AnyGround3D: Towards Grounding Any 3D Object in the Wild via 2D-to-3D Lifting

Yingping Liang[1] and Wenxuan Guo[1]†

Tsinghua University, Beijing, China
liangyingping@gmail.com, gwx22@mails.tsinghua.edu.cn

Abstract. 3D visual grounding aims to localize language-referred objects in 3D scenes. However, existing methods rely on limited annotated datasets under controlled environments and thus struggle to generalize in the wild. Therefore, the scarcity of scalable 3D grounding supervision remains a major challenge. In this paper, we propose **AnyGround3D**, a scalable framework that learns to ground arbitrary 3D objects in the wild via 2D-to-3D lifting. Instead of requiring manual 3D annotations, our approach lifts diverse Internet images into structured 3D supervision through multi-level 2D-to-3D lifting. First, we perform scene-level 2D-to-3D lifting to reconstruct metric point clouds from monocular images, providing geometric context. Second, we conduct object-level 2D-to-3D lifting to recover complete 3D object representations with fine-gained 3D bounding boxes and neural language descriptions. To further enhance the representation robustness and open-vocabulary generalization of the point encoder, we introduce feature-level lifting that distills semantic representations from 2D foundation models into the 3D encoder. Together, these lifting processes enable large-scale synthetic 3D grounding supervision across diverse real-world scenes. Extensive experiments demonstrate improved in-the-wild generalization across multiple 3D grounding benchmarks, without relying on human-annotated 3D grounding data.

Keywords: 3D Visual Grounding · Data Scaling · Feature Learning

1 Introduction

3D visual grounding aims to accurately localize language-referred objects within 3D scenes [10,34,20] and serves as a fundamental capability for embodied perception, robotics, and open-world 3D understanding [32,44,51,7,31,45,43]. Recent advances have demonstrated promising progress on curated 3D benchmarks [4,25,38], where dense annotations and structured scene reconstructions provide explicit supervision for learning cross-modal alignment.

Despite these advances, existing 3D grounding approaches remain fundamentally constrained by their training data [4,52,20,24]. Most current methods rely

⋆ † Corresponding author.

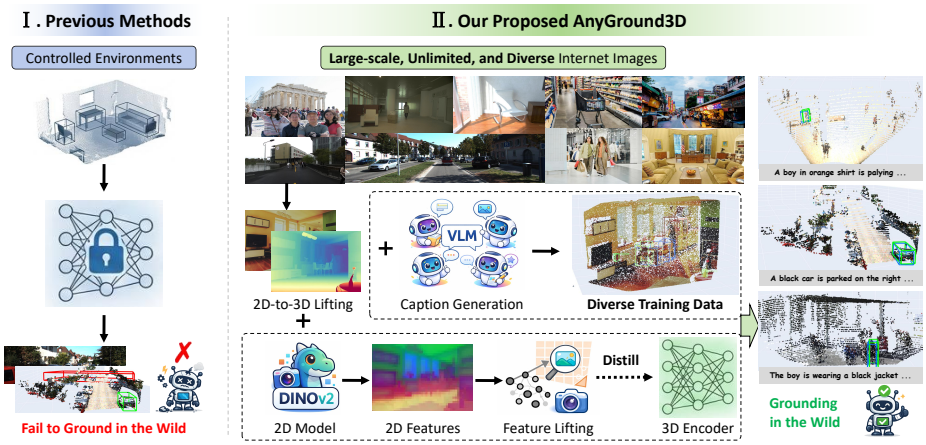

Fig. 1. Prior approaches rely on data collected in controlled 3D environments with structured annotations. AnyGround3D instead learns from large-scale and diverse Internet images via 2D-to-3D lifting, enabling scalable in-the-wild 3D grounding.

on limited-scale captured datasets with 3D bounding boxes and language descriptions. Such datasets are expensive to construct, limited in diversity, and biased toward static and clean environments. As a result, models trained under these settings often struggle to generalize to in-the-wild scenarios characterized by complex object layouts and diverse object types [34,10]. The scarcity of scalable 3D grounding supervision thus remains a major bottleneck.

In contrast, large-scale Internet images provide a diverse source of visual data. Compared to curated 3D datasets, Internet imagery covers significantly broader scene categories, object distributions, and real-world conditions [49,22]. However, leveraging such data for 3D grounding is non-trivial. Internet images lack explicit 3D structure, object-level 3D annotations, and grounded language supervision. Consequently, existing 3D grounding frameworks are unable to directly exploit the scale and realism offered by unconstrained visual data.

In this work, we address these challenges by proposing **AnyGround3D**, a scalable framework that learns robust 3D visual grounding via monocular scene lifting. As shown in Figure 1, instead of relying on manual 3D grounding annotations, AnyGround3D lifts Internet images into structured 3D supervision through multi-level 2D-to-3D lifting. First, we perform scene-level 2D-to-3D lifting, reconstructing metric point clouds from monocular images to recover geometric context. Second, we conduct object-level 2D-to-3D lifting, transforming 2D object instances into complete 3D object representations and synthesizing neural language descriptions for grounding supervision. Third, we introduce feature-level lifting, which distills semantic representations from 2D foundation models into the 3D encoder, enhancing representation robustness and open-vocabulary generalization. Together, these lifting processes enable large-scale synthetic 3D grounding supervision across diverse real-world scenes.

Specifically, AnyGround3D establishes a scalable lifting-based data generation pipeline that transforms unconstrained Internet images into geometrically structured 3D scenes with object-level associations. By combining geometric reconstruction, object-level lifting, and description generation, our framework automatically constructs synthetic 3D grounding pairs without requiring manual 3D annotations. This design naturally accommodates diverse object categories and open-vocabulary queries, facilitating open-world generalization.

Built upon this scalable supervision, we further adopt a cross-dimensional pretraining strategy for the 3D encoder. By distilling semantic representations from DINOv2 [30] into the lifted 3D feature space, the encoder inherits strong 2D semantic priors while maintaining 3D geometric consistency. This cross-dimensional alignment enhances scene understanding across diverse environments and provides a more informative initialization for 3D grounding training.

To comprehensively evaluate generalization, we conduct experiments across multiple 3D grounding benchmarks spanning in-the-wild scenes, diverse object categories, and zero-shot settings. Experimental results demonstrate that AnyGround3D consistently improves in-the-wild robustness and cross-domain generalization, without relying on human-annotated 3D grounding data.

In summary, our main contributions are as follows:

- We propose **AnyGround3D**, a scalable framework for in-the-wild 3D visual grounding that learns from synthetic supervision via 2D-to-3D lifting.
- We introduce a lifting-based 3DVG data generation pipeline that transforms large-scale diverse Internet images into structured 3D scenes with automatically synthesized grounding supervision.
- We present a feature-level cross-dimensional pretraining strategy that distills 2D foundation models into a 3D encoder, improving semantic robustness and open-vocabulary generalization.
- Experimental results demonstrate the strong in-the-wild and zero-shot performance of our method across diverse 3D grounding benchmarks, without relying on manually annotated 3D grounding data.

2 Related Work

3D Visual Grounding. 3D visual grounding (3DVG) aims to localize language-referred objects in 3D scenes by aligning linguistic expressions with geometric representations. Existing methods are predominantly trained on indoor RGB-D benchmarks such as ScanRefer [4] and Nr3D [1], built upon reconstructed datasets like ScanNet [5]. Early approaches follow proposal-based pipelines that match detected 3D objects with language features [53,47], while more recent works adopt transformer-based end-to-end architectures [15,11]. To enhance structural modeling and efficiency, MCLN [34] introduces collaborative geometric-language feature learning, and TSP3D [10] proposes text-guided sparse voxel pruning to reduce irrelevant spatial computation. SeeGround [21] and ReasonGrounder [27] propose zero-shot 3DVG frameworks but still focus on indoor scenes. Recent works [12,13,46,50] further explore leveraging vision-language

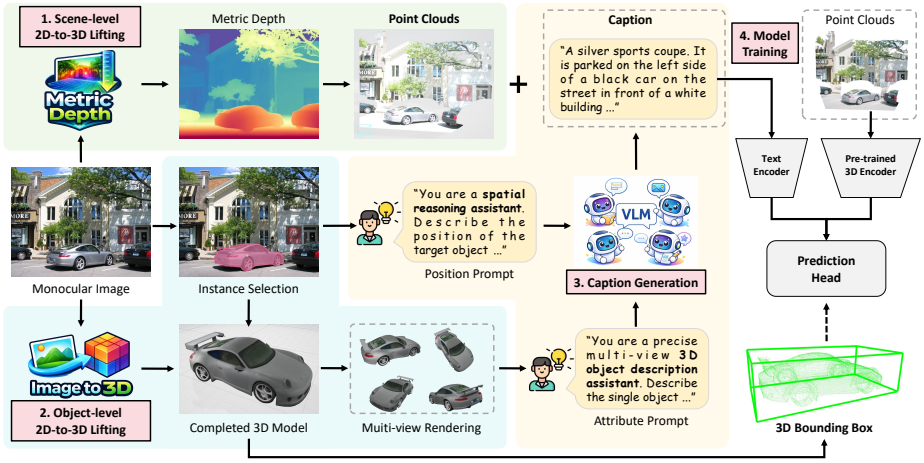

Fig. 2. Overview of our proposed data generation and model training process.

models (VLMs) for 3DVG. However, they typically rely on pre-detected or ground-truth 3D bounding boxess, limiting their ability to generalize to unseen environments without predefined object proposals. Therefore, current 3DVG methods remain limited in scale, diversity, and real-world variability. In contrast, we address this by leveraging lifting-based supervision from large-scale Internet images to enable scalable in-the-wild 3D grounding.

3D Grounding in the Wild. Grounding language in outdoor 3D scenes introduces challenges such as large spatial scales, sparse point clouds, and diverse object distributions [17,18,42,36,37,35]. Talk2Car [6], built on nuScenes [3], pioneers referential grounding for autonomous driving scenarios. STRefer [24] extends grounding to embodied outdoor settings with synchronized RGB and LiDAR streams. Mono3DVG [52] investigates monocular 3D grounding without explicit depth sensors. KITTI360Pose [16], based on KITTI-360 [9], focuses on text-to-position grounding with templated expressions, while Talk2LiDAR [26] and CityRefer [29] explore multi-sensor and city-scale scenarios. LabelAny3D [48] proposes a scalable pipeline for labeling arbitrary RGB images and corresponding 3D bounding boxes. However, it focuses on object-level 3D box annotation without associated language descriptions. More recently, 3EED [20] introduces embodied outdoor grounding with richer scene dynamics and long-range spatial reasoning. However, it still relies on manually collected multi-sensor data and human-annotated grounding supervision, resulting in limited scalability and domain coverage. In contrast, our method leverages scalable 2D-to-3D lifting from diverse Internet imagery, enabling large-scale synthetic 3D grounding supervision beyond platform-specific data collection.

3 Method

In this section, we first introduce the pipeline of our scalable data generation and grounding training process, as shown in Figure 2. Then, we present the feature-level lifting strategy for encoder pre-training, as shown in Figure 3. Finally, we describe the model training protocol and implementation details used in both pre-training and grounding optimization.

3.1 Scalable 3D Grounding Data Generation

Given a collection of unconstrained Internet images, our goal is to automatically construct synthetic 3D grounding supervision without manual 3D annotations. Starting from a monocular RGB image, we progressively lift 2D observations into structured 3D representations through scene-level, object-level, and language-guided lifting processes. Formally, let an input image be denoted as $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$. Our data generation pipeline produces a set of 3D grounding samples:

$$\mathcal{G} = \{(\mathcal{P}, \mathbf{b}, \mathbf{t})\},$$

where $\mathcal{P}$ denotes the reconstructed 3D point cloud of the scene, $\mathbf{b} \in \mathbb{R}^7$ represents a 3D bounding box parameterized by center, size, and orientation, and $\mathbf{t}$ is the associated language description.

Scene-level 2D-to-3D Lifting. We first recover the geometric structure of the scene from the monocular image. A depth estimator predicts a metric depth map $\mathbf{D} \in \mathbb{R}^{H\times W}$, where each scalar $d_{u,v}$ corresponds to the depth value at pixel location (u, v). Given camera intrinsics $\mathbf{K} \in \mathbb{R}^{3\times 3}$, each pixel $\mathbf{p}_{u,v} = (u, v, 1)^\top$ is back-projected into 3D space as:

$$\mathbf{x}_{u,v} = d_{u,v}\,\mathbf{K}^{-1}\mathbf{p}_{u,v},$$

where $\mathbf{x}_{u,v} \in \mathbb{R}^3$ denotes the 3D point in camera coordinates.

Aggregating all valid pixels yields a scene-level point cloud $\mathcal{P} = \{\mathbf{x}_{u,v}\}$. This process lifts the 2D image into a geometrically structured 3D representation that captures spatial layout and depth relationships.

Object-level 2D-to-3D Lifting. To obtain object-level grounding targets, we first apply an instance segmentation model to extract 2D object masks $\mathbf{M}_k \in \{0, 1\}^{H\times W}$, where k indexes object instances.

We further apply a single-image 3D reconstruction model [41] to recover a complete object-level 3D mesh $\mathcal{O}_k$. A 3D bounding box is then estimated from $\mathcal{O}_k$ and parameterized as:

$$\mathbf{b}_k = (\mathbf{c}_k, \mathbf{s}_k, \theta_k),$$

where $\mathbf{c}_k \in \mathbb{R}^3$ denotes the box center, $\mathbf{s}_k \in \mathbb{R}^3$ represents box dimensions, and $\theta_k \in \mathbb{R}$ is the heading angle.

Language Supervision Synthesis. To construct grounding queries, we generate language descriptions conditioned on both geometric context and object attributes. Specifically, we prompt a vision-language model with multi-dimensional

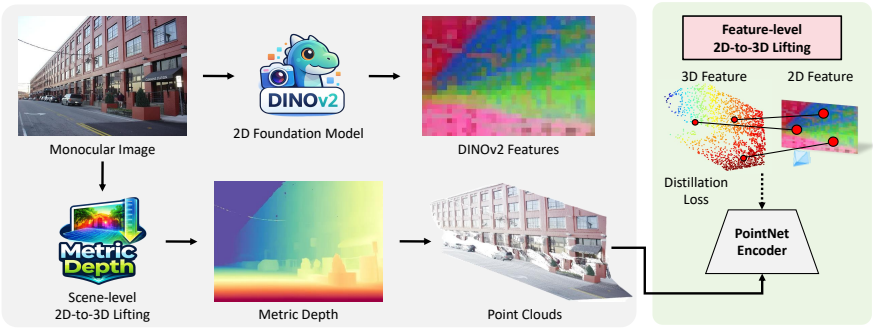

Fig. 3. Overview of our proposed encoder pre-training process.

object information to generate natural referring expressions:

$$\mathbf{t}_k = f_{\text{VLM}}(\mathbf{I}, \mathcal{O}_k, \mathbf{b}_k),$$

where $\mathbf{t}_k$ is a textual sequence describing the object's attributes and spatial relations within the scene.

By combining scene-level geometry $\mathcal{P}$, object-level bounding boxes $\mathbf{b}_k$, and synthesized language descriptions $\mathbf{t}_k$, we obtain large-scale 3D grounding supervision without manual annotation. Across diverse Internet images, this lifting-based pipeline enables scalable construction of synthetic 3D grounding pairs that span varied object categories and real-world environments. This paradigm expands the availability of training data for 3D grounding.

3.2 Feature-level Lifting for Encoder Pre-training

While scene-level and object-level lifting provide large-scale synthetic 3D grounding supervision, the quality of the learned representation largely depends on the robustness of the 3D encoder. To enhance semantic generalization and open-vocabulary capability, we introduce a feature-level 2D-to-3D lifting strategy that transfers strong 2D foundation priors into the 3D feature space via distillation.

Lifting 2D Features to 3D Points. Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, a pre-trained 2D foundation model (e.g., DINOv2 [30]) produces a dense semantic feature map $\mathbf{F}^{2D} \in \mathbb{R}^{H\times W\times C}$, where each feature vector $\mathbf{f}^{2D}_{u,v} \in \mathbb{R}^C$ corresponds to the pixel location (u, v).

Using the back-projection, each pixel (u, v) is mapped to a 3D point $\mathbf{x}_{u,v} \in \mathbb{R}^3$. We associate the 2D semantic feature with its corresponding 3D point, yielding a lifted point-wise feature set:

$$\mathcal{F}^{2D\to 3D} = \{(\mathbf{x}_{u,v}, \mathbf{f}^{2D}_{u,v})\}.$$

This establishes a correspondence between geometric coordinates in 3D space and high-level semantic representations from the 2D foundation model.

3D Encoder Representation. Let the 3D encoder be denoted as $g_\theta(\cdot)$ parameterized by θ. Given the reconstructed point cloud:

$$\mathcal{P} = \{\mathbf{x}_i\}_{i=1}^{N},$$

the encoder produces point-wise 3D features:

$$\mathbf{f}_i^{3D} = g_\theta(\mathcal{P})_i, \quad \mathbf{f}_i^{3D} \in \mathbb{R}^C.$$

Cross-Dimensional Feature Distillation. To transfer semantic knowledge from 2D to 3D space, we align the 3D point features with the lifted 2D semantic features. For each valid correspondence i, we minimize a feature regression loss:

$$\mathcal{L}_{\text{distill}} = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{f}_i^{3D} - \mathbf{f}_i^{2D} \right\|_2^2,$$

where $\mathbf{f}_i^{2D}$ denotes the 2D feature associated with the same 3D point.

By encouraging consistency between lifted 2D semantic features and 3D geometric representations, the encoder inherits rich semantic priors from large-scale 2D foundation models while maintaining spatial awareness. This cross-dimensional alignment enhances scene understanding across diverse environments and provides a stronger initialization for 3D grounding training.

3.3 Implementation Details

Data Sources. For scalable data generation across diverse environments, we collect images from the COCO [23] and KITTI [8] datasets, which span a wide range of object categories and real-world outdoor scenes.

Scene-level and Object-level Lifting. For scene-level geometry estimation, following LabelAny3D [48], we adopt the MoGe [39] model to obtain affine-invariant depth and camera intrinsics, and align it to metric depth predicted by Depth Pro [2] to recover real-world scale. The completed object is reconstructed into a 3D mesh using TRELLIS [41], which predicts canonical-pose meshes with normalized scale. To localize each object in the scene, we estimate its pose via dense correspondence matching between the real image and rendered mesh views using MASt3R [19]. Metric scale is recovered via depth alignment between rendered and real depth maps. The reconstructed mesh is then transformed into the scene coordinate frame. The reconstructed scene point clouds are uniformly downsampled to 30,000 points for storage efficiency. Bounding boxes containing fewer than 20 points are discarded to remove unreliable object proposals.

Language Supervision. Object-centric language descriptions are generated using the vision-language model Qwen2.5-VL-32B-Instruct [14]. We prompt the model with both positional cues and attribute-oriented instructions to synthesize grounding expressions describing object appearance and spatial relations.

Encoder Pre-training. Our grounding model is implemented in PyTorch and follows the training schedule and architectural design of 3EED [20], with modifications to accommodate our synthetic supervision. For cross-dimensional feature

lifting, we use DINOv2-large [30] as the 2D foundation model to extract dense semantic features. All input images are resized to a resolution of 560×560.
Grounding Model Training. To evaluate on 3EED dataset and Mono3DRefer dataset, we follow the model architecture and experimental setup of 3EED. The reconstructed point cloud $\mathcal{P}$ is uniformly down-sampled to 16,384 points. The points are encoded by a PointNet++ [33] backbone trained from scratch, whose final layer produces 1,024 visual tokens. An MLP head predicts an objectness score for each token, and the top 256 tokens are selected as object proposals and fed into a six-layer Transformer decoder for cross-modal grounding. For evaluation on WildRefer dataset, we follow the model architecture and experimental setup of WildRefer. The reconstructed point cloud $\mathcal{P}$ is down-sampled to 30,000 points. All experiments are conducted using four NVIDIA RTX 3090 GPUs.
Training Data Composition. Our model is trained jointly on a mixture of synthetic grounding data generated from COCO and KITTI, together with the training set of existing 3D grounding datasets including 3EED [20], WildRefer [24], and ScanRefer [4].

4 Experiments

In this section, we first introduce the datasets and evaluation metrics. Then, detailed comparisons are conducted with the state-of-the-art methods. Finally, ablations and discussions are performed.

4.1 Evaluation Datasets

We evaluate AnyGround3D on multiple in-the-wild 3DVG benchmarks, including 3EED [20], WildRefer [24], and Mono3DRefer [52]. These datasets exhibit distinct object distributions and scene complexities, providing a comprehensive testbed for evaluating 3D visual grounding in real-world scenarios.
3EED [20]. 3EED features RGB images and LiDAR scans collected from heterogeneous platforms and dynamic outdoor environments, including vehicles, drones, and quadruped robots. We report Acc@0.25 and Acc@0.5n.
WildRefer [24]. WildRefer focuses on large-scale daily-life outdoor scenes with human-centric activities and dense language annotations. The dataset includes two subsets, STRrefer and LifeRefer, emphasizing long-range and fine-grained grounding, respectively.
Mono3DRefer [52]. Mono3DRefer targets autonomous driving scenarios and provides 2D images with corresponding 3D bounding boxes. We reconstruct point cloud inputs from the original LiDAR scans and evaluate models without any specific fine-tuning on this benchmark.

4.2 Performance on 3EED

Table 1 presents the quantitative results on the 3EED benchmark, which includes diverse platforms such as vehicle-mounted, drone-based, and quadruped-mounted sensors. All methods are trained on the official 3EED training split.

Table 1. Performance on 3EED [20] benchmark featuring RGB and LiDAR data from vehicle, drone, and quadruped platforms. Acc@25/50 are given in percentage (%). * indicates the one-stage version without pretrained 3D decoder, as BUTD-DETR and EDA can run without pre-detection inputs.

Method	Vehicle		Drone		Quad		Union	
	@25	@50	@25	@50	@25	@50	@25	@50
3D-SPS [28]	59.14	35.90	40.12	6.04	38.87	17.54	48.88	22.10
BUTD-DETR* [15]	63.41	40.88	44.20	8.28	43.14	20.94	51.41	24.80
EDA* [40]	65.50	41.80	46.00	8.60	44.00	21.50	52.46	25.02
MCLN [34]	67.14	42.20	46.58	8.92	44.80	21.32	52.21	25.64
TSP3D [10]	66.94	42.55	45.71	44.58	46.66	24.30	52.31	26.03
WildRefer [24]	51.51	10.12	50.27	**9.85**	45.36	20.29	49.27	13.11
3EED [20]	80.86	50.11	53.45	9.75	53.31	24.08	63.84	29.66
AnyGround3D	**80.88**	**51.41**	**53.58**	9.81	**61.60**	**26.68**	**66.76**	**31.34**

Compared with prior grounding models, AnyGround3D achieves competitive or superior performance across all platforms. On the vehicle subset, our method achieves 80.88% / 51.41% at Acc@0.25/0.5, performing on par with the 3EED baseline while maintaining a clear margin over earlier approaches such as TSP3D, MCLN, and EDA. Notably, improvements of our proposed method are more pronounced on challenging drone and quadruped settings. On the quadruped platform, AnyGround3D achieves 61.60% / 26.68%, substantially outperforming the original 3EED model and previous methods. This indicates stronger robustness under viewpoint changes and sensor distribution shifts.

On the union split combining all platforms, our method achieves 66.76% / 31.34%, establishing the best overall performance among compared approaches. These results demonstrate that leveraging scalable lifting-based supervision improves cross-platform generalization without relying on platform-specific architectural modifications.

4.3 Performance on WildRefer

Table 2 reports the performance on the WildRefer benchmark, which emphasizes large-scale human-centric daily-life scenarios. All methods are trained on the official WildRefer training split for fair comparison.

WildRefer presents a distinct distribution shift compared to vehicle-centric or indoor datasets, as the majority of expressions focus on persons and fine-grained spatial relations in dynamic outdoor environments. Despite this distribution bias, AnyGround3D achieves competitive performance across both STRrefer and LifeRefer subsets. On the union split, our method consistently maintains strong localization accuracy under both IoU thresholds, demonstrating robustness to long-range references and complex human-object interactions. These results indicate that the lifting-based supervision and feature-level distillation improve

Table 2. Performance on WildRefer benchmarks [24], which focus on large-scale human-centric dailylife scenarios accompanied with abundant 3D object and natural language annotations.

Method	STRrefer		LifeRefer		Union	
	@0.25	@0.5	@0.25	@0.5	@0.25	@0.5
3D-SPS [28]	44.47	42.40	28.01	18.20	36.24	30.30
BUTD-DETR* [15]	56.66	45.12	32.46	12.82	44.56	28.97
EDA* [40]	57.41	45.59	29.32	12.25	43.37	28.92
MCLN [34]	58.81	47.20	33.04	13.30	44.42	31.39
TSP3D [10]	59.02	47.76	33.10	15.58	47.18	31.66
WildRefer [24]	62.01	54.97	38.89	18.42	50.45	36.70
3EED [20]	52.98	44.10	35.35	15.81	44.16	29.95
AnyGround3D	**63.79**	**59.10**	**39.56**	**18.90**	**51.67**	**39.00**

Table 3. Generalization performance of in-the-wild methods without specific fine-tuning on Mono3DRefer [52] benchmarks, which focus on autonomous driving scenarios. Note that the original data only provide 3D bounding boxes and 2D images. We construct the point cloud inputs using the original LiDAR scans.

Method	Training Data	Val set		Test set	
		@0.25	@0.5	@0.25	@0.5
WildRefer [24]	STRefer+LifeRefer	14.46	1.39	14.75	1.34
3EED [20]	3EED Dataset	25.74	12.19	25.85	13.46
AnyGround3D	**Ours**	**58.90**	**49.59**	**62.71**	**54.79**

semantic generalization beyond object-category priors, enabling stable grounding performance even in human-centric scenarios.

4.4 Qualitative Results

Figure 4 presents qualitative comparisons on representative in-the-wild scenes. The examples cover both vehicle-centric and human-centric scenarios, involving long-range references and fine-grained spatial descriptions. Previous methods such as WildRefer and 3EED are prone to failure under large viewpoint variations or long-distance references, leading to spatially shifted or semantically inconsistent predictions.

In contrast, AnyGround3D produces more accurate 3D localization results that align better with the described object attributes and spatial relations. Notably, in challenging cases involving distant targets or subtle relational cues, our model maintains stable grounding performance. This improvement suggests that the proposed lifting-based supervision and cross-dimensional feature distillation enhance spatial reasoning and semantic generalization.

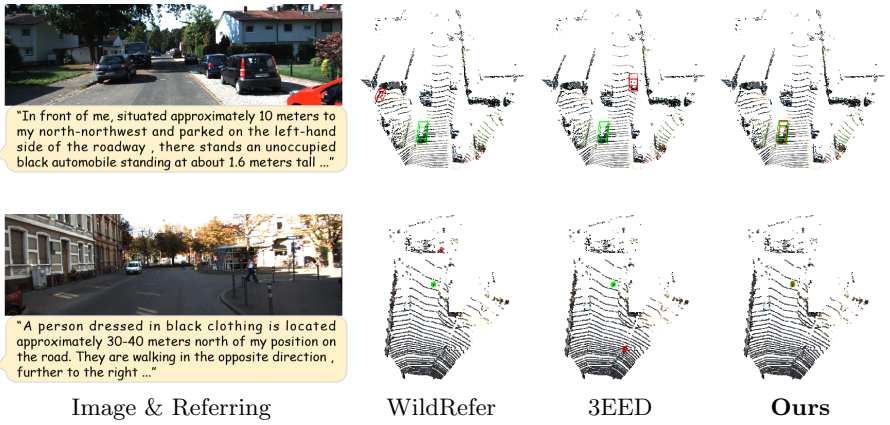

Fig. 4. Qualitative comparison of 3D visual grounding results on complex in-the-wild scenes. Green boxes denote correct grounding results, while red boxes indicate the predictions. Compared with prior methods, our AnyGround3D produces more accurate and spatially consistent localization under long-range and fine-grained descriptions.

4.5 Performance on Mono3DRefer

Table 3 evaluates the generalization on Mono3DRefer dataset. Mono3DRefer focuses on autonomous driving scenarios and differs significantly from 3EED and WildRefer in both object distribution and sensor configuration. For evaluation, we reconstruct point cloud inputs from the original LiDAR scans.

Methods trained on other benchmarks exhibit severe performance degradation under this distribution shift. In particular, the 3EED-trained model achieves only 16.01% / 5.23% on the validation set at Acc@0.25/0.5. Besides, WildRefer-trained models completely fail to generalize, likely because WildRefer predominantly contains pedestrian instances, resulting in a strong object-category bias that does not transfer to our more diverse target domain.

In contrast, AnyGround3D achieves 58.90% / 49.59% on the validation set and 62.66% / 54.88% on the test set, substantially outperforming prior approaches. The large margin under both IoU thresholds demonstrates that scalable lifting-based supervision significantly improves cross-domain robustness. These results highlight the effectiveness of our data scaling and cross-dimensional feature lifting strategy in mitigating domain bias and enabling open-world 3D grounding.

5 Discussions

In this section, we first conduct ablation studies to analyze the contribution of each component in our framework. Then, we examine data scaling beyond curated benchmarks and discuss how lifted supervision improves scene and category coverage. Next, we analyze language distribution and category bias across

Table 4. Ablation study on key components of AnyGround3D.

Ablation	3EED	
	Acc@0.25 ↑	Acc@0.5 ↑
Full (AnyGround3D)	**66.76**	**31.34**
w/o Encoder pre-train (no 2D→3D distillation)	66.34	31.31
w/o Caption synthesis (no VLM descriptions)	64.10	28.59
w/o Generated data (no 2D-lifted 3D data)	65.21	31.11

datasets. Finally, we study the effect of feature-level lifting through visualization and representation comparison.

5.1 Ablations

Table 4 presents ablation results on the 3EED benchmark to quantify the contribution of each component in AnyGround3D.

Effect of 2D→3D Encoder Pre-training. Removing encoder pre-training (i.e., disabling 2D-to-3D feature distillation) leads to a consistent, though moderate, performance drop from 66.76% to 66.34% at Acc@0.25. This indicates that transferring dense semantic priors from a strong 2D foundation model improves feature initialization and representation quality in 3D space. The gain is more evident under stricter localization metrics, suggesting improved spatial alignment between visual and geometric representations.

Effect of Caption Synthesis. Removing VLM-based caption synthesis results in the most significant degradation, dropping to 64.10% / 28.59%. The decrease of 2.66% at Acc@0.25 and 2.75% at Acc@0.5 highlights the importance of enriched language supervision. Synthetic object-centric descriptions introduce diverse attribute cues and spatial relations that are underrepresented in curated benchmarks. Without this linguistic augmentation, the model exhibits reduced robustness to fine-grained and long-range expressions.

Effect of Generated 2D-Lifted 3D Data. Excluding generated 2D-lifted 3D data reduces performance to 65.21% / 31.11%. While the drop is smaller compared to removing caption synthesis, it consistently affects both metrics. This suggests that scalable lifted supervision enhances scene diversity and object coverage, improving generalization under heterogeneous outdoor settings.

5.2 Data Scaling Beyond Curated Benchmarks

Table 5 highlights the scale and distribution differences among existing 3D grounding datasets. Most prior benchmarks are either indoor-focused or platform-specific, with limited scene diversity and strong category bias. Although recent datasets increase annotation volume, they remain constrained by data collection settings and sensor configurations.

In contrast, our lifting-based pipeline leverages large-scale Internet imagery to construct structured 3D grounding supervision without requiring manual 3D

Table 5. Training dataset comparison of existing 3D visual grounding benchmarks.

Scenario	Dataset	Source	Type	#Scenes	#Objects	#Captions
Indoor	ScanRefer	Real	RGB-D	0.8K	11K	51K
	Nr3D	Real	RGB-D	0.7K	4K	41K
	Sr3D	Real	RGB-D	0.8K	5K	83K
	SceneVerse	Real+CAD	RGB-D	68K	1.5M	2.5M
Outdoor	Mono3DRefer	Real	RGB	2K	8K	41K
	KITTI360Pose	Real	Lidar	–	15K	43K
	CityRefer	Real	Lidar	–	5K	35K
	STRefer	Real	Lidar+RGB	0.6K	3K	5K
	LifeRefer	Real	Lidar+RGB	3K	11K	25K
	Talk2LiDAR	Real	Lidar+RGB	6K	–	59K
	Talk2Car-3D	Real	Lidar+RGB	5K	–	10K
	3EED	Real	Lidar+RGB	0.2K	12K	22K
	Ours	Real	RGB-D	**18K**	**65K**	**130K**

Fig. 5. Word cloud comparison across different 3D grounding datasets. ScanRefer primarily focuses on indoor furniture and room-level objects, 3EED contains predominantly vehicle-related descriptions, and WildRefer mainly consists of person-centric expressions. In contrast, our generated data exhibits substantially broader semantic diversity, covering varied object categories, attributes, and spatial relations.

annotations. As shown in Table 5, our generated dataset substantially increases scene coverage and object diversity compared to existing outdoor benchmarks. This scalability plays a critical role in improving cross-domain robustness, as evidenced by the strong zero-shot performance on Mono3DRefer.

5.3 Language Distribution and Category Bias

Figure 5 illustrates the distribution of grounding expressions across datasets. Existing benchmarks exhibit clear semantic concentration: ScanRefer primarily describes indoor furniture and room-level objects, 3EED is largely vehicle-centric, and WildRefer emphasizes person-centric interactions. Such biases may encourage models to rely on dataset-specific priors rather than generalizable representations.

By constructing supervision from diverse Internet scenes, our generated data presents broader semantic coverage, including varied object categories, attributes, and spatial relations. This diversity reduces reliance on narrow object priors and

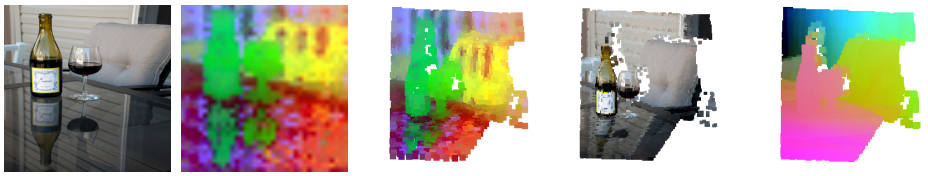

Fig. 6. Feature visualization. The original 3D features primarily capture geometric structure, leading to ambiguities for text-guided grounding. In contrast, the lifted 3D features, enriched by 2D foundation model representations, incorporate high-level semantic cues and appearance priors learned from large-scale image data.

contributes to improved generalization across heterogeneous platforms and environments.

5.4 Effect of Feature-level Lifting

Figure 6 visualizes the difference between original 3D features and lifted 3D features after cross-dimensional distillation. The original 3D encoder primarily captures geometric structure, which may be insufficient for resolving fine-grained text queries involving appearance attributes or semantic context. After feature-level lifting, the 3D features inherit high-level semantic cues from 2D foundation models, resulting in more discriminative representations aligned with language descriptions. This observation supports the hypothesis that transferring 2D semantic priors into 3D space improves grounding robustness, particularly under occlusion, viewpoint variation, and domain shift.

6 Conclusion

In this work, we propose AnyGround3D, a scalable framework for 3D visual grounding that alleviates the reliance on manually annotated 3D datasets. By leveraging monocular Internet images, we construct large-scale synthetic 3D grounding supervision through scene-level geometric lifting, object-level structural lifting, and feature-level semantic lifting. Furthermore, this unified lifting pipeline bridges 2D foundation models and structured 3D representations, enabling effective cross-dimensional knowledge transfer. Extensive experiments across in-the-wild 3D visual grounding benchmarks demonstrate the strong robustness and generalization of our method, without dataset-specific fine-tuning. The results suggest that scalable 2D-to-3D lifting provides a practical path toward open-world 3D grounding beyond curated and controlled environments.

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: ReferIt3D: Neural listeners for fine-grained 3D object identification in real-world scenes. In: European Conference on Computer Vision. pp. 422–440. Springer (2020)
2. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: International Conference on Learning Representations (2025)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuScenes: A multimodal dataset for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020)
4. Chen, D.Z., Chang, A.X., Nießner, M.: ScanRefer: 3D object localization in RGB-D scans using natural language. In: European Conference on Computer Vision. pp. 202–221. Springer (2020)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5828–5839 (2017)
6. Deruyttere, T., Vandenhende, S., Grujicic, D., Van Gool, L., Moens, M.F.: Talk2Car: Taking control of your self-driving car. arXiv preprint arXiv:1909.10838 (2019)
7. Fong, W.K., Mohan, R., Hurtado, J.V., Zhou, L., Caesar, H., Beijbom, O., Valada, A.: Panoptic nuScenes: A large-scale benchmark for LiDAR panoptic segmentation and tracking. IEEE Robotics and Automation Letters **7**, 3795–3802 (2022)
8. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
9. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3354–3361 (2012)
10. Guo, W., Xu, X., Wang, Z., Feng, J., Zhou, J., Lu, J.: Text-guided sparse voxel pruning for efficient 3d visual grounding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3666–3675 (2025)
11. Huang, S., Chen, Y., Jia, J., Wang, L.: Multi-view transformer for 3D visual grounding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15524–15533 (2022)
12. Huang, T., Zhang, Z., Tang, H.: 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. arXiv preprint arXiv:2507.23478 (2025)
13. Huang, X., Wu, J., Xie, Q., Han, K.: 3drs: Mllms need 3d-aware representation supervision for scene understanding. In: Conference on Neural Information Processing Systems (2025)
14. Hui, B., Yang, J., Cui, Z., Yang, J., Liu, D., Zhang, L., Liu, T., Zhang, J., Yu, B., Lu, K., et al.: Qwen2. 5-coder technical report. arXiv preprint arXiv:2409.12186 (2024)
15. Jain, A., Gkanatsios, N., Mediratta, I., Fragkiadaki, K.: Bottom up top down detection transformers for language grounding in images and point clouds. In: European Conference on Computer Vision. pp. 417–433. Springer (2022)
16. Kolmet, M., Zhou, Q., Ošep, A., Leal-Taixé, L.: Text2Pos: Text-to-point-cloud cross-modal localization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6687–6696 (2022)

17. Kong, L., Ren, J., Pan, L., Liu, Z.: LaserMix for semi-supervised LiDAR semantic segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21705–21715 (2023)
18. Kong, L., Xu, X., Ren, J., Zhang, W., Pan, L., Chen, K., Ooi, W.T., Liu, Z.: Multi-modal data-efficient 3D scene understanding for autonomous driving. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(5), 3748–3765 (2025)
19. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: European conference on computer vision. pp. 71–91. Springer (2024)
20. Li, R., Dong, Y., Hu, T., Liang, A., Liu, Y., Lu, D., Pan, L., Kong, L., Liang, J., Liu, Z.: 3EED: Ground everything everywhere in 3D. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 38 (2025)
21. Li, R., Li, S., Kong, L., Yang, X., Liang, J.: Seeground: See and ground for zero-shot open-vocabulary 3d visual grounding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3707–3717 (2025)
22. Liang, Y., Fu, Y., Hu, Y., Shao, W., Liu, J., Zhang, D.: Flow-anything: Learning real-world optical flow estimation from large-scale single-view images. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(10), 8435–8452 (2025)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
24. Lin, Z., Peng, X., Cong, P., Zheng, G., Sun, Y., Hou, Y., Zhu, X., Yang, S., Ma, Y.: WildRefer: 3D object localization in large-scale dynamic scenes with multi-modal visual data and natural language. In: European Conference on Computer Vision. pp. 456–473. Springer (2024)
25. Liu, D., Liu, Y., Huang, W., Hu, W.: A survey on text-guided 3-d visual grounding: Elements, recent advances, and future directions. IEEE Transactions on Neural Networks and Learning Systems (2025)
26. Liu, Y., Sun, B., Zheng, G., Wang, Y., Wang, J., Wang, F.Y.: Talk to parallel LiDARs: A human-LiDAR interaction method based on 3D visual grounding. arXiv preprint arXiv:2405.15274 (2024)
27. Liu, Z., Wang, Y., Zheng, S., Pan, T., Liang, L., Fu, Y., Xue, X.: Reasongrounder: Lvlm-guided hierarchical feature splatting for open-vocabulary 3d visual grounding and reasoning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3718–3727 (2025)
28. Luo, J., Fu, J., Kong, X., Gao, C., Ren, H., Shen, H., Xia, H., Liu, S.: 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16454–16463 (2022)
29. Miyanishi, T., Kitamori, F., Kurita, S., Lee, J., Kawanabe, M., Inoue, N.: CityRefer: Geography-aware 3D visual grounding dataset on city-scale point cloud data. In: Advances in Neural Information Processing Systems. vol. 36 (2023)
30. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
31. Pan, Y., Gao, B., Mei, J., Geng, S., Li, C., Zhao, H.: SemanticPOSS: A point cloud dataset with large quantity of dynamic instances. In: IEEE Intelligent Vehicles Symposium. pp. 687–693 (2020)

32. Pendleton, S.D., Andersen, H., Du, X., Shen, X., Meghjani, M., Eng, Y.H., Rus, D., Ang, M.H.: Perception, planning, control, and coordination for autonomous vehicles. Machines **5**(1), 6 (2017)
33. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
34. Qian, Z., Ma, Y., Lin, Z., Ji, J., Zheng, X., Sun, X., Ji, R.: Multi-branch collaborative learning network for 3D visual grounding. In: European Conference on Computer Vision. pp. 381–398. Springer (2024)
35. Razani, R., Cheng, R., Taghavi, E., Bingbing, L.: Lite-HDSeg: LiDAR semantic segmentation using lite harmonic dense convolutions. In: IEEE International Conference on Robotics and Automation. pp. 9550–9556 (2021)
36. Tang, H., Liu, Z., Li, X., Lin, Y., Han, S.: TorchSparse: Efficient point cloud inference engine. In: Conference on Machine Learning and Systems (2022)
37. Tang, H., Yang, S., Liu, Z., Hong, K., Yu, Z., Li, X., Dai, G., Wang, Y., Han, S.: TorchSparse++: Efficient training and inference framework for sparse convolution on gpus. In: IEEE/ACM International Symposium on Microarchitecture (2023)
38. Wang, A., Gong, Z., Chang, A.X.: Vigil3d: A linguistically diverse dataset for 3d visual grounding. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 30453–30475 (2025)
39. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5261–5271 (2025)
40. Wu, Y., Cheng, X., Zhang, R., Cheng, Z., Zhang, J.: EDA: Explicit text-decoupling and dense alignment for 3D visual grounding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19231–19242 (2023)
41. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21469–21480 (2025)
42. Xiao, A., Huang, J., Guan, D., Cui, K., Lu, S., Shao, L.: PolarMix: A general data augmentation technique for LiDAR point clouds. Advances in Neural Information Processing Systems **35**, 11035–11048 (2022)
43. Xiao, A., Huang, J., Guan, D., Zhan, F., Lu, S.: Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In: AAAI Conference on Artificial Intelligence. pp. 2795–2803 (2022)
44. Xiao, A., Huang, J., Guan, D., Zhang, X., Lu, S., Shao, L.: Unsupervised point cloud representation learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(9), 11321–11339 (2023)
45. Xiao, A., Huang, J., Xuan, W., Ren, R., Liu, K., Guan, D., El Saddik, A., Lu, S., Xing, E.P.: 3D semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9382–9392 (2023)
46. Yang, J., Chen, X., Madaan, N., Iyengar, M., Qian, S., Fouhey, D.F., Chai, J.: 3d-grand: A million-scale dataset for 3d-llms with better grounding and less hallucination. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29501–29512 (2025)
47. Yang, Z., et al.: SAT: 2D semantics assisted training for 3D visual grounding. In: IEEE/CVF International Conference on Computer Vision. pp. 1856–1866 (2021)

48. Yao, J., Redoy, R.M., Elbaum, S., Dwyer, M.B., Cheng, Z.: Labelany3d: Label any object 3d in the wild. In: Neural Information Processing Systems (2025)
49. Yingping, L., Yutao, H., Wenqi, S., Ying, F.: Learning dense feature matching via lifting single 2d image to 3d space. In: Proceedings of IEEE International Conference on Computer Vision. pp. 6621–6631 (2011)
50. Zantout, N., Zhang, H., Kachana, P., Qiu, J., Chen, G., Zhang, J., Wang, W.: Sort3d: Spatial object-centric reasoning toolbox for zero-shot 3d grounding using large language models. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2201–2208. IEEE (2025)
51. Zeng, C., Wang, W., et al.: Self-supervised learning for point cloud data: A survey. Expert Systems with Applications **237**, 121354 (2024)
52. Zhan, Y., Yuan, Y., Xiong, Z.: Mono3DVG: 3D visual grounding in monocular images. In: AAAI Conference on Artificial Intelligence. vol. 38, pp. 6988–6996 (2024)
53. Zhao, L., et al.: 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In: IEEE/CVF International Conference on Computer Vision. pp. 2928–2937 (2021)