Multi-Attribute Interactions Matter for 3D Visual Grounding

Can Xu[1], Yuehui Han[1], Rui Xu[1], Le Hui[2], Jin Xie[3,4*], Jian Yang[1]
[1]PCA Lab, Nanjing University of Science and Technology, Nanjing, China
[2]School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
[3]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[4]School of Intelligence Science and Technology, Nanjing University, Suzhou, China
volcanox123@gmail.com; {hanyh,xu_ray,csjxie,csjyang}@njust.edu.cn; huile@nwpu.edu.cn

Abstract

3D visual grounding aims to localize 3D objects described by free-form language sentences. Following the detection-then-matching paradigm, existing methods mainly focus on embedding object attributes in unimodal feature extraction and multimodal feature fusion, to enhance the discriminability of the proposal feature for accurate grounding. However, most of them ignore the explicit interaction of multiple attributes, causing a bias in unimodal representation and misalignment in multimodal fusion. In this paper, we propose a multi-attribute aware Transformer for 3D visual grounding, learning the multi-attribute interactions to refine the intra-modal and inter-modal grounding cues. Specifically, we first develop an attribute causal analysis module to quantify the causal effect of different attributes for the final prediction, which provides powerful supervision to correct the misleading attributes and adaptively capture other discriminative features. Then, we design an exchanging-based multimodal fusion module, which dynamically replaces tokens with low attribute attention between modalities before directly integrating low-dimensional global features. This ensures an attribute-level multimodal information fusion and helps align the language and vision details more efficiently for fine-grained multimodal features. Extensive experiments show that our method can achieve state-of-the-art performance on ScanRefer and Sr3D/Nr3D datasets. The code is publicly available at https://github.com/volcanoXC/MA2TransVG.

1. Introduction

3D Visual Grounding (VG) aims to locate the most relevant object in a given point cloud scene based on a language description. As a cornerstone of wide applications [6, 8, 14, 28, 43], such as vision-language navigation and autonomous robots, 3D VG has attracted increasing attention from both academia and industry. However, limited by the complexity of free-form language description and the irregularity of sparse point clouds, achieving accurate 3D visual grounding remains an open issue.

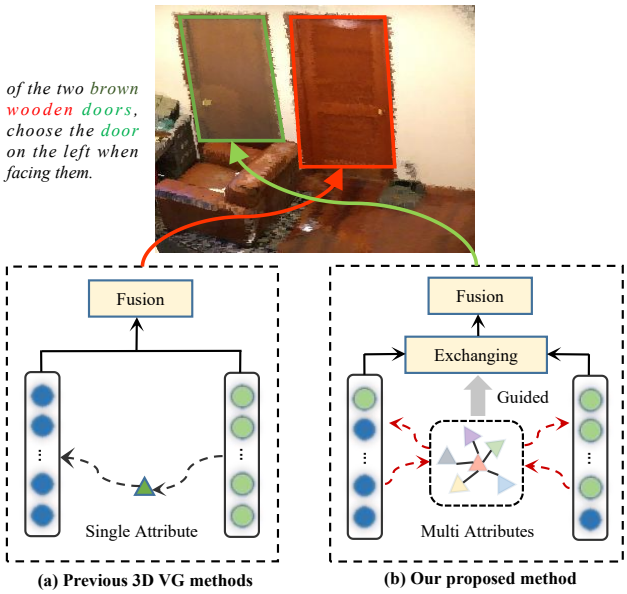

Figure 1. Different from prior works, we capture the multi-attribute interactions to guide the unimodal representation and multimodal fusion for more accurate grounding.

Existing 3D VG methods mainly follow a detection-then-matching pipeline. Specifically, a pool of proposals is first produced by the pre-trained general 3D object detectors [11, 29, 37]. Then, the linguistic description is associated with the visual feature of each proposal to generate multimodal object representation, which will be used to predict their referring confidence score for grounding. Since such a two-stage paradigm aims to collect enough semantic cues to distinguish the target object, the

quality of unimodal feature representation and multimodal feature fusion always play a critical role, which decides the discriminativeness of each proposal and their matching difficulty with the target label. Most methods focus on enhancing the feature from various perspectives, such as sparse convolution [47], 2D image assistance [5, 27, 46, 52], coarse proposal refinement [13, 22, 50], or segmentation/captioning collaboration [3, 10, 18, 26]. As shown in Fig. 1(a), recent works [7, 18, 19, 35] enhance the feature by embedding object attributes, such as color, name and language-conditioned correlation (*i.e.*, distance, view, direction), showing a promising future in advancing grounding performance.

Though prior works made significant progress, we argue few of them explore the multi-attribute interactions in either unimodal representation or multimodal fusion, which affects the grounding accuracy, especially for cases with multiple objects. As the CASE in Fig. 1, the model embeds the most salient attribute (object name of 'door') and learns other potential attributes (*e.g.*, 'brown', 'wooden', 'left') depending on the self-attention mechanism. Since the loss supervises the final prediction rather than the intermediate attention, the model may be encouraged to focus on some main attributes instead of all attributes, which weakens the discrimination of intra-modal representation. Meanwhile, when fusing the low-dimensional global feature between modalities via cross-attention, attention to 'brown' and 'wooden' should be encouraged for more discriminative language expression of the object, while they may be suppressed in the proposal representation to maximize the proposal's distinctiveness. Unfortunately, this would bring a misalignment in multimodal features, causing the model to select the auxiliary object (right door) rather than the primary object (left door).

To alleviate the bias in unimodal representation and misalignment in multimodal fusion, we propose a multi-attribute aware Transformer (MA^2TransVG) for 3D visual grounding, which enhances the unimodal representation with intrinsic attribute interaction and refines the cross-modal information fusion to the attribute level (See Fig. 1(b)). Concretely, we develop an attribute causal analysis module (ACAM), which quantifies the effect of each attribute (*i.e.*, learned attentions) and counterfactual interference (*i.e.*, wrong attention) for the final prediction (*i.e.*, grounding score). By maximizing the causal effect, powerful supervision is provided to correct the biased attribute and adaptively assign more proper attention to all attributes. Guided by the learned interaction, we further propose an exchanging-based multimodal fusion module (EMFM) to replace tokens with low attribute attention between modalities, which helps align more vision and language details and generate fine-grained multimodal features for accurate grounding. Our contributions include:

- We propose a multi-attribute aware Transformer for 3D object grounding, which explicitly models the multi-attribute interactions to prevent imbalanced and misaligned visual-textual representation.
- We propose the ACAM module to quantify the causal effect of multiple intrinsic attributes with the final prediction, which helps adaptively assign more proper attention to all attributes for a better intra-modal understanding.
- We propose the EMFM module to perform multimodal fusion based on multi-attribute exchanging, which helps align more attribute-level details between modalities for fine-grained multimodal features.
- We evaluate the proposed method on Nr3D/Sr3D and ScanRefer benchmarks, which outperforms all state-of-the-art methods by a large margin.

2. Related work

3D Visual Grounding. The interest in 3D VG has grown rapidly in recent years, and existing approaches include one-stage [21, 31] and two-stage methods. The two-stage methods [13, 16–18, 46, 47, 50] decouple the grounding into language-irrelevant object detection and cross-modal matching. Recent works focus on improving the intra-modal representation and modeling the inter-modal relationship. For example, Graph-based approaches [1, 18, 47] infer spatial relations among proposals by connecting each proposal with its top-N nearest neighbors. Transformer-based approaches [7, 17, 19, 21, 31, 46] leverage the attention mechanism to concentrate the object attributes on proposals that are more essential to the referring. In these methods, only the most salient attribute is embedded and the fusion between modalities occurs in the global feature space, resulting in a biased intra-modal representation and inadequate inter-modal fusion. Instead, we capture the multi-attribute interactions and refine the cross-modal learning at the attribute level.

Causal Inference in Vision. Causality analysis [32] has been successfully used in several areas, such as adversarial learning [24], reinforcement learning [23] and graph neural networks [42, 51]. In vision tasks [30, 40, 44], causality inference also works as an effect tool to discover hidden causal structures and confront data biases. For example, Rao *et al.* [38] learn more effective attention based on causal inference for fine-grained image categorization, person and vehicle re-identification. As the first attempt, we introduce the causality analysis into 3D VG to quantify the effect of each attribute, which helps model the complex interactions of various attributes to enhance the unimodal feature.

Transformers in Vision-and-Language. Transformer shows a powerful ability to learn multimodal features in various tasks, such as image captioning [9, 10, 48], vision-and-language navigation [28, 43], and vision question-answering [15, 53]. For 3D VG, different from most

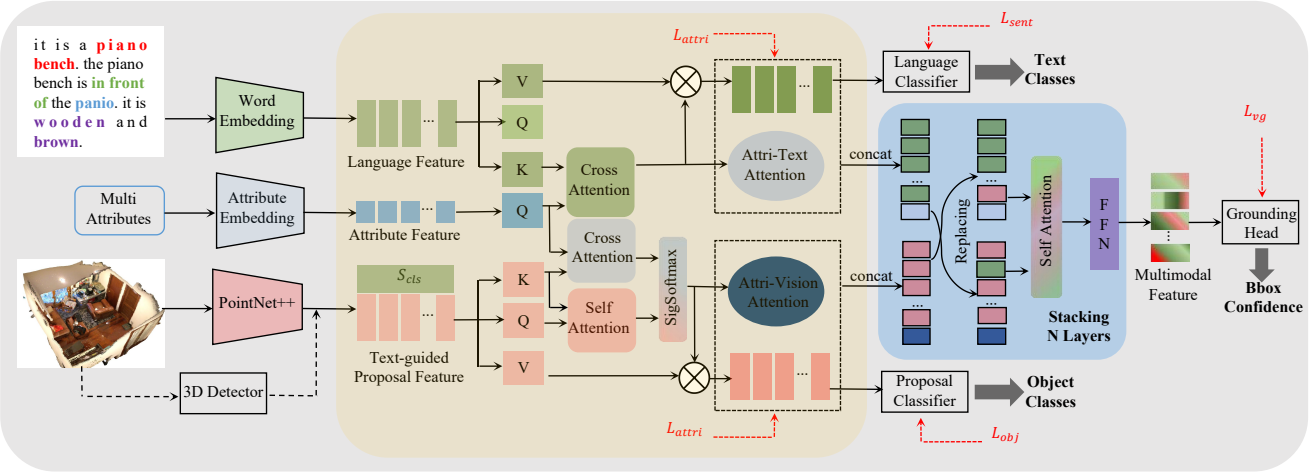

Figure 2. The framework of proposed MA^2TransVG. The embedding of pre-detected proposals, attributes, and language are extracted and then cross-attend in ACAM. By applying the causal effect analysis via L_{attri}, ACAM models the multi-attribute interactions for final prediction to generate the Attribute-Vision/Text attention, which can enhance the proposal and text embedding. The attention further guides the attribute-level feature exchanging in EMFM, which helps fuse the multimodal feature for accurate grounding.

methods that directly fuse the global unimodal features by stacking cross-attention layers [7, 17, 19, 21, 31, 46], we refine the cross-modal fusion to the attribute level, which helps suppress the misalignment between modalities to capture a finer-grained matching relationship.

3. Method

3.1. Overview

Fig. 2 shows an overview of proposed MA^2TransVG. Given a 3D point cloud with corresponding description, text features $\boldsymbol{T} \in \mathbb{R}^{(m+1)\times d}$, proposal features $\boldsymbol{O} \in \mathbb{R}^{n\times d}$ and attribute features $\boldsymbol{F} \in \mathbb{R}^{n\times d}$ are extracted (Sec. 3.2). The three streams then are fed into the attribute causal analysis module (ACAM) to capture the multi-attribute interactions (*i.e.*, Attribute-Vision attention and Attribute-Text attention) to enhance initial text and proposal feature. The attention is supervised by an attribute loss L_{attri}, which is designed by estimating the causal effect of attributes for the final prediction (Sec. 3.3). The enhanced proposal embedding $\boldsymbol{O}^*$ and text embedding $\boldsymbol{T}^*$ are further passed to the exchanging-based multimodal fusion module (EMFM), where their tokens with small attribute attention are replaced with the average of all the tokens in the other modality (Sec. 3.4). With the fused multimodal embedding $\boldsymbol{E} \in \mathbb{R}^{n\times d}$, we predict the final grounding results.

3.2. Input Modal Representation

Text Encoding. Given a sentence $\boldsymbol{S}$ with m word tokens, we use a pre-trained BERT model [20] to produce the text features $(t_{cls}, t_1, \ldots, t_m)$. $t_i \in \mathbb{R}^d$ is the feature of each token with a dimensionality of d and $t_{cls} \in \mathbb{R}^d$ is a special token for text classification.

Object Encoding. Given a point cloud scene $\boldsymbol{P} \in \mathbb{R}^{K\times 6}$ of K points described by XYZ coordinates and RGB colors, we use the GroupFree [29] detector to obtain an object proposal list $\boldsymbol{o} = (o_1, o_2, \ldots, o_n)$ containing n proposals and extract corresponding proposal feature $\boldsymbol{O} \in \mathbb{R}^{n\times d}$ based on PointNet++ [36]. Since we mainly focus on the object referred in the text, each proposal feature $O_i \in \mathbb{R}^d$ will be added with the t_{cls} before feeding into ACAM for attribute-vision causal effect analysis.

Attributes Encoding. Motivated by [3], for the proposal pool $\boldsymbol{o}$ with n objects, the multi-view RGB features $\boldsymbol{f_c} \in \mathbb{R}^{n\times 128}$ that potentially contain the color and materials attribute is first extracted. Then, we use the 3-dimensional box center and 24-dimensional corner coordinates to describe the object size attribute $\boldsymbol{f_s} \in \mathbb{R}^{n\times 27}$. For the location attribute, we further consider the distance and angle of each pair of objects (o_i, o_j). Connecting the object center point $(x_{c_i}, y_{c_i}, z_{c_i})$ and $(x_{c_j}, y_{c_j}, z_{c_j})$, the distance is their Euclidean distance d_{ij}. Similar to [7], the horizontal angle θ_h and vertical angle θ_v can be calculated as:

$$\begin{aligned} \theta_h &= \arctan 2((y_{c_i} - y_{c_j})/(x_{c_i} - x_{c_j}) \\ \theta_v &= \arcsin((z_{c_i} - z_{c_j})/d_{ij}) \end{aligned} \tag{1}$$

The location attribute $f_l^{ij} \in \mathbb{R}^5$ between any pair of objects is define, as:

$$f_l^{ij} = [d_{ij}, \sin(\theta_h), \cos(\theta_h), \sin(\theta_v), \cos(\theta_v)] \tag{2}$$

For each proposal o_i, the location attribute $\boldsymbol{f_l^i} \in \mathbb{R}^5$ can be calculated by averaging all f_l^{ij} connected with o_i, and the final attribute embedding can be generated by concatenating

all three attributes, as $\boldsymbol{f}^i = [\boldsymbol{f}_{\boldsymbol{l}}^i, \boldsymbol{f}_{\boldsymbol{c}}^i, \boldsymbol{f}_{\boldsymbol{s}}^i]$. We map $\boldsymbol{f}^i \in \mathbb{R}^d$ to the same dimension of text feature using a FC layer.

3.3. Attribute Causal Analysis Module

The attribute causal analysis module (ACAM) models the multi-attribute interactions with vision/text as Attribute-Vision/Text attention based on a well-designed Attribute-aware Transformer. To capture more influential and diverse interactions, ACAM further guides the learning process with an extra loss L_{attri}, which measures the causal effect of learned attribute-level attentions for the final prediction.

Attributes-Text Attention Learning. Given attribute embedding $\boldsymbol{F} \in \mathbb{R}^{n\times d}$, Attributes-Text attention $\boldsymbol{A_T}$ helps select related attribute features from the initial text feature $\boldsymbol{T} \in \mathbb{R}^{(m+1)\times d}$, thereby the text representation can be enhanced to better describe the target object. We define $\boldsymbol{A_T} \in \mathbb{R}^{n\times(m+1)}$ as the cross-attention between the Attribute embedding $\boldsymbol{F}$ and text embedding $\boldsymbol{T}$, which can be computed with a standard Transformer [41].

In order to enlarge the model capacity for diverse Attribute-Text correlations, we also introduce multi parallel attention layers ℓ as different transformation heads, and the multi-head attentions will be fused with the standard self-attention [41]. As shown in Fig. 2, $\boldsymbol{A_T}$ then guides to generate the enhanced text feature $\boldsymbol{T^*} \in \mathbb{R}^{n\times d}$, which will be used to train the language classifier with the supervision of the language classification loss L_{sent}, and also serves as input for subsequent EMFM for multi-modal fusion.

Attributes-Vision Attention Learning. Instead of directly calculating the cross-attention between proposal features $\boldsymbol{O}$ and attribute features $\boldsymbol{F}$, we first guide the proposal features with language token t_{cls}, as we prefer to care about specific categories of objects mentioned in the text [7]. Specifically, for the text-guided proposal feature, we adopt the similar Transformer structure in Attributes-Text attention learning to calculate its self-attention $\boldsymbol{A_V^s} \in \mathbb{R}^{n\times n}$, and cross-attention $\boldsymbol{A_V^c} \in \mathbb{R}^{n\times n}$ with $\boldsymbol{F}$.

Similar to [7], we then integrate the self-attention of proposals $\boldsymbol{A_V^s}$ with the cross attribute-proposal attention $\boldsymbol{A_V^c}$ using a sigmoid softmax (sigsoftmax) fusion operation to calculate the Attributes-Vision attention $\boldsymbol{A_V}$, as:

$$A_v = \frac{\sigma(\boldsymbol{A_V^c}(i,j))\exp(\boldsymbol{A_V^s}(i,j))}{\sum_{j=1}^{n}\sigma(\boldsymbol{A_V^c}(i,j))\exp(\boldsymbol{A_V^s}(i,j))} \tag{3}$$

where $\sigma(\,)$ is the sigmod function. $\boldsymbol{A_V^s}(i,j)$ and $\boldsymbol{A_V^c}(i,j)$ are the element in i-th row and j-th column in attention matrix $\boldsymbol{A_V^s}$ and $\boldsymbol{A_V^c}$, respectively. The same multi-head attention strategy is used in Attribute-Vision attention $\boldsymbol{A_V}$. In a similar manner, the enhanced proposal embedding $\boldsymbol{O}^*$ can be obtained by performing a matrix multiplication operation between $\boldsymbol{O}$ and $\boldsymbol{A_V}$, which will serve as the input of proposal classifier and EMFM module.

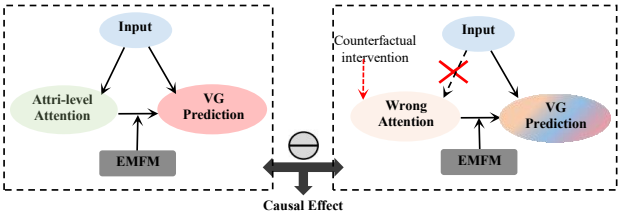

Figure 3. Visualization of derivating causal effects through counterfactual causality. We subtract the counterfactual prediction from the original prediction to analyze the effects of learned attention and maximize them in the training process.

Strengthen Interaction With L_{attri}. Though multiple attributes have been embedded to guide the language and vision representations via Attribute-Text/Vision attention, how these attributes interact with each other to affect predictions remains a black box. To address this issue, our core idea lies in measuring the causal effect of $\boldsymbol{A_V}$ and $\boldsymbol{A_T}$ for the final prediction. We design the attribute loss L_{attri} based on the counterfactual analysis [33], providing extra supervision to encourage the network to explore more influential interactions.

As shown in Fig. 3, taking the Attribute-vision attention as an example, we conduct counterfactual intervention by imagining non-existent attention $\bar{\boldsymbol{A}}_{\boldsymbol{V}}$ to replace the initial $\boldsymbol{A_V}$. We keep the vision representation $\boldsymbol{O}$ unchanged, which will be multipled with $\bar{\boldsymbol{A}}_{\boldsymbol{V}}$ to calculate the counterfactual representation $\bar{\boldsymbol{O}}^*$. $\bar{\boldsymbol{O}}^*$ are then fused with $\boldsymbol{T}$ to obtain the counterfactual prediction $\bar{\boldsymbol{Y}} \in \mathbb{R}^{n\times 1}$, as:

$$\bar{\boldsymbol{Y}} = do(\boldsymbol{A_V} = \bar{\boldsymbol{A}}_{\boldsymbol{V}}, \boldsymbol{T}) = G(\bar{\boldsymbol{O}}^*, \boldsymbol{T}) \tag{4}$$

where $G(\,)$ is the grounding operation. In practice, the exact form of how counterfactual is achieved is not limited, and our starting point here is just to set the ball rolling. In our case, random attention, uniform attention, shuffle attention or reversed attention are used as the counterfactuals.

Following [33, 34, 40], we define the causal effect $\boldsymbol{Y_e}$ of $\boldsymbol{A_V}$ as the difference between the observed prediction $\boldsymbol{Y}$ and the counterfactual prediction $\bar{Y}$, as:

$$\boldsymbol{Y_e} = \mathbb{E}_{\bar{A}_V\sim\gamma}(\boldsymbol{Y} - \bar{\boldsymbol{Y}}) \tag{5}$$

where γ means the distribution of counterfactual attention.

Since the causal effect can be interpreted as how the attention improves the final prediction compared to wrong attentions, we use $\boldsymbol{Y_e}$ to measure the learned attention, $i.e.$, the multi-attribute interactions. Based on this observation, we further design an attribute loss L_{attri}, as:

$$L_{attri} = L_{ce}(\boldsymbol{Y_e}, \boldsymbol{y}) \tag{6}$$

where L_{ce} is the cross-entropy loss and $\boldsymbol{y}$ is the GT label. Similarly, we can obtain the causal effect of Attribute-text attention and compute its attribute loss. By optimizing L_{attri} from $\boldsymbol{A_T}$ and $\boldsymbol{A_V}$, the model is expected to:

(1) improve the prediction based on correct attentions as much as possible, which helps discover more discriminative attributes; (2) penalize the prediction based on wrong attentions, which forces the model adaptively capture more attributes to avoid biased unimodal embeddings.

3.4. Exchanging-based Multimodal Fusion Module

We propose the exchanging-based multimodal fusion module (EMFM) to communicate attribute-level details of objects between both input vectors for accurate grounding. Specifically, since text/visual encoders tend to focus on different attributes for better unimodal embeddings, directly fusing unimodal embeddings using cross-attention will hurt the distinguishability of multimodal features. To narrow the semantic misalignment between modalities, we select and exchange tokens with low attention score using the average of all tokens from another modal, which helps align attribute-level information between modalities sufficiently and eliminate semantic ambiguity in multimodal representations of proposals.

Exchanging With Attributes Interactions. As shown in Fig. 3, we implement the EMFM by two Transformer encoders with shared parameters, where the shallow μ layers are set as regular Transformer encoder layers, followed by η exchanging layers. Since $\boldsymbol{A_V}$ and $\boldsymbol{A_T}$ have modeled the multi-attribute interactions, they are taken as references to select tokens for exchanging. Meanwhile, $\boldsymbol{A_V}$ and $\boldsymbol{A_T}$ will be updated using the self-attention learned from unimodal embedding $\boldsymbol{T}^*$ and $\boldsymbol{O}^*$. Considering the interdependence between the attribute attention and unimodal embedding, we concatenate them for the subsequent calculation, as:

$$\begin{aligned} \boldsymbol{T}^{\boldsymbol{e}}(0) &= \text{Concat}[\boldsymbol{A_T}, \boldsymbol{T}^*] \\ \boldsymbol{O}^{\boldsymbol{e}}(0) &= \text{Concat}[\boldsymbol{A_V}, \boldsymbol{O}^*] \end{aligned} \tag{7}$$

where $\boldsymbol{T}^{\boldsymbol{e}}(0)$ and $\boldsymbol{O}^{\boldsymbol{e}}(0)$ are embedding for textual and visual modalities at layer 0, respectively.

At layer $\mu+1$ with intermediate embedding $\boldsymbol{T}^{\boldsymbol{e}}(\mu+1)$, inspired by [4, 25], we select tokens with a θ-proportion of the smallest attention scores to $\boldsymbol{A_T}$ and replace their embedding vectors with the average embedding of all the tokens in $\boldsymbol{O}^{\boldsymbol{e}}(\mu+1)$. For instance, the selected k-th token of $\boldsymbol{T}^{\boldsymbol{e}}(\mu+1)$ can be updated by:

$$\boldsymbol{T}^{\boldsymbol{e}}(\mu+1)[k,:] = \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{O}^{\boldsymbol{e}}(\mu+1)[j,:] + \boldsymbol{T}^{\boldsymbol{e}}(\mu+1)[k,:] \tag{8}$$

The updated embeddings are used to calculate the multi-head self-attention [41], which will guide the exchanging operation at the next layer, *e.g.*, $\boldsymbol{A_T}$ of the $(\mu+1)$-th layer are updated by the self-attention of the previous layer μ.

Multimodal Fusion. We also apply the residual connection between adjacent exchanging layers to reduce the information loss caused by replacement. The exchanging process continues until reaching the η layer to generate $\boldsymbol{O}^{\boldsymbol{e}}(\eta)$ and $\boldsymbol{T}^{\boldsymbol{e}}(\eta)$, which will be fed into the feed-forward network (FFN) with a normalization layer and a fully connected layer to derive the final fusion embedding $\boldsymbol{E} \in \mathbb{R}^{n\times n}$.

Grounding Head. We use a two-layer feed-forward neural network as the object grounding head. Given the multimodal representation $\boldsymbol{E}$, the grounding head will predict the probability $\boldsymbol{p}_i \in \mathbb{R}^{n\times 1}$ for each object proposal, where the proposal with the maximum probability is selected as the target in the inference stage.

3.5. Training Objectives

Following the previous works [1, 5, 17], the loss of MA^2TransVG consists of the primary 3D object grounding loss L_{vg}, language classification loss L_{sent}, proposal classification loss L_{obj}, and two attribute losses L_{attri} for vision/text attention causal analysis, as:

$$L = L_{vg} + L_{sent} + L_{obj} + \lambda L_{attri} \tag{9}$$

where λ is the hyper-parameter to balance the loss.

4. Experiment

4.1. Datasets and Evaluation Metrics

We evaluate the proposed MA^2TransVG on ScanRefer [5] and Sr3D/Nr3D [1] datasets. ScanRefer offers 51,583 descriptions of 11,046 objects from 800 ScanNet [12] scenes. A target object is labeled as 'unique' if it is the only object of its class in the scene; otherwise, it is labeled as 'multiple'. For ScanRefer, we use Acc@0.25 and Acc@0.5 as evaluation metrics, which measure the fraction of language queries whose predicted box overlaps the ground truth box with 3D IoU higher than 0.25/0.5. Compared to ScanRefer where objects need to be detected and then matched, Sr3D/Nr3D datasets provide object masks for each scene and only need to recognize the classes of the proposals to choose the target object. Nr3D consists of 41,503 descriptions from 707 scenes with human annotation, while Sr3D contains 83,572 simple machine-generated descriptions. Each scene in Sr3D/Nr3D can also be divided into 'easy' and 'hard' depending on whether there are more than two instances, and can be divided into 'view-dependent' and 'view-independent' according to whether the referring expression is dependent on the camera view. Following [1], the accuracy is used to verify the model.

4.2. Implementation Details

The text encoding module is initialized from the first three layers of BERT [20], and the object encoding module PointNet++ [36] samples 1024 points for all the objects. We generate an initial set of $n = 256$ proposals using [29]. We set the dimension of unimodal representation $d = 768$ and use a three-layer transformer with 12 heads for

Table 1. Comparison of 3D VG results with the state-of-the-art on ScanRefer.

Method	Venue	Unique (19%)		Multiple (81%)		Overall	
		0.25	0.5	0.25	0.5	0.25	0.5
ScanRefer [5]	ECCV20	67.6	46.2	32.1	21.3	40.0	26.1
ReferIt3D [1]	ECCV20	53.8	37.5	21.0	12.8	26.4	16.9
TGNN [18]	AAAI21	68.6	56.8	29.8	23.2	37.4	29.7
InstanceRefer [47]	ICCV21	77.5	66.8	31.3	24.8	40.2	32.9
3DVG [50]	ICCV21	77.2	58.5	38.4	28.7	45.9	34.5
FFL-3DOG [13]	ICCV21	78.8	67.9	35.2	25.7	41.3	34.0
MVT [19]	CVPR22	77.7	66.5	31.9	25.3	40.8	33.3
3D-SPS [31]	CVPR22	81.6	64.8	39.5	29.6	47.7	36.4
ViL3DRef [7]	NeurIPS22	81.6	68.6	40.3	30.7	47.9	37.7
BUTD-DETR [22]	ECCV22	/	/	/	/	52.2	39.8
EDA [45]	CVPR23	85.8	68.6	49.1	37.6	54.6	42.3
M3DRef-CLIP [49]	ICCV23	/	**77.2**	/	36.8	/	44.7
3DJCG [3]	CVPR22	78.8	61.3	40.1	30.1	47.6	36.1
UniT3D [10]	ICCV23	82.8	73.1	36.4	31.1	42.3	39.1
3DRefTR-HR [26]	ICCV23	86.0	70.9	49.6	38.3	55.0	43.2
3DRefTR-SP [26]	ICCV23	86.1	71.0	50.1	38.7	55.5	43.5
Ours	/	**86.3**	74.1	**53.8**	**41.4**	**57.9**	**45.7**

the Attribute-Vision/Text attention learning. For a fair comparison, we keep the architecture parameters as same as previous works [7, 19, 46]. In ACAM, we use random attention to initialize the counterfactual attention. In EMFM, considering the balance of raw intra-modal knowledge and inter-modal communication, we empirically set $\theta = 10\%$. The start layer μ and end layer η are set to 1 and 4, respectively. The hyper-parameter of the whole loss in Eq. (9) is $\lambda = 0.5$. All models are trained on 4 NVIDIA RTX Titan GPUs, which are optimized using the AdamW algorithm with a batch size of 128 and a learning rate of 0.0005 with cosine decay scheduling.

4.3. Comparison with State-of-the-Art Methods

ScanRefer. In Tab. 1, we compare MA^2TransVG with existing works on the ScanRef dataset and observe the following results: **i)** Compared to pure 3D VG works such as BUTD-DETR [22], 3D-SPS [31], ViL3DRef [7], and EDA [45], our method achieves state-of-the-art performance of 57.9% and 45.7%, by a substantial margin with an overall improvement over 3.3% and 3.4%. **ii)** To facilitate the 3D grounding, previous works [5, 31, 46, 50] learn better point cloud features with the supplementary of 2D images while most recent works [3, 10, 26] learn better language-visual alignment by connecting 3D dense captioning or segmentation tasks. However, we outperform second-place 3DRefTR-SP [26] with an overall improvement over 2% for both IoU = 0.25 and IoU = 0.5 setting. This superiority illustrates that mining the multi-attribute interactions help capture more efficient and discriminative multi-modal representation to distinguish objects. **iii)** Our MA^2TransVG can reach a remarkable accuracy of 53.8% and 41.4% for 'Multiple' scenes, which verifies the effectiveness of our attribute interaction for understanding finer-grained language-visual contextual dependency.

Nr3D&Sr3D. Tab. 2 reports the grounding accuracy on Nr3D and Sr3D datasets. For a fair comparison, all the methods use the ground-truth object proposal with no ground-truth labels. In Nr3D, though the language descriptions are more complex to cause additional challenges for text understanding and cross-modal alignment, our method still outperforms the EDA [45], which is equipped with text component decoupling and dense matching between two modalities. Compared to EDA, our MA^2TransVG not only achieves a similar decoupling function for both descriptions and point clouds via attribute representation, but also considers their correlationships for prediction during Attribute-Text/Vision causal attention learning, which contributes to the great improvements by +13.1% in overall accuracy. Compared to other recent competitor ViL3DRef [7], 3DRefTR-SR [26], and M3DRef-CLIP [49], we also achieve a consist improvement of overall accuracy in all split settings by +0.8%, +2.6% and +15.8%, respectively. In Sr3D, our method reaches an accuracy of 73.9%, which greatly surpasses the best competitor ViL3DRef [7] 1.1% and outperforms other state-of-the-art methods [26, 45] over 5%. Similar to ViL3DRef [7] that guided the grounding with languaged-conditioned spatial relation, we not only embed the spatial attributes but also explicitly consider other attributes in the unimodal representation and further perform exchanging-based fusion. All the results show that multi-attribute interactions can help capture more semantic clues for accurate grounding.

Table 2. Comparison of 3D VG results with the state-of-the-art on Sr3D/Nr3D.

Method	Venue	Nr3D					Sr3D				
		Easy	Hard	View Dep	View Indep	Overall	Easy	Hard	View Dep	View Indep	Overall
ReferIt3D [1]	ECCV20	43.6	27.9	32.5	37.1	35.6	44.7	31.5	39.2	40.8	40.8
TGNN [18]	AAAI21	44.2	30.6	35.8	38.0	37.3	48.5	36.9	45.8	45.0	45.0
InstanceRefer [47]	ICCV21	46.0	31.8	34.5	41.9	38.8	51.1	40.5	45.4	48.1	48.0
3DVG [50]	ICCV21	48.5	34.8	34.8	43.7	40.8	54.2	44.9	44.6	51.7	51.4
FFL-3DOG [13]	ICCV21	48.2	35.0	37.1	44.7	41.7	/	/	/	/	/
TransRefer3D [17]	ACMM21	48.5	36.0	36.5	44.9	42.1	60.5	50.2	49.9	57.7	57.4
LanguageRefer [39]	CoRL21	51.0	36.6	41.7	45.0	43.9	58.9	49.3	49.2	56.3	56.0
SAT [46]	ICCV21	56.3	42.4	46.9	50.4	49.2	61.2	50.0	49.2	58.3	57.9
BUTD-DETR [22]	ECCV22	/	/	/	/	43.3	/	/	/	/	52.1
LAR [2]	NeurIPS22	58.4	42.3	47.4	52.1	48.9	63.0	51.2	50.0	59.1	59.4
3D-SPS [31]	CVPR22	58.1	45.1	48.0	53.2	51.5	56.2	65.4	49.2	63.2	62.6
MVT [19]	CVPR22	61.3	49.1	54.3	54.3	55.4	66.9	58.8	58.4	64.7	64.5
M3DRef-CLIP [49]	ICCV23	55.6	43.4	42.3	52.9	49.4	/	/	/	/	/
EDA [45]	CVPR23	/	/	/	/	52.1	/	/	/	/	68.1
3DRefTR-SR [26]	ICCV23	/	/	/	/	52.6	/	/	/	/	68.5
ViL3DRef [7]	NeurIPS22	70.2	57.4	62.0	64.5	64.4	74.9	67.9	63.8	73.2	72.8
Ours	/	**71.1**	**57.6**	**62.5**	**65.4**	**65.2**	**76.0**	**69.3**	**64.5**	**73.8**	**73.9**

4.4. Ablation Studies

Effectiveness of Each Component. We quantitatively investigate the contribution of each component in our method in Tab. 3. The baseline model simply extracts and fuses the object and language feature through a standard Transformer structure, where the modal relationships only implicitly depend on the self/cross-attention. Taking results on Nr3D as an example, when enhancing the feature with $\boldsymbol{A_T}$ or $\boldsymbol{A_V}$, we yield an accuracy of 55.2% and 51.9%, which can increase to 59.7% when combining two attentions. This indicates that embedding the multi-attribute within each single modal is the basis of accurate cross-modal understanding. Furthermore, we achieve a gain of 2.7% after introducing causal effect analysis, which helps adaptively capture more reliable multi-attribute interactions for feature enhancement. Exchanging tokens further improves the accuracy to 65.2% because multimodal knowledge exchanging boosts the quality of the learned embeddings for each modality. All the results demonstrate the effectiveness of each component in our proposed MA^2TransVG.

Effectiveness of Multi-attribute. To verify the effectiveness of multi-attribute interactions, we conduct ablation on attribute quantity and multi-attribute interactions. In Tab. 3. We adopt commonly used attributes (location f_l, size f_s, color f_c) with simple embedding method and achieve SOTA performance. The accuracy is 44.9% on Nr3D when generating attribute embedding only using f_l and f_s, which increases to 46.1% (**+1.2%**) and 60.2% (**+15.3%**) after adding f_c and interactions, respectively. Results show that our accuracy gain mainly comes from the multi-attribute interactions rather than embedding more attitudes.

Table 3. Ablation studies of each component on Nr3D and Sr3D.

f_l	f_s	f_c	A_T	A_V	L_{attri}	Exg	Nr3D	Sr3D
✓	✓						44.9	55.6
✓	✓		✓	✓	✓		60.2	69.2
✓	✓		✓	✓		✓	59.3	68.4
✓	✓		✓	✓	✓	✓	61.8	71.2
✓	✓	✓					46.1	57.8
✓	✓	✓	✓				55.2	64.6
✓	✓	✓		✓			51.9	62.2
✓	✓	✓	✓	✓			59.7	68.5
✓	✓	✓	✓	✓	✓		62.4	72.0
✓	✓	✓	✓	✓		✓	62.0	71.3
✓	✓	✓	✓	✓	✓	✓	65.2	73.9

Table 4. Comparison with various counterfactual attentions.

Acc@0.25	random attention	uniform attention	reversed attention	shuffle attention
ScanRefer	57.8	57.4	56.9	57.5
Nr3D	65.2	64.8	64.5	65.0
Sr3D	73.9	73.3	73.1	73.5

Effects of Counterfactual Attention. We implement different strategies to generate counterfactual attention including random attention, uniform attention, reversed attention, and shuffle attention, and report their result on Nr3D in Tab. 4. We see all the counterfactual attention strategies achieve a similar performer gain in three datasets,

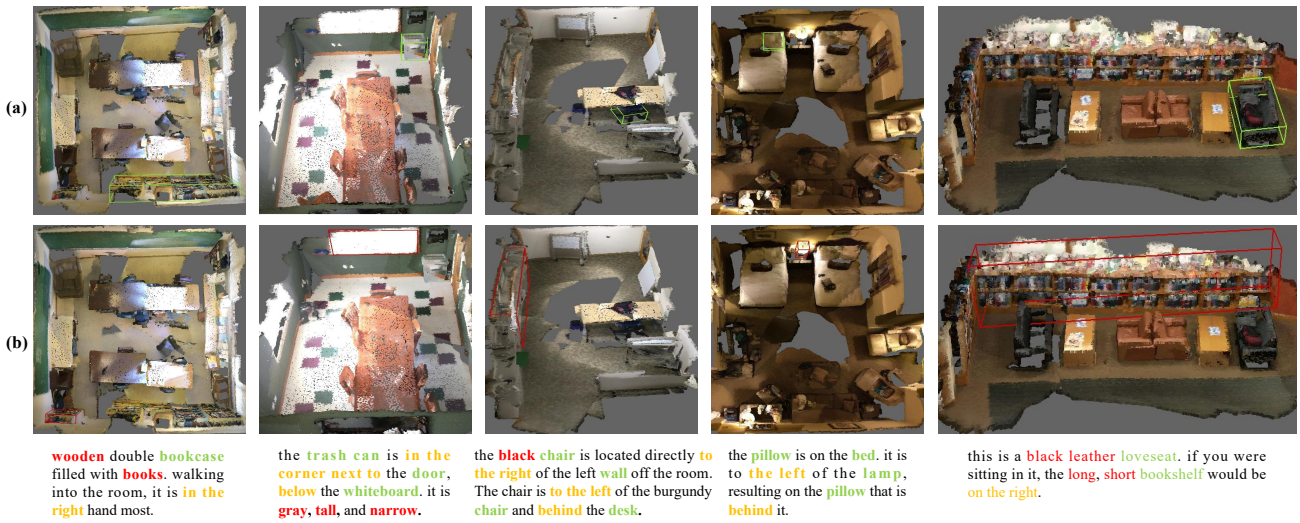

Figure 4. The visualization results. (a) and (b) present the correct/wrong prediction of MA^2TransVG with/without attribute interaction.

while random attention can obtain the best accuracy. We think it is because random attention helps try more diverse combinations of attribute interactives and can provide a more effective signal to supervise the attention.

Effects of Exchanging Operation. We conduct a sensitivity analysis of hyper-parameters in our EMFM module, including: **i)** the exchange proportion θ of tokens. As shown in Table. 5, the model performance increases and then drops with the increase of θ, which achieves a peak value of 65.2% when $\theta = 10\%$. The results show that too small or large θ will adversely affect the inter-modal fusion, which may bring inadequate alignment or attenuate the intra-modal knowledge. **ii)** the start layer μ and end layer η for multimodal exchanging. Since the default number of layers in the regular Transformer is 6, we first fix the start layer μ to 1 and investigate the value of η from $1, 2, 3, 4, 5$, where the accuracy improves quickly from 62.4% to 65.2%. We then fix the end layer η to 5, where the accuracy decreases when μ increases from 1 to 5. All the results verify the rationality of our hyper-parameters settings.

4.5. Visualization and Limitation

We show the visualization results in Fig. 4. The model may fail to recognize the spatial relationship among objects (the last three columns) or be significantly affected by other more salient objects (the first two columns) without considering the explicit multi-attribute interactions. Our MA^2TransVG can better perceive discriminative clues such as the appearance or spatial correlation of objects for multimodal feature enhancement and accurate grounding (the first row). We also notice the accuracy still remains improvement room for 'View-dependent' descriptions compared to the 'View-independent' or 'Easy', especially in Sr3D (In Tab. 2). In the future, we will extend more attribute embedding strategies, *e.g.*, better encoding absolute position or aggregating point clouds from various views.

Table 5. Hyper-parameter Sensitivity Analysis on Nr3D.

θ	0%	5%	10%	15%	20%
	64.4	64.7	65.2	63.9	61.4
η	1	2	3	4	5
$\mu = 1$	62.4	63.9	64.8	65.2	64.2
μ	1	2	3	4	5
$\eta = 5$	64.2	63.3	62.9	62.6	62.4

5. Conclusion

In this work, we propose an attribute-aware Transformer (MA^2TransVG) for 3D object grounding. As the first attempt to quantify the contribution of each attribute for final grounding, MA^2TransVG first designs a newly attribute causal effect analysis module and an attribute loss, which help model the multi-attribute interactions of objects for a better visual/textual modal understanding. Based on the learned multi-attribute interactions, the exchanging-based multimodal fusion module further dynamically aligns more discriminative details between modalities for a fine-grained multimodal feature. The proposed model outperforms the SOTA methods on Nr3D/Sr3D and ScanRefer datasets.

Acknowledgement

This work was supported by the National Science Fund of China (Grant Nos. 62276144, 62306238) and the Fundamental Research Funds for the Central Universities.

References

[1] 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. Springer, 2020. 2, 5, 6, 7

[2] EslamMohamed Bakr, Yasmeen Alsaedy, and Mohamed Elhoseiny. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. In *NeurIPS*, 2022. 7

[3] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *CVPR*, pages 16464–16473, 2022. 2, 3, 6

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 5

[5] 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. Springer, 2020. 2, 5, 6

[6] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *NeurIPS*, 34:5834–5847, 2021. 1

[7] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *NeurIPS*, 35:20522–20535, 2022. 2, 3, 4, 6, 7

[8] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *CVPR*, pages 16537–16547, 2022. 1

[9] 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

[10] Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *ICCV*, pages 18109–18119, 2023. 2, 6

[11] Bowen Cheng, Lu Sheng, Shaoshuai Shi, Ming Yang, and Dong Xu. Back-tracing representative points for voting-based 3d object detection in point clouds. In *CVPR*, pages 8963–8972, 2021. 1

[12] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 5

[13] 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 *CVPR*, pages 3722–3731, 2021. 2, 6, 7

[14] Chen Gao, Xingyu Peng, Mi Yan, He Wang, Lirong Yang, Haibing Ren, Hongsheng Li, and Si Liu. Adaptive zone-aware hierarchical planner for vision-language navigation. In *CVPR*, pages 14911–14920, 2023. 1

[15] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In *CVPR*, pages 14773–14783, 2023. 2

[16] 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. 2

[17] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *ACM MM*, pages 2344–2352, 2021. 2, 3, 5, 7

[18] 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. 2, 6, 7

[19] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *CVPR*, pages 15524–15533, 2022. 2, 3, 6, 7

[20] Kenton Lee Jacob Devlin, Ming-Wei Chang and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, page 2, 2019. 3, 5

[21] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Looking outside the box to ground language in 3d scenes. *arXiv preprint arXiv:2112.08879*, 2021. 2, 3

[22] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *ECCV*, pages 417–433. Springer, 2022. 2, 6, 7

[23] Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. *NeurIPS*, 31, 2018. 2

[24] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. In *ICLR*, 2018. 2

[25] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022. 5

[26] Haojia Lin, Yongdong Luo, Xiawu Zheng, Lijiang Li, Fei Chao, Taisong Jin, Donghao Luo, Chengjie Wang, Yan Wang, and Liujuan Cao. A unified framework for 3d point cloud visual grounding. *arXiv preprint arXiv:2308.11887*, 2023. 2, 6, 7

[27] Haolin Liu, Anran Lin, Xiaoguang Han, Lei Yang, Yizhou Yu, and Shuguang Cui. Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images. In *CVPR*, pages 6032–6041, 2021. 2

[28] Rui Liu, Xiaohan Wang, Wenguan Wang, and Yi Yang. Bird's-eye-view scene graph for vision-language navigation. In *ICCV*, pages 10968–10980, 2023. 1, 2

[29] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *ICCV*, pages 2949–2958, 2021. 1, 3, 5

[30] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *CVPR*, pages 6979–6987, 2017. 2

[31] 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. 2, 3, 6, 7

[32] Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014. 2

[33] Judea Pearl. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. 2022. 4

[34] Stephen Powell. The book of why: The new science of cause and effect. pearl, judea, and dana mackenzie. 2018. hachette uk. *Journal of MultiDisciplinary Evaluation*, 14(31):47–54, 2018. 4

[35] Mihir Prabhudesai, Hsiao-Yu Fish Tung, Syed Ashar Javed, Maximilian Sieb, Adam W Harley, and Katerina Fragkiadaki. Embodied language grounding with 3d visual feature representations. In *CVPR*, pages 2220–2229, 2020. 2

[36] 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, 5

[37] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, pages 9277–9286, 2019. 1

[38] Yongming Rao, Guangyi Chen, Jiwen Lu, and Jie Zhou. Counterfactual attention learning for fine-grained visual categorization and re-identification. In *ICCV*, pages 1025–1034, 2021. 2

[39] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, pages 1046–1056. PMLR, 2022. 7

[40] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, pages 3716–3725, 2020. 2, 4

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2022. 4, 5

[42] Hongjun Wang, Jiyuan Chen, Lun Du, Qiang Fu, Shi Han, and Xuan Song. Causal-based supervision of attention in graph neural network: A better and simpler choice towards powerful attention. *arXiv preprint arXiv:2305.13115*, 2023. 2

[43] Hanqing Wang, Wei Liang, Luc Van Gool, and Wenguan Wang. Dreamwalker: Mental planning for continuous vision-language navigation. In *ICCV*, pages 10873–10883, 2023. 1, 2

[44] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, pages 10760–10770, 2020. 2

[45] 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. 6, 7

[46] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *ICCV*, pages 1856–1866, 2021. 2, 3, 6, 7

[47] 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 *CVPR*, pages 1791–1800, 2021. 2, 6, 7

[48] 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

[49] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *ICCV*, pages 15225–15236, 2023. 6, 7

[50] 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. 2, 6, 7

[51] Tong Zhao, Gang Liu, Daheng Wang, Wenhao Yu, and Meng Jiang. Learning from counterfactual links for link prediction. In *Int. Conf. on Machine Learning*, pages 26911–26926. PMLR, 2022. 2

[52] Dave Zhenyu Chen, Qirui Wu, Matthias Niener, and Angel X Chang. D3net: A speaker-listener architecture for semi-supervised dense captionin and visual grounding in rgb-d scans. In *ECCV*, 2022. 2

[53] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *ICCV*, pages 2911–2921, 2023. 2