

EG-3DVG: Expression and Geometry Aware Grounding Decoder for 3D Visual Grounding

GwangWook Park¹ Hyo-Jun Lee² Jong-Hyeon Baek¹ Hanul Kim³ Yeong Jun Koh^{1*}

¹Chungnam National University ²Kangwon National University

³Seoul National University of Science and Technology

ccocco0609@o.cnu.ac.kr, gywns6298@kangwon.ac.kr, jonghyeon9587@o.cnu.ac.kr
 hukim@seoultech.ac.kr, yjkoh@cnu.ac.kr

Abstract

Despite recent progress in 3D visual grounding, existing methods still struggle with three core challenges: 1) cross-modal misalignment that prevents textual cues from being reliably delivered to visual representations, 2) intra-class confusion arising from insufficient understanding of fine-grained expression cues, and 3) geometric reasoning errors caused by inaccurate aggregation of spatially relevant visual features. We propose EG-3DVG, a unified framework that addresses these issues through an expression and geometry aware grounding decoder. The decoder integrates two complementary attention modules—position-guided expression cross-attention (PECA) for reliable text–vision alignment and geometry-aware masked attention (GMA) for selective aggregation of geometry-consistent visual cues. To further distinguish semantically similar instances, we introduce expression-aware contrastive learning (ECL), which strengthens the alignment between the target object token and expression-relevant words. Extensive experiments on ScanRefer and SR3D/NR3D demonstrate that EG-3DVG achieves state-of-the-art performance in both 3D bounding box localization and mask prediction, validating the effectiveness of our geometry- and expression-aware design. Code is available at <https://github.com/Gwan9Wook/EG3DVG>.

1. Introduction

Understanding 3D environments through natural language is crucial for advancing human-computer interaction, with broad applications in virtual and augmented reality, autonomous robotics, and self-driving technologies. Recently, 3D Visual Grounding (3DVG) [11, 15, 16, 27, 30, 31, 33, 35, 36, 38] has gained significant attention as a promising solution to facilitate natural language-based understanding of 3D environments. The objective of 3DVG is to localize a

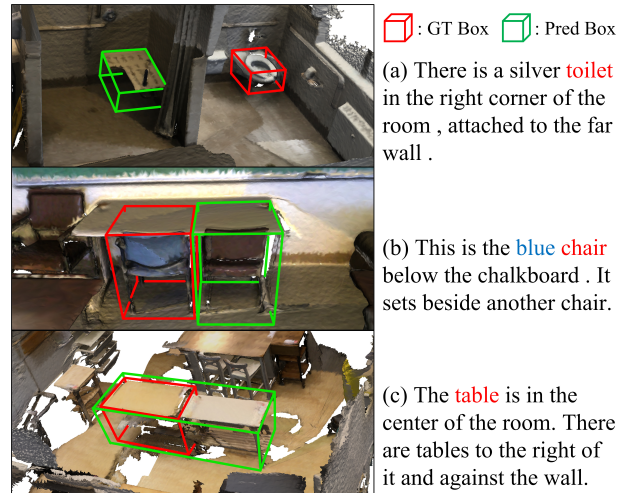


Figure 1. Illustration of failure examples on ScanRefer [6]: (a) cross-modal misalignment, (b) intra-category confusion, and (c) geometric reasoning error.

target object within a 3D point cloud scene based on a textual description that depicts the object’s attributes, such as its appearance (e.g. “blue”, “red”), category (e.g. “desk”, “chair”), and spatial relationships with other objects (e.g. “next to the table”, “on the left of the sofa”).

General 3DVG methods represent the target object mentioned in text descriptions using 3D bounding boxes [11, 16, 27, 30, 33, 38] or semantic masks [26, 27, 35, 36]. Specifically, the model first encodes the input 3D point cloud and the text description into visual and text tokens, respectively. These tokens are then fused via transformer-based architectures [32] to capture cross-modal interactions. A decoder module transforms the visual tokens into candidate 3D bounding boxes or segmentation masks. In parallel, the text tokens can be syntactically parsed into main, attribute, pronoun, relationship, and auxiliary tokens using off-the-shelf tools [29, 37]. Finally, the target object is identified

*Corresponding author

by selecting the candidate associated with the visual token that achieves the highest text-alignment score [27, 33, 38], which is computed by considering its contextual similarity with multiple types of text tokens (e.g., main, attribute, pronoun, and relationship tokens).

Despite recent advances in 3D visual grounding, existing models still suffer from three representative types of failure, as illustrated in Figure 1. (1) Cross-modal misalignment: The model often fails to accurately localize the referred object because textual semantics are not reliably conveyed to visual features through cross-attention, leading to incomplete text–vision alignment. (2) Intra-class confusion: The model often mislocalizes among semantically similar instances due to its limited ability to capture fine-grained expression cues in the input text. (3) Geometric reasoning error: The model misinterprets the 3D spatial configuration of surrounding points, including those from other objects or background regions, resulting in inaccurate localization of the referred target.

Motivated by these observations, we propose EG-3DVG, a novel 3D visual grounding framework that integrates the proposed modules within an expression and geometry aware grounding decoder to explicitly resolve these failure cases. To alleviate cross-modal misalignment, we introduce the position-guided expression cross-attention (PECA) module, which enriches text tokens with 3D positional cues and enables more reliable alignment between textual semantics and visual representations. To address intra-class confusion, we propose expression-aware contrastive learning (ECL), where the model learns to associate fine-grained expression cues with the correct visual instance by contrasting the target object against semantically similar negatives within the same scene. To reduce geometric reasoning errors, we develop the geometry-aware masked attention (GMA) module, which selectively aggregates spatially consistent visual information by exploiting 3D geometric relations between candidates and surrounding points. Extensive experiments on ScanRefer and SR3D/NR3D benchmarks demonstrate that EG-3DVG achieves state-of-the-art results in both 3D bounding box localization and mask prediction, validating the effectiveness of each proposed component. To summarize, the main contributions of this work are as follows:

- PECA injects 3D positional cues into text tokens to achieve spatially consistent text–visual alignment.
- ECL enhances the model’s ability to distinguish same-category objects with different expressions.
- GMA selectively aggregates visual information based on geometric relations between candidate tokens and surrounding points.
- EG-3DVG achieves superior performance on benchmarks in both 3D bounding box and mask prediction.

2. Related Works

Vision and Language: With growing interest in multi-modal learning [17, 28], various methods combining vision and language have recently been introduced. Early methods have usually focused on 2D vision-language tasks, such as image captioning [13, 19], visual question answering [3, 40], and visual grounding [18, 34, 41]. More recently, vision-and-language methods are increasingly being explored to understand 3D environments. 3D open-vocabulary object detection or segmentation [22, 24] enriches the diversity of text queries by aligning visual and textual feature spaces, rather than classifying into fixed categories. 3D dense captioning [7, 43] focuses on generating natural language descriptions for each object in a 3D scene, while 3D scene generation [12] aims to synthesize realistic 3D scenes from text prompts. These methods enable fine-grained semantic understanding of spatial environments, bridging the gap between textual and visual context. **3D Visual Grounding:** 3D Visual Grounding (3DVG) is the task of localizing a target object described by a text description within a 3D scene. 3DVG is commonly divided into two sub-tasks: 3D Referring Expression Comprehension (3DREC), which aims to predict a 3D bounding box, and 3D Referring Expression Segmentation (3DRES), which generates a semantic mask of the target object.

In 3DRES field, X-RefSeg3D [26] constructs a cross-modal relational graph that connects object nodes and textual cues, and uses graph reasoning to resolve relation expressions. 3D-STMN [36] introduces a superpoint-based end-to-end framework that includes the matching between superpoint and text to enable dense supervision and efficient inference for 3D-RES. In 3DREC field, multi-view based approaches [11, 15] project 3D scenes into multiple rendered views and aggregate view-specific features to capture spatial context. To reduce computational cost, VPP-Net [30] predicts informative viewpoints from a single input and performs grounding based on the predicted view. TSP3D [10] presents a lightweight framework that real-time grounding via text-guided sparse voxel pruning. MCLN [27] introduces a multi-branch architecture that jointly performs 3D object localization and semantic mask prediction. While existing methods have made progress, they still suffer from three major challenges, as illustrated in Figure 1.

3. Proposed Method

3.1. Model Overview

Given an input text description consisting of N_{text} words and a 3D scene represented by N_{point} points, the proposed EG-3DVG predicts a mask $\mathbf{m} \in \mathbb{R}^{N_{\text{point}}}$ and a 3D bounding box $\mathbf{b} \in \mathbb{R}^6$ corresponding to the target object described in the text description. Figure 2 illustrates the overall architecture of the proposed EG-3DVG.

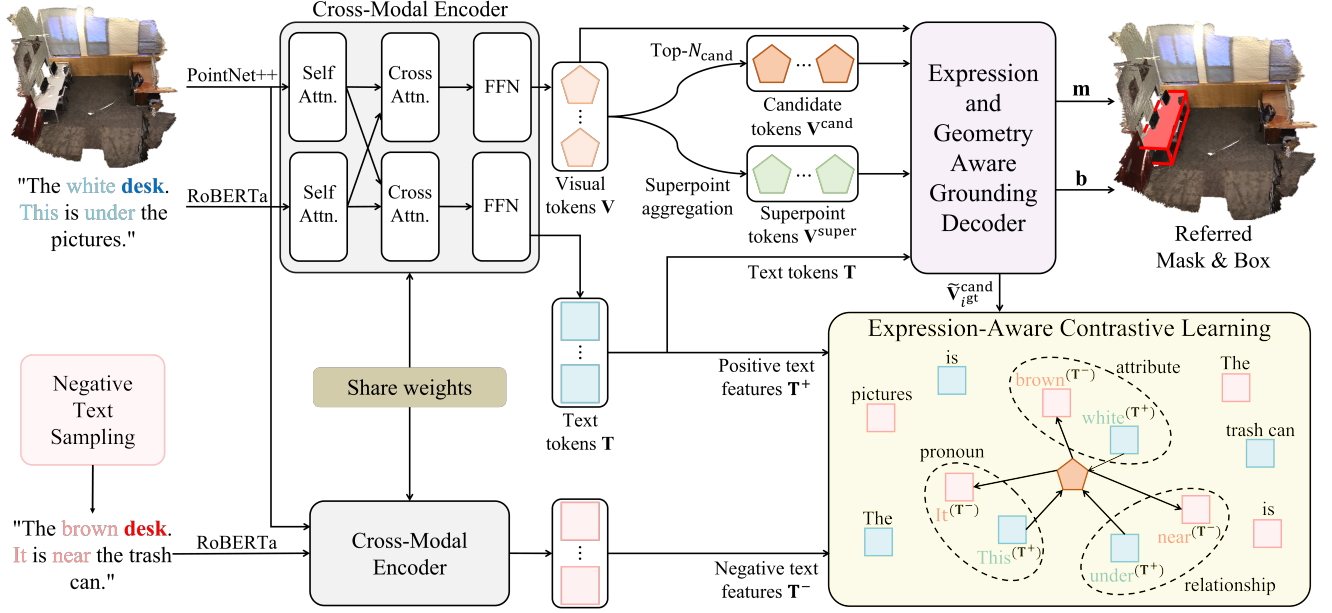


Figure 2. Overview of EG-3DVG. In expression and geometry aware grounding decoder, position-guided expression cross-attention injects positional information to each text tokens and geometry-aware masked attention incorporates geometric relationship into the attention mechanism. Expression-aware contrastive learning encourages the selected visual token to align with expression-relevant words (e.g. attribute, pronoun, relationship) from the text description while contrasting against other texts describing same-category distractors.

3.2. Cross-Modal Encoder

Text and Visual Token Generation: Following [16, 27, 30, 33, 38], we use pre-trained RoBERTa [20] and PointNet++ [25] to extract text tokens and visual tokens from the input text description and 3D point cloud, respectively. The text and visual tokens are then projected into $\mathbf{T} \in \mathbb{R}^{N_{\text{text}} \times C}$ and $\mathbf{V} \in \mathbb{R}^{N_{\text{vis}} \times C}$, through a cross-modal encoder, where C is the feature dimension. The cross-modal encoder consists of two transformer branches that separately process text and visual tokens, as shown in Figure 2. To facilitate cross-modal interaction, a cross-attention module in each branch uses opposite-modality tokens as key-value, while retaining processed tokens as queries. For the visual branch, 3D coordinates of each point are incorporated as positional embeddings in attention computation.

Candidate Visual Token Generation: Each visual token predicts a candidate score of the target object through a multilayer perceptron supervised with the KPS loss [16, 21, 27, 30, 33, 38]. Based on the candidate scores of the visual tokens, we select the top- N_{cand} tokens as candidate tokens $\mathbf{V}^{\text{cand}} \in \mathbb{R}^{N_{\text{cand}} \times C}$, which are subsequently used for 3D bounding box and mask prediction in the expression and geometry aware grounding decoder (Section 3.3).

Superpoint Token Generation: To obtain a more informative visual representation, we project the N_{vis} visual tokens onto a superpoint resolution of size N_{super} using a superpoint aggregation process [27], where $N_{\text{super}} \ll N_{\text{point}}$. Let $\{\mathbf{p}_s^{\text{super}} \in \mathbb{R}^3\}_{s=1}^{N_{\text{super}}}$ denote the set of 3D coordinates

of the superpoint centers. For each superpoint s , the K nearest visual tokens are selected based on the spatial distance between $\mathbf{p}_s^{\text{super}}$ and each visual token coordinate to compose $\mathbf{V}_s^{\text{knn}} \in \mathbb{R}^{K \times C}$ and $\mathbf{P}_s^{\text{knn}} \in \mathbb{R}^{K \times 3}$, where each row of $\mathbf{V}_s^{\text{knn}}$ is the visual token feature and that of $\mathbf{P}_s^{\text{knn}}$ contains the corresponding 3D position. Then, we encode the relative positional relationship between the s -th superpoint and its K nearest tokens by:

$$\tilde{\mathbf{V}}_s^{\text{knn}} = \mathbf{V}_s^{\text{knn}} + \text{FFN}(\mathbf{1}_K(\mathbf{p}_s^{\text{super}})^\top - \mathbf{P}_s^{\text{knn}}) \quad (1)$$

where $\mathbf{1}_K \in \mathbb{R}^K$ is a vector of ones and $\text{FFN}(\cdot)$ denotes feed forward networks consisting of two fully connected layers with ReLU activation. We then apply channel-wise max pooling over $\tilde{\mathbf{V}}_s^{\text{knn}}$ to obtain the aggregated superpoint feature $\mathbf{v}_s \in \mathbb{R}^C$, and stack all $\mathbf{v}_s$ to form the final superpoint tokens $\mathbf{V}^{\text{super}} = [\mathbf{v}_1 \dots \mathbf{v}_{N_{\text{super}}}] \in \mathbb{R}^{N_{\text{super}} \times C}$.

3.3. Expression and Geometry Aware Grounding Decoder

The expression and geometry aware grounding decoder (EGD) refines candidate representations by jointly exploiting textual and visual cues to accurately localize and segment the referred object. It consists of two complementary attention modules: the position-guided expression cross-attention (PECA), which aligns visual tokens with linguistic expressions in a spatially consistent manner, and the geometry-aware masked attention (GMA), which selectively aggregates geometry-consistent visual cues to en-

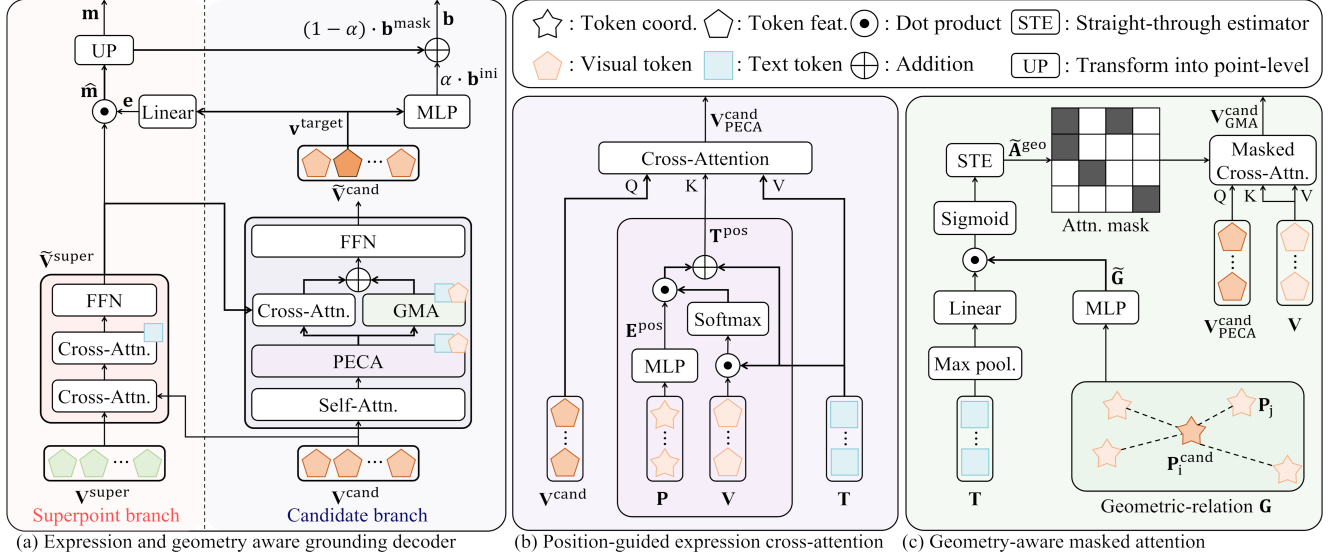


Figure 3. A diagram of (a) the expression and geometry aware grounding decoder, (b) position-guided expression cross-attention (PECA), and (c) geometry-aware masked attention (GMA). The top-right symbols on the cross-attention, PECA, and GMA blocks indicate the source tokens used to refine their input tokens.

hance candidate tokens with object-level representations.

Token Refinement: As shown in Figure 3(a), EGD consists of two branches: a superpoint token branch and a candidate token branch. The superpoint token branch refines $\mathbf{V}^{\text{super}}$ into $\tilde{\mathbf{V}}^{\text{super}}$ through cross-attention with the candidate tokens $\mathbf{V}^{\text{cand}}$, cross-attention with the text tokens $\mathbf{T}$, and subsequent feed-forward networks. Since the superpoint tokens $\mathbf{V}^{\text{super}}$ are derived solely from visual features, they lack both the contextual semantics of text tokens and the object-specific cues of candidate tokens. Therefore, this branch enriches semantics of $\mathbf{V}^{\text{super}}$ by integrating textual semantics from $\mathbf{T}$ and object-aware cues from $\mathbf{V}^{\text{cand}}$.

The candidate token branch refines $\mathbf{V}^{\text{cand}}$ into $\tilde{\mathbf{V}}^{\text{cand}}$ by incorporating contextual cues from text tokens, visual tokens, and superpoint features through the proposed position-guided expression cross-attention (PECA) and geometry-aware masked attention (GMA). Given the candidate tokens $\mathbf{V}^{\text{cand}}$, this branch first produces $\mathbf{V}^{\text{cand}}_{\text{PECA}}$ through self-attention and PECA, and then further refines $\mathbf{V}^{\text{cand}}_{\text{PECA}}$ into $\mathbf{V}^{\text{cand}}_{\text{super}}$ and $\mathbf{V}^{\text{cand}}_{\text{GMA}}$ through two parallel modules, cross-attention with the refined superpoint tokens $\tilde{\mathbf{V}}^{\text{super}}$ and GMA, respectively. The final refined candidate tokens $\tilde{\mathbf{V}}^{\text{cand}}$ are obtained by

$$\tilde{\mathbf{V}}^{\text{cand}} = \text{FFN}(\mathbf{V}^{\text{cand}}_{\text{super}} + \mathbf{V}^{\text{cand}}_{\text{GMA}}). \quad (2)$$

Mask and 3D Bounding Box Prediction: To estimate the target object’s mask $\mathbf{m}$ and 3D bounding box $\mathbf{b}$, we select the target object token $\mathbf{v}^{\text{target}} \in \mathbb{R}^C$ as the candidate with the highest text-alignment score [27, 33, 38] among all refined candidate tokens $\tilde{\mathbf{V}}^{\text{cand}} \in \mathbb{R}^{N_{\text{cand}} \times C}$. For mask prediction, we project the target object token $\mathbf{v}^{\text{target}}$ into

a mask embedding $\mathbf{e} \in \mathbb{R}^C$, and then generate the target mask over superpoints as $\hat{\mathbf{m}} = \tilde{\mathbf{V}}^{\text{super}} \mathbf{e}$. Finally, the superpoint-level mask $\hat{\mathbf{m}}$ is transformed into the point-level mask $\mathbf{m} \in \mathbb{R}^{N_{\text{point}}}$ by assigning each superpoint’s mask probability to all points it contains.

For 3D bounding box prediction, we first project the selected target token $\mathbf{v}^{\text{target}}$ to an initial bounding box $\mathbf{b}^{\text{ini}} \in \mathbb{R}^6$ via a multilayer perceptron (MLP). To incorporate spatial consistency with the predicted mask, we also construct a mask-derived bounding box $\mathbf{b}^{\text{mask}}$ that tightly encloses the predicted point-level mask $\mathbf{m}$. The final bounding box $\mathbf{b}$ is then obtained by a weighted sum of the two boxes:

$$\mathbf{b} = \alpha \mathbf{b}^{\text{ini}} + (1 - \alpha) \mathbf{b}^{\text{mask}} \quad (3)$$

where we set the weighting hyperparameter α to 0.5.

Position-Guided Expression Cross-Attention (PECA): To inject the contextual information of text tokens into visual tokens, the cross-attention operation where the visual tokens serve as queries and the text tokens act as key-value pairs is typically adopted. In this cross-attention, accurate attention weights between visual and text tokens is crucial for effective feature fusion. However, conventional cross-attention between visual and text tokens often produces sub-optimal alignment, as the two modalities contain inherently unbalanced spatial cues. Specifically, each visual token in $\mathbf{V}^{\text{cand}}$ encodes 3D positional information derived from the input point cloud (through PointNet++ feature extraction), whereas text tokens lack such explicit spatial context. To bridge this gap, we introduce the position-guided expression cross-attention (PECA) module, which injects 3D positional information into text tokens, enabling spatially

consistent attention weight computation between visual and text tokens.

Figure 3(b) illustrates the overall architecture of the proposed PECA module. Let $\mathbf{P} \in \mathbb{R}^{N_{\text{vis}} \times 3}$ denote the 3D coordinates corresponding to the visual tokens $\mathbf{V}$. The positional embeddings $\mathbf{E}^{\text{pos}} \in \mathbb{R}^{N_{\text{vis}} \times C}$ are then projected from $\mathbf{P}$ using an MLP. These positional embeddings are interfused into the text tokens $\mathbf{T}$ as follows:

$$\mathbf{T}^{\text{pos}} = \mathbf{T} + \text{Softmax}\left(\frac{\mathbf{T} \cdot \mathbf{V}^\top}{\sqrt{C}}\right) \mathbf{E}^{\text{pos}}. \quad (4)$$

Then, we fuse the text tokens into the candidate tokens using cross-attention, where $\mathbf{V}^{\text{cand}}$ serves as the query, $\mathbf{T}^{\text{pos}}$ as the key, and $\mathbf{T}$ as the value, producing $\mathbf{V}_{\text{PECA}}^{\text{cand}}$ as the output. Since $\mathbf{V}^{\text{cand}}$ include positional embeddings from the previous process, PECA thus guides a spatially consistent fusion between visual and textual features.

Geometry-Aware Masked Attention (GMA): Although the candidate tokens $\mathbf{V}_{\text{PECA}}^{\text{cand}}$ contain semantically aligned features, they are insufficient to fully represent the geometric extent of the target object. Many visual tokens within the object’s bounding box are not selected as candidates, leading to incomplete spatial coverage. To incorporate this missing geometric information, we exploit the full set of visual tokens $\mathbf{V}$ to enrich $\mathbf{V}_{\text{PECA}}^{\text{cand}}$. However, directly attending to all visual tokens introduces noise, since many tokens belong to irrelevant objects or background regions. To alleviate this issue, we propose the geometry-aware masked attention (GMA) module in Figure 3(c), which selectively aggregates geometric cues from $\mathbf{V}$ based on their spatial relationships with each candidate token.

Let $\mathbf{P}^{\text{cand}} \in \mathbb{R}^{N_{\text{cand}} \times 3}$ denote the 3D coordinates of candidate tokens. The geometric relation tensor $\mathbf{G} \in \mathbb{R}^{N_{\text{cand}} \times N_{\text{vis}} \times 4}$ is then defined as

$$\mathbf{G}(i, j) = [\mathbf{P}_i^{\text{cand}} - \mathbf{P}_j, \|\mathbf{P}_i^{\text{cand}} - \mathbf{P}_j\|_2] \quad (5)$$

where i indexes candidate tokens ($i \in [1, N_{\text{cand}}]$) and j indexes visual tokens ($j \in [1, N_{\text{vis}}]$). Each element $\mathbf{G}(i, j)$ thus represents the relative coordinate offset and Euclidean distance between the i -th candidate token and the j -th visual token. We project $\mathbf{G}$ into the geometric feature tensor $\tilde{\mathbf{G}} \in \mathbb{R}^{N_{\text{cand}} \times N_{\text{vis}} \times C}$ using an MLP.

To emphasize text-guided geometric relationships, we compute a geometry-based attention prior $\mathbf{A}^{\text{geo}} \in \mathbb{R}^{N_{\text{cand}} \times N_{\text{vis}}}$ from the relation tensor $\tilde{\mathbf{G}}$ as

$$\mathbf{A}^{\text{geo}} = \text{Sigmoid}\left(\mathcal{F}^{-1}\left(\mathcal{F}(\tilde{\mathbf{G}}) \cdot \text{MaxPool}(\mathbf{T})^\top\right)\right) \quad (6)$$

where $\mathcal{F}(\cdot)$ denotes spatial flattening and $\mathcal{F}^{-1}(\cdot)$ is its inverse reshaping operator. Here, $\text{MaxPool}(\mathbf{T}) \in \mathbb{R}^{1 \times C}$ summarizes the expression-related text embeddings, serving as a global linguistic context to modulate the geometric affinity in $\tilde{\mathbf{G}}$. The resulting $\mathbf{A}^{\text{geo}}$ thus provides a soft

geometry-aware attention map indicating how strongly each visual token should contribute to a given candidate based on their spatial proximity and text relevance. Then, we generate the binary attention mask $\tilde{\mathbf{A}}^{\text{geo}} \in \{0, 1\}^{N_{\text{cand}} \times N_{\text{vis}}}$ from $\mathbf{A}^{\text{geo}}$ using a trainable thresholding scheme based on the straight-through estimator (STE) [5]. The final output of GMA $\mathbf{V}_{\text{GMA}}^{\text{cand}} \in \mathbb{R}^{N_{\text{cand}} \times C}$ is then obtained by masked-attention:

$$\mathbf{V}_{\text{GMA}}^{\text{cand}} = \mathbf{V}_{\text{PECA}}^{\text{cand}} + \text{Softmax}\left(\frac{\mathbf{V}_{\text{PECA}}^{\text{cand}} \mathbf{V}^\top}{\sqrt{C}} + \log(\tilde{\mathbf{A}}^{\text{geo}})\right) \mathbf{V}. \quad (7)$$

We apply learnable linear projections to all query, key, and value tokens before attention computation, but omit them in equations for simplicity. Here, $\log(\tilde{\mathbf{A}}^{\text{geo}})$ acts as an additive attention mask [32], where non-related visual tokens are assigned $-\infty$ to suppress their contributions. Thus, GMA enables each candidate token to selectively aggregate spatially consistent visual cues, suppressing interference from geometrically distant or irrelevant tokens.

3.4. Expression-Aware Contrastive Learning

As shown in Figure 1(b), when multiple objects of the same category are present in a scene, called “multiple” setting in 3DVG, the performance significantly decreases due to intra-category misidentification. In such cases, the visual token corresponding to the target object should accurately capture the descriptive expressions provided in the input text description, including attribute, pronoun, and relationship words. To address this challenge, we introduce an expression-aware contrastive learning that encourages the visual token of the target object to align with the expression words from the input text description, while remaining distinct from negative text samples, as illustrated in Figure 2.

Negative Text Sampling: For each text description in a “multiple” scene, we construct negative text samples by selecting other text descriptions within the same scene that refer to different objects of the same category as the target. These text descriptions serve as negative supervision for contrastive learning, providing semantically similar but referentially different contexts.

Expression-Aware Contrastive Loss: We first construct the negative text features $\mathbf{T}^- \in \mathbb{R}^{N^- \times C}$ by randomly selecting one negative description from the sampled descriptions and processing it through RoBERTa followed by the cross-modal encoder. From the resulting token embeddings, we retain only those corresponding to expression-related words: attribute, pronoun, and relationship words. We use off-the-shelf language parsing tools [29, 37] to categorize these word categories. Similarly, we extract the positive text features $\mathbf{T}^+ \in \mathbb{R}^{N^+ \times C}$ for the input text description, by filtering the text embeddings $\mathbf{T}$ to include only the expression-related tokens. We then define the expression-aware contrastive loss, which encourages the visual token correspond-

Method	Input	Unique (~19%)		Multiple (~81%)		Overall	
		Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
<i>Two-stage Model</i>							
ScanRefer [6]	3D+2D	76.33	53.51	32.73	21.11	41.19	27.40
ReferIt3D [2]	3D	53.8	37.5	21.0	12.8	26.4	16.9
TGNN [14]	3D	68.61	56.80	29.84	23.18	37.37	29.70
InstanceRefer [42]	3D	77.45	66.83	31.27	24.77	40.23	32.93
FFL-3DOG [9]	3D	78.80	67.94	35.19	25.70	41.33	34.01
BUTD-DETR [16]	3D	82.88	64.98	44.73	33.97	50.42	38.60
EDA [38]	3D	85.76	68.57	49.13	37.64	54.59	42.26
VPP-Net [30]	3D	86.05	67.09	50.32	39.03	55.65	43.29
G ³ -LQ [33]	3D	<u>88.09</u>	72.73	51.48	<u>40.80</u>	56.90	45.58
ConcreteNet [31]	3D	86.40	<u>82.05</u>	42.41	38.39	50.61	<u>46.53</u>
MCLN [27]	3D	86.89	72.73	<u>51.96</u>	40.76	<u>57.17</u>	45.53
EG-3DVG (Ours)	3D	90.84	84.57	52.87	46.71	58.54	52.36
<i>Single-stage Model</i>							
BUTD-DETR [16]	3D	81.47	61.24	44.20	32.81	49.76	37.05
EDA [38]	3D	86.40	69.42	48.11	36.82	53.83	41.70
G ³ -LQ [33]	3D	<u>88.59</u>	<u>73.28</u>	50.23	39.72	55.95	44.72
MCLN [27]	3D	84.43	68.36	49.72	38.41	54.30	42.64
TSP3D [10]	3D	87.25	71.41	<u>51.04</u>	<u>42.37</u>	<u>56.45</u>	<u>46.71</u>
EG-3DVG (Ours)	3D	91.54	83.79	51.09	44.16	57.13	50.07

Table 1. 3DREC results on ScanRefer [6]. The unique refers to cases where the target instance is the only one of its category, and the multiple indicates situations where there is at least one more object of the target’s category. The best results are **bold**, and the second best results are underlined.

ing to the target object to be aligned with $\mathbf{T}^+$ while being dissimilar from $\mathbf{T}^-$.

$$L_{\text{cont}} = -\log \left(\frac{\frac{w^+}{N^+} \sum_{i=1}^{N^+} \exp \left(\frac{\tilde{\mathbf{v}}_{i^{\text{gt}}}^{\text{cand}} (\mathbf{T}_i^+)^{\top}}{\rho} \right)}{\frac{w^+}{N^+} \sum_{i=1}^{N^+} \exp \left(\frac{\tilde{\mathbf{v}}_{i^{\text{gt}}}^{\text{cand}} (\mathbf{T}_i^+)^{\top}}{\rho} \right) + \frac{w^-}{N^-} \sum_{j=1}^{N^-} \exp \left(\frac{\tilde{\mathbf{v}}_{i^{\text{gt}}}^{\text{cand}} (\mathbf{T}_j^-)^{\top}}{\rho} \right)} \right) \quad (8)$$

where $\tilde{\mathbf{v}}_{i^{\text{gt}}}^{\text{cand}}$ is i^{gt} -th row vector of $\tilde{\mathbf{V}}^{\text{cand}}$, and i^{gt} is the index of candidate whose predicted bounding box has the highest IoU with the ground-truth. This feature corresponds to the target and is used as its visual representation. We set the temperature scaling factor to $\rho = 0.07$, and the positive and negative weights to $w^+ = 1$ and $w^- = 2$, respectively.

3.5. Loss Functions

Following [11, 27, 30, 33, 35, 36, 38], we apply 3DREC loss L_{rec} for bounding box prediction and 3DRES loss L_{res} for semantic mask prediction. Also, we use KPS loss L_{kps} [16, 21, 27, 30, 33, 38], and the proposed expression-aware contrastive loss L_{cont} . The total loss is defined as

$$L = \lambda_{\text{rec}} L_{\text{rec}} + \lambda_{\text{res}} L_{\text{res}} + \lambda_{\text{kps}} L_{\text{kps}} + \lambda_{\text{cont}} L_{\text{cont}} \quad (9)$$

where, scaling weights λ_{rec} , λ_{res} , λ_{kps} , and λ_{cont} are set to 0.14, 0.14, 8, and 0.1 respectively.

4. Experimental Results

4.1. Dataset and Evaluation Metrics

We evaluate the performance of the proposed EG-3DVG on two representative 3D visual grounding benchmarks, ScanRefer [6] and SR3D/NR3D [2]. ScanRefer is built on 800 3D scenes from ScanNet [8], and includes 51,583 manually annotated natural language expressions referring to 11,046 respective target objects. SR3D/NR3D is also built on 707 ScanNet scenes, where SR3D includes 83,572 machine-generated descriptions and NR3D contains 41,503 human annotated descriptions. Following previous works, we define the two sub-tasks of 3D visual grounding as follows: 3D Referring Expression Comprehension (3DREC) for 3D bounding box prediction [11, 16, 27, 30, 33, 38], and 3D Referring Expression Segmentation (3DRES) for mask prediction [26, 27, 35, 36]. We evaluate 3DREC performance using Acc@0.25IoU and Acc@0.5IoU between predicted and ground-truth 3D bounding boxes, while 3DRES performance is evaluated using mean IoU, Acc@0.25IoU, and Acc@0.5IoU between predicted and ground-truth masks.

4.2. Implementation Details

The numbers of visual tokens N_{vis} and candidate visual tokens N_{cand} are set to 1,024 and 256, respectively. The maximum number of text tokens N_{text} is set to 256. The feature dimension for both modalities is fixed to $C = 288$. The

Method	Input	Acc@0.25	Acc@0.5	mIoU
TGNN [14]	3D	37.50	31.40	27.80
X-RefSeg3D [26]	3D	40.33	37.77	29.94
3D-STMN [36]	3D	54.60	39.80	39.50
RG-SAN [35]	3D	61.67	44.92	44.66
MCLN [27]	3D	58.70	<u>50.70</u>	<u>44.72</u>
EG-3DVG (Ours)	3D	<u>59.77</u>	53.77	47.28

Table 2. 3DRES results on ScanRefer [6].

cross-modal encoder is repeated three times, while EGD is repeated six times. RoBERTa is kept frozen during training. All experiments are conducted on two NVIDIA RTX A6000 GPUs with a batch size of 8 per GPU. We use the AdamW optimizer with a learning rate of 1×10^{-5} for PointNet++ and 2×10^{-4} for the remaining layers.

4.3. Main Results

Quantitative Results: Tables 1 and 2 compare the proposed EG-3DVG with existing 3DREC methods [10, 11, 16, 27, 30, 33, 38] and 3DRES methods [26, 27, 35, 36] on ScanRefer. The scores of other methods are from their respective papers. In Table 1, the single-stage models directly estimate the target object’s bounding box from the input scene, as in the proposed EG-3DVG. The two-stage methods [16, 27, 33, 38] first perform 3D object detection using off-the-shelf networks [21] and auxiliary use the detections in the grounding decoder. For the two-stage setting, we further extend EG-3DVG to use 3D detections as source features for the candidate branch. A detailed description of this extension is provided in the supplementary material. Unique refers to cases where the target is the only one of its category, while multiple indicates situations where there is at least one more object of the target’s category.

In Table 1, EG-3DVG significantly outperforms existing methods. Compared to the previous state-of-the-art method, TSP3D [10], EG-3DVG achieves gains of +0.68 and +3.36 in Acc@0.25IoU and Acc@0.5IoU, respectively. In the two-stage setting, EG-3DVG further surpasses MCLN [27] by +1.37 in Acc@0.25IoU and +6.83 in Acc@0.5IoU. In Table 2, EG-3DVG achieves the highest 3DRES performance, with margins of +1.07 Acc@0.25IoU, +3.07 Acc@0.5IoU, and +2.56 mIoU compared to MCLN. Table 3 further reports the 3DVG performance of EG-3DVG on the SR3D/NR3D datasets. The hard case corresponds to the multiple setting. EG-3DVG surpasses the previous 3DREC model, G³-LQ [33], by a margin of +2.1 in SR3D Acc@0.25 and +1.5 in NR3D Acc@0.25, while outperforming the 3DRES method, MCLN [27], by +9.1 in SR3D mIoU and +1.9 in NR3D mIoU.

Qualitative Results: Figure 4 presents qualitative comparisons between MCLN [27] and EG-3DVG. We observe that EG-3DVG is robust to the three types: cross-modal mis-

Method	SR3D		NR3D	
	Overall	Hard	Overall	Hard
<i>3DREC</i>				
ReferIt3D [2]	39.8	-	35.6	27.9
TGNN [14]	45.0	36.9	37.3	30.6
InstanceRefer [42]	48.0	40.5	38.8	31.8
3DVG-Transformer [44]	51.4	44.9	40.8	34.8
SAT [39]	57.9	50.0	49.2	42.4
3DRefTransformer [1]	47.0	38.3	39.0	32.0
3D-SPS [23]	62.6	65.4	51.5	45.1
BUTD-DETR [16]	67.0	63.2	54.6	48.4
LAR [4]	59.4	51.2	48.9	42.3
EDA [38]	68.1	62.9	52.1	46.1
VPP-Net [30]	68.7	-	56.9	-
G ³ -LQ [33]	<u>73.1</u>	<u>66.3</u>	58.4	<u>50.7</u>
MCLN [27]	68.4	-	<u>59.8</u>	-
EG-3DVG (Ours)	75.2	66.6	59.9	52.4
<i>3DRES</i>				
TGNN [14]	20.2	21.9	19.1	17.4
3D-STMN [36]	34.4	30.1	27.6	23.0
RG-SAN [35]	39.6	<u>35.2</u>	28.6	<u>24.0</u>
MCLN [27]	<u>49.8</u>	-	<u>46.1</u>	-
EG-3DVG (Ours)	58.9	52.0	48.0	41.9

Table 3. 3DREC and 3DRES results on SR3D/NR3D. For 3DREC, we assess performance using Acc@0.25. For 3DRES, we assess performance using mIoU. The best results are **bold**, and the second best results are underlined.

alignment (Figures 4(a,b)), intra-category confusion (Figure 4(c,d)), and geometric reasoning error (Figures 4(e)). These results indicate that the proposed PECA, ECL, and GMA effectively address each failure type in Figure 1.

4.4. Ablation Study

Table 4 reports the ablation study according to different combinations of position-guided expression cross-attention (PECA), geometry-aware masked attention (GMA), and expression-aware contrastive learning (ECL). The models without PECA and GMA use standard cross-attention with text tokens and visual tokens, respectively. As in Table 4, removing any of these components results in a clear drop in both 3DREC and 3DRES performance. This validates the effectiveness of each module: PECA guides spatially consistent fusion between visual and textual features, GMA enhances the model’s understanding of geometric relationships among visual tokens, and ECL improves model’s ability to distinguish fine-grained linguistic expressions.

In Table 5, we compare the proposed attention mask A^{geo} in the proposed GMA with other alternative methods. First, we use a fixed-radius mask, where the attention mask is defined as 1 for visual tokens within a fixed radius around each candidate and 0 otherwise. The radius is set

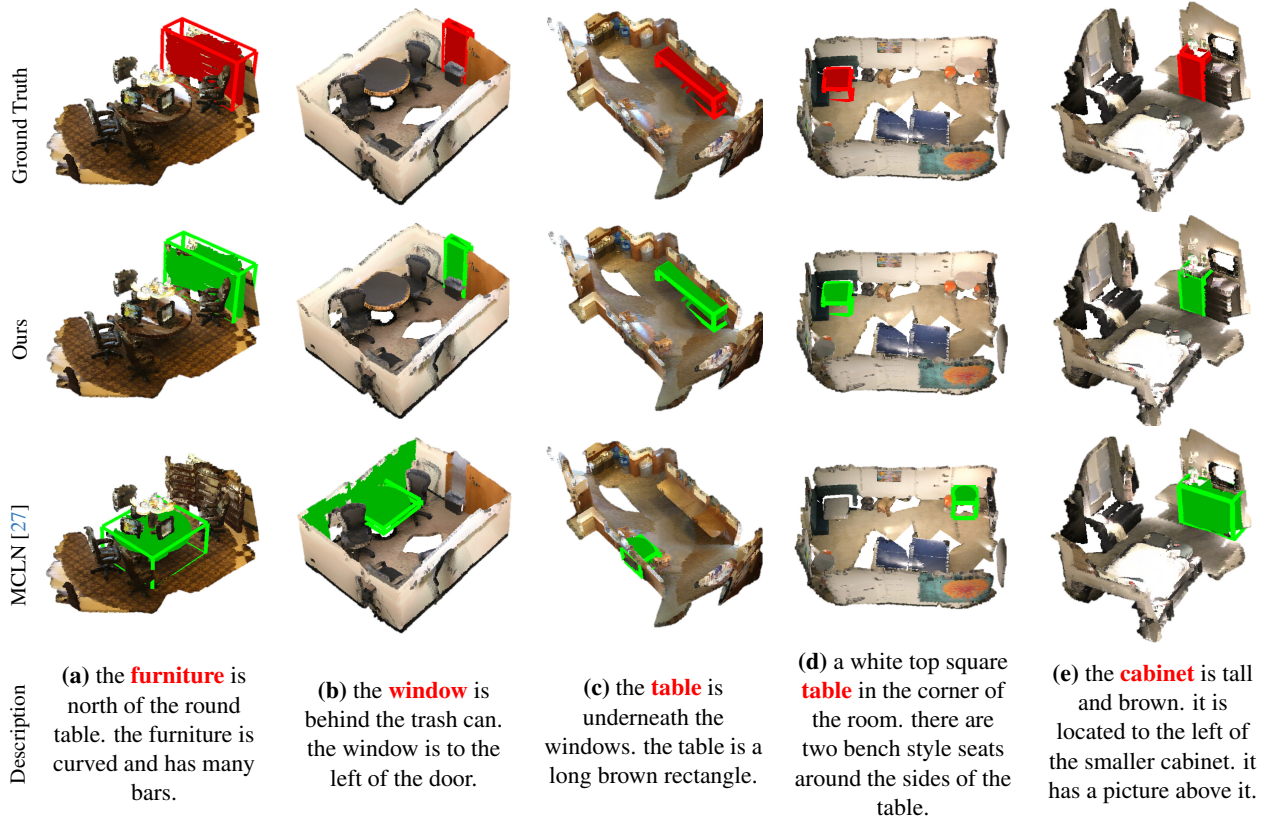


Figure 4. Qualitative results on ScanRefer: (a-b) cross-modal misalignment, (c-d) intra-category confusion, (e) geometric reasoning error.

PECA	GMA	ECL	3DREC		3DRES mIoU
			Acc@0.25	Acc@0.5	
			57.04	49.62	44.19
✓			57.73	50.92	45.69
	✓		57.05	50.55	46.19
		✓	57.56	50.85	45.84
✓	✓		58.06	51.84	46.57
✓		✓	58.18	50.95	46.34
	✓	✓	57.86	51.64	46.59
✓	✓	✓	58.54	52.36	47.28

Table 4. Ablation study on different combinations of PECA, GMA, and ECL on ScanRefer.

to 0.75m, corresponding to the average object size in ScanRefer. This approach simply aggregates information from nearby tokens without considering their geometric relationships. Second, we directly project the geometric feature $\mathbf{G}$ into a softened attention map through an MLP followed by a sigmoid function. This method excludes text tokens from the attention mask generation, leading to the aggregation of spatially close but expression-irrelevant visual features. In contrast, the proposed GMA jointly explores geometric and textual features for attention mask, consistently achieving the highest 3DREC and 3DRES performance.

Method	3DREC		3DRES mIoU
	Acc@0.25	Acc@0.5	
Fixed radius	57.76	51.12	46.23
Geometric feature	58.16	51.20	46.70
GMA	58.54	52.36	47.28

Table 5. Ablation study according to the generation methods of attention mask in GMA.

5. Conclusion

In this work, we presented EG-3DVG, a 3D visual grounding framework that addresses three major limitations: cross-modal misalignment, intra-class confusion, and geometric reasoning errors. We introduced an expression and geometry aware grounding decoder that integrates two complementary modules, PECA and GMA, to achieve reliable text–vision alignment and geometry-consistent feature aggregation. In addition, we proposed an expression-aware contrastive learning to enhance the model’s ability to distinguish semantically similar instances based on fine-grained expression cues. Experimental results on ScanRefer and SR3D/NR3D demonstrate that EG-3DVG achieves state-of-the-art performance on both 3D bounding box localization and mask prediction.

Acknowledgements

This work was partly by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. RS-2024-00397293, No. RS-2024-00352566, and No. RS-2025-00559165).

References

- [1] Ahmed Abdelreheem, Ujjwal Upadhyay, Ivan Skorokhodov, Rawan Al Yahya, Jun Chen, and Mohamed Elhoseiny. 3dref-transformer: Fine-grained object identification in real-world scenes using natural language. In *WACV*, pages 3941–3950, 2022. 7
- [2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, pages 422–440, 2020. 6, 7
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, pages 23716–23736, 2022. 2
- [4] Eslam Bakr, Yasmeen Alsaedy, and Mohamed Elhoseiny. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. In *NeurIPS*, pages 37146–37158, 2022. 7
- [5] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 5
- [6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, pages 202–221, 2020. 1, 6, 7
- [7] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *CVPR*, pages 11124–11133, 2023. 2
- [8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828–5839, 2017. 6
- [9] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-form description guided 3d visual graph network for object grounding in point cloud. In *ICCV*, pages 3722–3731, 2021. 6
- [10] Wenxuan Guo, Xiuwei Xu, Ziwei Wang, Jianjiang Feng, Jie Zhou, and Jiwen Lu. Text-guided sparse voxel pruning for efficient 3d visual grounding. In *CVPR*, pages 3666–3675, 2025. 2, 6, 7
- [11] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *ICCV*, pages 15372–15383, 2023. 1, 2, 6, 7
- [12] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *ICCV*, pages 7909–7920, 2023. 2
- [13] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *CVPR*, pages 17980–17989, 2022. 2
- [14] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *AAAI*, pages 1610–1618, 2021. 6, 7
- [15] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *CVPR*, pages 15524–15533, 2022. 1, 2
- [16] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Kateřina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *ECCV*, pages 417–433, 2022. 1, 3, 6, 7
- [17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 2
- [18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetrm: modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021. 2
- [19] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *CVPR*, pages 17990–17999, 2022. 2
- [20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3
- [21] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *ICCV*, pages 2949–2958, 2021. 3, 6, 7
- [22] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. In *CVPR*, pages 1190–1199, 2023. 2
- [23] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *CVPR*, pages 16454–16463, 2022. 7
- [24] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, pages 815–824, 2023. 2
- [25] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 3
- [26] Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-refseg3d: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks. In *AAAI*, pages 4551–4559, 2024. 1, 2, 6, 7
- [27] Zhipeng Qian, Yiwei Ma, Zhekai Lin, Jiayi Ji, Xiaowu Zheng, Xiaoshuai Sun, and Rongrong Ji. Multi-branch collaborative learning network for 3d visual grounding. In *ECCV*, pages 381–398, 2024. 1, 2, 3, 4, 6, 7, 8

- [28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. [2](#)
- [29] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. [1](#), [5](#)
- [30] Xiangxi Shi, Zhonghua Wu, and Stefan Lee. Viewpoint-aware visual grounding in 3d scenes. In *CVPR*, pages 14056–14065, 2024. [1](#), [2](#), [3](#), [6](#), [7](#)
- [31] Ozan Unal, Christos Sakaridis, Suman Saha, and Luc Van Gool. Four ways to improve verbo-visual fusion for dense 3d visual grounding. In *ECCV*, pages 196–213, 2024. [1](#), [6](#)
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. [1](#), [5](#)
- [33] Yuan Wang, Yali Li, and Shengjin Wang. G^3 -lq: Marrying hyperbolic alignment with explicit semantic-geometric modeling for 3d visual grounding. In *CVPR*, pages 13917–13926, 2024. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#)
- [34] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *CVPR*, pages 11686–11695, 2022. [2](#)
- [35] Changli Wu, Qi Chen, Jiayi Ji, Haowei Wang, Yiwei Ma, You Huang, Gen Luo, Hao Fei, Xiaoshuai Sun, and Rongrong Ji. RG-SAN: Rule-guided spatial awareness network for end-to-end 3d referring expression segmentation. In *NeurIPS*, 2024. [1](#), [6](#), [7](#)
- [36] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *AAAI*, pages 5940–5948, 2024. [1](#), [2](#), [6](#), [7](#)
- [37] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *CVPR*, pages 6609–6618, 2019. [1](#), [5](#)
- [38] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *CVPR*, pages 19231–19242, 2023. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#)
- [39] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *ICCV*, pages 1856–1866, 2021. [7](#)
- [40] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *AAAI*, pages 3081–3089, 2022. [2](#)
- [41] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattrnet: Modular attention network for referring expression comprehension. In *CVPR*, pages 1307–1315, 2018. [2](#)
- [42] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *ICCV*, pages 1791–1800, 2021. [6](#), [7](#)
- [43] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *CVPR*, pages 8563–8573, 2022. [2](#)
- [44] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, pages 2928–2937, 2021. [7](#)