View-on-Graph: Zero-Shot 3D Visual Grounding via Vision-Language Reasoning on Scene Graphs

Yuanyuan Liu[1], **Haiyang Mei**[1, 2], **Dongyang Zhan**[1], **Jiayue Zhao**[1], **Dongsheng Zhou**[3], **Bo Dong**[4], **Xin Yang**[1*]
[1]Key Laboratory of Social Computing and Cognitive Intelligence, Dalian University of Technology
[2]Show Lab, National University of Singapore
[3]Dalian University
[4]Cephia AI
liuyy990415@gmail.com, haiyang.mei@outlook.com, zhandongyang325@gmail.com, 2420458958@mail.dlut.edu.cn, zhoudongsheng@dlu.edu.cn, bo.dong@cephia.ai, xinyang@dlut.edu.cn

Abstract

3D visual grounding (3DVG) identifies objects in 3D scenes from language descriptions. Existing zero-shot approaches leverage 2D vision–language models (VLMs) by converting 3D spatial information (SI) into forms amenable to VLM processing, typically as composite inputs such as specified-view renderings or video sequences with overlaid object markers. However, this *VLM* $\oplus$ *SI* paradigm yields entangled visual representations that compel the VLM to process entire cluttered cues, making it hard to exploit spatial–semantic relationships effectively. In this work, we propose a new *VLM* $\otimes$ *SI* paradigm that externalizes the 3D SI into a form enabling the VLM to incrementally retrieve only what it needs during reasoning. We instantiate this paradigm with a novel **View-on-Graph (VoG)** method, which organizes the scene into a multi-modal, multi-layer scene graph and allows the VLM to operate as an active agent that selectively accesses necessary cues as it traverses the scene. This design offers two intrinsic advantages: (i) by structuring 3D context into a spatially and semantically coherent scene graph rather than confounding the VLM with densely entangled visual inputs, it lowers the VLM's reasoning difficulty; and (ii) by actively exploring and reasoning over the scene graph, it naturally produces transparent, step-by-step traces for interpretable 3DVG. Extensive experiments show that VoG achieves state-of-the-art zero-shot performance, establishing structured scene exploration as a promising strategy for advancing zero-shot 3DVG.

Code — https://github.com/YYLiuDLUT/VoG

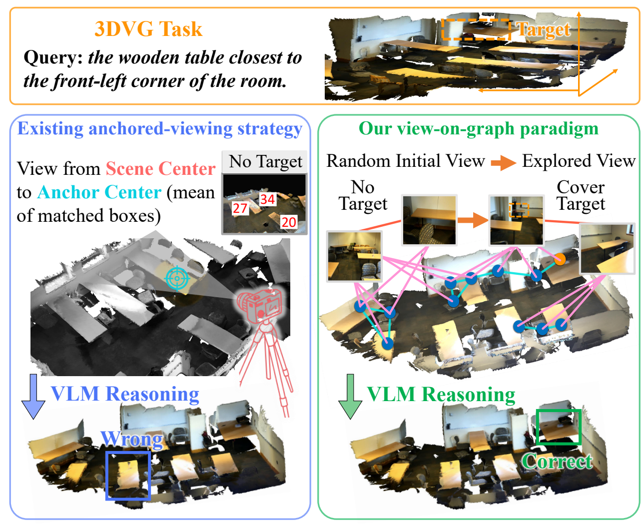

Figure 1: Comparison of zero-shot 3D visual grounding paradigms. **Left**: Passive, fixed-view processing paradigm. When the anchored-view observation misses the target, the VLM's reasoning over incomplete or misleading visual cues leads to failure. **Right**: Active, iterative exploration via our View-on-Graph (VoG) paradigm. By traversing the scene graph to navigate from misleading views toward informative observations, the VLM accurately locates the target and produces interpretable grounding traces.

Introduction

3D visual grounding (3DVG) aims to localize objects in 3D scenes from natural language descriptions, a key capability for applications such as augmented reality (Liu et al. 2023, 2025; Wei, Liu, and Luximon 2024), vision–language navigation (Chen et al. 2022b; Gong et al. 2024; Huang et al. 2022), and robotic perception (Chen et al. 2023; Hu et al. 2024; Kong et al. 2023). To solve this problem effectively, methods must integrate both textual understanding and spatial reasoning while coping with the complexity of 3D environments.

Recent work leverages 2D vision–language models (VLMs) for zero-shot 3DVG by converting 3D spatial information (SI) into forms amenable to VLM processing, typically as composite visual inputs such as specified-view renderings (Li et al. 2025) or video sequences (Qi et al. 2025) with overlaid object markers. We refer to this as the *VLM* $\oplus$ *SI* paradigm, where the VLM passively consumes the entire entangled SI, which forces it to wade through cluttered cues and makes it hard to exploit spatial–semantic re-

*Corresponding Author

lationships effectively. This naturally raises a question: *Can 3D spatial information be represented in a form that enables the VLM to incrementally retrieve only what it needs during its reasoning process?*

Much like how people use a search engine, we do not retain the entire internet in working memory. Instead, we retrieve relevant information step-by-step, progressively querying only what is necessary until the answer is found. Inspired by this principle of incremental retrieval, we design our zero-shot 3DVG approach to avoid overwhelming the VLM with the entire 3D SI upfront. Instead, we externalize the 3D SI into a structured, queryable scene graph (SG), enabling the VLM to operate as an active agent that selectively accesses spatial and semantic cues as it traverses the scene. This *VLM* $\otimes$ *SI* design inherently alleviates the reasoning difficulty associated with the entire entangled SI, maintains visual clarity, and supports more focused and effective reasoning.

To instantiate this paradigm, we design a novel **View-on-Graph (VoG)** method, which operates in two key stages. First, we transform the 3D spatial information into a multi-modal, multi-layer scene graph (MMMG) (Fig. 1). The MMMG consists of a view layer representing multi-view RGB images as nodes and an object layer representing detected 3D objects as nodes. Inter-layer edges connect views to the objects visible within them, while intra-layer edges link spatially adjacent viewpoints in the view layer and encode spatial relationships between objects in the object layer. Second, given a query, VoG enables the VLM to actively traverse this graph by alternating between exploration, where it moves along graph connections toward potentially informative viewpoints, and reasoning, where it cross-verifies observations and candidate objects from visited views. This iterative traversal continues until the target is confidently grounded or the search depth limit is reached, providing both targeted reasoning and interpretable grounding traces.

The advantage of this design is illustrated in Fig. 1. In our "VLM $\otimes$ SI" paradigm (right), the VLM can actively traverse the scene graph to move from less informative views toward more informative ones, even when the target is not visible from the starting viewpoint. This traversal capability allows the model to progressively refine its reasoning and achieve accurate grounding in large or visually complex scenes. In contrast, the existing "VLM $\oplus$ SI" paradigm (left) lacks such an exploration mechanism. When the anchored view misses critical visual evidence, the model is forced to reason over cluttered and potentially misleading cues, often leading to grounding failure.

In summary, this work makes the following contributions:

- We propose a new VLM $\otimes$ SI paradigm for zero-shot 3DVG, reframing the task as active and iterative scene exploration rather than passive processing of entire entangled 3D spatial information.
- We instantiate this paradigm with a novel View-on-Graph (VoG) method that structures 3D spatial information into a multi-modal, multi-layer scene graph and allows the VLM to autonomously traverse it during reasoning to identify critical cues for target grounding.
- We conduct extensive experiments validating the effectiveness of the VLM $\otimes$ SI paradigm and showing that VoG achieves state-of-the-art zero-shot 3DVG performance, while inherently providing interpretable and traceable grounding through its step-by-step exploration.

Related Work

3D Geometry-based Visual Grounding. 3DVG aims to localize target objects in unstructured point clouds conditioned on linguistic descriptions. Benchmarks such as ScanRefer (Chen, Chang, and Nießner 2020) and ReferIt3D (Achlioptas et al. 2020), built upon ScanNet (Dai et al. 2017), densely annotate object–sentence pairs for language-supervised 3D understanding. Early methods mainly rely on supervised training over geometric features and fall into two categories: two-stage pipelines (Yuan et al. 2021; Cai et al. 2022; Zhao et al. 2021), which generate proposals(Guan, Song, and Zhang 2024; Mei et al. 2021) before language matching, and single-stage pipelines (Wu et al. 2023; Wang, Li, and Wang 2024), which directly align 3D points with text in an end-to-end manner. Two-stage methods offer modularity but sacrifice efficiency and generalization, while single-stage ones depend less on proposals but remain limited by sparse geometry and missing appearance cues.

2D-enhanced Multi-modal Visual Grounding. To mitigate the visual limitations of point clouds, another line incorporates 2D imagery as auxiliary input (Zhang, Luo, and Lei 2024; Bakr, Alsaedy, and Elhoseiny 2022), enriching geometric context with high-resolution semantics. SAT (Yang et al. 2021) projects multi-view RGB features into 3D space and uses attention-based fusion for language alignment. OpenScene (Peng et al. 2023) fuses dense 2D semantics with 3D voxels, enabling reasoning about color and material. 3DLFVG (Zhang, Luo, and Lei 2024) transfers knowledge from 2D VLMs via cross-modal distillation, and LAR (Bakr, Alsaedy, and Elhoseiny 2022) models relations between projected 2D features and 3D points. While effective, these methods demand extra modalities and annotations, raising scalability and training complexity concerns.

Zero-/Few-shot Visual Grounding Methods. Recent trends shift toward few-shot (Wang et al. 2023) and zero-shot paradigms 3D-specific training (Kang et al. 2024; Zhu et al. 2024), aiming to eliminate the dependency on large-scale 3D training data. These approaches leverage powerful VLMs pre-trained on 2D image-text pairs to perform 3D grounding via cross-modal reasoning. For example, SeeGround (Li et al. 2025) introduces a render-and-reason framework, where 2D renderings of 3D scenes are used to query VLMs for localization, enabling open-vocabulary generalization. Despite their strong generalization capabilities, these methods passively feed the VLM with the entire entangled 3D spatial information, forcing it to reason over cluttered cues and rely on implicit spatial understanding. In contrast, we externalize 3D SI into a structured, queryable SG, enabling the VLM to incrementally retrieve only the spatial–semantic cues it needs during traversal.

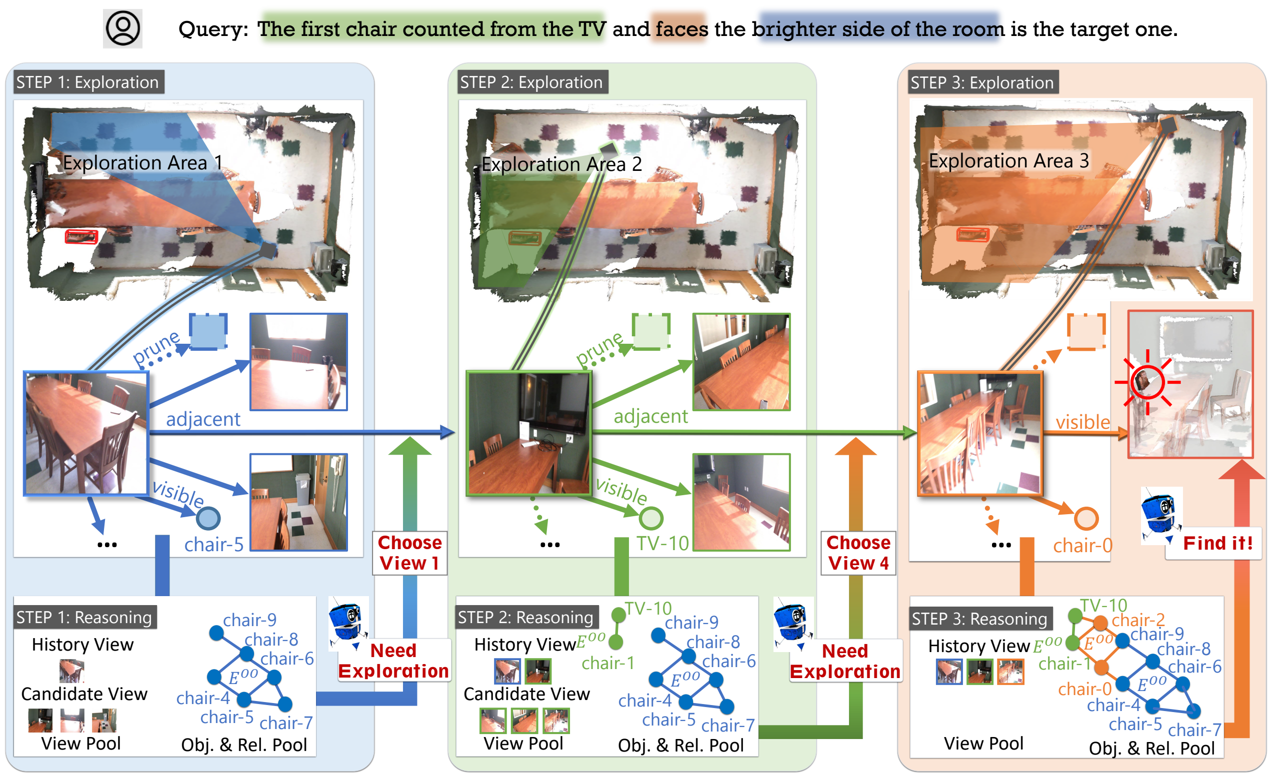

Figure 2: Workflow of VoG. Given the query, VoG identifies the target category (`chair`) and anchor (`TV`) as the search topic, then randomly selects an initial view where a chair is visible (View 0). From this view, it traverses neighboring nodes to form candidate nodes, which are fed to the VLM to decide whether further exploration is needed. The first exploration area reveals the "brighter side of the room" in the query, where multiple chairs are observed and added to the object pool. However, the anchor `TV` is missing, leaving the target unclear. The VLM then explores an area where the `TV` becomes visible (STEP 2) and observes "the first chair counted from the TV," adding `chair-1` to the pool. Ambiguity remains due to another nearby chair, prompting a final exploration to capture the full spatial layout. Once all query cues, including "faces the brighter side of the room" are confirmed, the VLM identifies `chair-1` as the target and terminates the search.

Methodology

We address zero-shot 3DVG by externalizing the 3D SI into a MMMG that separately encodes viewpoints, detected objects, and their spatial relations, rather than entangling them directly in the VLM's inputs. The 3DVG task is then reformulated as a viewpoint-guided traversal over this graph, where the VLM alternates between exploring informative viewpoints and reasoning over the accumulated spatial–semantic evidence (Fig. 2). The following sections detail the MMMG structure and the VoG framework.

Multi-Modal & Multi-Layered Scene Graph

SGs are widely used in robotics and scene reasoning tasks for efficient navigation, planning, and multi-view integration (Liu et al. 2022; Rana et al. 2023; Werby et al. 2024; Zhu et al. 2023a; Honerkamp et al. 2024; Koch et al. 2024a,b; Fang et al. 2025; Gao et al. 2024; Sun et al. 2023). We extend this idea to 3DVG by proposing MMMG that compactly represents each scene as:

$$G = \{V, O, E^{VV}, E^{VO}, E^{OO}\}, \qquad (1)$$

where V are view nodes with camera poses, O are object nodes with 3D boxes and semantics, and E^{VV}, E^{VO}, E^{OO} denote view–view connectivity, view–object visibility, and object–object spatial relations. These edges enable the VLM to reason across both visual and spatial modalities. The graph is built automatically from predicted 3D detections (Schult et al. 2022) using point-cloud distances and camera poses. This layered design keeps visual and spatial information explicitly separated, avoiding modality entanglement while aligning 2D view observations with structured, queryable 3D knowledge. Further details are provided in the Appendix. As illustrated in Fig. 2 (STEP 1–3), each image corresponds to a node v, which is linked to visible objects o via E^{VO} and connected to other views via E^{VV}. The exploration area represents the visible region of v.

View-On-Graph

Given the structured graph $\mathcal{M}_S$ and query q, VoG runs in three phases: initialization, exploration, and reasoning.

Initialization. Given a query, VoG leverages the underlying VLM to identify the target object O_T and candidate anchor objects $\{O_A\}$ (Li et al. 2025). VoG then treats the O_T as the search topic and traverses the entire $\mathcal{M}_S$ to collect all candidate views related to the topic. This yields an initial set of relevant views, from which one is randomly selected as the starting point. As illustrated in Fig. 2 (STEP 1), when the target is a `chair`, we randomly select the View 0 from the set of views in which chair is visible.

Exploration. At the beginning of the D-th iteration, each reasoning path P contains $D-1$ triples:

$$P_{D-1} = \{(e_d^{sub}, r_j^d, e_d^{obj})\}_{d=1}^{D-1}, \tag{2}$$

where e_d^{sub} and e_d^{obj} denote the subject and object nodes, respectively, and r_j^d represents the specific relation between them. e_d^{obj} and e_{d+1}^{sub} refer to the same node, forming P as a connected chain: $e_0 \to e_1 \to e_2 \to \cdots \to e_d$.

In the D-th reasoning step, we start from the current endpoint e_{D-1} and aim to identify the relevant candidate nodes $\mathcal{E}_D^{\text{cand}}$ among its neighbors with respect to the search topic. Since the underlying structure $\mathcal{M}_S$ is a multi-modal and multi-layered graph, these candidate nodes may reside in different layers. Accordingly, exploration proceeds in two complementary modes:

- **Intra-layer expansion**: Extend within the same layer (either view–view or object–object), filtering candidates based on semantic relevance to O_T and anchor objects $\{O_A\}$. This filtering preserves only nodes that contribute to the search topic, thus reducing ambiguity for the VLM:

$$\mathcal{E}_{D_v}^{\text{cand}} = \left\{(e_{D-1}, r, e_D) \mid r \in E^{VV},\ e_{D-1} \in \mathrm{V}\right\} \tag{3}$$

$$\mathcal{E}_{D_o}^{\text{cand}} = \left\{(e_{D-1}, r, e_D) \mid r \in E^{OO},\ e_{D-1} \in \mathrm{O}\right\} \tag{4}$$

- **Inter-layer transition**: Jump across layers from the current view node to its visible object nodes (no reverse), again filtering by semantic consistency:

$$\mathcal{E}_{D_o'}^{\text{cand}} = \left\{(e_{D-1}, r, e_D) \mid r \in E^{VO},\ e_{D-1} \in \mathrm{V}\right\} \tag{5}$$

After this exploration phase, the combined set of filtered candidates is:

$$\widetilde{\mathcal{E}}_D = \mathcal{E}_{D_v}^{\text{cand}} \cup \mathcal{E}_{D_o}^{\text{cand}} \cup \mathcal{E}_{D_o'}^{\text{cand}} \tag{6}$$

To avoid redundant cycles and ensure forward progress, all nodes already visited in previous iterations are excluded from $\widetilde{\mathcal{E}}_D$ before proceeding to the next reasoning step. As illustrated in Fig. 2 (Step 1), starting from View 0, we retain three relevant views containing chairs along with several visible chair object nodes. Irrelevant nodes are removed either due to sharing the same visual context or lacking semantic relevance to the topic. The remaining nodes form $\widetilde{\mathcal{E}}_D$.

Context Integration. After each exploration step, we integrate the context to form a complete reasoning input for the VLM through two operations:

- **Stitched View Construction:** We build an $S \times S$ visual grid I, placing previously explored views in the top-left and current candidate views in the remaining slots, with unused cells padded in white. This layout allows the VLM to visually compare historical and candidate views to assess redundancy or complementarity, and decide whether to continue view exploration or switch to object selection.
- **Accumulated Object Pool:** We maintain a growing pool of objects observed from explored views. These serve as global candidates, enabling the VLM to select the target object at any step and terminate reasoning early if confident.

$$\mathcal{E}_D^{\text{cand}} = \left(\sum_{d=1}^{D-1} \mathcal{E}_{d_o}^{\text{cand}}\right) + \widetilde{\mathcal{E}}_D \tag{7}$$

Reasoning. At each reasoning step, the VLM receives the grid image I and the candidate set $\mathcal{E}_D^{\text{cand}}$. The system prompts the VLM to decide whether the current exploration area should be further expanded in the view space or transitioned to the object space. When a view candidate is chosen, the system repeats the "Exploration" and "Reasoning" steps, progressively refining and constraining the exploration area until it naturally converges to the object space or the maximum reasoning depth D_{max} is reached. If the depth limit is reached without object selection, a forced global reasoning step is performed over the accumulated multi-view and multi-object evidence.

As illustrated in Fig. 2, VoG first explores the "brighter side of the room" finding multiple chairs but no `TV`, leaving the target ambiguous. It then moves to a view where the `TV` appears, spotting "the first chair counted from the TV" as `chair-1`. A nearby chair still causes uncertainty, so VoG explores once more to capture the full spatial layout. With all cues ("faces") confirmed, `chair-1` is identified as the target and the process ends.

Overall, the VoG procedure consists of: (i) an initial description parsing step; (ii) iterative expansion/contraction of the exploration area over D reasoning rounds; and (iii) a final global reasoning step if necessary — resulting in at most $D+2$ VLM calls.

Experiments

Experimental Settings

Datasets. We evaluate on two widely used 3D visual grounding benchmarks. **ScanRefer** (Chen, Chang, and Nießner 2020) contains 51,500 descriptions across 800 scenes with queries requiring either unique object localization or discrimination among same class distractors. **Nr3D** (Achlioptas et al. 2020) comprises 41,503 precise descriptions, categorized into "Easy" or "Hard" depending on distractor count, and labeled as view-dependent or view-independent. These datasets cover both point-cloud only grounding and bounding-box based grounding scenarios, offering diverse and challenging settings to evaluate the spatial representational capacity of the VoG.

Implementation Details. We adopt Qwen2-VL-2B (Wang et al. 2024) as the VLM. The MMMG viewpoint layer is built from ScanNet RGB images, using K-Means clustering camera poses into $M = 16$ representative views, with a maximum reasoning depth $D_{\text{max}} = 4$. All experiments are zero-shot on 8× NVIDIA H200 GPUs, yielding an average latency of 0.27 s/query (2.17 s/query per GPU) for Qwen2-VL-2B and 1.51 s/query (12.10 s/query per GPU) for Qwen2-VL-72B. Further details are provided in the Appendix.

Method	Venue	Agent	Unique		Multiple		Overall	
			Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
Fully-Supervised								
ScanRefer (Chen, Chang, and Nießner 2020)	ECCV'20	—	67.6	46.2	32.1	21.3	39.0	26.1
InstanceRefer (Yuan et al. 2021)	ICCV'21	—	77.5	66.8	31.3	24.8	40.2	32.9
3DVG-T (Zhao et al. 2021)	ICCV'21	—	77.2	58.5	38.4	28.7	45.9	34.5
BUTD-DETR (Jain et al. 2022)	ECCV'22	—	84.2	66.3	46.6	35.1	52.2	39.8
EDA (Wu et al. 2023)	CVPR'23	—	85.8	68.6	49.1	37.6	54.6	42.3
3D-VisTA (Zhu et al. 2023b)	ICCV'23	—	81.6	75.1	43.7	39.1	50.6	45.8
G3-LQ (Wang, Li, and Wang 2024)	CVPR'24	—	88.6	73.3	50.2	39.7	56.0	44.7
MCLN (Qian et al. 2024)	ECCV'24	—	86.9	72.7	52.0	40.8	57.2	45.7
ConcreteNet (Unal et al. 2024)	ECCV'24	—	86.4	82.1	42.4	38.4	50.6	46.5
Weakly-Supervised								
WS-3DVG (Wang et al. 2023)	ICCV'23	—	—	—	—	—	27.4	22.0
Zero-Shot								
LERF (Kerr et al. 2023)	ICCV'23	CLIP	—	—	—	—	4.8	0.9
OpenScene (Peng et al. 2023)	CVPR'23	CLIP	20.1	13.1	11.1	4.4	13.2	6.5
LLM-G (Yang et al. 2024)	ICRA'24	GPT-3.5	—	—	—	—	14.3	4.7
LLM-G (Yang et al. 2024)	ICRA'24	GPT-4 Turbo	—	—	—	—	17.1	5.3
ZSVG3D (Yuan et al. 2024)	CVPR'24	GPT-4 Turbo	63.8	58.4	27.7	24.6	36.4	32.7
SeeGround (Li et al. 2025)	CVPR'25	Qwen2-VL-2B	59.9	55.0	20.7	18.4	30.2	27.2
SeeGround (Li et al. 2025)	CVPR'25	Qwen2-VL-72B	75.7	68.9	**34.0**	30.0	44.1	39.4
VoG	**Ours**	Qwen2-VL-2B	69.1	63.1	25.3	21.9	35.9	31.9
VoG	**Ours**	Qwen2-VL-72B	**78.6**	**71.5**	**34.0**	**30.4**	**44.8**	**40.3**

Table 1: Evaluations of 3DVG on *ScanRefer* validation set. Results are reported for "Unique" (scenes with a single target object) and "Multiple" (scenes with distractors of the same class) subsets, along with overall performance.

Comparative Study

ScanRefer. Tab. 1 shows that our VoG framework sets a new state-of-the-art in the *zero-shot* setting. **(1) Large model advantage:** VoG-72B surpasses all existing zero-shot methods across all metrics, and even outperforms the weakly-supervised WS-3DVG. **(2) Small model efficiency:** VoG-2B, despite being **36**$\times$ **smaller** than SeeGround-72B, achieves **81.4%** of its Acc@0.25 performance and outperforms most prior zero-shot baselines. **(3) Competitiveness with supervision:** While trained with no 3DVG annotations, VoG-72B achieves results on par with several fully-supervised methods.

Nr3D. Tab. 2 further confirms our findings on this complementary benchmark. VoG-72B consistently leads all zero-shot baselines across *Easy*, *Hard*, *Dep.*, and *Indep.* splits, while narrowing the gap to fully-supervised methods (Achlioptas et al. 2020; Chen et al. 2022a; Huang et al. 2021; Chang et al. 2024).Meanwhile, VoG-2B maintains strong zero-shot competitiveness despite its compact size.

Visual Grounding Traceability

A unique advantage of VoG is its fully **traceable** reasoning process. Rather than outputting only the final grounded object, VoG reveals its sequence of exploratory steps. As shown in Fig. 3, the initial view contains a confusing chair with similar appearance ("black, rolling, high back"), yet VoG refrains from prematurely selecting it, showing that it actively aligns visual cues with the textual description (e.g., rejects mismatches such as "tucked beneath the desk"). It then shifts attention to related scene elements (e.g., nearby desks) and refines its hypothesis before converging on the correct object. This coherent, interpretable search path turns VoG from a "black-box" predictor into a transparent decision-maker, facilitating error diagnosis, trust, and insight into how appearance, attributes, and spatial relations guide grounding.

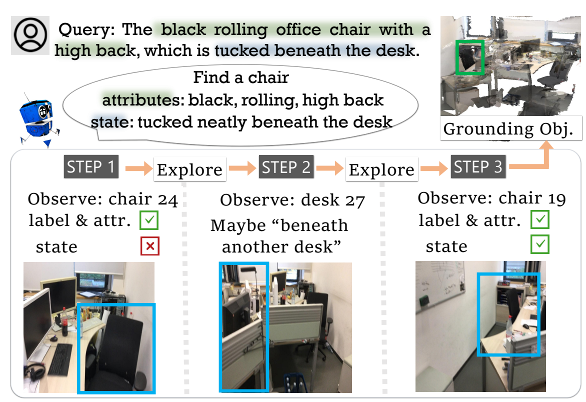

Figure 3: Traceable reasoning paths of VoG. Starting from an initial view showing a visually similar but mismatched chair, VoG rejects it due to missing the "tucked beneath the desk" cue. It then explores related scene elements (`desk`) and progressively refines its hypothesis. The search converges when both appearance and spatial state match, resulting in correct grounding.

Method	Venue	Easy	Hard	Dep.	Indep.	Overall
Fully-Supervised						
ReferIt3DNet	ECCV'20	43.6	27.9	32.5	37.1	35.6
TGNN	AAAI'21	44.2	30.6	35.8	38.0	37.3
InstanceRefer	ICCV'21	46.0	31.8	34.5	41.9	38.8
3DVG-T	ICCV'21	48.5	34.8	34.8	43.7	40.8
BUTD-DETR	ECCV'22	60.7	48.4	46.0	58.0	54.6
MiKASA	CVPR'24	69.7	59.4	65.4	64.0	64.4
ViL3DRel	—	70.2	57.4	62.0	64.5	64.4
Weakly-Supervised						
WS-3DVG	ICCV'23	27.3	18.0	21.6	22.9	22.5
Zero-Shot						
ZSVG3D-GPT4	CVPR'24	46.5	31.7	36.8	40.0	39.0
SeeGround-2B	CVPR'25	31.8	16.6	22.8	24.5	23.9
SeeGround-72B	CVPR'25	51.3	35.6	38.8	45.5	43.1
VoG-2B	**Ours**	40.3	22.6	30.4	31.5	31.1
VoG-72B	**Ours**	**58.9**	**37.2**	**39.4**	**52.1**	**47.6**

Table 2: Performance on Nr3D validation set. Queries are labeled as *Easy* (one distractor) or *Hard* (multiple distractors), and as *View-Dependent* or *View-Independent* based on viewpoint requirements for grounding.

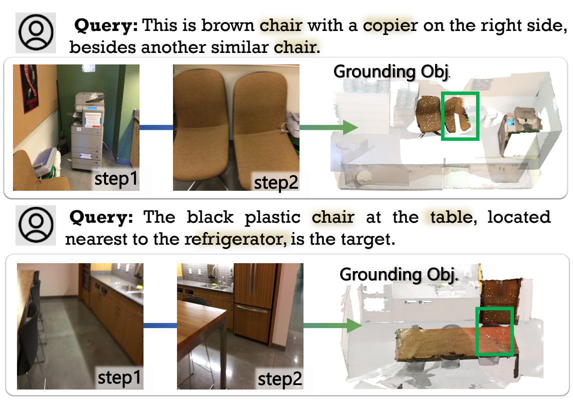

Figure 4: Qualitative grounding results. (**Top**) The initial view reveals a brown chair matching part of the query description ("copier machine to the right"), but VoG explores further to verify spatial cues ("beside another similar chair") before confirming the target. (**Bottom**) Starting from "at the table", VoG explores to confirm "nearest to the refrigerator" before final grounding.

Ablation Study

We perform various ablation studies to understand the importance of different factors in VoG. We conduct our ablation studies on ScanRefer, use the open-source Qwen2-VL-2B as the VLM, and the performance is measured using overall Acc@0.25.

(1) Do different Tree Structures affect 3DVG's performance? We study the impact of MMMG's structural components by disabling multi-round reasoning and feeding all information in a single round. As shown in Fig. 5, removing any component consistently harms performance, with the absence of all structure (S7) causing the largest drop, confirming that MMMG is crucial for spatial reasoning. Note that in S5*, we annotate each image with a global ID, same as (Qi et al. 2025), to assist the VLM in identifying the objects. This global ID marker serves as a form of View-Object relations, so the conclusions drawn from this configuration should be considered with reservation.

Variants	S	Round	Graph Traversal	Pool	Acc
R1(ours)	S1	✓	✓	✓	35.9
R2	S1	✓	✗	✓	30.3
R3	S1	✓	✓	✗	32.7
R4	S1	✓	✗	✗	32.6
R5	S1	✗	✗	N/A	32.1
R6	S4	✓	N/A	✓	27.9
R7	S4	✓	N/A	✗	31.7
R8	S4	✗	N/A	N/A	31.6

Table 3: Ablation study on the VoG modules. **S** denotes the scene structure used in Fig. 5, **Round** indicates multi-step reasoning. **Pool** denotes context integration.

(2) How well do VoG modules complement each other? We analyze two fixed structures, Full Structure (S1) and Minimal Structure (S4), to evaluate the contribution of each module (Tab. 3). To implement multi-round reasoning without graph traversal and pool (R4 and R7), we randomly select $M/\text{max\_step}$ view nodes in each round to ensure all information is seen within the D_{max} steps. Note that, N/A means modules cannot be performed.

- Graph Traversal Effects: By filtering key information in each round, graph traversal prevents VoG from over-exploring irrelevant regions (R1 vs R2). Without it, noisy candidates are repeatedly explored, causing most reasoning paths to fail even reaching D_{max} (R2 vs R4, R6 vs R7);
- Context Integration Effects: Enabling VoG to reuse promising candidates rather than starting from scratch each round (R1 vs R3). Without it, well-filtered candidates from graph traversal are discarded after each round, leading to the loss of valuable exploration history context.
- Complementarity of Graph Traversal & Context Integration: When combined both modules (R1), graph traversal ensures that only high-quality candidates are preserved, while context integration carries them forward across rounds. This combination enables coherent, focused exploration throughout reasoning (Fig. 4).
- Multi-round Reasoning vs. One-round: Multi-round reasoning remains more effective than providing all information at once, as it allows incremental retrieval and reasoning at each step (R4 vs R5, R7 vs R8).

(3) How does search depth affect VoG? We evaluate `max_step` from 1 to 5 (Fig. 6 (a)). Accuracy improves with depth, but the gain saturates beyond 4 since most questions require reasoning depths no greater than 3. Considering the linear cost growth, we set 4 as the max step.

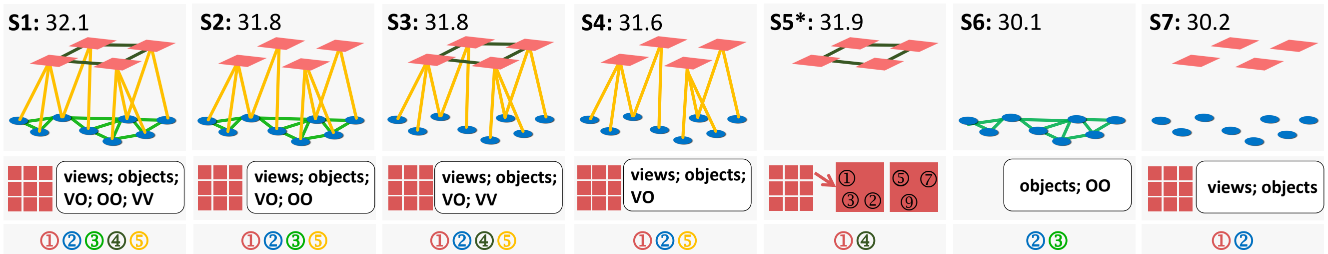

Figure 5: Graph Structure Ablation. S1: Full structure. S2–S4: Remove one type of edge while keeping others. S5*: Keep only images input with global object IDs. S6: Keep only text input. S7: Remove all structure. The first row shows the graph structure configurations and the corresponding accuracy.

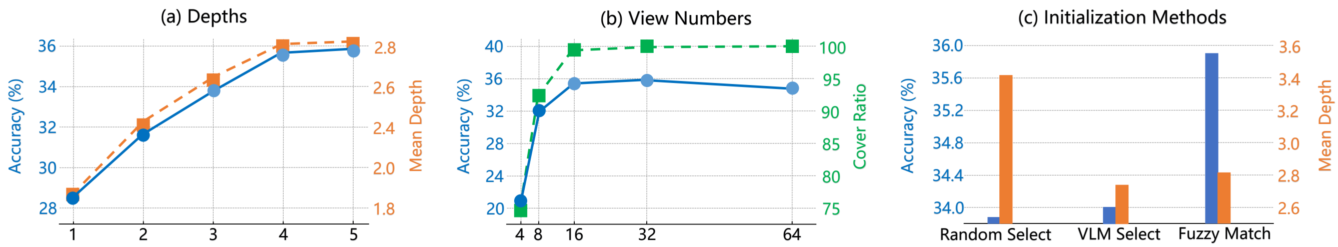

Figure 6: Ablation on reasoning depth, view number, and initialization methods.

(4) How does graph size affect VoG? The MMMG size depends on the number of view and object nodes. Since all objects are included, we focus on view nodes. Too few views yield low coverage and poor accuracy (Fig. 6 (b)). Increasing views improves both metrics, but beyond 32, excessive candidates per round make selection harder and slightly degrade performance. In VoG, the view images set as 16.

(5) How does path initialization affect VoG? The initial view determines VoG's starting point. The closer it is to the target, the fewer reasoning steps are needed. Our method selects it from fuzzy-matching the target and anchor in the VLM-processed description. Compared with random selection and direct VLM choice. As shown in Fig. 6 (c), our method shows a certain degree of error correction, achieving competitive results even with random initialization. Direct VLM selection causes a significant accuracy drop, most searches stop at step 1 because the chosen view already contains most query-related information, discouraging further exploration and causing frequent misclassification.

Figure 7: VoG efficacy drops in axis-symmetric scenes where Query 2 refers to the same target as Query 1, but in Step 2 selects a visually similar yet opposite-side view, resulting in grounding failure.

Limitations

VoG struggles in fully axis-symmetric scenes. In the example of Fig. 7, although *Query 2* refers to the same target as *Query 1*, VoG selects a visually similar but opposite-side view at Step 2, ultimately leading to grounding failure. Such symmetric layouts are inherently challenging for existing methods. In this scene, VoG achieves an average accuracy of 32.8% whereas SeeGround (Li et al. 2025) attains 23.4%. This suggests a future work of enhancing the scene graph with stronger spatial encoding, such as geometric constraints or pose-aware relational edges, to better preserve spatial continuity and resolve symmetric ambiguities.

Conclusion

We introduced View-on-Graph (VoG), a framework that redefines zero-shot 3DVG from the existing $VLM \oplus SI$ formulation to an interactive $VLM \otimes SI$ paradigm. By structuring 3D spatial information into a multi-modal, multi-layer scene graph and enabling the VLM to perform active and interactive exploration, VoG alleviates reasoning difficulty and enhances interpretability through step-by-step traceable grounding. Extensive experiments demonstrate the effectiveness of this $VLM \otimes SI$ paradigm and show that VoG achieves state-of-the-art zero-shot 3DVG performance, providing clear evidence that explicit scene structuring and interactive exploration can advance zero-shot 3DVG.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grants 62441216 and 62502068.

References

Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Bakr, E.; Alsaedy, Y.; and Elhoseiny, M. 2022. Look around and refer: 2D synthetic semantics knowledge distillation for 3D visual grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Cai, D.; Zhao, L.; Zhang, J.; Sheng, L.; and Xu, D. 2022. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chang, C.-P.; Wang, S.; Pagani, A.; and Stricker, D. 2024. MiKASA: Multi-key-anchor & scene-aware transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Chen, R.; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; and Wang, W. 2023. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, S.; Guhur, P.-L.; Tapaswi, M.; Schmid, C.; and Laptev, I. 2022a. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in Neural Information Processing Systems (NeurIPS)*.

Chen, S.; Guhur, P.-L.; Tapaswi, M.; Schmid, C.; and Laptev, I. 2022b. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fang, Y.; Zhang, L.; Wei, Z.; Wu, Z.; Wang, S.; Liu, C.; and Yang, X. 2025. ScaleGraph: A Scalable Self-supervised Framework for Cross-domain Zero-shot Graph Learning. *Pattern Recognition*.

Gao, G.; Liu, W.; Chen, A.; Geiger, A.; and Schölkopf, B. 2024. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gong, Z.; Hu, T.; Qiu, R.; and Liang, J. 2024. From cognition to precognition: A future-aware framework for social navigation. *arXiv preprint arXiv:2409.13244*.

Guan, H.; Song, C.; and Zhang, Z. 2024. GRAMO: geometric resampling augmentation for monocular 3D object detection. *Frontiers of Computer Science*.

Honerkamp, D.; Büchner, M.; Despinoy, F.; Welschehold, T.; and Valada, A. 2024. Language-grounded dynamic scene graphs for interactive object search with mobile manipulation. *IEEE Robotics and Automation Letters*.

Hu, T.; Jiao, J.; Xu, Y.; Liu, H.; Wang, S.; and Liu, M. 2024. Dhp-mapping: A dense panoptic mapping system with hierarchical world representation and label optimization techniques. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Huang, P.-H.; Lee, H.-H.; Chen, H.-T.; and Liu, T.-L. 2021. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Huang, Z.; Shangguan, Z.; Zhang, J.; Bar, G.; Boyd, M.; and Ohn-Bar, E. 2022. Assister: Assistive navigation via conditional instruction generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Jain, A.; Gkanatsios, N.; Mediratta, I.; and Fragkiadaki, K. 2022. Bottom up top down detection transformers for language grounding in images and point clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Kang, W.; Qu, M.; Kini, J.; Wei, Y.; Shah, M.; and Yan, Y. 2024. Intent3D: 3D Object Detection in RGB-D Scans Based on Human Intention. *arXiv preprint arXiv:2405.18295*.

Kerr, J.; Kim, C. M.; Goldberg, K.; Kanazawa, A.; and Tancik, M. 2023. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Koch, S.; Hermosilla, P.; Vaskevicius, N.; Colosi, M.; and Ropinski, T. 2024a. Sgrec3d: Self-supervised 3d scene graph learning via object-level scene reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Koch, S.; Vaskevicius, N.; Colosi, M.; Hermosilla, P.; and Ropinski, T. 2024b. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kong, L.; Liu, Y.; Li, X.; Chen, R.; Zhang, W.; Ren, J.; Pan, L.; Chen, K.; and Liu, Z. 2023. Robo3d: Towards robust and reliable 3d perception against corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Li, R.; Li, S.; Kong, L.; Yang, X.; and Liang, J. 2025. 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 (CVPR)*.

Liu, Y.; Long, C.; Zhang, Z.; Liu, B.; Zhang, Q.; Yin, B.; and Yang, X. 2022. Explore contextual information for 3d scene graph generation. *IEEE Transactions on Visualization and Computer Graphics*.

Liu, Z.; Yang, B.; Luximon, Y.; Kumar, A.; and Li, J. 2023. RayDF: neural ray-surface distance fields with multi-view consistency. *arXiv preprint arXiv:2310.19629*.

Liu, Z.; Ye, W.; Luximon, Y.; Wan, P.; and Zhang, D. 2025. Unleashing the potential of multi-modal foundation models and video diffusion for 4d dynamic physical scene simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mei, H.; Liu, Y.; Wei, Z.; Zhou, D.; Wei, X.; Zhang, Q.; and Yang, X. 2021. Exploring dense context for salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*.

Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. 2023. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi, Z.; Zhang, Z.; Fang, Y.; Wang, J.; and Zhao, H. 2025. GPT4Scene: Understand 3D Scenes from Videos with Vision-Language Models. *arXiv preprint arXiv:2501.01428*.

Qian, Z.; Ma, Y.; Lin, Z.; Ji, J.; Zheng, X.; Sun, X.; and Ji, R. 2024. Multi-branch Collaborative Learning Network for 3D Visual Grounding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Rana, K.; Haviland, J.; Garg, S.; Abou-Chakra, J.; Reid, I.; and Suenderhauf, N. 2023. Sayplan: Grounding large language models using 3d scene graphs for scalable robot task planning. *arXiv preprint arXiv:2307.06135*.

Schult, J.; Engelmann, F.; Hermans, A.; Litany, O.; Tang, S.; and Leibe, B. 2022. Mask3d: Mask transformer for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*.

Sun, J.; Xu, C.; Tang, L.; Wang, S.; Lin, C.; Gong, Y.; Ni, L. M.; Shum, H.-Y.; and Guo, J. 2023. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. *arXiv preprint arXiv:2307.07697*.

Unal, O.; Sakaridis, C.; Saha, S.; and Van Gool, L. 2024. Four ways to improve verbo-visual fusion for dense 3d visual grounding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Y.; Li, Y.; and Wang, S. 2024. Gˆ 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 (CVPR)*.

Wang, Z.; Huang, H.; Zhao, Y.; Li, L.; Cheng, X.; Zhu, Y.; Yin, A.; and Zhao, Z. 2023. Distilling coarse-to-fine semantic matching knowledge for weakly supervised 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wei, X.; Liu, Z.; and Luximon, Y. 2024. Sir: Multi-view inverse rendering with decomposable shadow for indoor scenes. *arXiv e-prints*.

Werby, A.; Huang, C.; Büchner, M.; Valada, A.; and Burgard, W. 2024. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

Wu, Y.; Cheng, X.; Zhang, R.; Cheng, Z.; and Zhang, J. 2023. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, J.; Chen, X.; Qian, S.; Madaan, N.; Iyengar, M.; Fouhey, D. F.; and Chai, J. 2024. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

Yang, Z.; Zhang, S.; Wang, L.; and Luo, J. 2021. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yuan, Z.; Ren, J.; Feng, C.-M.; Zhao, H.; Cui, S.; and Li, Z. 2024. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuan, Z.; Yan, X.; Liao, Y.; Zhang, R.; Wang, S.; Li, Z.; and Cui, S. 2021. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhang, Y.; Luo, H.; and Lei, Y. 2024. Towards CLIP-driven Language-free 3D Visual Grounding via 2D-3D Relational Enhancement and Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhao, L.; Cai, D.; Sheng, L.; and Xu, D. 2021. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhu, C.; Wang, T.; Zhang, W.; Chen, K.; and Liu, X. 2024. Scanreason: Empowering 3d visual grounding with reasoning capabilities. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhu, Y.; Zhuang, F.; Zhang, X.; Qi, Z.; Shi, Z.; Cao, J.; and He, Q. 2023a. Combat data shift in few-shot learning with knowledge graph. *Frontiers of Computer Science*.

Zhu, Z.; Ma, X.; Chen, Y.; Deng, Z.; Huang, S.; and Li, Q. 2023b. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.