Mono3DVG-EnSD: Enhanced Spatial-aware and Dimension-decoupled Text Encoding for Monocular 3D Visual Grounding

Yuzhen Li[1], Min Liu[1]*, Zhaoyang Li[1], Yuan Bian[1], Xueping Wang[2], Erbo Zhai[1], Yaonan Wang[1]

[1]School of Artificial Intelligence and Robotics, Hunan University, Changsha, Hunan, China
[2]College of Information Science and Engineering, Hunan Normal University, Changsha, Hunan, China
{zzrs, liu_min, zhaoyli, yuanbian, wang_xueping, zhaierbo, yaonan}@hnu.edu.cn

Abstract

Monocular 3D Visual Grounding (Mono3DVG) is an emerging task that locates 3D objects in RGB images using text descriptions with geometric cues. However, existing methods face two key limitations. Firstly, they often over-rely on high-certainty keywords that explicitly identify the target object while neglecting critical spatial descriptions. Secondly, generalized textual features contain both 2D and 3D descriptive information, thereby capturing an additional dimension of details compared to singular 2D or 3D visual features. This characteristic leads to cross-dimensional interference when refining visual features under text guidance. To overcome these challenges, we propose Mono3DVG-EnSD, a novel framework that integrates two key components: the CLIP-Guided Lexical Certainty Adapter (CLIP-LCA) and the Dimension-Decoupled Module (D2M). The CLIP-LCA dynamically masks high-certainty keywords while retaining low-certainty implicit spatial descriptions, thereby forcing the model to develop a deeper understanding of spatial relationships in captions for object localization. Meanwhile, the D2M decouples dimension-specific (2D/3D) textual features from generalized textual features to guide corresponding visual features at same dimension, which mitigates cross-dimensional interference by ensuring dimensionally-consistent cross-modal interactions. Through comprehensive comparisons and ablation studies on the Mono3DRefer dataset, our method achieves state-of-the-art (SOTA) performance across all metrics. Notably, it improves the challenging Far(Acc@0.5) scenario by a significant +13.54%.

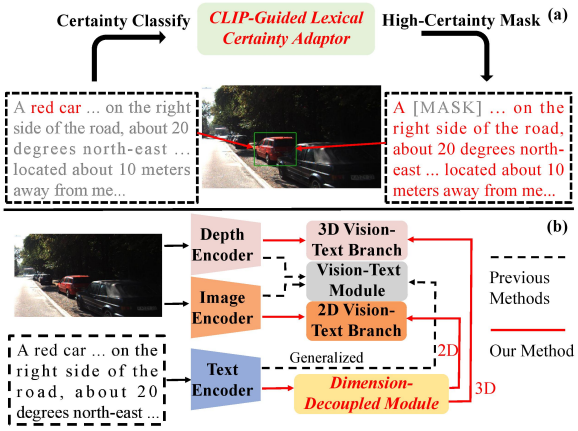

Figure 1: (a) Existing methods over-rely on high-certainty words (e.g., "red car") within descriptions (left), causing spatial description neglect. Our CLIP-LCA dynamically masks such high-certainty words during training, forcing the text encoder to comprehend spatial descriptors; (b) Previous methods employ generalized textual features (containing both 2D and 3D information) to refine both 3D visual (depth) features and 2D visual features, causing cross-dimensional interference. We propose D2M to decouple dimension-specific text features (2D and 3D separately) for dimensionally-consistent cross-modal interactions.

Introduction

The ability to locate objects through linguistic instructions constitutes a fundamental capability for human-robot interaction systems. While 2D visual grounding methods (Huang et al. 2022a; Zheng et al. 2025) have achieved significant progress in image understanding, their inherent limitation in depth perception restricts their capacity to interpret spatial relationships. Therefore, researchers have adopted multi-modal fusion techniques, which integrate complementary sensor data to enhance robust scene perception. For instance, RGB-D methods (Tan, Yang, and Wang 2024; Chen, Chang, and Nießner 2020) are prevalent in indoor scene understanding and LiDAR-camera fusion (Lin et al. 2024) is widely used in outdoor robotic perception. Despite their effectiveness, the broader adoption of these methods is still limited by the high expenses of RGB-D and LiDAR sensors.

Monocular 3D object detection (Li, Jia, and Shi 2024a) enables the estimation of 3D spatial location of objects from a single image. However, it often fails to capture the semantic context of the 3D environment. This limits its applications in human interaction systems, like robotics, where understanding natural language instructions is crucial for instruction-guided object localization. To bridge this gap, researchers have introduced monocular 3D visual grounding (mono3DVG) (Zhan, Yuan, and Xiong 2024a), a novel task that uses linguistic captions with geometry information

*Corresponding Author

to accurately localize 3D objects from a single RGB image. Unlike traditional 3D visual grounding methods that rely on point clouds or depth maps as 3D representations, the monocular approach operates solely on RGB images. To compensate for the absence of explicit 3D geometry in visual data, the mono3DVG task supplements captions with geometry descriptions (e.g., "10 meters away").

Compared to traditional 3D visual grounding, mono3DVG offers a more efficient and economical alternative. However, existing mono3DVG methods face two critical challenges. Firstly, captions in the mono3DVG not only provide enriched 3D spatial descriptors but also contain high-certainty keywords whose identification directly enables precise target localization. This leads the model to rely on simple keyword matching rather than developing a robust understanding of the underlying spatial semantics, thereby compromising its generalization capability in complex scenarios. As illustrated in Fig.1 (a), the keyword "red car" in the text description is sufficient to accurately locate the corresponding object in the image. This phenomenon may lead the text encoder to over-rely on high-certainty words while neglecting critical spatial descriptors such as "on the right side", "20 degrees north-east" and "10 meters away". Consequently, the model exhibits poor performance in scenarios where precise object localization requires a comprehensive understanding of spatial relationships. Secondly, previous methods suffer from cross-dimensional interference caused by the interaction between textual and visual features, as presented in Fig.1 (b). These methods employ separate image and depth encoders to extract 2D and 3D visual (depth) features respectively, while the text encoder integrates 2D descriptions, 3D descriptions, and common semantic information from captions into generalized textual features. The generalized textual features typically contain multi-dimensional semantic information, whereas visual features extracted from image encoder or depth encoder capture only single-dimensional details. When employing generalized textual features to refine 2D or 3D visual features, the textual features with multi-dimensional information may introduce irrelevant dimensional noise into single-dimensional visual features. For example, 2D attributes in generalized textual features (e.g., "red", "shirt") can distort 3D visual feature encoding, leading to inaccurate depth estimation.

To address the issue of high-certainty words in captions that explicitly identify corresponding objects, we propose a CLIP-Guided Lexical Certainty Adapter (CLIP-LCA), which dynamically adjusts the certainty levels of these keywords to encourage the text encoder to capture spatial information. During the training phase, we calculate a similarity score between each word in the textual description and the corresponding target region by leveraging CLIP's (Radford et al. 2021) image-text representation alignment capabilities. Based on these scores, we categorize the words into high-certainty and low-certainty classes. Words identified as high-certainty are masked during training, which balances attention of the text encoder between high-certainty keywords and low-certainty spatial descriptors. CLIP-LCA enhances the text encoder's ability to understand spatial relationships in descriptions, enabling more comprehensive sentence comprehension and consequently improving localization accuracy.

To resolve the cross-dimensional interference problem, we propose a Dimension-Decoupled Module (D2M). The D2M framework first employs two parallel cross-attention modules with distinct learnable embeddings to decouple generalized textual features into dimension-specific representations. Specifically, a 2D learnable embedding interacts with the generalized textual features to capture coarse 2D textual features, while a parallel 3D learnable embedding simultaneously learns coarse 3D textual features. To further enhance the dimension-specific information of the two coarse textual features without additional data support, we propose a dual-branch reverse cross-attention module. In the 2D branch, the coarse 2D textual features serve as queries (Q) and values (V), while the coarse 3D textual features act as keys (K). The low-attention regions between Q and K indicate dimensional discrepancies, thereby identifying 2D-specific details that require enhancement. The enhancement process first inverts low-attention weights and then applies them through matrix multiplication to the V. The 3D branch operates symmetrically when reversing the qkv configuration to enhance 3D part. D2M generates refined 2D-specific and 3D-specific textual features that provide more precise dimension-specific guidance for corresponding dimensional visual features.

The primary contributions of this paper can be summarized as follows:

- We propose the CLIP-Guided Lexical Certainty Adapter (CLIP-LCA), which enhances text encoder understanding of spatial relationships by dynamically adjusting word certainty. This mitigates over-reliance on high-certainty keywords while improving attention to spatial descriptors.
- We introduce the Dimension-Decoupled Module (D2M) that resolves cross-dimensional interference by employing disentangled 2D and 3D textual features to guide corresponding visual features, enabling dimensionally-consistent cross-modal refinement.
- Experimental results demonstrate that our Mono3DVG-EnSD achieves state-of-the-art performance across all metrics, with a significant +13.54% improvement on the challenging Far(Acc@0.5) metric compared to previous methods.

Related Work

Monocular 3D Object Detection

Monocular 3D object detection focuses on predicting 3D bounding boxes from a single image and can be divided into two approaches depending on the use of supplementary data. One group of methods operates exclusively on monocular images, like M3D-RPN (Brazil and Liu 2019), which introduces a 3D region proposal network and utilizes depth-aware convolutions. Methods like SMOKE (Liu, Wu, and Tóth 2020) and FCOS3D (Wang et al. 2021) adopt a key-point estimation strategy inspired by CenterNet (Zhou,

Wang, and Krähenbühl 2019), where size and location of 3D boxes are derived from heatmap peaks. To address the challenge of feature interference, MonoLSS (Li, Jia, and Shi 2024b) proposes a Learnable Sample Selection (LSS) module that dynamically filters noise features, thereby improving 3D representation learning. MonoPair (Chen et al. 2020) improves localization accuracy by modeling spatial relationships between neighboring objects, which is beneficial for occluded instances. MonoEF (Zhou et al. 2021) introduces a novel approach to estimate camera extrinsics using vanishing points and horizon line detection, followed by a feature rectification module to correct distortions in latent representations. MonoCon (Liu, Xue, and Wu 2022) employs an auxiliary learning strategy during training, where monocular contextual features are aligned with 3D bounding box properties, and the auxiliary head is discarded during inference to maintain computational efficiency. MonoDDE (Li et al. 2022) utilizes depth cues to generate multiple depth hypotheses per object. Another line of research enhances 3D detection by integrating auxiliary data like depth maps, point clouds, or CAD models. D4LCN (Ding et al. 2020) proposes depth-aware convolutions that adapt receptive fields according to inferred depth. DID-M3D (Peng et al. 2022) improves depth estimation by separating instance-specific depths into attribute and visual components. ROI-10D (Manhardt, Kehl, and Gaidon 2019) predicts 3D boxes via dense depth estimation. For LIDAR-based methods, CaDDN (Reading et al. 2021) projects LIDAR-derived depth maps into a monocular network, then converts features into a bird's-eye-view space for detection. Beyond depth maps and LIDAR data, AutoShape (Liu et al. 2021b) utilizes keypoints extracted from CAD models to address limitations caused by sparse supervision.

2D Visual Grounding

Visual grounding, a task building upon object detection, seeks to create accurate correspondences between textual descriptions and specific areas in images. Initial methodologies in this domain primarily employed two-stage pipelines (Liu et al. 2019a; Yang, Li, and Yu 2019), dividing the process into distinct phases: region proposal generation and cross-modal alignment. In the first phase, an independent object detector (e.g., Faster R-CNN) produces candidate regions without considering textual context, which may lead to semantic inconsistencies due to the absence of language-aware guidance. To mitigate cross-modal semantic misalignment, MattNet (Yu et al. 2018) parses textual descriptions into structured elements to enable detailed vision-language interaction. In contrast to region-based methods, proposal-free techniques (Chen et al. 2018; Liao et al. 2020) employ spatial feature fusion, allowing direct region prediction through thorough multimodal understanding. FAOA (Yang et al. 2019) exemplifies this approach, combining fused visual-textual representations with YOLOv3 (Redmon and Farhadi 2018) for end-to-end visual grounding. Building on success of transformers across multiple domains, several transformer-based architectures have been adapted for visual grounding. The first attempt to leverage transformers in this task was proposed by TransVG (Deng et al. 2021). Referring Transformer (Li et al. 2022) employs phrase-based queries to jointly perform region localization and segmentation in a unified framework. Many transformer-based methods (Zheng et al. 2025; Liu et al. 2025a,b) achieved success.

3D Visual Grounding

The establishment of benchmark datasets for 3D visual grounding began with the introduction of Referit3D (Achlioptas et al. 2020) and ScanRefer (Chen, Chang, and Nießner 2020). Earlier works followed a two-stage pipeline similar to 2D approaches. PointNet++ (Qi et al. 2017) relies on a pre-trained detector for proposal generation and feature extraction. To enrich semantic understanding, SAT (Yang et al. 2021) incorporates 2D object semantics for improved model training. For handling complex descriptions and localizing objects in point clouds, Feng et al. (Feng et al. 2021) propose three specialized modules. Recent advances in 3D visual grounding leverage transformer-based architectures, including 3DVG-Trans (Zhao et al. 2021), LanguageRefer (Roh et al. 2022), and Multi-View Trans (Huang et al. 2022b). Unified frameworks [5][6] have been proposed to jointly address visual grounding and dense captioning within a single model. Liu et al. (Liu et al. 2021a) proposed a novel 3D visual grounding task using RGB-D images. Unlike prior indoor-focused studies that primarily predicted furniture objects, Lin et al. (Lin et al. 2024) extended the scenario to outdoor environments by incorporating both 2D images and 3D point clouds. The practical deployment of LiDAR and RGB-D based systems faces challenges due to hardware cost and availability constraints. As a solution, Mono3DVG (Zhan, Yuan, and Xiong 2024b) establishes a monocular alternative that combines geometrically detailed text with single RGB images to predict 3D bounding boxes.

Method

Building upon the baseline framework (Zhan, Yuan, and Xiong 2024b), we begin by reviewing its core architecture as the foundation for our methods. To extract generalized text features (T_t), the framework employs RoBERTa-base (Liu et al. 2019b) followed by a linear projection layer. Similarly, multi-scale visual features (V^*_{2D}) are extracted using ResNet-50 (He et al. 2016) with a linear layer. We follow the method presented in (Zhang et al. 2023) by implementing a lightweight depth predictor to derive 3D visual features V^*_{3D}. We then employ a dual-branch encoder architecture, comprising a visual encoder for 2D feature alignment and a depth encoder for 3D feature interaction. The visual encoder (Eq. 1) leverages a multi-scale deformable attention (MSDA) layer for efficient feature encoding. Subsequently, it incorporates 2D-specific textual information (T_{2D}) through a cross-attention module, followed by feature refinement via a feed-forward network (FFN) layer. The depth encoder (Eq. 2) employs a multi-head self-attention layer followed by a feed-forward network (FFN) to encode geometric (3D visual) embeddings, which then interact with 3D-specific textual features (T_{3D}) through a cross-attention layer.

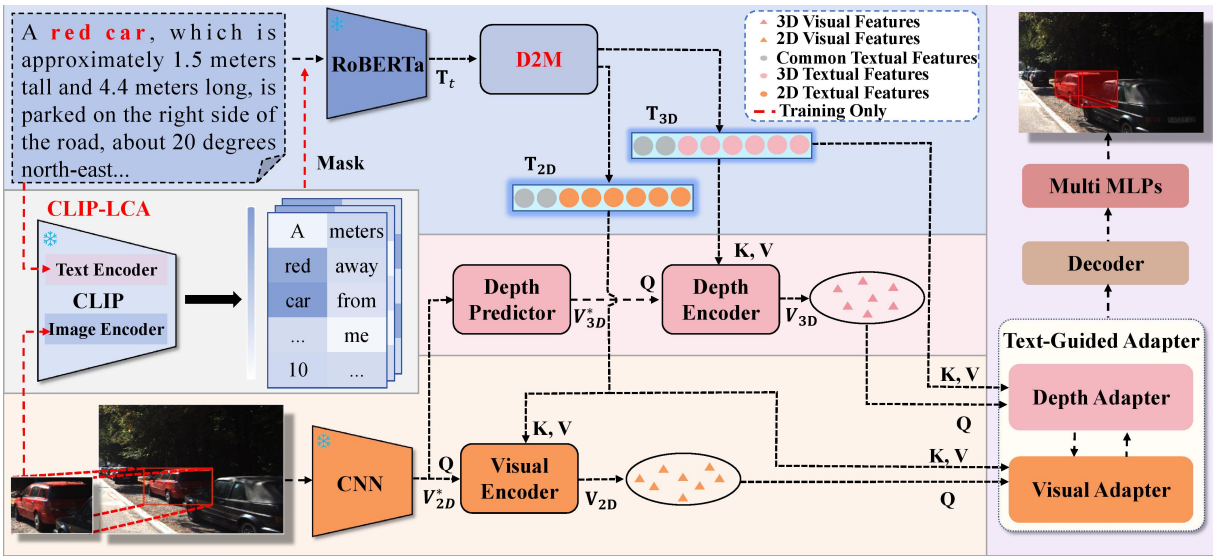

Figure 2: Framework Overview. The architecture integrates multiple feature extraction modules: RoBERTa for generalized textual features (T_t), visual encoder for 2D visual features (V_{2D}), and depth encoder for 3D visual features (V_{3D}). CLIP-LCA dynamically adjusts the certainty level of textual descriptions during training. D2M decomposes generalized textual features (T_t) into 2D-specific (T_{2D}) and 3D-specific textual features (T_{3D}). The adapter refines dimension-specific features of target objects, followed by decoder and multi-MLPs head for 2D-3D attribute prediction.

$$V_{2D} = FFN(MHCA(MSDA(V^*_{2D}), T_{2D})), \quad (1)$$

$$V_{3D} = MHCA(FFN(MHSA(V^*_{3D})), T_{3D}). \quad (2)$$

The adapter module is designed to refine both visual and geometric features. Specifically, the visual features interact with 2D-specific textual features, while the geometric features engage with their 3D-specific textual counterparts. The decoder employs a progressive fusion strategy, sequentially injecting geometric features, generalized textual representations, and 2D visual features into a learnable query via sequential cross-modal attention layers, as detailed in (Zhan, Yuan, and Xiong 2024b).

CLIP-Guided Lexical Certainty Adapter

To prevent the model from over-relying on explicit lexical cues, CLIP-LCA applies a dynamic strategy to mask high-certainty words during training, as illustrated in Fig.2. Specifically, CLIP-LCA begins by cropping the target region according to the ground-truth annotations and encoding it with the visual branch of CLIP (ViT-B/16) to obtain visual features. Simultaneously, each word in the caption is independently processed by CLIP's text encoder to generate textual features. Then, we compute the similarity between the textual feature of each word and the target visual feature. Based on these similarity scores, a k-means clustering algorithm (k = 2) is employed to partition the words into high-certainty (high similarity scores) and low-certainty (low similarity scores) categories. The cropped visual content primarily preserves the intrinsic 2D attributes of the object, such as its class name, color, shape. Consequently, words in the textual description corresponding to the object name or color exhibit high similarity with the visual features. In contrast, spatial descriptors that refer to the relative position of the object and its relationship to other objects show low similarity, since the cropping operation removes contextual information about the spatial relationships of the object within the full image. To enhance the text encoder's understanding of spatial information, we mask words categorized as high-certainty with "***". By masking explicit object cues, this strategy encourages the model to rely on implicit spatial semantics for grounding the corresponding object. Benefiting from this enhanced spatial understanding, the model effectively integrates explicit textual cues (high-certainty words) during inference, enabling more accurate and robust grounding performance.

Dimension-Decoupled Module

To generate dimension-specific text features for dimensionally-consistent refinements, we propose a Dimension-Decoupled module. The detailed architecture of this module is illustrated in Figure 3. The proposed D2M framework consists of two parallel branches: a 2D branch and a 3D branch. Each branch begins with a learnable query that interacts with the generalized textual features T_t through a cross-attention layer, followed by a feed-forward neural network (FFN) mapping layer, as formulated below.

$$H_{2D} = MHCA(L_{2D}, T_t), \quad (3)$$

$$H_{3D} = MHCA(L_{3D}, T_t), \quad (4)$$

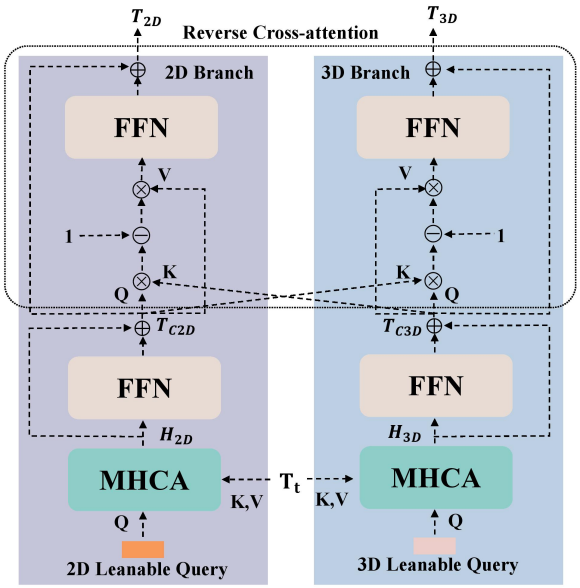

Figure 3: The detailed architecture of Dimension-Decoupled Module. The left part illustrates the procedure for decoupling 2D-specific textual features T_{2D}, while the right section presents the corresponding workflow for 3D-specific textual features T_{3D}.

$$T_{C2D} = FFN(H_{2D}) + H_{2D}, \tag{5}$$

$$T_{C3D} = FFN(H_{3D}) + H_{3D}. \tag{6}$$

The learnable queries for two dimensions, denoted as L_{2D} and L_{3D}. Through the Eqs. (3)-(6), we derive both 2D and 3D textual features at a coarse level, denoted as T_{C2D} and T_{C3D}. To refine the coarse textual features, we propose a reverse cross-attention module, detailed in Equations (7)-(8).

$$T_{2D} = T_{C2D} + FFN(Softmax(1 - T_{C2D}T_{C3D})T_{C2D}), \tag{7}$$

$$T_{3D} = T_{C3D} + FFN(Softmax(1 - T_{C3D}T_{C2D})T_{C3D}). \tag{8}$$

The core objective of the reverse cross-attention module is to compute low-attention regions between Query and Key features and subsequently enhance the information corresponding to these low-attention regions. This is achieved by first computing the attention map between the Queries and Keys, then inverting this map by subtracting values from 1, followed by softmax normalization. This inversion shifts attention toward originally low-attention parts, thereby enhancing the contribution of these features during subsequent processes with Values. The D2M framework provides two separate textual features, each encoded with either 2D-specific or 3D-specific details. These textual features guide the subsequent refinements of their corresponding visual features, ensuring dimensionally-consistent cross-modal refinement.

Loss Function

The prediction head utilizes multiple MLPs to estimate both two-dimensional and three-dimensional attributes. The 2D prediction branch outputs classification, 2D box size, and the projected 3D center. For 3D prediction, it outputs 3D box size, orientation, and depth values. The loss function for 2D predictions is formulated as:

$$L_{2D} = \lambda_1 L_{classs} + \lambda_2 L_{lrtb} + \lambda_3 L_{GIoU} + \lambda_4 L_{xy3D}, \tag{9}$$

where λ_{1-4} are set to (2,5,2,10) following MonoDETR (Zhang et al. 2023). The classification loss L_{class} employs the Focal loss function (Lin et al. 2017) for nine classes prediction. L_{lrtb} and L_{xy3D} utilize the L1 loss function. The 2D bounding box regression is optimized using the GIoU loss (Rezatofighi et al. 2019), denoted as L_{GIoU}. The loss for 3D part is expressed as:

$$L_{3D} = L_{size3D} + L_{orien} + L_{depth}, \tag{10}$$

where L_{size3D}, L_{orien} and L_{depth} represent the 3D IoU oriented loss (Ma et al. 2021), MultiBin loss and Laplacian aleatoric uncertainty loss (Chen, Chang, and Nießner 2020). To supervise the depth map prediction, we employ a Focal loss, represented as L_{dmap}. The total loss function is then defined as follows:

$$L_{overall} = L_{2D} + L_{3D} + L_{dmap}. \tag{11}$$

Experiments

Mono3DRefer Dataset

For our experiments, we adopt the Mono3DRefer dataset (Zhan, Yuan, and Xiong 2024b), which includes 2,025 images sampled from the KITTI dataset (Geiger, Lenz, and Urtasun 2012). These images are annotated with 41,140 captions with a vocabulary of 5,271 words. In terms of description quantity, Mono3DRefer provides a comparable scale of descriptions as ScanRefer and Nr3D, excluding the template-generated Sr3D dataset. More importantly, Mono3DRefer supports object annotations at significantly longer distances, with a maximum range of 102 meters. In contrast, RGB-D sensors typically operate within 10 meters, while LiDAR systems are generally constrained to around 30 meters.

Implementation Details

We conduct all experiments on an NVIDIA GeForce RTX 3090 GPU, training the model for 60 epochs with the AdamW optimizer under the following configuration: a batch size of 10, initial learning rate 10^{-4}, weight decay of 10^{-4}, and a dropout rate of 0.1. For evaluation, we adopt the 3D IoU thresholds of 0.25 and 0.5, following previous research (Lin et al. 2024), to assess accuracy across 9 scenarios.

To ensure fair comparisons, we employ the same baseline model as Mono3DVG (Zhan, Yuan, and Xiong 2024b). **Two-stage Methods Evaluation**: (1) CatRand randomly selects category-matched ground truth boxes as predictions.

Method	Type	Unique		Multiple		Overall	
		Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
CatRand	Two-Stage	100	100	24.47	24.43	38.69	38.67
Cube R-CNN + Rand	Two-Stage	32.76	14.61	13.36	7.21	17.02	8.60
Cube R-CNN + Best	Two-Stage	35.29	16.67	60.52	32.99	55.77	29.92
ZSGNet + backproj	One-stage	9.02	0.29	16.56	2.23	15.14	1.87
FAOA + backproj	One-stage	11.96	2.06	13.79	2.12	13.44	2.11
ReSC + backproj	One-stage	11.96	0.49	23.69	3.94	21.48	3.29
TransVG + backproj	Tran.-based	15.78	4.02	21.84	4.16	20.70	4.14
Mono3DVG-TR	Tran.-based	57.65	33.04	65.92	46.85	64.36	44.25
Mono3DVG-EnSD(Ours)	Tran.-based	**66.67**	**42.65**	**70.17**	**55.22**	**69.51**	**52.85**

Table 1: Comparison of model performance across Unique, Multiple, and Overall scenarios. Underlined results indicate better performance than our bolded accuracy.

Method	Type	Near/Easy		Medium/Moderate		Far/Hard	
		Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
CatRand	Two-Stage	31.16/47.29	31.05/47.26	35.49/33.92	35.49/33.92	52.11/30.83	52.11/30.74
Cube R-CNN + Rand	Two-Stage	17.40/21.12	11.45/11.41	18.01/17.85	8.15/8.01	14.91/10.56	6.38/5.18
Cube R-CNN + Best	Two-Stage	67.76/59.66	41.45/33.05	60.69/60.56	30.35/33.45	34.72/46.25	17.01/22.52
ZSGNet + backproj	One-stage	24.87/21.33	0.59/3.35	16.74/13.87	3.71/0.63	2.15/7.57	0.07/0.84
FAOA + backproj	One-stage	18.03/17.51	0.53/3.43	15.64/12.18	3.95/1.34	4.86/8.83	0.62/0.90
ReSC + backproj	One-stage	33.68/27.90	0.59/5.71	24.03/19.23	6.15/1.97	4.24/14.41	1.25/1.02
TransVG + backproj	Tran.-based	29.34/28.88	0.86/6.95	25.05/16.41	8.02/2.75	4.17/12.91	0.97/1.38
Mono3DVG-TR	Tran.-based	64.74/72.36	53.49/51.80	75.44/69.23	55.48/48.66	45.07/49.01	15.35/29.91
Mono3DVG-EnSD(Ours)	Tran.-based	**67.80/80.47**	**59.21/62.10**	**79.71/71.06**	**62.97/55.77**	**54.31/52.85**	**28.89/37.42**

Table 2: Performance comparisons for near-medium-far scenarios and easy-moderate-hard scenarios. Underlined results indicate better performance than our bolded accuracy.

(2) (Cube RCNN (Brazil et al. 2023) + Rand) randomly samples prediction from proposal generated by Cube RCNN. (3) (Cube RCNN (Brazil et al. 2023) + Best) selects the optimal box match for each proposal to quantify the upper-bound performance of two-stage methods. **One-stage methods**: We adapt four state-of-the-art 2D visual grounding methods (ZSGNet (Sadhu, Chen, and Nevatia 2019), FAOA (Yang et al. 2019), ReSC (Yang, Li, and Yu 2020), TransVG (Deng et al. 2021)) to 3D area via back-projection for comparative evaluation. We analyze methods performance across three dimensions: (1) We evaluate the approach under '**unique**' and '**multiple**' scenarios. The unique scenario involves single-object cases, while the multiple scenario contains multiple objects sharing the same category label; (2) To estimate performance across different distance ranges, we categorize evaluation metrics into three depth ranges: **Near** (0-15m), **Medium** (15-35m) and **Far** (>35m). The "near" scenario presents fewer challenges due to clearer observations, while the "far" often involves higher ambiguity. The medium range serves as an intermediate zone that represents the practical scenario in real-world applications; (3) To evaluate the impact of visual occlusion and truncation on performance, we categorize the cases into three levels based on occlusion degree and truncation ratio: **Easy** cases with no occlusion and truncation ratio below 0.15, **Moderate** cases including no/partial occluded objects and truncation ratio between 0.15-0.3, and **Hard** containing severely occluded objects or truncation ratio exceeding 0.3.

Quantitative Comparisons and Analyses

As shown in Table 1, CatRand achieves 100% accuracy in the "unique" scenario but drops to 24% in the "multiple" scenario. Similarly, Cube R-CNN Rand performs better on "unique" than on "multiple" scenarios. When an image contains only a single object, providing the category label is sufficient. However, for images with multiple objects, additional information beyond the label is required to eliminate localization ambiguities, such as spatial or distinct descriptions. As presented in Table 2, the CatRand method demonstrates better performance in the 'far' scenario compared to other distance range scenarios. However, both our method and baseline approaches exhibit progressively lower accuracy as depth increases. The "far" scenario contains fewer ambiguous instances, enabling CatRand's random ground truth selection to achieve better performance. In contrast, other methods highly depend on predicted bounding box accuracy, which often leads to imprecise depth and 3D extent, particularly for distinct objects. Our proposed model achieves superior performance in nearly all distance-based scenarios, except for a slight performance gap under the CatRand sampling condition on the "far" scenario. Experimental results on the easy-moderate-hard metrics show that our framework achieves state-of-the-art performance. In summary, our method achieves consistent ac-

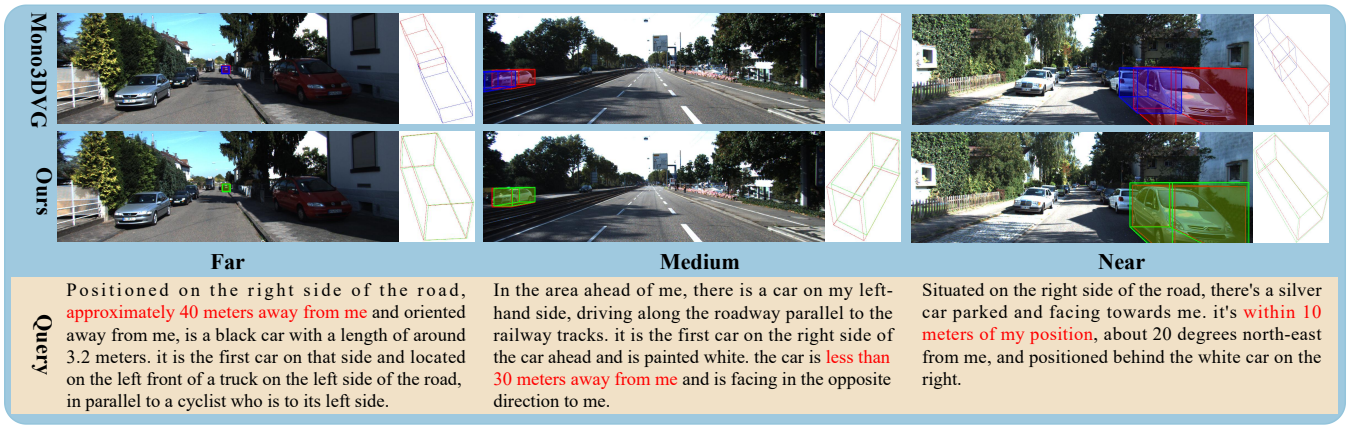

Figure 4: Visualization of 3D bounding box predictions from Mono3DVG-TR and our Mono3DVG-EnSD. Ground truth (red), our predictions (green), and Mono3DVG-TR predictions (blue) are shown for comparison.

curacy improvements across all evaluation scenarios. Most notably, under the strict acc@0.5 criterion, we observe substantial gains of +13.54% for "far" scenarios and +10.30% for "easy" scenarios. Meanwhile, our method achieves performance improvements of 9.61%, 8.37%, and 8.60% in "unique", "multiple", and "overall" scenarios (acc@0.5).

Qualitative Analysis

Figure 4 provides a qualitative comparison of 3D visual grounding results between the baseline Mono3DVG-TR and our proposed Mono3DVG-EnSD, demonstrating the effectiveness of our approach. For clear visual analysis, we select no occlusion and fully visible vehicle samples on three distance intervals: near (<15m), medium (15-35m), and far (>35m) scenarios. The qualitative comparison reveals that while both methods perform competitively in estimating orientation and predicting 3D object dimensions, our method shows superior performance in depth positioning accuracy. The experimental results validate that our Dimension-Decoupled module effectively maintains dimensional consistency during text-visual feature interaction, thereby enabling more accurate refinement of 3D visual features.

Ablation Studies

Ablation experiments were conducted on the Mono3DRefer dataset to assess the proposed CLIP-Guided Lexical Certainty Adapter (CLIP-LCA) and Dimension-Decoupled Module (D2M) component. Their performance is quantified using the standard Acc@0.25 and Acc@0.5 metrics on the "Overall" scenario, as summarized in Table 3. Row 1 corresponds to the baseline model's performance. Row 2 reflects results employing exclusively the CLIP-LCA method, while Row 3 illustrates outcomes integrating only the D2M. Comparison indicates that both components independently yield significant accuracy gains (CLIP-LCA: +2.21%/+5.04%; D2M: +3.75%/+6.83%), confirming their individual contributions to monocular 3D visual grounding. The final row presents the combined CLIP-LCA+D2M implementation, where synergistic improvement exceeds the individual improvements (+5.15%/+8.60%). Due to space limitations, we primarily present the ablation studies of our two key modules in the "Overall" scenario. It is worth noting that our modules consistently demonstrate accuracy improvements across all other scenarios as well.

CLIP-LCA	D2M	Acc@0.25	Acc@0.5
-	-	64.36	44.25
✓	-	66.57	49.29
-	✓	68.11	51.08
✓	✓	69.51	52.85

Table 3: The ablation studies of our proposed methods on the "Overall" scenario.

Conclusion

Existing mono3DVG methods face two major limitations. Firstly, the text encoder heavily rely on high-certainty words for localization, limiting their ability to comprehend spatial descriptions. Secondly, cross-dimensional interference occurs during the interaction between textual and visual features. To address these challenges, we propose Mono3DVG-EnSD, a novel framework designed to enhance dimensional alignment and understanding ability of spatial descriptions. The CLIP-Guided Lexical Certainty Adapter (CLIP-LCA) dynamically suppress high-certainty keywords, enabling the model to effectively comprehend implicit spatial cues in captions for more robust localization. The proposed Dimension-Decoupled Module (D2M) first disentangles generalized textual features into distinct 2D-specific and 3D-specific textual features. By then guiding 2D or 3D visual features with their corresponding dimensional textual features, the D2M ensures dimensional consistency for the cross-modal interactions. Our experimental results demonstrate that Mono3DVG-EnSD achieves performance improvements across all nine scenarios compared to the baseline. Through ablation experiments, we further confirm that both the CLIP-LCA and the D2M independently contribute to these performance gains.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62425305, U22B2050 and 62221002, in part by the Science and Technology Innovation Program of Hunan Province under Grant 2023RC1048.

References

Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 422–440. Springer.

Brazil, G.; Kumar, A.; Straub, J.; Ravi, N.; Johnson, J.; and Gkioxari, G. 2023. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13154–13164.

Brazil, G.; and Liu, X. 2019. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9287–9296.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, 202–221. Springer.

Chen, X.; Ma, L.; Chen, J.; Jie, Z.; Liu, W.; and Luo, J. 2018. Real-time referring expression comprehension by single-stage grounding network. *arXiv preprint arXiv:1812.03426*.

Chen, Y.; Tai, L.; Sun, K.; and Li, M. 2020. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12093–12102.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. TransVG: End-to-End Visual Grounding with Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1749–1759.

Ding, M.; Huo, Y.; Yi, H.; Wang, Z.; Shi, J.; Lu, Z.; and Luo, P. 2020. Learning depth-guided convolutions for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops*, 1000–1001.

Feng, M.; Li, Z.; Li, Q.; Zhang, L.; Zhang, X.; Zhu, G.; Zhang, H.; Wang, Y.; and Mian, A. 2021. 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*, 3722–3731.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, 3354–3361. IEEE.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, J.; Qin, Y.; Qi, J.; Sun, Q.; and Zhang, H. 2022a. Deconfounded visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 998–1006.

Huang, S.; Chen, Y.; Jia, J.; and Wang, L. 2022b. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15524–15533.

Li, Z.; Jia, J.; and Shi, Y. 2024a. MonoLSS: Learnable sample selection for monocular 3D detection. In *2024 International Conference on 3D Vision (3DV)*, 1125–1135. IEEE.

Li, Z.; Jia, J.; and Shi, Y. 2024b. MonoLSS: Learnable sample selection for monocular 3D detection. In *2024 International Conference on 3D Vision (3DV)*, 1125–1135. IEEE.

Li, Z.; Qu, Z.; Zhou, Y.; Liu, J.; Wang, H.; and Jiang, L. 2022. Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2791–2800.

Liao, Y.; Liu, S.; Li, G.; Wang, F.; Chen, Y.; Qian, C.; and Li, B. 2020. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10880–10889.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, Z.; Peng, X.; Cong, P.; Zheng, G.; Sun, Y.; Hou, Y.; Zhu, X.; Yang, S.; and Ma, Y. 2024. Wildrefer: 3d object localization in large-scale dynamic scenes with multi-modal visual data and natural language. In *European Conference on Computer Vision*, 456–473. Springer.

Liu, D.; Zhang, H.; Wu, F.; and Zha, Z.-J. 2019a. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4673–4682.

Liu, H.; Lin, A.; Han, X.; Yang, L.; Yu, Y.; and Cui, S. 2021a. Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6032–6041.

Liu, J.; Liu, Y.; Shang, F.; Liu, H.; Liu, J.; and Feng, W. 2025a. Improving Generalization in Federated Learning with Highly Heterogeneous Data via Momentum-Based Stochastic Controlled Weight Averaging. In *Forty-second International Conference on Machine Learning*.

Liu, J.; Shang, F.; Zhu, K.; Liu, H.; Liu, Y.; and Liu, J. 2025b. FedAdamW: A Communication-Efficient Optimizer with Convergence and Generalization Guarantees for Federated Large Models. *arXiv preprint arXiv:2510.27486*.

Liu, X.; Xue, N.; and Wu, T. 2022. Learning auxiliary monocular contexts helps monocular 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1810–1818.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b.

Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Wu, Z.; and Tóth, R. 2020. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 996–997.

Liu, Z.; Zhou, D.; Lu, F.; Fang, J.; and Zhang, L. 2021b. Autoshape: Real-time shape-aware monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15641–15650.

Ma, X.; Zhang, Y.; Xu, D.; Zhou, D.; Yi, S.; Li, H.; and Ouyang, W. 2021. Delving into localization errors for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4721–4730.

Manhardt, F.; Kehl, W.; and Gaidon, A. 2019. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2069–2078.

Peng, L.; Wu, X.; Yang, Z.; Liu, H.; and Cai, D. 2022. Didm3d: Decoupling instance depth for monocular 3d object detection. In *European Conference on Computer Vision*, 71–88. Springer.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Reading, C.; Harakeh, A.; Chae, J.; and Waslander, S. L. 2021. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8555–8564.

Redmon, J.; and Farhadi, A. 2018. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Roh, J.; Desingh, K.; Farhadi, A.; and Fox, D. 2022. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, 1046–1056. PMLR.

Sadhu, A.; Chen, K.; and Nevatia, R. 2019. Zero-shot grounding of objects from natural language queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4694–4703.

Tan, Z.; Yang, W.; and Wang, Z. 2024. Reimagining 3D Visual Grounding: Instance Segmentation and Transformers for Fragmented Point Cloud Scenarios. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, MMAsia '23. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702051.

Wang, T.; Zhu, X.; Pang, J.; and Lin, D. 2021. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 913–922.

Yang, S.; Li, G.; and Yu, Y. 2019. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4644–4653.

Yang, S.; Li, G.; and Yu, Y. 2020. Relationship-embedded representation learning for grounding referring expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8): 2765–2779.

Yang, Z.; Gong, B.; Wang, L.; Huang, W.; Yu, D.; and Luo, J. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4683–4693.

Yang, Z.; Zhang, S.; Wang, L.; and Luo, J. 2021. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1856–1866.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1307–1315.

Zhan, Y.; Yuan, Y.; and Xiong, Z. 2024a. Mono3dvg: 3d visual grounding in monocular images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6988–6996.

Zhan, Y.; Yuan, Y.; and Xiong, Z. 2024b. Mono3dvg: 3d visual grounding in monocular images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6988–6996.

Zhang, R.; Qiu, H.; Wang, T.; Guo, Z.; Cui, Z.; Qiao, Y.; Li, H.; and Gao, P. 2023. Monodetr: Depth-guided transformer for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9155–9166.

Zhao, L.; Cai, D.; Sheng, L.; and Xu, D. 2021. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2928–2937.

Zheng, S.; Zhao, P.; Zheng, Z.; He, P.; Cheng, H.; Cai, Y.; and Huang, Q. 2025. Look Around Before Locating: Considering Content and Structure Information for Visual Grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1656–1664.

Zhou, X.; Wang, D.; and Krähenbühl, P. 2019. Objects as points. *arXiv preprint arXiv:1904.07850*.

Zhou, Y.; He, Y.; Zhu, H.; Wang, C.; Li, H.; and Jiang, Q. 2021. Monocular 3d object detection: An extrinsic parameter free approach. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7556–7566.