

UZ3DVG: Unaided Zero-Shot 3D Visual Grounding with Generated Language Conditions

Wenbin Tan^{1,2}, Jiawen Lin¹, Yuan Xie³, Yachao Zhang^{1*}, Yanyun Qu^{1*}

¹Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen University, Xiamen, China

²School of Big Data, Tongren University, Tongren, China

³School of Computer Science and Technology, East China Normal University, Shanghai, China

wbtan@stu.xmu.edu.cn, yachaozhang@xmu.edu.cn, yyqu@xmu.edu.cn

Abstract

Zero-Shot 3D Visual Grounding (Zero-Shot 3DVG) aims to localize target objects in 3D scenes from natural language descriptions without relying on instance-wise description annotations. Existing methods rely on extra 2D images during inference and/or require multi-turn interactions with large language models (LLMs) or vision-language models (VLMs), which increase latency, computational cost, and deployment complexity. To overcome these limitations, we propose Unaided Zero-shot 3D Visual Grounding with generated language conditions (UZ3DVG), which is fed with 3D point clouds and textual descriptions only during inference and does not depend on external models. This is a new training paradigm: a VLM is employed solely to produce object-wise descriptions (pseudo labels) and reasoning chains for training a lightweight 3DVG model with robust spatial reasoning. Specifically, the introduced Open-Vocabulary Multi-Source Spatial Annotation and Reasoning Chain Generator processes RGB-D images or 3D-projected 2D images from open-world scenes to generate spatial pseudo-labels and reasoning chains for training. Then, we propose Reasoning Chain Distillation, which transfers reasoning knowledge extracted by a large teacher network to a lightweight student network. To represent both global and local geometric relationships, the Geometry-Aware Spatial Modeling (GeoSM) module is introduced to align textual reasoning with 3D spatial structures. Experiments show that UZ3DVG achieves SOTA zero-shot performance on ScanRefer and NR3D, with inference speeds up to 7.69 FPS, approximately 38 times faster than SOTA methods. The code is available at <https://github.com/tanwb/UZ3DVG>.

*Corresponding Author.

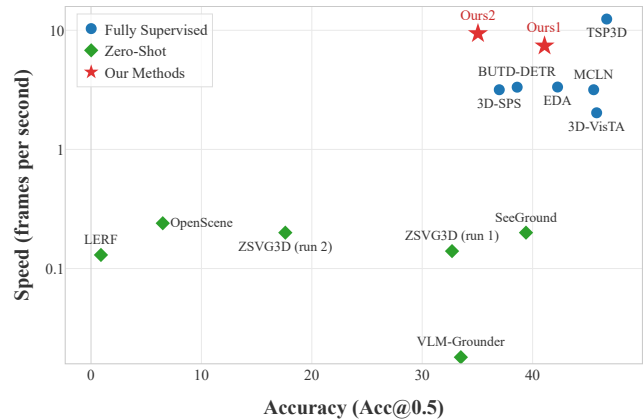
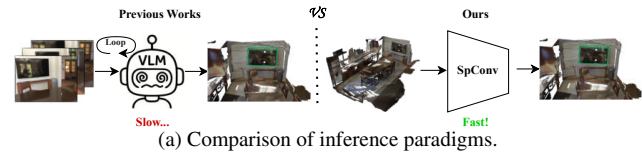


Figure 1. Comparison with zero-shot 3DVG methods on ScanRefer [4]. UZ3DVG surpasses existing approaches in accuracy and inference speed, and is the first zero-shot 3DVG framework to rely solely on 3D point clouds at inference, without 2D images or VLM interaction. Inference speed is measured following TSP3D [10].

1. Introduction

3D visual grounding (3DVG)[32, 40, 46, 51, 57] aims to localize target objects in 3D scenes based on natural language descriptions. This task holds broad application potential in domains such as indoor robotic visual navigation[42], virtual reality (VR)[44], and augmented reality (AR)[7].

Existing 3DVG methods, such as EDA[45], MCLN[28], and TSP3D[10], primarily focus on the closed-set setting[2, 4, 15, 23, 34, 49, 50, 58]. Although these methods have achieved significant performance improvements, their train-

ing heavily relies on large-scale human-annotated paired data (i.e., language descriptions and 3D objects) [2, 4]. This high annotation cost severely limits the generalization ability and deployment of these models in novel scenes.

Recent studies explore zero-shot 3DVG [17, 19, 47, 51], including ZSVG3D [51] and SeeGround [17]. These methods integrate large language models (LLMs) [25] or vision-language models (VLMs) [12] into the 3DVG inference pipeline, leveraging their semantic understanding and reasoning capabilities. However, they often call LLMs or VLMs during inference or perform multi-round interactions, as illustrated in Fig. 1, leading to high latency, computational overhead, increased costs, and deployment difficulties. Current zero-shot 3DVG methods typically run at less than the real-time threshold of 6 FPS [10].

To address these challenges in zero-shot 3DVG, we propose Unaided Zero-shot 3D Visual Grounding with generated language conditions (UZ3DVG). To the best of our knowledge, UZ3DVG is the first zero-shot 3DVG model that, during inference, uses only 3D point cloud inputs without 2D imagery or external LLMs/VLMs. Its core idea comprises two components: (i) constructing high-quality 3D pseudo-labels, and (ii) supervising a lightweight model with pseudo-labels and reasoning chains generated by a VLM.

We design the Open-Vocabulary Multi-Source Spatial Annotation and Reasoning Chain Generator (MSAG-RC) to extract RGB images and depth information, enabling effective transformation from 2D imagery to 3D spatial annotations. MSAG-RC uses open-vocabulary 2D detectors (e.g., Grounding DINO [20]) and segmentation models such as SAM2 [29] to obtain object masks and categories. By combining these with depth maps and camera parameters, the 2D segmentation results are back-projected into 3D space to recover object-scale information. The projected 3D data is encoded and fed into a VLM via prompt engineering to generate spatial pseudo-labels and reasoning chains, which serve as supervision to address the absence of spatial annotations and reasoning guidance in zero-shot training.

To transfer reasoning knowledge and logical structures within the chains to a compact model, we employ Reasoning Chain Distillation (RCD). A large teacher network extracts such knowledge from the generated chains and distills it into a lightweight student network. Due to the multi-relational complexity of the chains, mapping target information to sparse point clouds remains challenging. To mitigate this, we propose the Geometry-Aware Spatial Modeling (GeoSM) module, which encodes global context and local geometric relationships. GeoSM processes aligned visual features to provide positional references at both levels, thereby enhancing the correspondence between textual reasoning chains and 3D spatial structures.

Experimental results show that UZ3DVG achieves $\text{Acc}@0.25/0.5$ of 45.42%/41.08% on ScanRefer, surpassing

SeeGround [17] by 1.32%/1.68%. On the more challenging Multiple subset of ScanRefer, it achieves $\text{Acc}@0.25/0.5$ of 40.49%/36.33%, outperforming SeeGround [17] by 6.49%/6.33%. The inference speed reaches 7.69 FPS, about 38 times faster than SeeGround’s 0.2 FPS. These results confirm UZ3DVG’s robustness in complex scenes as well as its superior accuracy and efficiency.

In summary, our contributions are:

- We propose a novel zero-shot 3DVG paradigm that, at inference, relies solely on 3D point clouds as input, without using 2D images or interacting with VLMs, and employs an efficient network, retaining high speed and flexible deployment.
- To address the gap between 2D foundation models and 3D pseudo labels, we design the MSAG-RC to embed multi-modal information from RGB and depth into 3D space, and leverage the zero-shot capability of a VLM to generate spatial pseudo labels and reasoning chains. We further introduce Reasoning Chain Distillation (RCD) to alleviate the inherent shift between foundation model knowledge and 3D visual grounding.
- UZ3DVG achieves state-of-the-art performance on both the ScanRefer and NR3D, with an inference speed approximately 38 times faster than that of existing methods.

2. Related Work

3D Visual Grounding. Most current 3DVG research[13, 18, 27, 28, 43, 46, 55] follows supervised paradigms, requiring large-scale manual 3D labels from densely annotated datasets such as ScanRefer[4] and ReferIt3D[2], which align language descriptions with target objects in 3D scenes.

Representative works[5, 6, 10] address challenges including text–point cloud alignment[33, 45], geometric feature extraction[39, 40], and contextual cue modeling[57], using approaches such as Transformer-based attention[11, 57, 58], contrastive learning[1], graph-based relational modeling[9, 13, 36, 50], and text-guided voxel pruning[10], generally improving accuracy via manual 3D label supervision. Weak supervision was also explored[54, 56], but not in 3D visual grounding.

However, reliance on manual annotation limits generalization to unseen environments. To address this, we propose a zero-shot open-vocabulary 3DVG paradigm that uses automatically generated 3D pseudo labels to enhance adaptability and robustness in diverse 3D scenes.

Zero-Shot Open-Vocabulary 3DVG. Advances in large language models (LLMs) [25] and vision–language models (VLMs) [12, 31, 35] have greatly improved multimodal reasoning and scene understanding in images [6, 8, 14, 37]. However, existing LLMs/VLMs cannot directly process raw 3D point clouds, making zero-shot 3DVG inefficient and costly. Recent efforts integrate LLMs or VLMs into 3DVG inference, such as ZSVG3D [51], Agent3D-Zero [53], See-

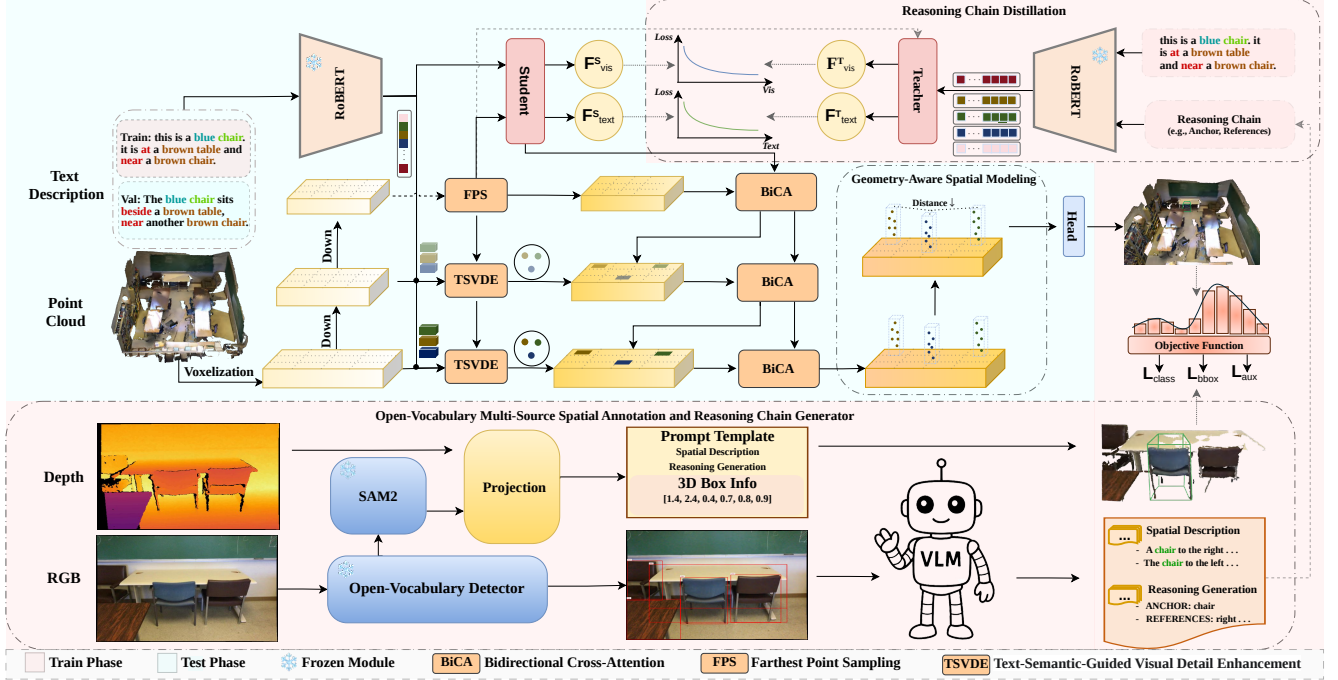


Figure 2. Overview of the UZ3DVG framework with training and inference phases. In training, MSAG-RC converts RGB-D images into spatial description pseudo-labels and reasoning chains to guide 3DVG model training, and RCD distills reasoning knowledge and logic from a teacher network to a lightweight student network head. Reasoning-enhanced text features are aligned with 3D backbone features (SpConv) and augmented via TSVDE, then processed by GeoSM to model global context and local geometry before the prediction head outputs localization results. During inference, only point clouds and text inputs are used, without MSAG-RC and RCD-teacher.

Ground [17], VLM-Grounder [47], and OpenEQA [24], which employ strategies like 3D-to-2D rendering, multi-view projection, and embodied-agent interaction.

Although effective, these methods rely on VLM-based inference, causing high latency and computational overhead. Our work aims to address these speed and deployment challenges with UZ3DVG, which relies solely on 3D point clouds during inference, without VLM interaction.

3. Methodology

3.1. Overview

To tackle slow inference, and deployment complexity caused by 2D reliance and multi-turn LLM/VLM interactions in zero-shot 3DVG, we introduce UZ3DVG. The network architecture, illustrated in Figure 2, consists of a training phase and an inference phase. During the training phase, both branches operate simultaneously, beginning with RGB-depth images processed by the Open-Vocabulary Multi-Source Spatial Annotation and Reasoning Chain Generator (MSAG-RC) to produce spatial description pseudo-labels and corresponding reasoning chains for guiding the 3DVG model. To enhance spatial reasoning ability and accelerate inference, these pseudo-labels and reasoning chains are first fed into a larger teacher net-

work to extract reasoning knowledge and logic, which are then transferred to a lightweight student network through Reasoning Chain Distillation (RCD). The spatial description pseudo-labels undergo feature extraction before entering the student network, while the Text-Semantic-guided Visual Detail Enhancement (TSVDE) module samples and augments details related to textual semantics from low-level visual features to help the model locate representations of reasoning logic. Following cross-modal text-visual alignment, the aligned visual features are passed to the Geometry-Aware Spatial Modeling (GeoSM) module to explicitly model global information and local geometric relationships within the 3D scene, assisting the student model in reasoning and improving its localization performance. Notably, the pseudo-labels, 3D information, and reasoning chains generated by MSAG-RC are used exclusively during training, whereas in the inference phase only the actual localization descriptions and 3D point clouds are required to achieve 3D visual localization.

3.2. Multi-Source Annotation and Reasoning Chain

To address annotation scarcity and insufficient structured reasoning supervision in zero-shot 3D localization, and to effectively bridge the representational gap between 2D image inputs and 3D spatial localization annotations, we pro-

pose the Open-Vocabulary Multi-Source Spatial Annotation and Reasoning Chain Generator (MSAG-RC). This approach relies solely on 2D inputs obtained from RGB-depth images or projected 3D point clouds, enabling the joint generation of 3D semantic labels and spatial reasoning chains without manual annotations.

Open-Vocabulary 3D Bounding Box Generation. Given an RGB-D video sequence, we uniformly sample frames at a fixed interval to obtain the i -th RGB image $I_i \in \mathbb{R}^{H \times W \times 3}$ and its corresponding depth map $D_i \in \mathbb{R}^{H \times W}$. We first apply an open-vocabulary 2D object detector (e.g., Grounding DINO [20]) to predict coarse 2D bounding boxes $\mathbf{b}_{i,j}^{2D} \in \mathbb{R}^4$. To obtain more accurate object boundaries, we further employ SAM2 [29] to convert each valid $\mathbf{b}_{i,j}^{2D}$ into a pixel-level segmentation mask:

$$M_{i,j} = \text{SAM2}(I_i, \mathbf{b}_{i,j}^{2D}) \in \{0, 1\}^{H \times W}. \quad (1)$$

After filtering invalid or noisy depth values, the valid masked pixel set is defined as

$$\Omega_{i,j} = \{(u, v) \mid M_{i,j}[u, v] = 1, D_i[u, v] \in \mathcal{R}\}, \quad (2)$$

where $\mathcal{R}$ denotes the valid depth range after outlier removal. For all valid pixels $(u, v) \in \Omega_{i,j}$, the corresponding depth values are given by $z_i(u, v) = D_i[u, v]$. We back-project all such pixels into 3D with the depth intrinsic matrix K_d and camera pose $\mathbf{T}_{c2w}^{(i)}$, and further transform them into the axis-aligned scene coordinate system with $\mathbf{A}$:

$$\tilde{\mathbf{P}}_{i,j} = \mathbf{A} \mathbf{T}_{c2w}^{(i)} (D_i[u, v] K_d^{-1}[u, v, 1]^\top). \quad (3)$$

An axis-aligned 3D pseudo box $\mathbf{b}_{i,j}^{3D} = [\mathbf{c}_{i,j}, \mathbf{s}_{i,j}] \in \mathbb{R}^6$ is computed from the aligned point cloud extrema $\mathbf{p}_{i,j}^{\min} = \min(\tilde{\mathbf{P}}_{i,j})$ and $\mathbf{p}_{i,j}^{\max} = \max(\tilde{\mathbf{P}}_{i,j})$ as

$$\mathbf{c}_{i,j} = \frac{\mathbf{p}_{i,j}^{\max} + \mathbf{p}_{i,j}^{\min}}{2}, \quad \mathbf{s}_{i,j}^{\text{raw}} = \mathbf{p}_{i,j}^{\max} - \mathbf{p}_{i,j}^{\min}. \quad (4)$$

Multi-Source Spatial Annotation. After obtaining object-level 3D geometry, we design a symbolic-logic-based spatial prompt constructor to transform geometric attributes into structured prompts for a VLM. For the j -th valid object in frame i , let its category be $\kappa_{i,j}$, its 3D pseudo box be $\mathbf{b}_{i,j}^{3D} = [\mathbf{c}_{i,j}, \mathbf{s}_{i,j}]$, and its target-specific spatial context be $\mathcal{C}_{i,j}$. Together with the annotated RGB image $\tilde{I}_{i,j}$, the VLM generates a natural language spatial description:

$$d_{i,j} = \text{VLM}(\tilde{I}_{i,j}; \mathbf{c}_{i,j}, \mathbf{s}_{i,j}, \kappa_{i,j}, \mathcal{C}_{i,j}), \quad (5)$$

where $\tilde{I}_{i,j}$ is the RGB image with the target object highlighted by a 2D bounding box and identifier, and $d_{i,j}$ is a human readable room-level spatial description.

Reasoning Chain Generation. To produce structured supervision for lightweight model reasoning, we craft prompts

that instruct the VLMs to emulate a robot localizing targets from object relationships in the pseudo 3D scene. Notably, this pseudo 3D scene is the same one used to generate $d_{i,j}$. After generating the description, the VLM is immediately prompted to produce the corresponding reasoning chain. The predefined template requires the reasoning chain to contain the structured components Anchor, References, and Reasoning. To further improve pseudo-label quality, prompt engineering requires the VLM to verify that $\mathcal{R}_{i,j}$ is consistent with $d_{i,j}$ while correcting localization failures in both. After self-verification, the VLM returns:

$$(d_{i,j}, \mathcal{R}_{i,j}, \gamma_{i,j}) = \mathcal{G}(\tilde{I}_{i,j}, \mathbf{b}_{i,j}^{3D}, \kappa_{i,j}, \mathcal{C}_{i,j}), \quad (6)$$

where $\mathcal{G}$ denotes the prompt-constrained VLM generation function, $\mathcal{R}_{i,j}$ is the structured reasoning chain, and $\gamma_{i,j}$ is the self-assigned confidence score; see Suppl. for details.

The final outputs are 3D spatial semantic labels, reasoning chains, and confidence scores. The entire process is automated through prompt engineering, requiring no manual labeling and enabling 3D localization annotation from RGB-D sequences in open environments.

3.3. Reasoning Chain Distillation

To efficiently transfer reasoning chain knowledge from the teacher model to the student model while maintaining fast inference speed, we design the Reasoning Chain Distillation module (RCD). This module adopts a dual path distillation scheme, in which the teacher network simultaneously guides the student network with textual and visual features enhanced by reasoning chains. Under distillation loss constraints, the reasoning logic embedded in the chain is internalized into the student’s spatial reasoning capability.

RCD-Teacher. The teacher network consists of three tightly connected layers: (i) reasoning chain parsing and encoding, (ii) noise filtering and feature enhancement, and (iii) cross-modal alignment and distillation output.

First, given a reasoning chain $\mathcal{R}$ and its corresponding open-vocabulary spatial annotation text features $\mathbf{T}_{\text{ovsd}}$, the teacher network begins by extracting the *Anchor*, *Reference*, and *Reasoning* components from the structured reasoning chain. These components are separately encoded by a frozen RoBERTa encoder [21] and yield the initial component-level representation $\mathbf{C}_{\text{chain}}^{\text{tea}}$, which constitutes the reasoning chain parsing and encoding stage.

This is followed by the noise filtering and feature enhancement stage. To suppress noisy components while retaining those semantically aligned with $\mathbf{T}_{\text{ovsd}}$, the teacher applies a quality gating mechanism to $\mathbf{C}_{\text{chain}}^{\text{tea}}$ conditioned on the global text context, yielding the gated component representation

$$\mathbf{C}_{\text{chain}}^{\text{filt}} = \text{Gate}(\mathbf{C}_{\text{chain}}^{\text{tea}}, \mathbf{T}_{\text{ovsd}}). \quad (7)$$

To model logical dependencies among the gated components, we apply self-attention: $\mathbf{H}_c^{\text{tea}} = \text{SA}_c(\mathbf{C}_{\text{chain}}^{\text{filt}})$. The

resulting component features are first grounded to the original text features through cross-attention, and are then injected back into the textual stream through a cross-attention module followed by gated fusion, producing the enhanced textual representation $\mathbf{T}_{\text{fus}}^{\text{tea}}$.

Finally, in the cross-modal alignment and distillation output stage, the fused textual representation $\mathbf{T}_{\text{fus}}^{\text{tea}}$ is aligned with teacher visual features $\mathbf{V}^{\text{tea}}$, which are derived from the same source visual input as the student features, through three layers of bidirectional cross-modal attention. Specifically, self-attention is first applied to both streams, followed by cross-attention:

$$(\mathbf{V}_{\text{out}}^{\text{tea}}, \mathbf{T}_{\text{out}}^{\text{tea}}) = \text{CA}(\text{SA}_v(\mathbf{V}^{\text{tea}}), \text{SA}_t(\mathbf{T}_{\text{fus}}^{\text{tea}})). \quad (8)$$

The resulting aligned features $(\mathbf{V}_{\text{out}}^{\text{tea}}, \mathbf{T}_{\text{out}}^{\text{tea}})$ serve as the distillation targets for guiding the student network.

RCD-Student. Under the guidance of the teacher, the student network is designed to identify reasoning-relevant features in localization descriptions, extract the underlying reasoning logic, and improve localization accuracy through a lightweight pseudo reasoning mechanism. It is structurally analogous to the teacher but has reduced capacity, comprising three stages: (i) pseudo reasoning signal generation, (ii) reasoning aware feature enhancement, and (iii) cross modal alignment.

Given the language spatial description features $\mathbf{T}_{\text{ovsd}}$, we apply masked global average pooling to obtain a global context: $\mathbf{t}_{\text{pool}}^{\text{stu}} = \text{GAP}(\mathbf{T}_{\text{ovsd}})$. This pooled feature is fed into a lightweight query generator and an attention-based extractor to produce three pseudo reasoning components $\mathbf{R}_{\text{comp}}^{\text{stu}}$, which mimic the three components in the teacher reasoning chain, namely Anchor, Reference, and Reasoning. These components are further aggregated into a global student reasoning representation $\mathbf{R}^{\text{stu}}$.

To align the student representation with the teacher representation, we apply a reasoning distillation objective that enforces consistency at both the component and global levels, defined as

$$\mathcal{L}_r = \mathcal{L}_r^{\text{comp}} + \mathcal{L}_r^{\text{global}}, \quad (9)$$

where $\mathcal{L}_r^{\text{comp}}$ and $\mathcal{L}_r^{\text{global}}$ denote student-teacher alignment over the reasoning components and the aggregated global representations, respectively.

The student then enhances its textual representation by concatenating $\mathbf{R}^{\text{stu}}$ with $\mathbf{T}_{\text{ovsd}}$, transforming the fused features with an MLP, and injecting the result into the original textual stream with confidence weighted residual fusion. The enhanced textual features are subsequently aligned with visual features through a lightweight bidirectional cross modal encoder with two layers, yielding $\mathbf{V}_{\text{out}}^{\text{stu}}$ and $\mathbf{T}_{\text{out}}^{\text{stu}}$.

To transfer the teacher’s reasoning dynamics, we define reasoning gains as

$$\mathbf{G}^{\text{tea}} = \mathbf{T}_{\text{out}}^{\text{tea}} - \mathbf{T}_{\text{ovsd}}, \quad \mathbf{G}^{\text{stu}} = \mathbf{T}_{\text{out}}^{\text{stu}} - \mathbf{T}_{\text{ovsd}}, \quad (10)$$

and align them via

$$\mathcal{L}_g = \mathcal{L}_{\text{cos}} + \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{mag}}, \quad (11)$$

where $\mathcal{L}_{\text{cos}}$, $\mathcal{L}_{\text{mse}}$, and $\mathcal{L}_{\text{mag}}$ respectively constrain gain direction, normalized gain magnitude, and overall enhancement strength. Additionally, visual features are distilled with a foreground-background hybrid loss:

$$\mathcal{L}_v = \mathcal{L}_{\text{fg}}(\mathbf{V}_{\text{out}}^{\text{stu}}, \mathbf{V}_{\text{out}}^{\text{tea}}) + \mathcal{L}_{\text{bg}}(\mathbf{V}_{\text{out}}^{\text{stu}}, \mathbf{V}_{\text{out}}^{\text{tea}}), \quad (12)$$

where $\mathcal{L}_{\text{fg}}$ and $\mathcal{L}_{\text{bg}}$ denote foreground matching and background matching, respectively.

Collectively, the above losses enable efficient reasoning chain distillation.

Text-Semantic-Guided Visual Detail Enhancement. To enrich visual features with fine-grained details from text-relevant regions and improve reasoning performance, we introduce the Text-Semantic-Guided Visual Detail Enhancement module (TSVDE). During upsampling, the module leverages textual semantics and reasoning-enhanced visual features to guide feature refinement. Specifically, it applies similarity-based sampling to emphasize regions closely aligned with textual semantics, thereby enhancing the fine-grained details required by the reasoning chain (see Suppl. Sec. 1.1 for details).

3.4. Geometry-Aware Spatial Modeling

Reasoning chains often contain spatial predicates (e.g., “left of”, “above”) relying on geometric relationships. However, visual features from sparse convolutions (SpConv) lack explicit spatial-relation modeling, which limits reasoning-to-visual mapping.

To address this limitation, we propose Geometry-Aware Spatial Modeling (GeoSM) to integrate global and local geometry. First, to capture the global 3D spatial context, we normalize the coordinates $\mathbf{p}_i$ corresponding to the aligned visual features, following ideas from 3DGraphLLM [52], and compute absolute positional embeddings via sinusoidal encoding:

$$\hat{\mathbf{p}}_i = 2 \cdot \frac{\mathbf{p}_i - \mathbf{p}_{\min}}{\mathbf{p}_{\max} - \mathbf{p}_{\min} + \varepsilon} - 1, \quad (13)$$

where $\mathbf{p}_{\min}$ and $\mathbf{p}_{\max}$ are point cloud bounds, and ε ensures stability. The encoded positions capture global layout cues, yielding the positional feature $\mathbf{f}_i^{\text{pos}}$ after linear projection.

For local geometry, we build K -nearest-neighbor neighborhoods. For each point $\mathbf{p}_i$, the relative displacement is defined as $\Delta \mathbf{p}_{ij} = \mathbf{p}_j - \mathbf{p}_i$. We aggregate neighbor features, concatenate them with $\Delta \mathbf{p}_{ij}$, and pass the combined representation through an MLP followed by a max pooling layer to obtain local relational feature $\mathbf{f}_i^{\text{rel}}$.

Finally, concatenating $\mathbf{f}_i^{\text{pos}}$ and $\mathbf{f}_i^{\text{rel}}$ yields $\mathbf{g}_i = [\mathbf{f}_i^{\text{pos}}, \mathbf{f}_i^{\text{rel}}]$, which is then added back to the original visual feature through a residual connection, thereby providing comprehensive spatial context for reasoning.

Table 1. Evaluation of 3DVG methods on the ScanRefer [4] validation set. Results marked with $\dagger\dagger$ are computed on a subset of 250 samples (250/9508 $\approx$ 2.6%) from the validation set. Inference speed follows TSP3D [10] (batch size = 1, NVIDIA 3090 GPU). NR: Not Released.

Method	Venue	Supervision	Unique		Multiple		Overall		FPS
			Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	
3D-SPS [23]	CVPR'22	Fully	84.12	66.72	40.32	29.82	48.82	36.98	3.17
BUTD-DETR [15]	ECCV'22	Fully	82.88	64.98	44.73	33.97	50.42	38.60	3.33
EDA [45]	CVPR'23	Fully	85.76	68.57	49.13	37.64	54.59	42.26	3.34
3D-VisTA [58]	ICCV'23	Fully	81.60	75.10	43.70	39.10	50.60	45.80	2.03
VPP-Net [32]	CVPR'24	Fully	86.05	67.09	50.53	39.03	55.65	43.29	NR
G3-LQ [40]	CVPR'24	Fully	88.59	73.28	50.23	39.72	55.95	44.72	NR
MCLN [28]	ECCV'24	Fully	86.89	72.73	51.96	40.76	57.17	45.53	3.17
TSP3D [10]	CVPR'25	Fully	–	–	–	–	56.45	46.71	12.43
PC-CrossDiff [33]	AAAI'26	Fully	86.28	75.08	53.59	43.12	58.47	47.89	4.07
WS-3DVG [41]	ICCV'23	Weakly	–	–	–	–	27.40	22.00	NR
<i>Mask3D-Based</i>									
ZSVG3D [51]	CVPR'24	Zero-Shot	63.80	58.40	27.70	24.60	36.40	32.70	0.14
SeeGround [17]	CVPR'25	Zero-Shot	75.70	68.90	34.00	30.00	44.10	39.40	0.20
Ours	–	Zero-Shot	73.57	68.15	40.49	36.33	45.42	41.08	7.69
<i>Non-Mask3D</i>									
VLM-G [47] $\dagger\dagger$	CORL'24	Zero-Shot	66.00	29.80	48.30	33.50	51.60	32.80	0.02
LERF [16]	ICCV'23	Zero-Shot	–	–	–	–	4.80	0.90	0.13
OpenScene [26]	CVPR'23	Zero-Shot	20.10	13.10	11.10	4.40	13.20	6.50	0.24
LLM-G [48]	ICRA'24	Zero-Shot	–	–	–	–	17.10	5.30	0.05
ZSVG3D [51]	CVPR'24	Zero-Shot	32.50	27.80	16.10	14.60	20.00	17.60	0.21
Ours	–	Zero-Shot	69.62	56.94	38.48	31.21	43.13	35.05	9.43

3.5. Objective Function

Following previous works [10, 22, 28, 45], the overall loss of UZ3DVG consists of the referring localization loss $\mathcal{L}_{\text{rec}}$ and the distillation losses, where $\mathcal{L}_{\text{rec}}$ includes the bounding box regression loss $\mathcal{L}_{\text{bbox}}$ and the classification loss $\mathcal{L}_{\text{cls}}$.

The teacher warm-up loss is provided in the Suppl. The term $\mathcal{L}_{\text{rec}}$ guides the learning of the referring localization task in the main model, while the distillation losses supervise the distillation of the reasoning chain. The total loss is defined as

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_g\mathcal{L}_g + \lambda_v\mathcal{L}_v + \lambda_r\mathcal{L}_r, \quad (14)$$

where λ_* denotes the balance coefficients.

4. Experiments

4.1. Experimental Setup and Datasets

To ensure fairness in comparative evaluation, we assess model accuracy on two widely used benchmark datasets: ScanRefer [4] and NR3D [2]. ScanRefer comprises 800 ScanNet scenes and a validation set of 9,508 natural language descriptions, each specifying a target object and its spatial context. Results are reported separately for

“Unique” and “Multiple” cases. To avoid data leakage, all RGB-D images in our experiments come from scenes disjoint from those in the ScanRefer validation set. NR3D, a subset of ReferIt3D [2], contains 7,904 validation samples, with evaluation distinguishing “Easy” and “Hard”, and annotations further labeling references as “View-Dependent” (Dep.) or “View-Independent” (Indep.).

Without loss of generality, in the MSAG-RC branch, we employ Qwen3-VL-Plus [3, 35] and Doubao-seed-1.6 [31] as VLMs. Details are provided in Suppl. Sec. S1.2. During inference, for a fair comparison with ZSVG3D [51] and SeeGround [17] when using Mask3D [30], we use Mask3D [30] only for post hoc bounding box refinement; more details are provided in Suppl. Sec. S1.3. The accuracies of LERF [16] and OpenScene [26] are taken from ZSVG3D [51] and SeeGround [17], and their speeds are evaluated using the official code.

4.2. Results and Analysis

Performance on ScanRefer. Table 1 presents a comparison between our method, UZ3DVG, and state-of-the-art zero-shot 3DVG approaches on the ScanRefer [4] validation set, including the “Unique” subset, the “Multiple” subset, and

Table 2. Nr3D [2] val (††: 250-query subset): direct inference with the same Table 1 checkpoint (trained on ScanRefer pseudo-labels), without Nr3D fine-tuning. Evaluation protocols follow ReferIt3D [2] and MCLN [28], with Mask3D refinement.

Method	Easy	Hard	Dep.	Indep.	Overall
<i>Supervision: Fully Supervised</i>					
ReferIt3DNet [2]	43.6	27.9	32.5	37.1	35.6
TGNN [13]	44.2	30.6	35.8	38.0	37.3
InstanceRefer [50]	46.0	31.8	34.5	41.9	38.8
3DVG-Trans [57]	48.5	34.8	34.8	43.7	40.8
BUTD-DETR[15]	–	–	–	–	43.3
AugRefer [38]	–	–	–	–	48.4
TSP3D [10]	–	–	–	–	48.7
MCLN [28]	–	–	–	–	59.8
PC-CrossDiff [33]	62.2	57.5	60.6	59.6	59.9
<i>Supervision: Weakly Supervised</i>					
WS-3DVG [41]	27.3	18.0	21.6	22.9	22.5
<i>Supervision: Zero-Shot</i>					
VLM-G ^{††} [47]	55.2	39.5	45.8	49.4	48.0
ZSVG3D [51]	46.5	31.7	36.8	40.0	39.0
SeeGround [17]	54.5	38.3	42.3	48.2	46.1
Ours	52.2	40.8	39.3	50.1	46.5

the overall results. Under the zero-shot setting, UZ3DVG achieves notable gains in localization accuracy and inference speed, outperforming the weakly supervised method WS-3DVG [41] in accuracy while remaining competitive with fully supervised approaches.

Without Mask3D-based boundary refinement, some methods (e.g., SeeGround [17]) cannot operate due to reliance on Mask3D segmentation outputs for bounding information. We thus compare UZ3DVG with ZSVG3D [51], the best-performing zero-shot method under non-Mask3D settings. UZ3DVG obtains 43.13% and 35.05% on Acc@0.25 and Acc@0.50, surpassing ZSVG3D by 23.13% and 17.45%, respectively, and reaching 9.43 FPS versus 0.21 FPS ($\sim 45\times$ speedup).

With Mask3D-based boundary refinement (Supplementary Sec. S1.3), UZ3DVG further improves to 45.42% and 41.08% in Acc@0.25 and Acc@0.5, exceeding SeeGround [17] by 1.32% and 1.68%, with 7.69 FPS compared to 0.20 FPS ($\sim 38\times$). On the more challenging “Multiple” subset, which requires stronger spatial discrimination, UZ3DVG scores 40.49% and 36.33%, outperforming SeeGround by 6.49% and 6.33%. Although certain fully supervised methods (e.g., MCLN [28] and TSP3D [10]) attain marginally higher accuracy due to extensive annotations, UZ3DVG delivers competitive zero-shot performance with substantial speed advantages, making it highly scalable for large-scale 3DVG tasks without manual labels.

Performance on NR3D. Table 2 demonstrates that UZ3DVG achieves competitive performance on the NR3D [2], with an overall accuracy of 46.5%, surpassing ZSVG3D [51] by 7.5% and SeeGround [17] by an additional 0.4%. Furthermore, UZ3DVG exhibits remarkable results in both the *Hard* and *View-Independent* categories. In the *Hard* setting, it achieves an accuracy of 40.8%, outperforming ZSVG3D [51] by 9.1% and SeeGround [17] by 2.5%. These improvements clearly demonstrate the robustness and strong generalization ability of our method.

Table 3. Ablation of the main modules on ScanRefer [4] dataset.

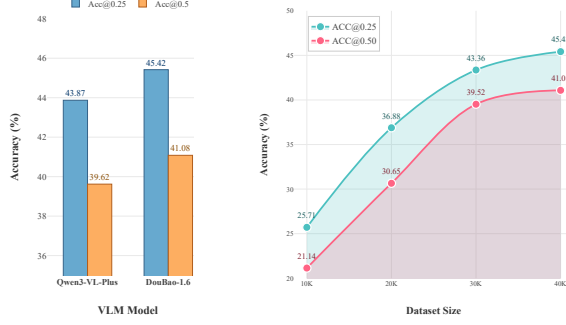
TSVDE	RCD	GeoSM	Acc@0.25	Acc@0.50
w/o	w/o	w/o	40.75	37.18
✓	w/o	w/o	41.83	38.21
✓	✓	w/o	43.26	39.65
✓	w/o	✓	43.15	39.47
✓	✓	✓	45.42	41.08

4.3. Ablation Studies

Effectiveness of the Proposed Components. We evaluate the contributions of TSVDE, RCD, and GeoSM to localization accuracy on the ScanRefer [4] dataset through ablation studies (Table 3), progressively removing each component from the full model. Without TSVDE, random sampling serves as the baseline; adding TSVDE improves IoU@0.25 by 1.08% and IoU@0.50 by 1.03%, showing that semantically guided point cloud sampling effectively supplements text-relevant visual details and facilitates the model’s localization reasoning. Incorporating RCD, which distills the reasoning logic of a vision-language model into the student network, yields gains of 2.51% and 2.47%, demonstrating effective integration of spatial reasoning knowledge for better understanding of spatial relations in text. Integrating GeoSM on top of TSVDE increases IoU@0.25 by 2.40% and IoU@0.50 by 2.29%, confirming that its fused global and local geometric representations strengthen spatial reasoning. Combining all three components results in overall improvements of 4.67% and 3.90%, highlighting their complementary effects and effectiveness.

Impact of Different VLMs Choice. Figure 3a shows that the use of different VLMs for generating open-vocabulary descriptions for model training results in variations in performance on the ScanRefer [4] dataset. With Qwen3-VL-Plus [3], localization accuracies reach 43.87% for Acc@0.25 and 39.62% for Acc@0.50, while the Doubao-seed-1.6 [31] achieves 45.42% and 41.08%, respectively, which are slightly higher. These results show that spatial-description label quality directly governs the model’s spatial reasoning learning and overall accuracy.

Effect of Pseudo-Label Quantity. Figure 3b shows the



(a) Accuracy for Different VLMs. (b) Effect of training data amount. Figure 3. Comparison of localization accuracy under different VLMs and training data scales.

effect of varying the number of open-vocabulary spatial description pseudo-labels on ScanRefer [4] performance. Starting from 10K pseudo-labels, increasing the number to 20K improves Acc@0.25 by 11.17% and Acc@0.50 by 9.51%, showing clear gains in the limited-label regime. Expanding to 30K brings further improvements of 6.48% and 8.87%. However, at 40K, the gains drop to 2.06% and 1.56%, suggesting that pseudo-label noise reduces the benefit of further increasing the number of labels.

Table 4. Comparison of accuracy on the ScanRefer [4] dataset using spatial descriptions generated directly from 2D boxes or enhanced by our MSAG-RC module, with or without reasoning chains. RC: Reasoning chains.

Module	Acc@0.25	Acc@0.50
2D Info	34.16	22.83
MSAG-RC w/o RC	43.15 (+8.99)	39.47 (+16.64)
MSAG-RC	45.42 (+2.27)	41.08 (+1.61)

Comparison of Different Spatial Descriptions. Table 4 compares spatial descriptions generated directly from 2D boxes with those enhanced by our MSAG-RC module. The 2D-only approach has limited ability to capture spatial relationships, leading to lower localization accuracy. By contrast, our MSAG-RC module projects 2D boxes into 3D and strengthens contextual information around the target, improving Acc@0.25 by 8.99% and Acc@0.50 by 16.64%. When reasoning chains are added to the paired spatial descriptions used for distillation, accuracy further increases by 2.27% and 1.61%, demonstrating that our MSAG-RC module facilitates better 3D visual grounding training.

4.4. Qualitative Analysis

From the first and second columns in Fig. 4, when the input description contains complex spatial relations, both ZSVG3D [51] and SeeGround [17] mislocalize the target. In scenarios where multiple same-category instances are present and referenced in the description (third column),

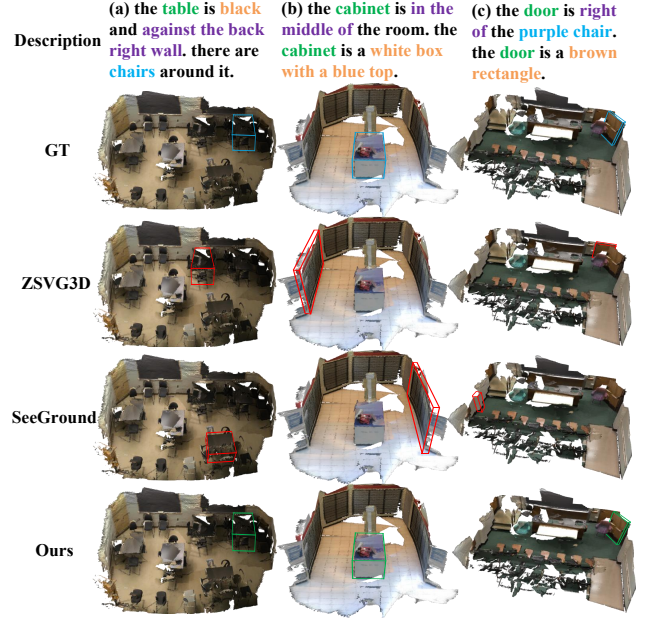


Figure 4. Qualitative comparison of our UZ3DVG with ZSVG3D [51] and the baseline SeeGround [17].

these methods can identify the category but fail to determine the correct target. In contrast, our method accurately localizes the intended target, demonstrating robust spatial reasoning. This capability stems from RCD, which transfers spatial reasoning logic, and the GeoSM module, which models both global scene context and local geometric information. They strengthen the mapping between reasoning chains and 3D structures, improving localization accuracy.

5. Conclusion

In this paper, we present Unaided Zero-shot 3D Visual Grounding with generated language conditions (UZ3DVG), a paradigm that performs inference solely from 3D point clouds, removing dependence on 2D inputs and external LLMs/VLMs that cause high latency, heavy computation, and deployment difficulties. To address the gap between 2D foundation models and 3D pseudo labels, we design the MSAG-RC to embed multi-modal information from RGB and depth into 3D space, and leverage the zero-shot capability of a VLM to generate spatial pseudo labels and reasoning chains. To further enhance reasoning, the RCD module distills chain knowledge into a student network in zero-shot settings. The GeoSM module captures global layouts and local geometric relations, thereby enhancing the correspondence between textual reasoning chains and 3D spatial structures, further improving localization accuracy. Experiments show that UZ3DVG achieves SOTA on ScanRefer and NR3D, runs at 7.69 FPS, and supports fully localized deployment, establishing a new paradigm for zero-shot 3DVG in open-world scenes.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62306165, the Science and Technology on Sonar Laboratory under Grant 2024-JCJQ-LB-32/07, the State Grid Corporation Headquarters Technology Project under Grant 52120025004V-463-FGS, the Fundamental Research Funds for the Central Universities under Grant 20720250031, the Guizhou Provincial Department of Science and Technology under Grant Qian Ke He QN [2025] 337, the Guizhou Provincial Basic Research Program (Natural Science) under Grant NO.MS (2026) 814, and the Tongren Municipal Science and Technology Bureau under Grant Tongshi [2025] 54.

References

- [1] 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*, pages 3941–3950, 2022. 2
- [2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes. In *Computer Vision – ECCV 2020*, pages 422–440, Cham, 2020. Springer International Publishing. 1, 2, 6, 7
- [3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, and et al. Qwen3-VL Technical Report, 2025. 6, 7
- [4] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D Object Localization in RGB-D Scans Using Natural Language. In *Computer Vision – ECCV 2020*, pages 202–221, Cham, 2020. Springer International Publishing. 1, 2, 6, 7, 8
- [5] Qi Chen, Changli Wu, Jiayi Ji, Yiwei Ma, Danni Yang, and Xiaoshuai Sun. IPDN: Image-enhanced Prompt Decoding Network for 3D Referring Expression Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):2132–2140, 2025. 2
- [6] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. CLIP2Scene: Towards Label-Efficient 3D Scene Understanding by CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023. 2
- [7] Shaveta Dargan, Shally Bansal, Munish Kumar, Ajay Mittal, and Krishan Kumar. Augmented Reality: A Comprehensive Review. *Archives of Computational Methods in Engineering*, 30(2):1057–1080, 2023. 1
- [8] Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A Memory-Augmented Multimodal Agent for Video Understanding. In *Computer Vision – ECCV 2024*, pages 75–92, Cham, 2025. Springer Nature Switzerland. 2
- [9] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-Form Description Guided 3D Visual Graph Network for Object Grounding in Point Cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3722–3731, 2021. 2
- [10] 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*, pages 3666–3675, 2025. 1, 2, 6, 7
- [11] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. TransRefer3D: Entity-and-Relation Aware Transformer for Fine-Grained 3D Visual Grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2344–2352, New York, NY, USA, 2021. Association for Computing Machinery. 2
- [12] Wenyi Hong, Weihang Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, Xiaotao Gu, Xiaohan Zhang, Guanyu Feng, Da Yin, Zihan Wang, Ji Qi, Xixuan Song, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Yuxiao Dong, and Jie Tang. CogVLM2: Visual Language Models for Image and Video Understanding, 2024. 2
- [13] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-Guided Graph Neural Networks for Referring 3D Instance Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2):1610–1618, 2021. 2, 7
- [14] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. OpenIns3D: Snap and Lookup for 3D Open-Vocabulary Instance Segmentation. In *Computer Vision – ECCV 2024*, pages 169–185, Cham, 2025. Springer Nature Switzerland. 2
- [15] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom Up Top Down Detection Transformers for Language Grounding in Images and Point Clouds. In *Computer Vision – ECCV 2022*, pages 417–433, Cham, 2022. Springer Nature Switzerland. 1, 6, 7
- [16] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language Embedded Radiance Fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 6
- [17] 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*, pages 3707–3717, 2025. 2, 3, 6, 7, 8
- [18] Haojia Lin, Yongdong Luo, Xiawu Zheng, Lijiang Li, Fei Chao, Taisong Jin, Donghao Luo, Yan Wang, Liujuan Cao, and Rongrong Ji. A Unified Framework for 3D Point Cloud Visual Grounding, 2023. 2
- [19] Jiawen Lin, Shiran Bian, Yihang Zhu, Wenbin Tan, Yachao Zhang, Yuan Xie, and Yanyun Qu. SeqVLM: Proposal-Guided Multi-View Sequences Reasoning via VLM for Zero-Shot 3D Visual Grounding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3094–3103, New York, NY, USA, 2025. Association for Computing Machinery. 2

- [20] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *Computer Vision – ECCV 2024*, pages 38–55, Cham, 2025. Springer Nature Switzerland. 2, 4
- [21] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019. 4
- [22] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-Free 3D Object Detection via Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021. 6
- [23] 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*, pages 16454–16463, 2022. 1, 6
- [24] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Alexander Sax, and Aravind Rajeswaran. OpenEQA: Embodied Question Answering in the Era of Foundation Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498, 2024. 3
- [25] OpenAI, Josh Achiam, Steven Adler, and et al. GPT-4 Technical Report, 2024. 2
- [26] Songyou Peng, Kyle Genova, Chiyu “Max” Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. OpenScene: 3D Scene Understanding With Open Vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023. 6
- [27] Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-RefSeg3D: Enhancing Referring 3D Instance Segmentation via Structured Cross-Modal Graph Neural Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4551–4559, 2024. 2
- [28] Zhipeng Qian, Yiwei Ma, Zhekai Lin, Jiayi Ji, Xiwu Zheng, Xiaoshuai Sun, and Rongrong Ji. Multi-branch Collaborative Learning Network for 3D Visual Grounding. In *Computer Vision – ECCV 2024*, pages 381–398, Cham, 2025. Springer Nature Switzerland. 1, 2, 6, 7
- [29] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollar, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *The Thirteenth International Conference on Learning Representations*, 2024. 2, 4
- [30] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3D: Mask Transformer for 3D Semantic Instance Segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223, 2023. 6
- [31] ByteDance Seed, Jiaze Chen, Tiantian Fan, Xin Liu, and et al. Seed1.5-Thinking: Advancing Superb Reasoning Models with Reinforcement Learning, 2025. 2, 6, 7
- [32] Xiangxi Shi, Zhonghua Wu, and Stefan Lee. Viewpoint-Aware Visual Grounding in 3D Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14056–14065, 2024. 1, 6
- [33] Wenbin Tan, Jiawen Lin, Fangyong Wang, Yuan Xie, Yong Xie, Yachao Zhang, and Yanyun Qu. PC-CrossDiff: Point-Cluster Dual-Level Cross-Modal Differential Attention for Unified 3D Referring and Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(11):9332–9340, 2026. 2, 6, 7
- [34] Ozan Unal, Christos Sakaridis, Suman Saha, and Luc Van Gool. Four Ways to Improve Verbo-visual Fusion for Dense 3D Visual Grounding. In *Computer Vision – ECCV 2024*, pages 196–213, Cham, 2025. Springer Nature Switzerland. 1
- [35] Peng Wang, Shuai Bai, Sinan Tan, and et al. Qwen2-VL: Enhancing Vision-Language Model’s Perception of the World at Any Resolution, 2024. 2, 6
- [36] Sai Wang, Yutian Lin, and Yu Wu. Omni-Q: Omni-Directional Scene Understanding for Unsupervised Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14261–14270, 2024. 2
- [37] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-Form Video Understanding with Large Language Model as Agent. In *Computer Vision – ECCV 2024*, pages 58–76, Cham, 2025. Springer Nature Switzerland. 2
- [38] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8006–8014, 2025. 7
- [39] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic Graph CNN for Learning on Point Clouds. *ACM Trans. Graph.*, 38(5):146:1–146:12, 2019. 2
- [40] Yuan Wang, Yali Li, and Shengjin Wang. G³-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*, pages 13917–13926, 2024. 1, 2, 6
- [41] 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*, pages 2662–2671, 2023. 6, 7
- [42] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical Open-

- Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024. 1
- [43] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3D-STMN: Dependency-Driven Superpoint-Text Matching Network for End-to-End 3D Referring Expression Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):5940–5948, 2024. 2
- [44] Dapeng Wu, Zhigang Yang, Puning Zhang, Ruyan Wang, Boran Yang, and Xinqiang Ma. Virtual-Reality Interpromotion Technology for Metaverse: A Survey. *IEEE Internet of Things Journal*, 10(18):15788–15809, 2023. 1
- [45] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. EDA: Explicit Text-Decoupling and Dense Alignment for 3D Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19231–19242, 2023. 1, 2, 6
- [46] Can Xu, Yuehui Han, Rui Xu, Le Hui, Jin Xie, and Jian Yang. Multi-Attribute Interactions Matter for 3D Visual Grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17253–17262, 2024. 1, 2
- [47] 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*, pages 3961–3985. PMLR, 2025. 2, 3, 6, 7
- [48] 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. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7694–7701, 2024. 6
- [49] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. SAT: 2D Semantics Assisted Training for 3D Visual Grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1856–1866, 2021. 1
- [50] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. 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*, pages 1791–1800, 2021. 1, 2, 7
- [51] 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*, pages 20623–20633, 2024. 1, 2, 6, 7, 8
- [52] Tatiana Zemskova and Dmitry Yudin. 3DGraphLLM: Combining Semantic Graphs and Large Language Models for 3D Scene Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8885–8895, 2025. 5
- [53] Sha Zhang, Di Huang, Jiajun Deng, Shixiang Tang, Wanli Ouyang, Tong He, and Yanyong Zhang. Agent3D-Zero: An Agent for Zero-Shot 3D Understanding. In *Computer Vision – ECCV 2024*, pages 186–202, Cham, 2025. Springer Nature Switzerland. 2
- [54] Yachao Zhang, Yanyun Qu, Yuan Xie, Zonghao Li, Shanshan Zheng, and Cuihua Li. Perturbed Self-Distillation: Weakly Supervised Large-Scale Point Cloud Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15520–15528, 2021. 2
- [55] Yachao Zhang, Runze Hu, Ronghui Li, Yanyun Qu, Yuan Xie, and Xiu Li. Cross-Modal Match for Language Conditioned 3D Object Grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7359–7367, 2024. 2
- [56] Yachao Zhang, Yuxiang Lan, Yuan Xie, Cuihua Li, and Yanyun Qu. Cross-Cloud Consistency for Weakly Supervised Point Cloud Semantic Segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(8):14452–14463, 2025. 2
- [57] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation Modeling for Visual Grounding on Point Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021. 1, 2, 7
- [58] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. 1, 2, 6