Language Conditioned Spatial Relation Reasoning for 3D Object Grounding

Anonymous Author(s)
Affiliation
Address
email

Abstract

Localizing objects in 3D scenes based on natural language requires understanding and reasoning about spatial relations. In particular, it is often crucial to distinguish similar objects referred by the text, such as "the left most chair" and "a chair next to the window". In this work we propose a language-conditioned transformer model for grounding 3D objects and their spatial relations. To this end, we design a spatial self-attention layer that accounts for relative distances and orientations between objects in input 3D point clouds. Training such a layer with visual and language inputs enables to disambiguate spatial relations and to localize objects referred by the text. To facilitate the cross-modal learning of relations, we further propose a teacher-student approach where the teacher model is first trained using ground-truth object labels, and then helps to train a student model using point cloud inputs. We perform ablation studies showing advantages of our approach. We also demonstrate our model to significantly outperform the state of the art on the challenging Nr3D, Sr3D and ScanRefer 3D object grounding datasets. Our code and pretrained models will become publicly available.

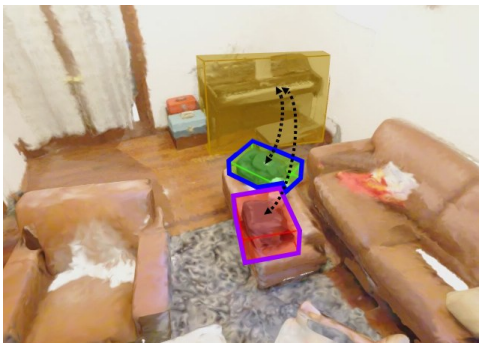
1 Introduction

In order to enable robots to carry out human instructions in the real world, it is important for them to understand natural language and to ground the mentioned objects in 3D environments. Therefore, research attention is shifting from object grounding in 2D images [1, 2, 3, 4, 5, 6] to the 3D object grounding task [7, 8], which localizes an object referred by a sentence in a 3D point cloud.

Language expressions often refer to objects by their relative spatial locations in 3D scenes. Figure 1 illustrates example scenes and corresponding sentences where object grounding requires disambiguation between objects of the same class. For instance, *"the backpack closest to the piano"* requires to compare relative distances among objects, while *"choose the door on the left when facing them"* requires to infer the correct viewpoint and understand relative directions. The complexity and diversity of spatial language make it extremely challenging for spatial reasoning in 3D scenes.

Given the critical role of the spatial language, many existing methods attempt to model spatial relations for 3D object grounding. Early work [9, 10, 11, 12] explicitly build 3D visual graphs based on distances between objects and apply graph neural networks to learn relationships. However, since only nearest neighbors are considered in the graph, it remains difficult to infer relationships between distant objects such as *"farthest"*. More recently, transformer architectures [13, 14, 15, 9, 16] have been used, as they have the potential to learn relations between pairs of objects with a multi-head self-attention mechanism [17]. Enabling transformers to better understand 3D spatial relations expressed by natural language, however, remains an open research problem.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.



(a) The backpack closest to the piano.

(b) Of the two brown wooden doors, choose the door on the left when facing them.

Figure 1: Example sentences that refer to objects in 3D scenes. The green box denotes the ground-truth object, the blue box is the prediction from our model, and the purple one is from a baseline model without explicit spatial reasoning and knowledge distillation.

In addition, 3D object grounding suffers more from the scarcity of training data compared to its 2D counterpart. The success of 2D object grounding models [18, 19] can be largely attributed to large-scale datasets with image-text pairs [1, 2, 20, 21], whereas the limited amount of 3D scene-language pairs increases the difficulty of 3D object reasoning. To address this challenge, Yang *et al*. [14] propose to use 2D-3D alignments to assist the training of 3D models, but require additional high-quality 2D images and camera parameters which are not always available.

In this work, we propose a **Vi**sion-and-**L**anguage **3D Rel**ation reasoning model (ViL3DRel) to tackle the above issues in 3D object grounding. We design a new spatial self-attention layer for the transformer architecture to enhance 3D spatial understanding. This layer encodes relative distances and orientations for all pairs of objects, and explicitly learns spatial attention to capture spatial relations referred by the the language expressions. We propose a sigmoid softmax function to effectively fuse the spatial attention with standard self-attention. Rotation augmentation is used to further improve view invariant reasoning. To alleviate the negative effect of noisy object features, we further propose a teacher-student training approach. The teacher and student share the same model architecture, but use different inputs. The teacher uses ground-truth object labels, which enables to better learn the relation reasoning. The learned relation knowledge is distilled to the student taking point cloud features as input.

To summarize, our contributions are three-fold:

- We propose a ViL3DRel model for the 3D object grounding task. It uses a new spatial self-attention to explicitly encode pairwise 3D spatial relations in the transformer layer for better language conditioned spatial understanding.
- We employ a teacher-student training strategy, which facilitates the cross-modal learning of relations. The student with point cloud inputs benefits from the relation reasoning model of the teacher trained with ground-truth object label input.
- We evaluate our ViL3DRel approach on the challenging Nr3D, Sr3D [9] and ScanRefer [7] benchmarks. Our model significantly outperforms state-of-the-art approaches, with 9.3, 8.3 absolute gains on Nr3D and Sr3D given ground-truth object proposals, and 4.47 points on ScanRefer using the same detected object proposals compared to the previous work [16] .

2 Related Work

2D and 3D Object Grounding. The object grounding task aims to localize objects in 2D images or 3D point clouds given a sentence. Existing approaches can be categorized into two groups, namely one- and two-stage frameworks. The one-stage methods densely fuse the text features with patch- or point-level visual representations to directly regress the bounding box [18, 22]. These approaches are flexible to detect various objects given the input sentence. The two-stage methods adopt a *detection-then-matching* pipeline [1, 23, 8, 7], where the detection stage generates object proposals, and then the matching stage selects the best proposal according to the sentence. The decoupled object perception and cross-modal matching make the two-stage methods easier to analyze [15].

The benchmark datasets for 3D object grounding are ReferIt3D (Nr3D and Sr3D) [8] and ScanRefer [7], which are all built upon scenes and object annotations in the ScanNet dataset [24]. Graph-based approaches [25, 26] are widely adopted in early works to infer spatial relations. The object graph is constructed by connecting each object with its top nearest neighbors [8, 10, 11] based on Euclidean distance. Inspired by the success of transformers [17], recent works [14, 9, 15, 16, 22] have adopted transformers for 3D object grounding. BEAUTY-DETR [27] and 3D-SPS [22] are one-stage methods. However, most works follow the two-stage framework with pre-detected object proposals. LanguageRefer [15] converts the cross-modal task into a language modeling problem with predicted object labels. SAT [14] adopts a multimodal transformer and transfers 2D semantics to assist the training of the 3D model. Multi-view transformer [16] aggregates object representations from multiple views to improve view robustness. Perhaps most similar to our work, 3DVG-Transformer [13] proposes a coordinate-guided attention module to encode spatial distances among objects. In contrast, we introduce a spatial self-attention module, which explicitly encodes both relative distances and relative orientations among objects. It also conditions spatial relations on the language and presents a more effective strategy for attention fusion.

Transformers in Vision-and-Language. Transformer-based architectures have led to significant improvements in various vision-and-language tasks such as text-video retrieval [28], visual grounding [18, 29], image captioning [30], vision question-answering [31] and vision-and-language navigation [32, 33]. Most of the methods project textual and visual inputs into a sequence of tokens, and use multimodal transformers to learn cross-modal semantic relationships. While relative positional encoding (RPE) [34] has been explored mostly separately in vision and language transformers, we here integrate RPE with a cross-modal transformer and show its benefits for resolving spatial object relations referred by text.

Knowledge Distillation. Knowledge distillation [35] is typically used to compress a large network (teacher) into a compact model (student). The common objective is to let the student model mimic the soft logits of the teacher model. Beyer *et al.* [36] show that a consistent and patient teacher is essential in knowledge distillation. Jiao *et al.* [37] further demonstrate that the intermediate representations learned by the teacher are beneficial. In contrast to distilling knowledge from a heavy model to a light one, our teacher network first learns cross-modal object relations using ground-truth object labels. We then transfer this knowledge to the target student network that uses noisy inputs.

3 Method

Given a sentence S, the goal of 3D object grounding is to detect an object referred in S by locating its 3D bounding box $B_T \in \mathbb{R}^6$ in a 3D point clouds P_{scene}. We assume $P_{scene} \in \mathbb{R}^{K\times 6}$ contains K points each represented by XYZ coordinates and RGB values. We follow the *detection-then-matching* framework [7, 8, 13, 14, 16] for 3D object grounding and assume to be given a list of object proposals $(O_1, \cdots, O_N)$ obtained via automatic 3D instance segmentation or ground-truth annotations (depending on the evaluation setup). Each object O_i is represented by a subset of K_i points $P_i \subset P_{scene}, P_i \in \mathbb{R}^{K_i\times 6}$. Our work is focused on interpreting spatial relations and selecting the target object $O_T, T \in [1, N]$ among N object proposals. We first present an overview of our ViL3DRel model in Section 3.1. We then introduce our language-conditioned spatial self-attention module and the teacher-student training approach in Sections 3.2 and 3.3 respectively.

3.1 Architecture Overview

Figure 2 shows an overview of our ViL3DRel model, consisting of four modules: text encoding, object encoding, multimodal fusion and a grounding head as described next.

Text encoding. Given the sentence S with M word tokens, we use a pre-trained BERT model [38] to encode S into a sequence of word features $(s_{cls}, s_1, \cdots, s_M), s_i \in \mathbb{R}^d$, where s_{cls} is a special classification token and d is the dimensionality of the feature.

Object encoding. For each object O_i, we first normalize the coordinates of its point cloud P_i into a unit ball, and then use PointNet++ [39] to compute the object feature $o_i^0 \in \mathbb{R}^d$. We also obtain the object center $c_i = [c_x, c_y, c_z] \in \mathbb{R}^3$ and the object size $z_i = [z_x, z_y, z_z] \in \mathbb{R}^3$ from object points P_i as the mean and the spatial extent of P_i respectively. We use a linear projection layer to obtain the absolute 3D location feature as $l_i = W_l[c_i; z_i] \in \mathbb{R}^d$.

Multimodal fusion. A stack of transformer layers [17] are applied to fuse the text and object modality features. Each transformer layer is composed of a spatial self-attention layer, a cross-attention layer and a feed-forward neural network (FFN). Our new spatial self-attention layer aims to improve the understanding of spatial relations among objects referred by the sentence. Assume o_i^l is the input

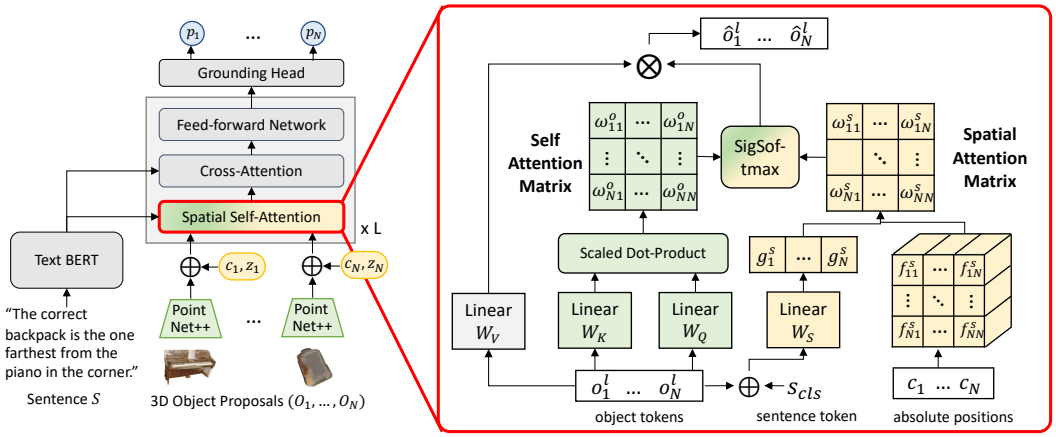


Figure 2: Left: overview of the ViL3DRel model; Right: language-conditioned spatial self-attention.

embedding for object proposal O_i before the i-th layer, we first add it with its absolute 3D location feature l_i to enhance the spatial information. The spatial self-attention then generates contextualized object representations $\hat{o}_i^l$. We will describe the proposed spatial self-attention in details in Section 3.2. The cross-attention layer takes $\hat{o}_i^l$ as queries and text features $(s_{cls}, s_1, \cdots, s_M)$ as keys and values to learn cross-modal relations. The final FFN uses two fully connected layers to encode each output tokens after the attention layer. For detailed explanation of Transformer attention and FFN layers please refer to [17].

Object grounding. We use a two-layer feed-forward neural network as the object grounding head to predict the target object. Given the output embedding o_i^L from the last multimodal fusion layer, the grounding head generates a scalar score for each object proposal O_i and applies softmax function to obtain the probability p_i. The object proposal with the maximum probability is predicted as target.

3.2 Spatial Self-Attention

The standard self-attention relies on token embeddings to learn relations among tokens. However, the input embeddings of object proposals are mixed with semantic features and absolute 3D locations, making it difficult to accurately infer spatial relations among objects such as relative distances or orientations in Figure 1. As spatial relations among objects referred by language are important to distinguish object instances, we propose to inject a language conditioned spatial attention that explicitly captures pairwise spatial relations to complement the standard self-attention. We first describe the standard self-attention and then introduce our proposed spatial self-attention as illustrated in the right part of Figure 2.

Given the feature matrix $X \in \mathbb{R}^{N \times d}$ for N object proposals, the self-attention mechanism first computes the query, key and value embeddings from X as $Q = XW_Q, K = XW_K, V = XW_V$ respectively, where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d_h}$ are learnable parameters and d_h is the dimensionality of the output embedding. It then calculates attention weights given the query and key embeddings and aggregates the value embeddings as follows:

$$\Omega^o = \text{softmax}\left(\frac{QK^T}{\sqrt{d_h}}\right); \quad \text{SelfAttn}(Q, K, V) = \Omega^o V, \tag{1}$$

where Ω^o is an $N \times N$ attention matrix, whose elements ω_{ij}^o is the attention weight between the i-th and j-th object proposal. In order to capture more diverse relations, multi-head self-attention is used where each head computes an independent $\text{SelfAttn}(Q, K, V)$ and the outputs from all heads are concatenated.

To model spatial relations among objects, we propose to use explicit pairwise spatial features $f_{ij}^s \in \mathbb{R}^5, i, j \in [1, N]$. For each pair of objects (O_i, O_j), we compute their Euclidean distance $d_{ij} = ||c_i - c_j||_2$ as well as horizontal and vertical angles θ_h, θ_v of the line connecting object centers c_i and c_j. We then define the pairwise spatial feature f_{ij}^s as:

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

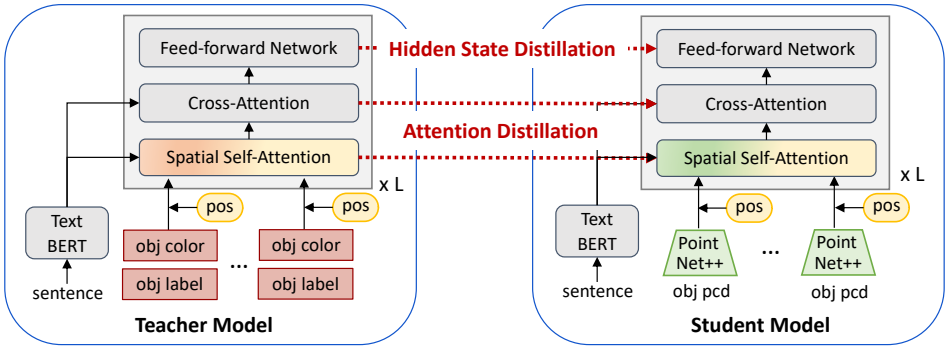


Figure 3: Teacher student learning with hidden state and attention distillation.

We automatically generate a language conditioned weight g_i^s to select relevant spatial relations for each object proposal O_i:

$$g_i^s = W_S^T(s_{cls} + o_i^l), \tag{3}$$

where $W_S \in \mathbb{R}^{d\times 5}$ is a learnable parameter and we omit the bias term for simplicity. We then define the spatial relevance for (O_i, O_j) as:

$$\omega_{ij}^s = g_i^s \cdot f_{ij}^s. \tag{4}$$

We modulate the self-attention in Eq (1) with the above language conditioned spatial relevancy using the sigmoid softmax (sigsoftmax) fusion function:

$$\omega_{ij} = \frac{\sigma(\omega_{ij}^s)\,\exp(w_{ij}^o)}{\sum_{l=1}^{N}\sigma(\omega_{il}^s)\,\exp(w_{il}^o)}, \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function. In this way, we compute the new self-attention matrix $\Omega = [\omega_{ij}]_{N\times N}$ which explicitly considers 3D relative spatial locations, absolute spatial locations and object appearances in relation reasoning. We also use the multi-head mechanism to support different types of spatial relations and compute multiple spatial attentions to fuse with the standard self-attention. We concatenate the outputs from all heads in the spatial self-attention layer.

3.3 Teacher-Student Training

Incorrect estimation of object classes from P_i can deteriorate the cross-modal alignment and spatial relation reasoning. To improve object grounding in point clouds previous work [14] relies on additional supervision in 2D images. Here, instead, we propose a new teacher-student training approach using no additional training data. The teacher and student models share the same transformer architectures except for the object encoding module. The teacher uses ground-truth object semantic features, while the student uses 3D point clouds. Since there is less cross-modal gap between the teacher's inputs and the sentence, the teacher model can excel at learning language relevant spatial relations among objects and aligning objects with the sentence. Such knowledge can then be distilled to the student model to improve its training. Figure 3 illustrates the teacher-student training pipeline. We describe the inputs of the teacher and the knowledge distillation objectives in the following.

Teacher's inputs. We use ground-truth class labels and dominant colors of the object as the object representations for the teacher model, since the color is the most widely used attributes in the sentence. Specifically, we encode the ground-truth object class label using pre-trained Glove word vectors [40]. To obtain the dominant colors of the object, we fit a Gaussian Mixture model on the RGB values of all points in the object, where the mixture component is set to 3. We linearly project the mean value in each component into a color embedding and use the mixture weights of the components to average all the color embeddings. The final object representation is the sum of the class label embedding and the averaged color embedding.

Knowledge distillation objectives. We transfer the attention weights and hidden states in the multimodal fusion module of the teacher model to the student model. The self-attentions learned by the teacher capture spatial relations among objects, and the cross-attentions measure cross-modal matching, which are both essential for 3D object grounding but hard to learn given noisy object features. We facilitate student learning by forcing it to mimic all the attention matrices Ω of the

teacher using the following objective function:

$$\mathcal{L}_{attn} = \frac{1}{LH} \sum_{l=1}^{L} \sum_{h=1}^{H} \text{MSE}(\Omega_{lh}^{S} - \Omega_{lh}^{T}), \tag{6}$$

where $\Omega_{\cdot}^{S}$ and $\Omega_{\cdot}^{T}$ are attention weights in student and teacher model respectively, $\text{MSE}(\cdot)$ is the mean square error function, L is the number of multimodal fusion layers and H is the number of heads in self-attention. In addition to distilling the relation knowledge in attention weights, we also distill the output embeddings of the transformer layers:

$$\mathcal{L}_{hidden} = \frac{1}{LN} \sum_{l=0}^{L} \sum_{i=1}^{N} \text{MSE}(o_i^{lS} - o_i^{lT}). \tag{7}$$

Training. We follow the previous works [8, 7] to use multiple auxiliary losses in training the transformer model, including a 3D object grounding loss $\mathcal{L}_{og}$, sentence classification loss $\mathcal{L}_{sent}$ and two object classification losses $\mathcal{L}_{obj}^{u}$ and $\mathcal{L}_{obj}^{m}$. $\mathcal{L}_{sent}$ relies on s_{cls} to predict the target object class from the sentence. $\mathcal{L}_{obj}^{u}$ and $\mathcal{L}_{obj}^{m}$ use unimodal object representation o_i^0 and multimodal fused object representation o_i^L respectively to predict the classes for input object proposals. Therefore, the overall training objective is as follows:

$$\mathcal{L} = \mathcal{L}_{og} + \mathcal{L}_{sent} + \mathcal{L}_{obj}^{u} + \mathcal{L}_{obj}^{m} + \lambda_a \mathcal{L}_{attn} + \lambda_h \mathcal{L}_{hidden} \tag{8}$$

where λ_a, λ_h are two hyper-parameters to balance the losses.

4 Experiments

4.1 Datasets

Nr3D dataset [8]. It contains 37,842 human written sentences that refer to annotated objects in the 3D indoor scene dataset ScanNet [24]. The dataset includes 641 scenes with 511 (resp. 130) scenes for training (resp. validation). It covers 76 target object classes. The annotated sentences are designed to refer to objects with multiple same-class distractors in the scene. The sentences are split into "easy" and "hard" subsets in evaluation, where the target object in "easy" subset only contains one same-class distractor in the scene while it contains multiple ones in the "hard" subset. According to whether the sentence requires a specific viewpoint to ground the referred object, the dataset can also be partitioned into "view depedent" and "view independent" subsets.

Sr3D dataset [8]. It uses templates to automatically generate sentences. The sentences only utilize spatial relations to distinguish objects of the same class. It has 1,018 training scenes and 255 validation scenes from ScanNet and 83,570 sentences in total. The dataset can be split in the same way as Nr3D during evaluation.

ScanRefer dataset [7]. It has 51,583 human written sentences for 800 scenes in ScanNet. We follow the official split and use 36,665 and 9,508 samples for training and validation respectively. According to whether the target object is a unique object class in the scene, the dataset can be divided into a "unique" and a "multiple" subset.

4.2 Experimental Setting

Evaluation Metrics. We evaluate models under two evaluation settings. One uses ground-truth object proposals, which is the default setting in the Nr3D and Sr3D datasets. The metric is the accuracy of selecting the target bounding box among the proposals. The other setting does not provide ground-truth object proposals and requires the model to regress a 3D bounding box, which is the default setting for the ScanRefer dataset. The evaluation metrics are acc@0.25 and acc@0.5, which is the percentage of correctly predicted bounding boxes whose IoU is larger than 0.25 or 0.5 with the ground-truth.

Implementation details. For the model architecture, we set the dimension $d = 768$ and use 12 heads for all the transformer layers. The text encoding module is a three-layer transformer initialized from BERT [38], and the multimodal fusion module contains four layers. The object encoding module PointNet++ [39] samples 1024 points for all the objects. These architecture parameters are the same as in previous work [14, 16] to ensure a fair comparison. We first train the PointNet++ for object classification on ScanNet, which achieves 61.9% accuracy on the validation set. Its weights remain fixed during the following training steps. Rotation augmentation is used to increase the viewpoint

Table 1: Grounding accuracy (%) on the Nr3D dataset with ground-truth object labels. Dist stands for Distance; Ort for Orientation; MHA for multi-head spatial attention; RotAug for Rotation Augmentation; sigs for sigsoftmax in Eq (5); and '-' means not applicable.

	Spatial Relation Reasoning				Rot	Color	Overall	ViewDep	ViewIndep
	Dist	Ort	MHA	Fusion	Aug				
R1	-	-	-	-	×	×	53.5	51.4	54.6
R2	-	-	-	-	×	✓	55.1	53.8	55.8
R3	-	-	-	-	✓	✓	62.4	58.3	64.5
R4	✓	×	✓	sigs	✓	✓	66.0	53.8	72.0
R5	×	✓	✓	sigs	✓	✓	71.3	68.5	72.6
R6	✓	✓	×	sigs	✓	✓	67.7	65.2	69.0
R7	✓	✓	✓	bias	✓	✓	55.4	46.8	59.6
R8	✓	✓	✓	ctx	✓	✓	56.4	50.8	59.1
R9	✓	✓	✓	sigs	✓	✓	**74.4**	**71.3**	**75.9**

invariance. The hyper-parameters in the loss function Eq (8) are set to $\lambda_a = 1$ and $\lambda_h = 0.02$. We train the model with a batch size of 128 and a learning rate of 0.0005 with warm-up of 5000 iterations and cosine decay scheduling. The AdamW algorithm [41] is used in the optimization. We train for 50 epochs for the teacher model and 100 epochs for the student model on Nr3D and ScanRefer datasets. The training epochs are reduced to half on the Sr3D dataset, as it is larger and easier to converge. All models are trained on a single NVIDIA RTX A6000 GPU.

4.3 Ablation Studies

We carry out extensive experiments on the Nr3D dataset to demonstrate the effectiveness of our proposed spatial self-attention and teacher-student training.

4.3.1 Spatial Relation Reasoning

We evaluate the proposed spatial self-attention with ground-truth object labels, which decouples the object perception from spatial relation reasoning. Table 1 presents the results on the Nr3D dataset.

Baselines. Our baseline R1 is the teacher model in the left part of Figure 3 without the object color, the proposed spatial self-attention and the rotation augmentation. Row R2 shows that the additional color information helps the object grounding, since the sentences often describe the color attributes. The rotation augmentation brings additional gains; it improves from 55.1% (R2) to 62.4% (R3). It is more effective for view independent sentences (+8.7%) than for view dependent sentences (+4.5%) because rotation augmentation mainly improves view invariance. The performance of R3 serves as a strong baseline. We will show the contributions of our proposed model on top of it.

Pairwise spatial features. In R4-R5, we compare different pairwise spatial features of Eq (2). R4 only uses the pairwise distances, while R5 only uses the pairwise orientations. We can see that the pairwise distances benefit the view independent sentences which contain distance-related spatial relations such as *"next to"* and *"farthest from"*, but do not improve the performance for view dependent sentences. In contrast, pairwise orientations significantly boost the performance for view dependent sentences (+10.2%). This shows that explicitly encoding the relative orientations facilitates the learning of view dependent spatial relations such as *"in front of"* and *"to the left of"*. The relative distance and orientation features are also complementary. Their combination achieves the best performance as shown in our full model (R9).

Multi-head spatial attention. The model in R6 uses single-head attention for the spatial relevance ω_{ij}^s in Eq (5) though keeps multi-head for the standard attention weight ω_{ij}^o. Compared to the full model R9 with multi-head spatial attention, such single-head spatial attention achieves significantly worse performance. This suggests that the multi-head is necessary to learn spatial relations.

Attention fusion. We utilize a sigmoid softmax function in Eq (5) to aggregate the spatial attention and the standard self-attention weights. We compare the proposed fusion mechanism with standard relative positional encoding methods [34] in vision transformers: bias mode and contextual mode. The bias mode adds the spatial attention weight as a bias term in the standard self-attention, while the contextual mode further considers the interaction with queries by injecting the pairwise spatial features into the key embeddings. From the results in R7 and R8, we can see that the standard relative

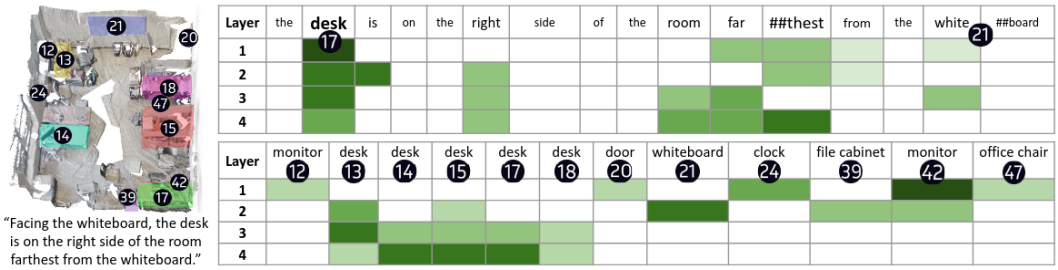


Figure 4: An example of the learned cross-attention (upper) and self-attention (bottom) at different transformer layers in the teacher model. The query is the target object proposal (desk 17). The darker color denotes higher attention weights from the query to the keys.

positional encoding methods fail to model 3D spatial relations among objects; they even perform worse than the strong baseline (R3), which doesn't rely on relative positional encoding. The proposed sigmoid softmax function is by far more effective to modulate the standard self-attention with spatial relations.

4.3.2 Teacher-Student Training

The teacher model with ground-truth object labels learns well how to perform spatial relation reasoning given the input sentence. Figure 4 provides a qualitative example of the learned cross- and self-attention weights at all multimodal fusion layers in the teacher model. For the cross-attention weights, we can see that the teacher first aligns the object proposal with the correct words in the text, and then gradually shifts its attentions to relation words to perform spatial relation reasoning. For the self-attention weights, in the first layer, the teacher mostly attends to nearby objects to be aware of the context; then in the second layer, it focuses more on the reference object mentioned in the sentence (*e.g.* , the whiteboard); finally, it concentrates on the same-class distractors to distinguish them. Such reasoning steps are promising to ease the training of the student model. In the following, we evaluate the proposed teacher-student training to distill relational knowledge from the teacher to a student with point cloud inputs extracted from the ground-truth object proposals.

Table 2: Grounding accuracy (%) on the Nr3D dataset with ground-truth object proposals.

	init.	$\mathcal{L}_{attn}$	$\mathcal{L}_{hidden}$	Overall
Teacher		-		74.4
Student	×	×	×	58.1
	✓	×	×	62.6
	×	✓	×	63.6
	×	×	✓	62.1
	×	✓	✓	**64.4**

Table 2 presents the performance of different student models. The first row in the student block does not use any knowledge from the teacher model. In the second row, we use the weights in the

Table 3: Grounding accuracy (%) on Nr3D and Sr3D datasets with ground-truth object proposals.

Method	Nr3D					Sr3D				
	Overall	Easy	Hard	View Dep	View Indep	Overall	Easy	Hard	View Dep	View Indep
ReferIt3D [8]	35.6	43.6	27.9	32.5	37.1	40.8	44.7	31.5	39.2	40.8
ScanRefer [7]	34.2	41.0	23.5	29.9	35.4	-	-	-	-	-
TGNN [10]	37.3	44.2	30.6	35.8	38.0	-	-	-	-	-
InstanceRefer [11]	38.8	46.0	31.8	34.5	41.9	48.0	51.1	40.5	45.4	48.1
FFL-3DOG [12]	41.7	48.2	35.0	37.1	44.7	-	-	-	-	-
3DVG-Trans [13]	40.8	48.5	34.8	34.8	43.7	51.4	54.2	44.9	44.6	51.7
TransRefer3D [9]	42.1	48.5	36.0	36.5	44.9	57.4	60.5	50.2	49.9	57.7
LanguageRefer [15]	43.9	51.0	36.6	41.7	45.0	56.0	58.9	49.3	49.2	56.3
SAT [14]	49.2	56.3	42.4	46.9	50.4	57.9	61.2	50.0	49.2	58.3
3D-SPS [22]	51.5	58.1	45.1	48.0	53.2	62.6	56.2	65.4	49.2	63.2
Multi-view [16]	55.1	61.3	49.1	54.3	55.4	64.5	66.9	58.8	58.4	64.7
ViL3DRel (Ours)	**64.4**	**70.2**	**57.4**	**62.0**	**64.5**	**72.8**	**74.9**	**67.9**	**63.8**	**73.2**

Table 5: Grounding accuracy (%) on ScanRefer with detected object proposals. VN and PG denote the VoteNet and PointGroup models pretrained on 18 classes of the ScanNet dataset, while Optimized denotes end-to-end training of the object detector on the ScanRefer dataset.

Method	Det	Unique		Multiple		Overall	
		acc@0.25	acc@0.5	acc@0.25	acc@0.5	acc@0.25	acc@0.5
ScanRefer [7]	Optimized	65.00	43.31	30.63	19.75	37.30	24.32
TGNN [10]		68.61	56.80	29.84	23.18	37.37	29.70
3DVG-Trans [13]		77.16	58.47	38.38	28.70	45.90	34.47
3D-SPS [22]		81.63	64.77	39.48	29.61	47.65	36.42
Non-SAT [14]	VN [43]	68.48	47.38	31.81	21.34	38.92	26.40
SAT [14]	VN [43]	73.21	50.83	37.64	25.16	44.54	30.14
InstanceRefer [11]	PG [42]	77.45	66.83	31.27	24.77	40.23	32.93
Multi-view [16]	PG [42]	77.67	66.45	31.92	25.26	40.80	33.26
ViL3DRel (Ours)	PG [42]	**81.58**	**68.62**	**40.30**	**30.71**	**47.94**	**37.73**
UpperBound	PG [42]	88.63	74.47	78.82	60.37	80.64	62.98

teacher model to initialize the student model since the teacher and student share the same architecture except for the input object representations. Such simple weight initialization already facilitates the training of the student, and achieves 4.5% gains. The attention distillation in Eq (6) and hidden state distillation in Eq (7) are both beneficial to improve the student model compared to the first row without knowledge distillation. It is more effective to distill the relation knowledge in the attention weight matrices. The combination of $\mathcal{L}_{attn}$ and $\mathcal{L}_{hidden}$ is helpful and achieves the best performance. We empirically find that combining weight initialization and the two distillation losses performs similarly as using the distillation losses alone.

4.4 Comparison with state-of-the-art Methods

In Table 3, we compare our ViL3DRel model with state-of-the-art methods on Nr3D and Sr3D datasets. All the compared works use ground-truth object proposals, but no ground-truth labels. Our model achieves significant improvements over the previous best method [16], with 9.3% and 8.3% absolute gains on Nr3D and Sr3D respectively. We outperform [13] by even larger margins on the two datasets. This work also use spatial information, but doesn't use a spatial attention module. Table 4 and Table 5 present results on ScanRefer dataset with ground-truth object proposals and detected object proposals, respectively. For fair comparison, we utilize a pre-trained PointGroup [42] model to generate object proposals, which is trained on 18 classes in ScanNet. To be noted, the upper block in Table 5 optimizes the detection stage or utilizes single-stage pipeline. These methods use more data to train the object proposals, so the comparison to these methods is not entirely fair. There is a large gap between the ground-truth and detected proposals (over 12% for our method), because the detected proposals might not include the target or reference objects in the sentence. However, our ViL3DRel model still outperforms the state of the art in both settings. Figure 1 compares our ViL3DRel model and a baseline without spatial self-attention and teacher-student training. We present more qualitative comparisons in the supplementary material.

Table 4: Grounding accuracy (%) on ScanRefer with ground-truth object proposals.

	Overall
ReferIt3D [8]	46.9
Non-SAT [14]	48.2
SAT [14]	53.8
ViL3DRel (Ours)	**59.8**

5 Conclusion

In this work, we propose a ViL3DRel model for 3D object grounding. It contains a newly designed spatial self-attention module to improve language conditioned spatial relation reasoning in the transformer layer, which explicitly considers relative distances and orientations among objects. A teacher-student training approach is further proposed to transfer relation knowledge from a teacher with ground-truth object labels to a student with point cloud inputs. The proposed model significantly outperforms the state of the art on Nr3D, Sr3D and ScanRefer datasets. Since our model belongs to the two-stage framework, it is limited by imperfect object proposals in the first detection stage. This paper has minimal ethical, privacy and safety concerns.

References

[1] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *ECCV*, pages 69–85. Springer, 2016.

[2] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016.

[3] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *CVPR*, pages 4185–4194, 2019.

[4] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *CVPR*, pages 10086–10095, 2020.

[5] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *CVPR*, pages 12538–12547, 2019.

[6] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *CVPR*, pages 9982–9991, 2020.

[7] 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.

[8] 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.

[9] 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.

[10] 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*, volume 35, pages 1610–1618, 2021.

[11] 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.

[12] 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.

[13] 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.

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

[15] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *CoRL*, pages 1046–1056. PMLR, 2021.

[16] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *CVPR*, 2022.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[18] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *ICCV*, pages 1780–1790, 2021.

[19] Sanjay Subramanian, Will Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *ACL*, 2022.

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

[21] 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. PMLR, 2021.

[22] 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*, 2022.

[23] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[24] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[25] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *ICLR*, 2018.

[26] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *TOG*, 38(5):1–12, 2019.

[27] 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.

[28] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473, 2019.

[29] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. *CVPR*, 2022.

[30] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, pages 10578–10587, 2020.

[31] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, volume 34, pages 13041–13049, 2020.

[32] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *NeurIPS*, 34, 2021.

[33] 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. *CVPR*, 2022.

[34] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *ICCV*, pages 10033–10041, 2021.

[35] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[36] Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. *arXiv preprint arXiv:2106.05237*, 2021.

[37] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. In *EMNLP Findings*, pages 4163–4174, 2020.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, 2019.

[39] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 30, 2017.

[40] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[41] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

[42] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *CVPR*, pages 4867–4876, 2020.

[43] 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the conclusion section.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the conclusion section.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] We will release the code, data and instructions upon acceptance.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See the experiment section.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See the supplementary material.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the experimental section.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]