KG-REFLECTGROUND: EVIDENCE-GUIDED ZERO-SHOT 3D VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot 3D visual grounding (3DVG) requires reliable spatial reasoning without task-specific training, yet achieving both high grounding accuracy and computational efficiency remains challenging. Existing LLM-based approaches generally incur lower inference costs but remain limited in capturing visually critical cues, whereas MLLM-based approaches provide stronger visual disambiguation at the expense of substantially higher computational overhead. To reconcile these competing requirements, we propose KG-ReflectGround, an efficient zero-shot 3DVG framework that adaptively allocates spatial reasoning and visual verification according to evidence gain. The framework constructs a query-aware sparse spatial knowledge graph over candidate 3D object instances, emphasizing discriminative relations to provide compact, query-relevant structural evidence. Relation-ordered evidence-gain reasoning progressively refines the candidate set and adaptively halts when additional relational evidence provides limited benefit, thereby lowering the computational cost. An evidence-aware verification policy assesses the sufficiency of structural evidence and selectively invokes MLLM-based visual verification only for ambiguous or appearance-dependent queries, allowing visual evidence to complement spatial reasoning where needed. Experiments on ScanRefer and Nr3D demonstrate strong zero-shot grounding performance with reduced multimodal computation. KG-ReflectGround achieves 52.0% Overall Acc@0.25 and 47.8% Overall Acc@0.5 on ScanRefer, with an average inference latency of 3.48 s per query.

1 INTRODUCTION

Understanding natural language instructions in 3D environments is essential for embodied agents and intelligent robots (Anderson et al., 2018; Shridhar et al., 2020). A robot may be asked to "find the red chair next to the table under the window," which requires jointly resolving object semantics, appearance attributes, and spatial relations. 3DVG addresses this problem by localizing the referred object in a 3D scene, providing an important interface between language understanding and robotic perception. Recent 3D scene understanding benchmarks have enabled systematic evaluation of this capability, such as ScanNet (Dai et al., 2017), together with language-guided grounding datasets including ScanRefer (Chen et al., 2020) and ReferIt3D (Achlioptas et al., 2020).

Most existing 3DVG methods rely on task-specific training with densely annotated language-object pairs and can be broadly grouped according to their primary modeling focus. (1) *Proposal- and selection-based methods* aim to identify the referred object through candidate evaluation and progressive selection. For example, 3D-SPS (Luo et al., 2022) performs referred-point progressive selection to focus on target-relevant regions, while 3DRefTransformer (Abdelreheem et al., 2022) models fine-grained object-language correspondence for candidate identification. However, such methods still depend heavily on large-scale annotated language-object pairs and task-specific optimization. With an emphasis on cross-modal representation learning, (2) *Cross-modal representation and alignment methods* seek stronger correspondence between language and 3D scenes MVT (Huang et al., 2022) aggregates multi-view information for language-guided grounding, whereas 3D-VisTA (Zhu et al., 2023) exploits large-scale vision-text pretraining to improve 3D visual-language alignment. Nevertheless, the learned cross-modal associations are often tied to the training distribution, which limits generalization to unseen scenes, novel object compositions, and diverse referring expressions. (3) *Context- and interaction-based methods* further enhance ground-

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

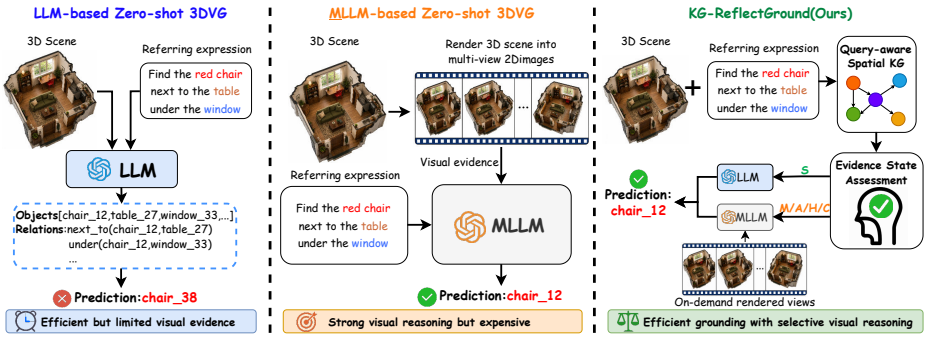


Figure 1: Comparison of zero-shot 3D visual grounding paradigms. LLM-based methods are efficient but use limited visual evidence; MLLM-based methods offer richer visual reasoning at higher cost. KG-ReflectGround combines query-aware spatial reasoning with selective visual verification to balance accuracy and efficiency.

ing by modeling object relations or suppressing irrelevant scene information. MCLN (Qian et al., 2024) introduces multi-branch collaborative learning to strengthen cross-modal interaction, while TSP3D (Guo et al., 2025) uses text-guided sparse voxel pruning to retain task-relevant 3D features more efficiently. Despite their strong performance on closed-set benchmarks, retraining or fine-tuning is typically required when transferring to new datasets or deployment domains. Zero-shot 3DVG therefore provides an attractive alternative by leveraging foundation models to perform grounding without task-specific training on target datasets.

As illustrated in Figure 1, LLM-based methods perform efficient language-guided reasoning over structured scene descriptions (Yang et al., 2023; Yuan et al., 2024), but lack direct visual evidence. MLLM-based methods improve appearance understanding by incorporating visual observations (Xu et al., 2025; Li et al., 2025), yet repeated rendering and multimodal inference introduce substantial computational overhead. Moreover, many methods either ignore explicit 3D structure or apply fixed-depth reasoning to all queries, wasting computation on simple cases while remaining insufficient for ambiguous ones.

To address these limitations, we propose KG-ReflectGround, an efficient zero-shot 3DVG framework based on evidence-gain-guided spatial graph reasoning. The framework consists of three main stages. *Query-aware Sparse Spatial Knowledge Graph Construction* builds a compact structural representation over candidate 3D object instances by pruning redundant spatial connections, preserving uncertainty in spatial relations, and emphasizing query-relevant evidence. *Relation-Ordered Evidence-Gain Reasoning* progressively refines the candidate set by prioritizing more discriminative spatial relations, since different constraints contribute unequally to candidate elimination. To avoid fixed-depth reasoning across queries of different difficulty, the framework dynamically estimates the benefit of further reasoning and terminates once additional relational evidence provides limited improvement. *Evidence-Aware Adaptive Verification* then evaluates whether structural evidence is sufficient and selectively invokes multimodal visual verification for ambiguous or appearance-dependent cases.

Experiments on ScanRefer and Nr3D demonstrate that KG-ReflectGround achieves strong zero-shot grounding performance while reducing unnecessary multimodal computation. Using Qwen3-VL-8B, our method obtains 52.0% Overall Acc@0.25 and 47.8% Overall Acc@0.5 on ScanRefer.

The main contributions are:

- We propose KG-ReflectGround, an efficient zero-shot 3DVG framework that combines explicit spatial reasoning with selective multimodal verification.
- We design a query-aware sparse spatial knowledge graph and relation-ordered evidence-gain reasoning with adaptive halting for efficient and interpretable grounding.
- We develop an evidence-aware verification strategy that selectively invokes MLLM reasoning only when structural evidence is insufficient.

- Extensive experiments on ScanRefer and Nr3D demonstrate that KG-ReflectGround achieves strong zero-shot grounding performance while maintaining high computational efficiency.

2 Related Work

Existing 3D visual grounding methods can be broadly divided into two categories according to whether task-specific grounding supervision is required. Supervised approaches learn language-object alignment from annotated 3D scenes, whereas zero-shot methods exploit foundation models to transfer semantic and visual knowledge without training on the target grounding datasets.

2.1 Supervised 3D Visual Grounding

3D visual grounding aims to localize a referred object in a 3D scene according to a natural-language expression. ScanRefer (Chen et al., 2020) and ReferIt3D (Achlioptas et al., 2020) established widely used benchmarks for evaluating language-guided object localization in reconstructed indoor environments. Early approaches focused primarily on cross-modal alignment between language descriptions and 3D object proposals. Subsequent methods, including 3D-SPS (Luo et al., 2022), 3D-VisTA (Zhu et al., 2023), MCLN (Qian et al., 2024), and TSP3D (Guo et al., 2025), improve grounding through stronger object representations, contextual reasoning, and multimodal interaction. Despite their effectiveness, these approaches require large-scale annotated language-object pairs and task-specific training, limiting generalization to unseen scenes and referring expressions.

2.2 Zero-Shot 3D Grounding with Foundation Models

Recent foundation-model-based approaches enable zero-shot 3D visual grounding without task-specific training on target datasets. LLM-based methods, such as LLM-Grounder (Yang et al., 2023) and ZSVG3D (Yuan et al., 2024), transform 3D scenes into textual or structured representations for language-guided reasoning. Recent methods further exploit explicit relational structures, such as Chain-of-Semantics (Shi et al., 2025) and VoG (Liu et al., 2026), to support structured spatial reasoning. These approaches are computationally efficient but depend heavily on scene textualization and lack direct visual evidence for appearance-related ambiguities.

Recent multimodal approaches improve grounding by incorporating visual observations. VLM-Grounder (Xu et al., 2025) and SeeGround (Li et al., 2025) leverage rendered views for stronger visual evidence, while UZ3DVG (Tan et al., 2026) explores generated language conditions. However, repeated visual rendering or similar reasoning procedures across queries may introduce unnecessary computation. In contrast, KG-ReflectGround performs query-aware graph reasoning with adaptive halting and selectively activates MLLM-based verification only when structural evidence is insufficient.

3 Method

3.1 Overview

Given a reconstructed 3D scene and a referring expression, KG-ReflectGround identifies the target object without grounding-specific training. As illustrated in Figure 2, the framework consists of three stages. First, it performs Query-aware Sparse Spatial Knowledge Graph Construction over candidate 3D object instances. Second, Relation-Ordered Evidence-Gain Reasoning progressively refines the candidate set and adaptively terminates through Evidence Gain & Adaptive Halting when further relational evidence provides limited benefit. Finally, Evidence-Aware Adaptive Verification selects LLM-based Reasoning or MLLM-based Verification for the final prediction.

3.2 Problem Formulation

Under the ground-truth-object setting, a 3D scene is represented by a set of annotated candidate object instances

$$\mathcal{O} = \{o_i\}_{i=1}^{N}, \qquad o_i = (c_i, \mathbf{b}_i, \mathbf{p}_i, \mathbf{s}_i, \mathbf{v}_i), \tag{1}$$

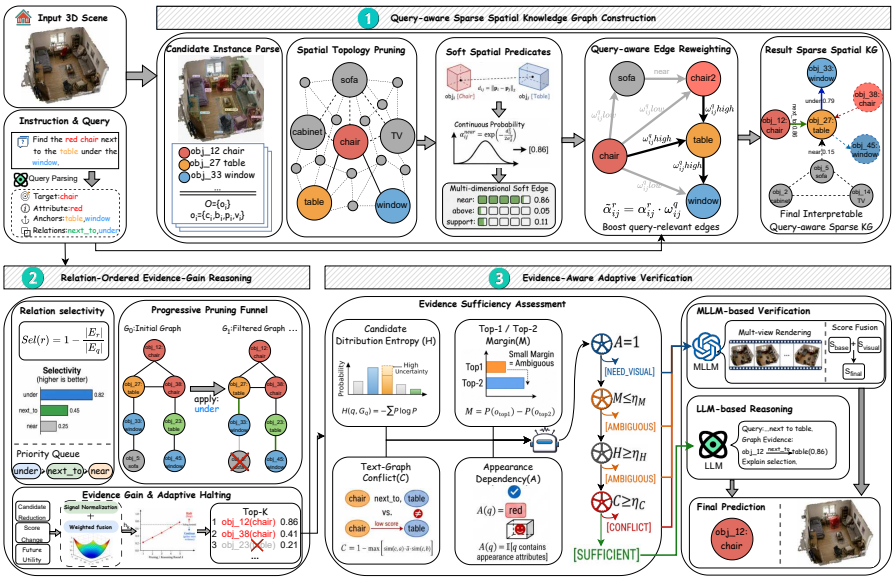


Figure 2: Overall framework of KG-ReflectGround. The framework constructs a query-aware sparse spatial knowledge graph over candidate object instances, performs relation-ordered evidence-gain-guided reasoning with adaptive halting, and applies an evidence-aware verification policy to select text-only LLM reasoning or MLLM-based visual verification.

where c_i, $\mathbf{b}_i$, $\mathbf{p}_i$, and $\mathbf{s}_i$ denote the semantic category, 3D bounding box, center position, and size of object o_i, respectively. Optional visual evidence is denoted by $\mathbf{v}_i$.

For a referring expression q, we first parse it into structured components using a lightweight LLM with a fixed zero-shot prompt:

$$\mathcal{P}(q) = \{q_{\text{tar}}, q_{\text{attr}}, \mathcal{A}_q, \mathcal{T}_q\}, \tag{2}$$

where q_{tar} denotes the target category, q_{attr} contains appearance attributes, $\mathcal{A}_q$ denotes anchor concepts, and $\mathcal{T}_q$ contains spatial-relation triplets. The same parsing prompt is used for all samples and contains no dataset-specific demonstrations.

3.3 Query-Aware Sparse Spatial Knowledge Graph

We construct a query-conditioned sparse spatial graph

$$G_q = (V, E_q, R), \quad V = O,$$

where R denotes spatial predicates and E_q contains query-relevant weighted edges. Unlike learned scene graph construction approaches, our graph is generated directly from object geometry and semantic compatibility without grounding-specific training, enabling zero-shot transfer across unseen scenes.

3.3.1 Spatial Topology Pruning

Instead of densely connecting all object pairs, each object is connected only to its K_{nn} nearest spatial neighbors:

$$\mathcal{N}_{K_{\text{nn}}}(o_i) = \text{TopK}_{o_j \in \mathcal{O} \setminus \{o_i\}} \left(-\|\mathbf{p}_i - \mathbf{p}_j\|_2\right). \tag{3}$$

This sparse topology removes redundant long-range connections while preserving local relational evidence relevant to language-guided grounding.

For each retained object pair, we construct a geometric descriptor

$$\mathbf{g}_{ij} = [\Delta\mathbf{p}_{ij}, d_{ij}, \text{IoU}_{xy}(\mathbf{b}_i, \mathbf{b}_j), \Delta z_{ij}, \mathbf{s}_i \oslash \mathbf{s}_j], \tag{4}$$

where $\Delta\mathbf{p}_{ij} = \mathbf{p}_j - \mathbf{p}_i$, $d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|_2$, IoU_{xy} denotes the overlap of the two boxes projected onto the horizontal plane, $\Delta z_{ij} = z_j - z_i$, and $\oslash$ denotes element-wise division.

3.3.2 SOFT SPATIAL PREDICATES

Hard relation assignment is sensitive to spatial boundary cases and geometric ambiguity. We therefore represent each relation as a soft distribution over possible spatial predicates:

$$\alpha_{ij}^{r} = \frac{\exp(\phi_r(\mathbf{g}_{ij})/T_r)}{\sum_{r' \in \mathcal{R}} \exp(\phi_{r'}(\mathbf{g}_{ij})/T_{r'})}, \tag{5}$$

where $\phi_r(\cdot)$ is a deterministic relation-specific geometric score computed from object geometry, and T_r controls distribution softness. No relation classifier is trained on ScanRefer or Nr3D, allowing ambiguous object pairs to retain multiple compatible spatial hypotheses.

3.3.3 QUERY-AWARE EDGE REWEIGHTING

The relevance of edge (o_i, r, o_j) to query q is defined as

$$\omega_{ij}^{q,r} = \lambda_t \, s(c_i, q_{\text{tar}}) + \lambda_a \, s(c_j, \mathcal{A}_q) + \lambda_r \, s(r, \mathcal{T}_q), \tag{6}$$

where $s(\cdot, \cdot)$ denotes cosine similarity computed from frozen vision-language embeddings (Radford et al., 2021). The final query-aware edge confidence is

$$\widetilde{\alpha}_{ij}^{r} = \alpha_{ij}^{r} \, \sigma\big(\omega_{ij}^{q,r}\big) \, . \tag{7}$$

This operation suppresses irrelevant relations and emphasizes edges aligned with the target, anchors, and queried predicates, yielding a compact query-aware representation for subsequent reasoning.

3.4 RELATION-ORDERED EVIDENCE-GAIN REASONING

A referring expression may contain multiple spatial constraints with different discriminative abilities. We define the evidence gain of a reasoning step as its expected benefit in reducing candidate ambiguity or improving candidate discrimination. Since not every relation provides the same gain, KG-ReflectGround processes discriminative relations first and continues reasoning only while further relations are expected to provide useful evidence.

3.4.1 RELATION SELECTIVITY ORDERING

To estimate the potential evidence gain of each relation, we define its selectivity as

$$\mathrm{Sel}(r) = 1 - \frac{|\mathcal{E}_r|}{|\mathcal{E}_q|}, \tag{8}$$

where $\mathcal{E}_r$ is the set of query-aware edges associated with relation r. We use $\mathrm{Sel}(r)$ as a heuristic for relation ordering: relations supported by fewer graph edges receive higher priority. Edge frequency does not directly measure candidate elimination; the reduction achieved after applying a relation is measured separately by G_k in Eq. (12).

The queried relations are ordered as

$$\boldsymbol{\pi} = \mathrm{Argsort}_{r \in \mathcal{T}_q} \big(-\mathrm{Sel}(r) \big). \tag{9}$$

The resulting order π places relations with higher selectivity at earlier reasoning stages.

3.4.2 PROGRESSIVE CANDIDATE REFINEMENT

Starting from candidates compatible with the target category, the candidate set is progressively refined by accumulating evidence from ordered relation constraints:

$$\mathcal{C}_k = \{ o_i \in \mathcal{C}_{k-1} \mid m(o_i, t_{\pi_k}, \mathcal{G}_q) \geq \eta_r \} \, , \tag{10}$$

where $m(\cdot)$ measures the compatibility between a candidate and the current relation triplet based on query-aware graph evidence.

Candidate evidence is accumulated over the processed relations:

$$s_i^k = s_i^{k-1} + \widetilde{\alpha}_{i\, a_k}^{r_k} s(c_i, q_{\text{tar}}), \tag{11}$$

where a_k is the matched anchor object for relation r_k. The accumulated score integrates evidence obtained from multiple relations, allowing candidates supported by complementary spatial constraints to progressively increase their confidence.

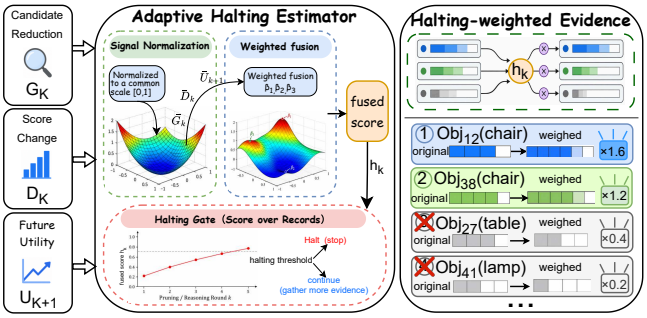


Figure 3: Evidence-gain estimation and adaptive halting. Candidate reduction, score change, and future utility determine the halting score.

Candidate reduction, score change, and future utility measure the benefit of each reasoning step. These signals are normalized and combined into a halting score. As new relations provide less useful evidence, the score rises toward the stopping threshold. Reasoning then stops, and the accumulated evidence is used to rank the remaining candidates. This lets the method avoid extra steps when the candidate ranking is sufficiently stable.

3.4.3 Evidence-Gain Estimation and Adaptive Halting

Instead of adopting a fixed reasoning depth, we estimate the evidence gain of additional graph reasoning from three complementary signals: candidate reduction, score change, and future utility. Candidate reduction is defined as

$$G_k = \frac{|\mathcal{C}_{k-1}| - |\mathcal{C}_k|}{|\mathcal{C}_{k-1}| + \epsilon}. \tag{12}$$

A high G_k indicates that the current reasoning step provides strong evidence by removing candidates inconsistent with the spatial constraints. Score change is measured by

$$D_k = \frac{1}{|\mathcal{C}_k|} \sum_{o_i \in \mathcal{C}_k} |s_i^k - s_i^{k-1}|. \tag{13}$$

A decreasing D_k indicates that newly introduced evidence no longer changes candidate preference significantly, suggesting diminishing returns from further reasoning.

The expected utility of the next relation is

$$U_{k+1} = \mathrm{Sel}(r_{\pi_{k+1}}) \, \mathrm{Cov}(r_{\pi_{k+1}}, \mathcal{C}_k), \tag{14}$$

where $\mathrm{Cov}(\cdot)$ measures the fraction of remaining candidates affected by the next relation. This term estimates the expected evidence gain of remaining relations before executing the next reasoning step.

Before fusion, the three utility signals are independently normalized to $[0, 1]$, yielding $\bar{G}_k$, $\bar{D}_k$, and $\bar{U}_{k+1}$. We then compute a halting score that estimates whether further reasoning is unlikely to provide sufficient additional evidence:

$$\begin{aligned} z_k &= \beta_1(1 - \bar{G}_k) + \beta_2(1 - \bar{D}_k) + \beta_3(1 - \bar{U}_{k+1}), \\ h_k &= \sigma(\gamma_h(z_k - b_h)) \, . \end{aligned} \tag{15}$$

Here, $\gamma_h > 0$ and b_h are fixed scale and offset parameters for the halting score. Reasoning terminates when $h_k \geq \tau$ or when no unprocessed relation remains. A high h_k indicates that the available discriminative evidence has largely been exhausted and that additional reasoning is unlikely to improve the current ranking.

3.5 Evidence-Aware Adaptive Verification

Unlike a learned routing network, our proposed verification policy does not introduce additional training parameters. Instead, it performs evidence sufficiency assessment directly from the intermediate grounding state. Graph reasoning is efficient but may remain insufficient when multiple candidates satisfy similar spatial constraints or when the query contains appearance-dependent attributes. Since MLLM-based verification introduces additional computational cost, KG-ReflectGround first

evaluates evidence sufficiency and activates visual verification only when structural reasoning cannot provide reliable grounding.

To quantify prediction uncertainty, we normalize candidate scores as

$$P_i = \frac{\exp(s_i/T_s)}{\sum_{o_j \in \mathcal{C}} \exp(s_j/T_s)}. \tag{16}$$

We assess evidence sufficiency using candidate entropy H, ranking margin M, text–graph conflict C, and appearance dependency A:

$$\begin{aligned} H &= -\sum_i P_i \log P_i, \qquad M = P_{(1)} - P_{(2)}, \\ C &= 1 - \max_{(i,r,j)} \left[s(c_i, a) \tilde{\alpha}^r_{ij} s(c_j, b) \right], \qquad A = \mathbb{I}[q_{\text{attr}} \neq \emptyset]. \end{aligned} \tag{17}$$

MLLM verification is invoked if $A = 1$, $H > \eta_H$, $M < \eta_M$, or $C > \eta_C$. It uses object-centered rendered views alongside the candidate set, matched relation paths, and graph confidences; otherwise, reasoning uses structural evidence alone. The final prediction is

$$S_i = S_i^{\text{text}} + \lambda_g S_i^{\text{graph}} + \mathbb{I}[z(q) = \text{MLLM}]\, \lambda_v S_i^{\text{visual}}, \qquad \hat{o} = \arg\max_{o_i \in \mathcal{C}} S_i. \tag{18}$$

4 Experiments

4.1 Experimental Setup

Detailed experimental settings, including the datasets, backbone, candidate objects, hyperparameters, and hardware, are provided in Appendix A.1.

4.2 Comparison with State-of-the-Art Methods

Results on ScanRefer. Table 1 compares KG-ReflectGround with supervised and zero-shot 3D visual grounding methods. Supervised methods are included as references because they use task-specific training. Using Qwen3-VL-8B, KG-ReflectGround achieves Acc@0.25/Acc@0.5 of 80.9/77.6, 40.1/37.7, and 52.0/47.8 on the Unique, Multiple, and Overall subsets, respectively. Compared with SeeGround using the same backbone, it improves Overall Acc@0.25 and Acc@0.5 by 4.6 and 3.7 percentage points. The gains on the Multiple subset, where same-category objects must be distinguished, support the value of query-aware spatial reasoning. KG-ReflectGround also outperforms the Qwen2-VL-72B variant, suggesting that its gains are not solely due to backbone size.

Table 1: Comparison on ScanRefer. We report Acc@0.25/Acc@0.5 under Unique, Multiple, and Overall settings. Best zero-shot results are shown in bold and second-best results are underlined.

Method	Venue	Sup.	Agent	Unique Acc@0.25/0.5	Multiple Acc@0.25/0.5	Overall Acc@0.25/0.5
Fully Supervised						
G3-LQ (Wang et al., 2024)	CVPR'24	Fully	–	88.6/73.3	50.2/39.7	56.0/44.7
AugRefer (Wang et al., 2025b)	AAAI'25	Fully	–	86.2/70.8	50.0/39.1	55.7/44.0
TSP3D (Guo et al., 2025)	CVPR'25	Fully	–	87.5/64.6	51.2/45.3	56.5/46.7
S^2-MLLM (Xu et al., 2026)	CVPR'26	Fully	LLaVA-V-7B	87.4/77.8	52.4/46.6	59.2/52.7
Weakly Supervised						
WS-3DVG (Wang et al., 2023)	ICCV'23	Weakly	–	–/–	–/–	27.4/22.0
Zero-Shot						
ZSVG3D (Yuan et al., 2024)	CVPR'24	Zero-Shot	GPT-4 turbo	63.8/58.4	27.7/24.6	36.4/32.7
VLM-Grounder (Xu et al., 2025)	CoRL'24	Zero-Shot	GPT-4o	66.0/29.8	48.3/33.5	51.6/32.8
UZ3DVG (Tan et al., 2026)	CVPR'26	Zero-Shot	Qwen2-VL-72B	69.6/56.9	38.5/31.2	43.1/34.0
VoG (Liu et al., 2026)	AAAI'26	Zero-Shot	Qwen2-VL-72B	69.6/51.9	34.0/30.4	44.8/40.3
SeeGround (Li et al., 2025)	CVPR'25	Zero-Shot	Qwen2-VL-72B	75.7/68.9	34.0/30.0	44.1/39.4
SeeGround (Li et al., 2025)	CVPR'25	Zero-Shot	Qwen3-VL-8B	74.7/70.8	38.1/35.1	47.4/44.1
KG-ReflectGround	Ours	Zero-Shot	Qwen2-VL-72B	77.5/72.3	38.5/35.2	48.6/45.1
KG-ReflectGround	Ours	Zero-Shot	Qwen3-VL-8B	**80.9/77.6**	**40.1/37.7**	**52.0/47.8**

Table 2: Comparison on Nr3D. We report accuracy (%) on the Easy, Hard, View-Dependent (Dep), View-Independent (Inde), and Overall subsets. Best zero-shot results are shown in bold and second-best results are underlined.

Method	Venue	Agent	Easy	Hard	Dep	Inde	Overall
Fully Supervised							
BUTD-DETR (Jain et al., 2022)	ECCV'22	–	60.7	48.4	46.0	58.0	54.6
DDPA-3DVG (Gu et al., 2025)	IJCAI'25	–	–	–	–	–	54.4
S^2-MLLM (Xu et al., 2026)	CVPR'26	–	–	–	–	–	50.6
Weakly Supervised							
WS-3DVG (Wang et al., 2023)	ICCV'23	–	27.3	18.0	21.6	22.9	22.5
Ross3D (Wang et al., 2025a)	ICCV'25	–	–	23.7	–	–	39.0
Zero-Shot							
ZSVG3D (Yuan et al., 2024)	CVPR'24	–	46.5	31.7	36.8	40.0	39.0
VLM-Grounder (Xu et al., 2025)	CoRL'24	–	55.2	39.5	45.8	49.4	48.0
VoG (Liu et al., 2026)	AAAI'26	–	58.9	37.2	39.4	52.1	47.6
SeeGround (Li et al., 2025)	CVPR'25	Qwen2-VL-72B	54.5	38.3	42.3	48.2	46.1
SeeGround (Li et al., 2025)	CVPR'25	Qwen3-VL-8B	53.4	40.2	41.2	49.5	46.6
KG-ReflectGround	Ours	Qwen3-VL-8B	**59.6**	**40.5**	**46.9**	**52.8**	**49.2**

Results on Nr3D. KG-ReflectGround achieves 49.2% Overall accuracy, with 59.6% on Easy and 40.5% on Hard expressions. It reaches 46.9% and 52.8% on View-Dependent and View-Independent expressions, respectively. The Hard results show the benefit of spatial reasoning for similar candidates, while the View-Dependent results support selective visual verification.

4.3 Ablation Studies

Table 3: Ablation of query-aware spatial graph construction on ScanRefer. Overall Acc@0.25 and Acc@0.5 are reported.

Sparse	Soft Pred.	Query Rew.	Acc@0.25	Acc@0.5
×	×	×	30.1	23.8
✓	×	×	36.3	28.6
✓	✓	×	47.4	42.9
✓	✓	✓	**52.0**	**47.8**

Table 3 shows the effect of adding graph components in sequence. Sparse topology removes redundant connections and raises Acc@0.25 from 30.1 to 36.3. Soft spatial predicates yield the largest further gain, reaching 47.4 by representing spatial relations more gradually than binary edges.

Query-conditioned edge reweighting further improves grounding by emphasizing relations involving the target object, anchor concepts, and queried predicates, reaching 52.0 Acc@0.25 and 47.8 Acc@0.5.

Table 4: Ablation of evidence-guided reasoning and evidence-aware verification on ScanRefer.

Variant	Acc@0.25	Acc@0.5	MLLM (%)
w/o Relation Ordering	49.8	46.2	35.1
w/o Candidate Entropy	50.2	46.7	27.8
w/o Ranking Margin	50.4	47.0	28.5
w/o Text-Graph Conflict	49.9	46.3	27.1
w/o Appearance Dependency	50.1	46.5	25.9
Full Model	**52.0**	**47.8**	**31.6**

Table 4 shows that removing relation ordering lowers Acc@0.5 from 47.8 to 46.2. Ablating candidate entropy or ranking margin also reduces accuracy, highlighting their role in assessing uncertainty. The full model achieves Acc@0.25/Acc@0.5 of 52.0/47.8 while invoking the MLLM for 31.6% of queries.

Per-query latency and grounding accuracy are reported in Appendix A.6.

Figure 4 shows candidate pruning and declining evidence signals across reasoning steps. At step 4, the halting score exceeds $\tau = 0.70$ with two candidates remaining, triggering the stopping decision.

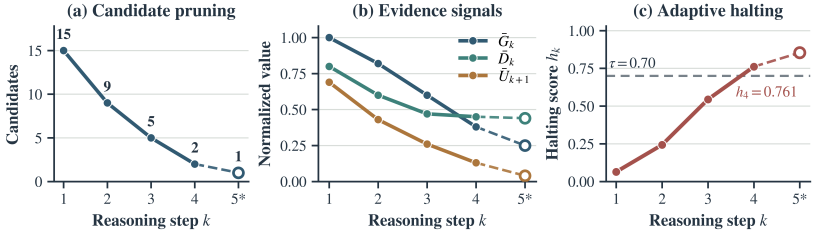


Figure 4: Reasoning dynamics for a ScanRefer query. Step 5^* shows a forced continuation after stopping at step 4.

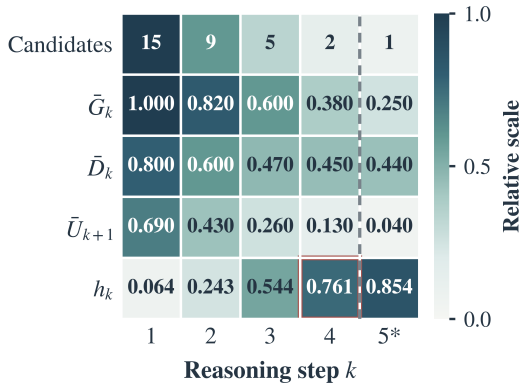


Figure 5: Candidate pruning and adaptive halting on a ScanRefer query. The outlined cell marks the step-4 stop; 5^* denotes forced continuation.

Figure 5 complements Figure 4 with an iteration-wise view of candidate pruning and adaptive halting. The candidate count falls from 15 to 9 and then to 5 in the first three steps, reaching 2 at step 4. Over the same steps, $\bar{G}_k$ decreases from 1.00 to 0.38 and $\bar{U}_{k+1}$ from 0.69 to 0.13, indicating diminishing utility from additional relations. The signals change at different rates: $\bar{D}_4$ remains at 0.45 even as future utility becomes small. At step 4, $h_4 = 0.761$ first exceeds $\tau = 0.70$. The method therefore stops when the combined evidence indicates limited value in further reasoning, although two candidates remain. Step 5^* shows a forced continuation after this stopping decision.

4.4 Qualitative Analysis

Figure 6 compares KG-ReflectGround with VLM-Grounder and SeeGround on four ScanRefer queries involving relations such as "in the middle," "opposite to," "in the center," and "between." Both baselines select nearby objects with plausible categories or appearances but fail to satisfy the specified spatial relations, especially when similar tables or chairs coexist. In contrast, KG-ReflectGround uses the queried relations and surrounding anchor objects to distinguish the target from these distractors. It produces the correct grounding in all four examples where the two baselines fail.

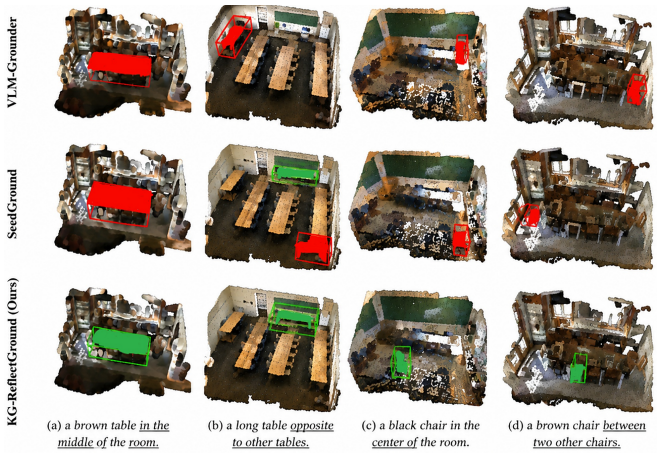


Figure 6: Qualitative grounding examples. Red and green boxes indicate incorrect and correct predictions, respectively.

These examples show that matching object categories or appearances alone is insufficient when several plausible candidates occur in the same scene. Relations to surrounding anchor objects provide the evidence needed to identify the intended table or chair.

5 Conclusion

This paper presents KG-ReflectGround, an efficient zero-shot 3D visual grounding framework that allocates spatial reasoning and visual verification according to evidence gain. The proposed method combines query-aware sparse graph construction, adaptive reasoning, and selective MLLM verification to reduce unnecessary multimodal computation while maintaining strong grounding accuracy. Experiments on ScanRefer and Nr3D validate the effectiveness of the framework, showing strong grounding performance with favorable computational efficiency. The gains stem from the complementary effects of query-aware spatial modeling, evidence-gain-guided adaptive reasoning, and selective visual verification, which together improve candidate discrimination and reduce inference cost.

ETHICS STATEMENT

This work studies zero-shot 3D visual grounding using the existing ScanRefer and Nr3D benchmarks. We conducted no new experiments involving human participants or animals and collected no new personal data. Our method uses annotated object instances and referring expressions from these benchmarks; it is evaluated on reconstructed indoor scenes rather than deployed in physical environments. Because indoor scans can contain details of private spaces, researchers using such data should follow the datasets' applicable access and usage conditions and avoid exposing sensitive scene content. Grounding errors may also affect downstream embodied systems, particularly for ambiguous instructions or visually similar objects. The reported benchmark results should therefore not be taken as evidence of reliability in safety-critical deployment. We have considered these limitations in presenting the method and its results, and the authors take responsibility for the ethical conduct and reporting of this research.

REPRODUCIBILITY STATEMENT

The method is specified in Section 3, including graph construction, relation ordering, candidate refinement, adaptive halting, and evidence-aware verification. Appendix A.1 describes the datasets, ground-truth-object evaluation setting, fixed model backbone, hyperparameters, and hardware. Appendix A.2 provides the complete inference procedure, while Appendices A.3–A.4 give supporting analyses. The main results and ablations are reported in Section 4; additional latency measurements and reasoning-depth comparisons appear in Appendices A.6–A.7. We use the same query-parsing prompt and training-free inference settings across ScanRefer and Nr3D, except for the component explicitly changed in each ablation. These descriptions are intended to make the evaluation protocol and implementation choices clear and to support independent reproduction of the reported results.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and editing parts of the manuscript, including improvements to wording, organization, and clarity. They were also used to provide feedback on methodological descriptions and the interpretation of results, to help check mathematical explanations and numerical consistency, and to assist with preparing and refining research visualizations. The authors reviewed the resulting text, calculations, and figures against the research implementation and experimental records before including them in the paper. Generative AI tools are also components of the proposed inference method, as described in Section 3; this methodological use is distinct from their use in preparing the manuscript. The authors take full responsibility for the final methods, experimental results, analysis, citations, and manuscript content.

REFERENCES

Ahmed Abdelreheem, Ujjwal Upadhyay, Ivan Skorokhodov, Rawan Al Yahya, Jun Chen, and Mohamed Elhoseiny. 3DRefTransformer: Fine-grained object identification in real-world scenes using natural language. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3941–3950. IEEE, 2022.

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

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683. IEEE, 2018.

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*, pp. 202–221. Springer, 2020.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5828–5839. IEEE, 2017.

Hongjie Gu, Jinlong Fan, Liang Zheng, Jing Zhang, and Yuxiang Yang. DDPA-3DVG: Vision-language dual-decoupling and progressive alignment for 3d visual grounding. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1044–1052, 2025.

Wenxuan Guo, Xiuwei Xu, Ziwei Wang, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Text-guided sparse voxel pruning for efficient 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3666–3675. IEEE, 2025.

Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15524–15533. IEEE, 2022.

Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pp. 417–433. Springer, 2022.

Rong Li, Shijie Li, Lingdong Kong, Xulei Yang, and Junwei Liang. Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3707–3717. IEEE, 2025.

Yuanyuan Liu, Haiyang Mei, Dongyang Zhan, Jiayue Zhao, Dongsheng Zhou, Bo Dong, and Xin Yang. View-on-graph: Zero-shot 3d visual grounding via vision-language reasoning on scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7386–7394, 2026.

Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3D-SPS: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16454–16463. IEEE, 2022.

Zhipeng Qian, Yiwei Ma, Zhekai Lin, Jiayi Ji, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Multi-branch collaborative learning network for 3d visual grounding. In *European Conference on Computer Vision*. Springer, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Jiaxin Shi, Mingyue Xiang, Hao Sun, Yixuan Huang, and Zhi Weng. Chain of semantics programming in 3d gaussian splatting representation for 3d vision grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24560–24569. IEEE, 2025.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10740–10749. IEEE, 2020.

Wenbin Tan, Jiawen Lin, Yuan Xie, Yachao Zhang, and Yanyun Qu. UZ3DVG: Unaided zero-shot 3d visual grounding with generated language conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9547–9557. IEEE, 2026.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9275–9286. IEEE, 2025a.

Xinyi Wang, Na Zhao, Zhiyuan Han, Dan Guo, and Xun Yang. Augrefer: Advancing 3d visual grounding via cross-modal augmentation and spatial relation-based referring. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8006–8014, 2025b.

Yuan Wang, Yali Li, and Shengjin Wang. G^3-LQ: Marrying hyperbolic alignment with explicit semantic-geometric modeling for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13917–13926. IEEE, 2024.

Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. Distilling coarse-to-fine semantic matching knowledge for weakly supervised 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2662–2671. IEEE, 2023.

Beining Xu, Siting Zhu, Zhao Jin, Junxian Li, and Hesheng Wang. S^2-MLLM: Boosting spatial reasoning capability of mllms for 3d visual grounding with structural guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2557–2569. IEEE, 2026.

Runsen Xu, Zhiwei Huang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. VLM-Grounder: A vlm agent for zero-shot 3d visual grounding. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 3961–3985, 2025.

Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F. Fouhey, and Joyce Chai. LLM-Grounder: Open-vocabulary 3d visual grounding with large language model as an agent. *arXiv preprint arXiv:2309.12311*, 2023.

Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20623–20633. IEEE, 2024.

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*, pp. 2911–2921. IEEE, 2023.

A APPENDIX

This appendix provides additional implementation details, complete inference procedures, theoretical analyses, and supplementary quantitative and qualitative results for KG-ReflectGround.

A.1 EXPERIMENTAL SETUP

Datasets. We evaluate KG-ReflectGround on ScanRefer and Nr3D. ScanRefer associates natural-language descriptions with objects in reconstructed ScanNet scenes and reports grounding results under the Unique, Multiple, and Overall settings. Nr3D contains human-generated referring expressions with substantial ambiguity among same-category objects.

Implementation setting. We use Qwen3-VL-8B as the fixed backbone for language reasoning and multimodal visual verification. Both datasets are evaluated under the ground-truth-object setting, using annotated ScanNet object instances as candidates. No grounding-specific detector, relation classifier, or routing network is trained on ScanRefer or Nr3D. The same training-free inference pipeline is applied across datasets.

Query parsing. Given a referring expression q, we parse it into $P(q) = \{q_{\mathrm{tar}}, q_{\mathrm{attr}}, A_q, T_q\}$ using a fixed zero-shot prompt. Here, q_{tar} denotes the target category, q_{attr} describes appearance attributes, A_q identifies anchor objects, and T_q specifies spatial relations. The same prompt template is used for all samples without dataset-specific demonstrations.

Sparse graph construction. We construct a query-aware sparse spatial graph over candidate objects. Each object is connected to its $K_{\mathrm{nn}} = 8$ nearest spatial neighbors, and pairwise relations are computed deterministically from object geometry. Query-conditioned edge reweighting uses $\lambda_t = 0.40$, $\lambda_a = 0.30$, and $\lambda_r = 0.30$ to balance target, anchor, and relation relevance.

Adaptive reasoning. The maximum reasoning depth is $K = 5$, and candidate refinement uses $\eta_r = 0.45$. Candidate reduction, score change, and future utility are independently normalized and fused with $\beta_1 = 0.40$, $\beta_2 = 0.30$, and $\beta_3 = 0.30$. Reasoning stops when the halting score reaches the threshold $\tau = 0.70$ or no unprocessed relation remains. The halting-score parameters are fixed at $\gamma_h = 7.38$ and $b_h = 0.517$.

Evidence-aware verification. Visual verification is triggered when $A = 1$, $H > \eta_H$, $M < \eta_M$, or $C > \eta_C$, where $\eta_H = 0.65$, $\eta_M = 0.15$, and $\eta_C = 0.45$. Otherwise, the final prediction is produced directly from structural reasoning.

Additional numerical settings. We set $T_r = 1.0$ for the spatial predicates in Eq. (5) and $T_s = 1.0$ for candidate-score normalization in Eq. (16). The final-score weights in Eq. (18) are $\lambda_g = \lambda_v = 1.0$. We use $\epsilon = 10^{-8}$ in Eq. (12).

Reproducibility and hardware. No model parameter is optimized using ScanRefer or Nr3D grounding annotations. The candidate construction, query-parsing prompt, graph construction rules, and reasoning and verification thresholds are shared across datasets. In each ablation, only the specified component is changed. All experiments are conducted on four NVIDIA A100 GPUs.

A.2 Complete Inference Algorithm

Algorithm 1 summarizes the complete inference procedure of KG-ReflectGround.

Algorithm 1 KG-ReflectGround Inference

Require: 3D object set $\mathcal{O}$, referring expression q
Ensure: Predicted target object $\hat{o}$

Parse query: $\mathcal{P}(q) = \{q_{\text{tar}}, q_{\text{attr}}, A_q, T_q\}$
Initialize candidate set $C_0 = \{o_i \in \mathcal{O} \mid c_i \text{ matches } q_{\text{tar}}\}$
Construct sparse K_{nn} spatial topology
for each retained object pair (o_i, o_j) **do**
    Compute geometric descriptor g_{ij}
    Compute soft spatial predicates α_{ij}^r
    Compute query-aware edge confidence $\tilde{\alpha}_{ij}^r$
end for
Compute relation selectivity $\text{Sel}(r)$
Sort queried relations into order π
Initialize candidate scores s_i^0
for $k = 1, \ldots, K$ **do**
    Select relation r_{π_k}
    Refine candidate set:

$$C_k = \{o_i \in C_{k-1} \mid m(o_i, t_{\pi_k}, G_q) \geq \eta_r\}$$

    Update candidate evidence scores s_i^k
    Compute candidate reduction G_k
    Compute score change D_k
    Estimate future utility U_{k+1}
    Normalize $\bar{G}_k, \bar{D}_k, \bar{U}_{k+1}$
    Compute $z_k \leftarrow \beta_1(1 - \bar{G}_k) + \beta_2(1 - \bar{D}_k) + \beta_3(1 - \bar{U}_{k+1})$ and $h_k \leftarrow \sigma(\gamma_h(z_k - b_h))$
    if $h_k \geq \tau$ or no unprocessed relation remains **then**
        break
    end if
end for
Compute evidence-sufficiency signals H, M, C, A
if $A = 1$ or $H > \eta_H$ or $M < \eta_M$ or $C > \eta_C$ **then**
    Perform MLLM-based visual verification
else
    Perform LLM-based structural reasoning
end if
Fuse available evidence scores
$\hat{o} = \arg\max_{o_i \in C} S_i$
return $\hat{o}$

A.3 Properties of Sparse Graph Reasoning

For a directed graph with N objects, retaining at most K_{nn} neighbors per object yields no more than $N \min(K_{\text{nn}}, N-1)$ edges, compared with $N(N-1)$ in a fully connected graph. Evaluating $|\mathcal{R}|$ predicates on these edges costs $O(NK_{\text{nn}}|\mathcal{R}|)$ after neighbor search.

Because Eq. (10) filters the previous candidate set, $C_k \subseteq C_{k-1}$ at every executed step. Removed candidates cannot re-enter at a later step; this does not imply that every retained candidate is correct.

A.4 Analysis of Evidence-Gain Adaptive Halting

Let the weighted lack-of-utility score be

$$z_k = \beta_1(1 - \bar{G}_k) + \beta_2(1 - \bar{D}_k) + \beta_3(1 - \bar{U}_{k+1}).$$

The halting score in Eq. (15) is

$$h_k = \sigma(\gamma_h(z_k - b_h)),$$

where $\sigma(x) = (1 + e^{-x})^{-1}$, $\gamma_h > 0$, and $\beta_1, \beta_2, \beta_3 > 0$.

The score is monotonically decreasing with respect to each normalized utility signal when the other signals are fixed. For example,

$$\frac{\partial h_k}{\partial \bar{G}_k} = -\gamma_h \beta_1 h_k(1 - h_k) < 0.$$

The same calculation applies to $\bar{D}_k$ and $\bar{U}_{k+1}$. Thus, as candidate reduction, score change, or expected future utility declines, its contribution to the halting score increases.

Because the sigmoid is strictly increasing, the stopping condition $h_k \geq \tau$ is equivalent to

$$z_k \geq b_h + \frac{1}{\gamma_h} \log \frac{\tau}{1 - \tau}.$$

With $\tau = 0.70$, $\gamma_h = 7.38$, and $b_h = 0.517$, the corresponding threshold on z_k is approximately 0.632. In the ScanRefer example shown in Figs. 4 and 5, $z_3 = 0.541$ gives $h_3 = 0.544 < 0.70$, whereas $z_4 = 0.674$ gives $h_4 = 0.761 > 0.70$. Reasoning therefore stops at step 4, although two candidates remain. The plotted step 5 is a forced continuation for analysis and is not part of the adaptive inference trajectory.

A.5 Evidence-Aware Verification Details

The verification policy combines four complementary signals that capture different failure modes of structural reasoning.

Candidate entropy. The entropy

$$H = -\sum_i P_i \log P_i$$

measures the uncertainty of the candidate distribution. A large entropy indicates that probability mass is distributed across several plausible objects and therefore no candidate is clearly preferred.

Ranking margin. The margin

$$M = P_{(1)} - P_{(2)}$$

measures the separation between the two most likely candidates. Even when the global entropy is moderate, a small margin indicates that the top two objects remain difficult to distinguish.

Text–graph conflict. The conflict score C measures whether the strongest graph-supported relation is sufficiently compatible with the referring expression. A large conflict score indicates that the current graph evidence cannot fully explain the language constraints.

Appearance dependency. The binary signal

$$A = \mathbb{I}[q_{\text{attr}} \neq \emptyset]$$

explicitly identifies queries containing appearance-dependent attributes. Such cases may require visual evidence even when the structural graph is otherwise confident.

These signals are complementary: H captures global candidate uncertainty, M focuses on local top-rank ambiguity, C measures language–structure inconsistency, and A detects an explicit need for appearance information.

Algorithm 2 Evidence-Aware Verification Policy

Require: Candidate scores $\{s_i\}$, query attributes q_{attr}, graph evidence G_q
Ensure: Verification route $z(q)$
Normalize candidate scores into $\{P_i\}$
Compute entropy $H = -\sum_i P_i \log P_i$
Compute ranking margin $M = P_{(1)} - P_{(2)}$
Compute text–graph conflict C
Set $A = \mathbb{I}[q_{\text{attr}} \neq \emptyset]$
if $A = 1$ **or** $H > \eta_H$ **or** $M < \eta_M$ **or** $C > \eta_C$ **then**
    $z(q) \leftarrow \text{MLLM}$
else
    $z(q) \leftarrow \text{LLM}$
end if
return $z(q)$

A.6 Efficiency Analysis

Table 5: Per-query latency and grounding accuracy of representative zero-shot 3DVG methods on ScanRefer.

Method	Latency (s)	Acc@0.25	Acc@0.5
LLM-Grounder	20.0	44.8	40.3
VLM-Grounder	50.0	51.6	32.8
ZSVG3D	7.14	36.4	32.7
SeeGround	5.00	47.4	44.1
KG-ReflectGround	**3.48**	**52.0**	**47.8**

KG-ReflectGround takes 3.48 s per query, compared with 5.00 s for SeeGround and 7.14 s for ZSVG3D, while achieving the highest Acc@0.25 and Acc@0.5 among the compared methods. This indicates a favorable accuracy–efficiency trade-off.

A.7 Additional Analysis of Adaptive Halting

Table 6 provides a detailed comparison between fixed-depth and adaptive reasoning. Using only one reasoning step yields the lowest computational cost but may terminate before sufficient spatial evidence has been accumulated. Increasing the fixed depth to three steps substantially improves grounding accuracy, indicating the importance of multi-relation reasoning.

Further increasing the reasoning depth to five steps, however, does not provide additional improvement. This suggests that once candidate rankings become stable, processing additional relations may introduce redundant evidence without improving discrimination. In contrast, the adaptive strategy achieves the best Overall Acc@0.25 and Acc@0.5 while requiring only 3.2 reasoning steps on average.

These results support the central motivation of evidence-gain-based reasoning: different queries require different amounts of relational evidence, and a fixed reasoning depth cannot optimally balance accuracy and computation across all samples.

Table 6: Comparison of fixed-depth reasoning and adaptive halting on ScanRefer. We report average reasoning steps, MLLM invocation rate, and Overall grounding accuracy.

Strategy	Setting	Avg. Steps	MLLM (%)	Acc@0.25	Acc@0.5
Fixed Depth	$K = 1$	1.0	23.8	47.8	44.3
Fixed Depth	$K = 3$	3.0	30.9	50.8	47.0
Fixed Depth	$K = 5$	5.0	38.7	50.7	46.9
Adaptive	$\tau = 0.70$	**3.2**	**31.6**	**52.0**	**47.8**

A.8 Additional Qualitative Results

Fig. 7 shows three ScanRefer queries for which SeeGround selects an incorrect object and KG-ReflectGround identifies the target. These scenes contain similar objects, making relations to surrounding objects important for resolving the referring expression. The per-query latency comparison is reported separately in Table 5.

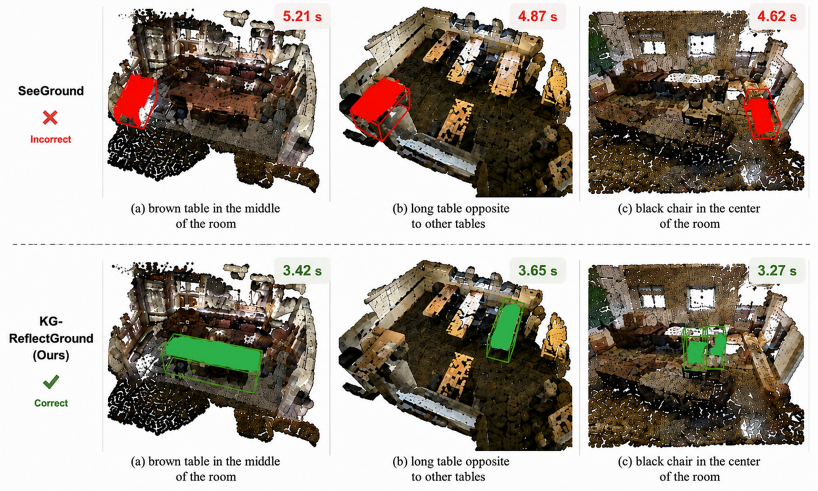


Figure 7: Comparison between SeeGround and KG-ReflectGround on ScanRefer.