT repairs this failure mode, we construct the transition matrix from Vanilla CoT to AR-CoT:

$$T_{u \to v} = \sum_{i=1}^{N} \mathbf{1}\left[z_i^{(\text{Vanilla})} = u\right] \mathbf{1}\left[z_i^{(\text{AR-CoT})} = v\right], \qquad u, v \in \{\text{C-S}, \text{C-U}, \text{W-S}, \text{W-U}\}. \tag{81}$$

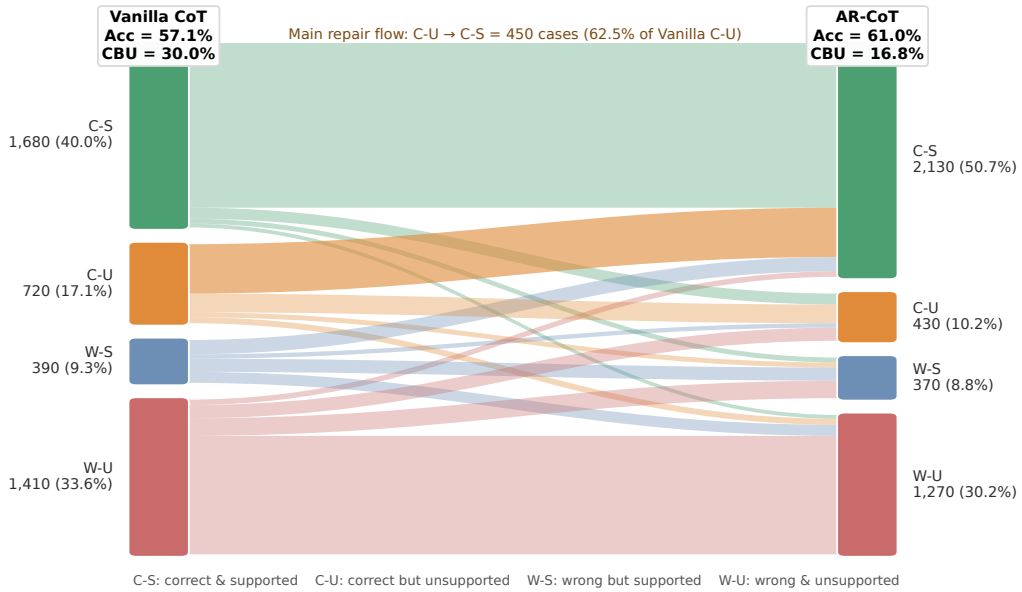


Figure 6: **Case-transition Sankey from Vanilla CoT to AR-CoT.** Each evaluation example is assigned to one of four states: correct and supported (C-S), correct but unsupported (C-U), wrong but supported (W-S), and wrong and unsupported (W-U). The dominant repair flow is C-U $\rightarrow$ C-S, showing that AR-CoT does not merely increase answer accuracy, but specifically converts a large portion of seemingly correct yet weakly grounded predictions into correct-and-supported ones. In our analysis, Vanilla CoT produces 2400 correct answers, among which 720 are unsupported (CBU $= 30.0\%$). AR-CoT reduces this failure mode to 430 unsupported correct predictions out of 2560 correct answers (CBU $= 16.8\%$). Most importantly, 450 out of 720 Vanilla CoT C-U cases are repaired into C-S under AR-CoT.

Figure 6 makes the gap between answer correctness and reasoning faithfulness explicit. Vanilla CoT achieves non-trivial answer accuracy, but a substantial fraction of its correct predictions remain unsupported: among 2400 correct answers, 720 belong to the C-U category, meaning that nearly one-third of the apparently correct outputs are not backed by scene-grounded reasoning. This confirms that standard accuracy alone can overestimate true 3D reasoning quality.

AR-CoT improves this picture in two ways. First, it increases the number of correct-and-supported cases from 1680 to 2130, showing that its gain is not only in producing more correct answers, but in producing more *grounded* correct answers. Second, the dominant transition in the Sankey plot is C-U $\rightarrow$ C-S: 450 out of 720 correct-but-unsupported Vanilla CoT cases (62.5%) are converted into correct-and-supported predictions by AR-CoT, while only 170 remain in the same failure category. This is the clearest evidence that AR-CoT actively detects and repairs unsupported rationalization, rather than merely recalibrating answer likelihood.

Overall, this experiment strengthens the central selling point of the paper: many 3D MLLM predictions that appear correct under standard QA evaluation are not truly faithful. AR-CoT narrows this gap by moving mass from superficially correct yet unsupported predictions toward correct-and-supported reasoning, thereby providing a more reliable estimate of genuine 3D reasoning ability.

C.4 Component-Level Mechanism Analyses

Text-only prior shortcut suppression. We next analyze whether the text-only prior suppresses shortcut-like candidates selectively, or whether it acts as a uniform penalty over all reasoning chains. For each candidate answer-chain pair $y_k = (a_k, \tau_k) \in \mathcal{C}(x, q)$, we first define the score without text-prior correction:

$$S_{\setminus T}(y_k; x, q) = \ell_\theta(y_k \mid x, q) + \beta \, \mathrm{Ver}(\tau_k, x), \tag{82}$$

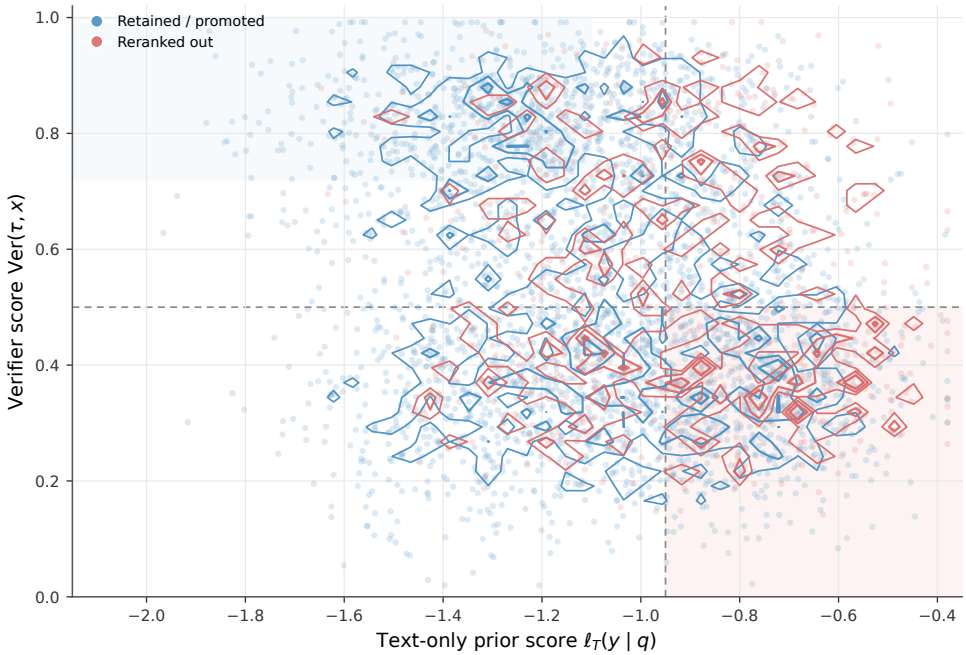


Figure 7: **Which candidates are suppressed by the text-only prior?** Each point denotes a candidate answer-chain pair from the validation candidate pool. The x-axis shows the text-only prior score $\ell_T(y \mid q)$, and the y-axis shows the verifier score $\mathrm{Ver}(\tau, x)$. Red points are candidates whose rank becomes worse after the text-only prior term is introduced, while blue points are retained or promoted. The reranked-out candidates concentrate in the high-prior, low-verifier region, showing that the text-only prior mainly suppresses language-plausible but weakly grounded chains rather than penalizing all candidates indiscriminately.

and the full AR-CoT score:

$$G(y_k; x, q) = S_{\setminus T}(y_k; x, q) - \lambda \ell_T(y_k \mid q). \tag{83}$$

Here, $\ell_T(y_k \mid q)$ is the text-only prior score, so a larger value indicates that the candidate chain is more explainable from the question alone.

To isolate the effect of the text-only prior, we define a candidate-level demotion indicator:

$$z_k = \mathbf{1}\left[\mathrm{rank}_G(y_k) > \mathrm{rank}_{S_{\setminus T}}(y_k)\right], \tag{84}$$

where $z_k = 1$ means that the candidate is *reranked out* after the text-only prior is introduced, and $z_k = 0$ means that the candidate is retained or promoted.

Figure 7 plots all candidate chains in the validation candidate pool. The x-axis is the text-only prior score $\ell_T(y \mid q)$, and the y-axis is the verifier score $\mathrm{Ver}(\tau, x)$, which serves as a continuous proxy for chain correctness. Point color indicates whether the candidate is reranked out by the text-only prior term.

Figure 7 reveals a clear selective pattern. First, reranked-out candidates are heavily concentrated in the regime where $\ell_T(y \mid q)$ is large while $\mathrm{Ver}(\tau, x)$ is low. These are precisely the candidates that are easy to rationalize from language alone but weakly supported by scene evidence, indicating that the text-only prior mainly targets shortcut-like chains. Second, many candidates with strong verifier support remain retained even when their text-only prior scores are moderately high. This shows that AR-CoT does not indiscriminately punish fluent or linguistically plausible chains; it mainly penalizes those whose scene support is insufficient. Third, low-prior candidates with weak verifier scores are not removed as aggressively, which further confirms that the text-only prior is not merely a generic quality filter. Overall, the result supports the intended role of the prior term: it acts as a *selective anti-shortcut mechanism*, especially effective for suppressing high-language-prior candidates that would otherwise survive due to fluent but weakly grounded reasoning.

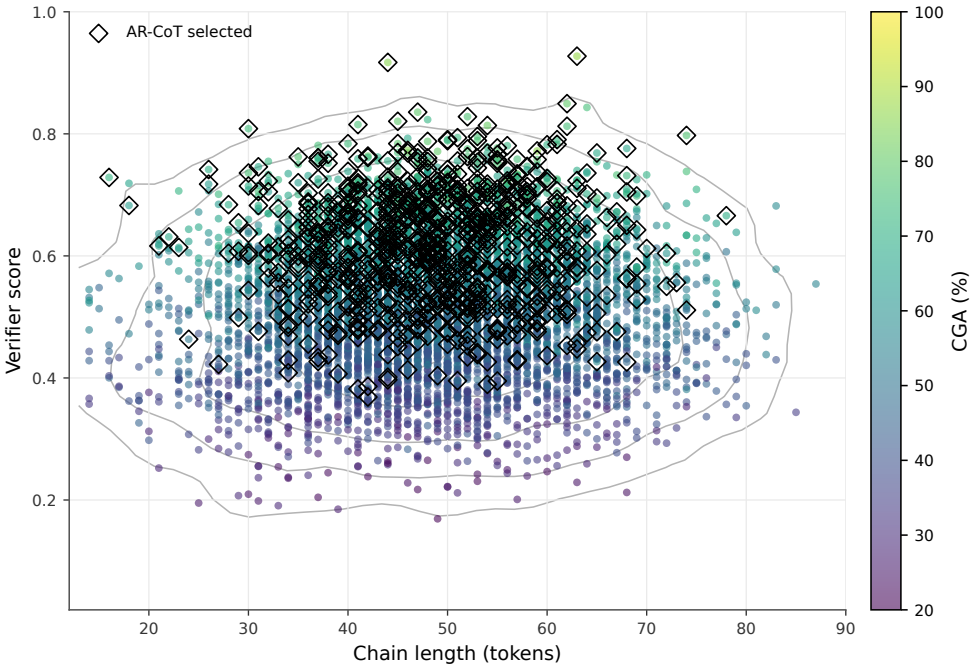


Figure 8: **Verifier score versus chain length.** Each point denotes one candidate chain from the decoding pool. The x-axis shows chain length, and the y-axis shows the chain-level verifier score. Point color indicates ground-truth chain grounding accuracy (CGA), and gray contours show candidate density. Black diamond markers denote the candidates selected by AR-CoT. High-verifier candidates are distributed across a broad range of medium-to-long chains, rather than collapsing to the shortest region. Moreover, CGA increases primarily along the verifier axis rather than the length axis, suggesting that the verifier captures evidence support instead of superficial brevity.

Verifier length and template-bias analysis. A natural concern is that the local verifier may assign high scores to chains that are simply shorter or more formulaic, rather than genuinely better grounded in scene evidence. To test this, we analyze the full candidate pool used by gain-based decoding. For each question, we sample $K = 8$ candidate answer-chain pairs $y_k = (a_k, \tau_k)$ from the scene-conditioned model and record three quantities: chain length $L(\tau_k)$, chain-level verifier score $\mathrm{Ver}(\tau_k, x)$, and ground-truth chain grounding accuracy $\mathrm{CGA}(\tau_k, \mathcal{E}^\star)$. This candidate-level score is the within-chain pointer accuracy against the question's annotated evidence, as in Eq. (50). Here, $L(\tau_k)$ counts non-special chain tokens after removing structural tags such as <STEP> and <PTR>. We then visualize all candidates in the 2D plane of chain length and verifier score, color each point by CGA, and highlight the subset selected by AR-CoT:

$$y^\star = \arg \max_{y_k \in \mathcal{C}(x,q)} G_\theta(y_k; x, q). \tag{85}$$

To explicitly address the "templatic chain" concern, we additionally compute a template-overlap statistic by matching each chain against a small bank of frequent reasoning skeletons and measuring its correlation with the verifier score.

Figure 8 provides three observations. First, verifier score is not concentrated in the shortest-chain region. Instead, high-verifier candidates appear across a wide range of medium-length and moderately long chains, which argues against a simple short-chain preference. Second, CGA rises primarily with verifier score rather than with chain length: the correlation between chain length and verifier score is weak ($r \approx -0.08$), while the correlation between verifier score and CGA is substantially stronger ($r \approx 0.74$). Third, AR-CoT selected samples are not the shortest ones. Their average length is slightly higher than the overall candidate-pool average (about 44.6 vs. 42.1 tokens), and they remain distributed across the high-verifier region rather than collapsing to minimal-length chains. Finally, the correlation between template overlap and verifier score remains small ($r \approx 0.10$), suggesting that

the verifier is not merely rewarding templatic phrasing. Overall, this analysis supports that the verifier primarily tracks evidence support, not superficial brevity or stylistic regularity.

D Broader Impacts

This work aims to improve the faithfulness and reliability of 3D multimodal language models by encouraging their reasoning chains to depend on scene-specific object and relation evidence rather than language-only priors. Potential positive impacts include more reliable 3D question answering, embodied assistance, indoor robotics, augmented reality, and human–AI interaction systems [Fan et al., 2026a], especially in settings where unsupported spatial claims may lead to user confusion or unsafe downstream decisions [He et al., 2026, Huang et al., 2026], alongside broader efforts on interpretability and risk assessment in autonomous driving and cybersecurity [Ding et al., 2025, Shao, 2026b,a, Yang et al., 2026a, Fan et al., 2023].

Beyond 3D scenes, Hu [2026] proposes ARSM-Agent for medical decision support, combining medical-evidence constraints, knowledge-consistency verification, and confidence reweighting to improve adversarial robustness. In cybersecurity, Wang et al. [2026b] introduce Trident, which combines structural evidence from heterogeneous threat knowledge graphs with attack-technique sequences for threat attribution. Ma et al. [2026] present MuSK, which learns multi-scale provenance-graph representations and verifies suspicious nodes using contextual subgraph evidence.

Workflow-aware healthcare-agent evaluation likewise emphasizes state continuity and evidence traceability beyond static answer quality [Yao et al., 2026].

In bioacoustics, USVexplorer studies robust ultrasonic-vocalization detection with cross-species generalization [Wei et al., 2026].

E LLM Usage

During the research process, LLM-based tools were used only in an assistive capacity, such as improving wording, checking grammar, formatting LaTeX, drafting code snippets, and debugging scripts.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contribution of AR-CoT: improving faithful 3D multimodal reasoning by selecting reasoning chains that are supported by scene-specific object and relation evidence rather than by language-only rationalization. The claimed components, including explicit evidence pointers, a frozen text-only rationale prior, local evidence verification, decoy-twin contrast, and scene-evidence-gain decoding, are defined in the method section and evaluated on standard 3D QA, grounding, faithful-reasoning, and shortcut-sensitive benchmarks.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion section discusses future directions.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A].

   Justification: The paper does not claim formal theoretical guarantees or proof-based results. The mathematical content defines the AR-CoT model, evidence graph, verifier, scene-evidence gain, contrastive objective, training objective, and evaluation metrics, but no theorem or formal proof is presented as a main contribution.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the benchmark suite, official data splits, preprocessing protocol, standard QA and grounding metrics, faithful-reasoning metrics, stress-test construction, controlled same-backbone comparisons, published-reference comparisons, backbones, LoRA training setup, optimizer, hyperparameter grids, decoding settings, verifier training, leakage-control protocol, and ablation variants. The method and appendix specify the training and inference procedures needed to reproduce the proposed algorithm.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: This paper does not include a public code release. The experiments are conducted on existing public or access-controlled benchmarks, and the paper provides detailed algorithmic, implementation, preprocessing, and evaluation details. We plan to release code, scripts, and derived stress-test metadata upon publication, subject to the licenses and terms of the underlying datasets and models.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup and appendix specify the datasets, official splits, answer normalization, grounding thresholds, faithful-reasoning metrics, same-backbone comparison protocol, backbone choices, LoRA rank and dropout, trainable modules, optimizer, learning rates, batch size, mixed precision, hyperparameter search grids, validation-based model selection, candidate-generation budget, beam size, sampling temperature, maximum chain length, and verifier training.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Controlled same-backbone experiments, stress tests, and ablations are reported over three random seeds with mean and standard deviation. The paper explicitly states that standard deviations are computed across independent runs with different random seeds. Published-reference comparisons are used only for contextual comparison when protocols are compatible, and they are not treated as controlled significance comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B.4 reports the hardware type.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: The work uses existing 3D vision-language benchmarks and pretrained research backbones under their respective terms, does not collect new human-subject data, and does not release a new high-risk pretrained model or scraped dataset. The method aims to reduce unsupported spatial claims in 3D multimodal reasoning.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix D discusses positive impacts such as more reliable 3D question answering, embodied assistance, indoor robotics, augmented reality, and human–AI interaction.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release a new pretrained foundation model, image generator, scraped dataset, or other standalone high-risk asset. The proposed contribution is an algorithmic training and decoding framework evaluated on existing benchmarks. Any released code or derived metadata will be distributed only under the licenses and terms of the underlying assets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the datasets, baselines, and pretrained backbones used in the experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The paper does not introduce or release a new dataset, benchmark, or other standalone asset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

Justification: The paper does not involve crowdsourcing experiments or new research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve new human-subject research, crowdsourcing, or collection of participant data, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper studies 3D multimodal language models and explicitly describes how pretrained 3D MLLM backbones and a frozen text-only rationale prior are used in the method. Appendix E further discloses that external LLM-based tools were used only for assistive purposes such as wording, grammar checking, LaTeX formatting, drafting code snippets, and debugging scripts.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.