ViewRefer: Grasp the Multi-view Knowledge for 3D Visual Grounding with GPT and Prototype Guidance

Ziyu Guo[1,2*] Yiwen Tang[1*] Renrui Zhang[1,2*]

Dong Wang[1] Zhigang Wang[1] Bin Zhao[1✉] Xuelong Li[1]

[1] Shanghai Artificial Intelligence Laboratory [2] The Chinese University of Hong Kong

{guoziyu, tangyiwen, wangdong, zhaobin}@pjlab.org.cn

Abstract

Understanding 3D scenes from multi-view inputs has been proven to alleviate the view discrepancy issue in 3D visual grounding. However, existing methods normally neglect the view cues embedded in the text modality and fail to weigh the relative importance of different views. In this paper, we propose ***ViewRefer****, a multi-view framework for 3D visual grounding exploring how to grasp the view knowledge from both text and 3D modalities. For the text branch, ViewRefer leverages the diverse linguistic knowledge of large-scale language models, e.g., GPT, to expand a single grounding text to multiple geometry-consistent descriptions. Meanwhile, in the 3D modality, a transformer fusion module with inter-view attention is introduced to boost the interaction of objects across views. On top of that, we further present a set of learnable multi-view prototypes, which memorize scene-agnostic knowledge for different views, and enhance the framework from two perspectives: a view-guided attention module for more robust text features, and a view-guided scoring strategy during the final prediction. With our designed paradigm, ViewRefer achieves superior performance on three benchmarks and surpasses the second-best by* ***+2.8%****,* ***+1.2%****, and* ***+0.73%*** *on Sr3D, Nr3D, and ScanRefer. Code will be released at https://github.com/ZiyuGuo99/ViewRefer3D.*

1. Introduction

The aim of visual grounding is to ascertain the precise location of an object from an image or a 3D scene according to given query texts. The field of 2D visual grounding [22, 42, 26] has undergone significant advances recently. Meanwhile, the developments in embodied agent [34, 9], vision-language navigation [60, 38], and autonomous driving [14] have also stimulated increasing attention on 3D visual grounding [1, 6].

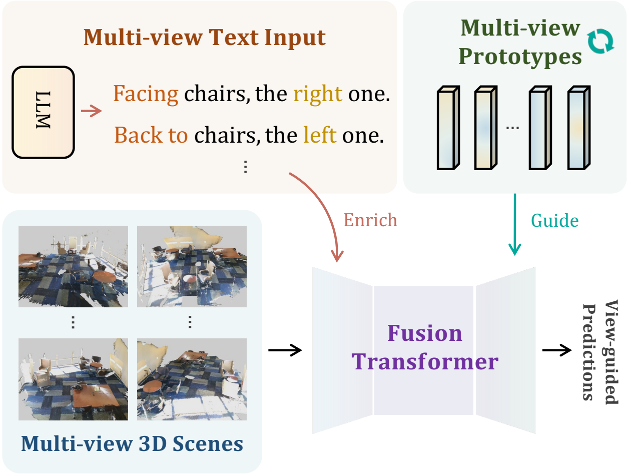


Figure 1: **The Paradigm of ViewRefer.** With our proposed LLM-expanded texts and multi-view prototypes, ViewRefer adopts a fusion transformer to effectively grasp view knowledge from multi-modal data, which achieves superior 3D grounding performance.

Inspired by the strategies in 2D counterparts, most 3D visual grounding methods adopt a two-stage pipeline, which first detect all object proposals in the scene and then ground the target ones. Unlike 2D images with fixed object positions, the large-scale 3D scenes consist of irregular-distributed point clouds with intricate spatial information, which is view-invariant and causes more challenges. As discussed in previous works [33, 20], one urging issue is *view discrepancy*, caused by the uncertain perspective between the intelligent agent (model) and the commander (grounding text giver). Given relative positions between objects, the text descriptions are supposed to change according to different viewpoints, e.g., when turning the view from "facing" into "back to", the "right" chair should be rectified as a "left" one. Unfortunately, the public available datasets [1, 6] for 3D visual grounding only provide one text query corresponding to point clouds of uncertain viewpoints.

[*] Equal contribution ✉ Corresponding author

To alleviate such potential misalignment, existing methods either manually align the 3D scenes to the paired texts [33], or simultaneously feed multiple views into the network for better view robustness [20]. However, these methods have two major limitations. Firstly, they only focus on solving the view dependence issue from the 3D modality, while neglecting the lack of view cues within text input. Secondly, for the multi-view input, they introduce no specifically designed modules to capture the view knowledge, which is yet significant to discriminate the relative importance of each view. Therefore, we ask: *Can we explicitly grasp the view knowledge from both text and 3D modalities to further boost the 3D grounding performance?*

To this end, we propose **ViewRefer**, a multi-view framework for 3D visual grounding, which captures sufficient view cues from both text and 3D modalities to understand the spatial inter-object relation. Our overall paradigm is shown in Figure 1. For the text modality, we leverage large-scale language models (LLMs) to expand the input grounding text with view-related descriptions. Such LLM-expanded texts can capture sufficient multi-view semantics inherited from LLMs' linguistic knowledge and perform better grounding performance for the target objects. For the 3D modality, we take as input the multi-view 3D scenes and adopt a fusion transformer for 3D-text feature interactions. In each block, we apply different attention mechanisms to exchange information across modalities, views, and objects, contributing to thorough multi-modal fusion. On top of that, we further introduce a set of learnable multi-view prototypes, which aims to capture the inter-view knowledge during training. The guidance of prototypes lies in two aspects. The first complements input text features with adaptive multi-view semantics, the second refines the final output by weighing the importance of different views. Both of them provide high-level guidance for multi-view visual grounding in ViewRefer.

To demonstrate the effectiveness of our approach, we evaluate its performance on three commonly used benchmarks, i.e., Sr3D [1], Nr3D [1] and ScanRefer [6], where ViewRefer consistently achieves superior performance, surpassing the second-best by **+2.8%**, **+1.2%**, **+0.73%**, respectively. The main contributions of our paper are summarized as follows:

- We propose ViewRefer, a multi-view framework for 3D visual grounding, which grasps view knowledge to alleviate the challenging view discrepancy issue.
- For the text and 3D modalities, we respectively introduce LLM-expanded grounding texts and a fusion transformer for capturing multi-view information.
- We present multi-view prototypes to provide high-level guidance to our framework, which contributes to superior 3D grounding performance.

2. Related Work

3D Visual Grounding. 3D Visual Grounding task aims to locate the targeted object in a 3D scene according to a natural language expression. As the baselines, Scanrefer [6] and Referit3D [1] first propose datasets for 3D visual grounding that contain object-annotation pairs on ScanNet [7]. Most recent works [15, 45, 57, 33, 43, 20, 40] adopt a two-stage pipeline that leveraging ground truth or a 3D object detector [24, 21] to obtain the object proposals, utilizing text and 3D encoder [29, 30, 50, 52] to extract features, and then grounding the target one after the feature fusion. Among them, InstanceRefer [45] simplifies the task by treating it as an instance-matching problem. LanguageRefer [33] transforms the multi-modal task into a language modeling problem by replacing the 3D features with predicted object labels. SAT [43] utilizes extra 2D semantics to enhance the multi-modal alignment. MVT [20] first attempts to alleviate the view discrepancy issue by starting from 3D modality to build a view-robust multi-modal representation. Different from the two-stage methods, 3D-SPS [28] treats the 3D visual grounding task as a keypoint selection problem, and first proposes a single-stage network to bridge the gap between detection and matching. Some of the works above point out the crucial view discrepancy issue in 3D visual grounding and propose some preliminary designs to address it. Unlike prior works that only focus on solving the issue from the 3D modality, we start by grasping view knowledge from both text and 3D modalities to address the challenging view discrepancy problem.

Multi-view Learning in 3D. Some recent studies [36, 3, 35, 11] in 3D vision concentrate on improving representation learning by generating 2D renderings from 3D under multiple viewpoints. MVCNN [36] generates a large number of 2D images and simply encodes them for 3D shape classification. PointCLIP [48] transfers 2D knowledge into the 3D domain via multi-view projection for zero-shot learning. SimpleView [11] proposes an effective multi-view framework for shape classification in 3D point cloud learning. Following the framework of [47], I2P-MAE [51] utilizes projected multi-view 2D depth maps to guide 3D point cloud pre-training. DETR3D [39], PETR [27], and related methods [17, 18] conduct 3D object detection based on multi-view images. These works have shown the effectiveness of multi-view representations in enhancing the performance and robustness of various 3D tasks. For our 3D visual grounding task, we introduce ViewRefer to effectively grasp the multi-view cues from multi-modal input data.

Multi-modality Learning in 3D. Multi-modality learning aims at learning from signals of multiple modalities at the same time, which obtains more robust performance

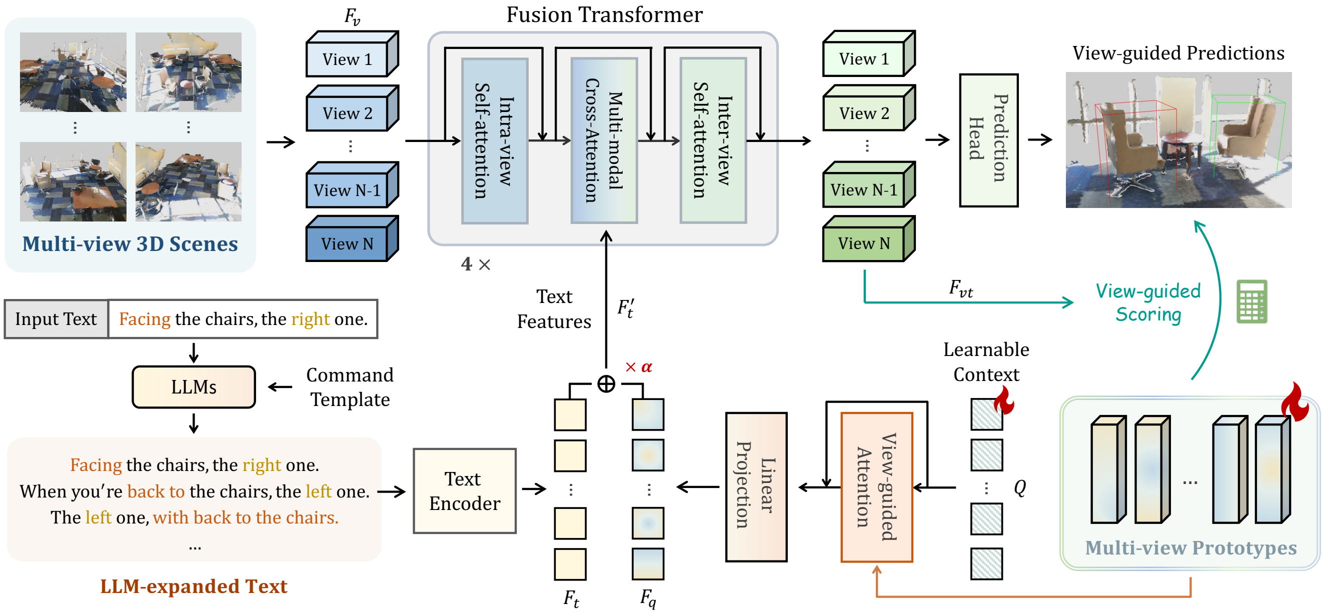


Figure 2: **Overall Pipeline of ViewRefer for 3D Visual Grounding.** We leverage LLMs to enrich the input texts, and introduce a fusion transformer with inter-view attention to enhance cross-view interaction. Upon that, we propose learnable multi-view prototypes to boost the multi-modal fusion via view-guided textual context and scoring strategy.

than single-modality learning. Recently, multiple multi-modality networks in 2D vision [32, 10, 55, 58, 59, 13, 25, 54, 31, 46, 49, 44, 37] show effective performance. Inspired by them, PointCLIP V1 [48] and its V2 variant [61] introduce multi-modality into 3D by adopting CLIP's pre-trained knowledge on 3D domain. What's more, [56, 53] conduct point-voxel joint-training design, and CrossPoint [2] proposes an image-point contrastive learning network. Also, via filter inflation, Image2point [41] utilizes pre-trained 2D knowledge for point cloud understanding. With the emergence of MAE [16], some works [51, 12, 5] combine multi-modality learning with MAE-based pre-training paradigm and achieve great representation capabilities. For 3D visual grounding, a natural multi-modality task in 3D, ViewRefer imports specific designs for both single-modal feature extraction and multi-modal fusion for precise cross-modal grounding.

3. Method

In this section, we illustrate the details of ViewRefer for 3D visual grounding. We first present our overall pipeline in Section 3.1. Then, in Section 3.2 and 3.3, we respectively elaborate on the proposed designs for text and 3D modalities, i.e., LLM-expanded grounding texts and fusion transformer. Finally, in Section 3.4, we introduce the multi-view prototypes into our framework.

3.1. Overall Pipeline

The whole framework of ViewRefer is shown in Figure 2. Given the point cloud of a 3D scene, we first rotate its coordinates as N different views and encode the N-view features $F_v \in \mathbb{R}^{N\times K\times D}$ following [20], where K and D denote object number and feature dimension, respectively. Meanwhile, for the input text, we propose to feed it into the pre-trained large-scale language models (LLMs) and obtain M expanded texts with view cues, denoted as T (Section 3.2). Then, we adopt BERT [8] as the text encoder to extract the LLM-expanded text features as $F_t \in \mathbb{R}^{M\times L\times D}$, where L denotes the max sequence length.

On top of that, the N-view 3D features and M view-guided text features are fed into fusion transformer with cascaded blocks for multi-modal interactions (Section 3.3). Each block sequentially contains the layers for intra-view self-attention, multi-model cross-attention, and inter-view self-attention. By this fusion transformer, the multi-view 3D features can sufficiently interact with each other and incorporate grounding information from the expanded text features. After this, the fused features, denoted as F_{vt}, are passed into a prediction head for multi-view grounding logits, which are finally aggregated across views as the output.

To better grasp the latent view cues, we propose a set of multi-view prototypes learnable during training, and leverage them to assist the grounding from two aspects (Section 3.4). First, we leverage the prototypes to produce view-guided context F_q to the text domain, which is then combined with text features as F_t' before feeding into the fusion transformer. Second, the prototypes are also utilized for weighing the importance of different views. By a view-guided scoring strategy, we further inject multi-view knowledge into the final logits for better grounding results.

We follow previous works [33, 20] to adopt three losses on ViewRefer, which are cross-entropy loss L_{ref} on the grounding predictions, L_{text} of text classification on the text encoder, and L_{3D} of 3D shape classification upon the object encoder. The whole loss function is formulated as

$$L = L_{ref} + \beta \cdot L_{text} + \gamma \cdot L_{3D}, \tag{1}$$

where β and γ denote the weights for L_{text} and L_{3D}. Both of them are set as 0.5 in our method.

3.2. LLM-expanded Grounding Texts

To fully exploit view knowledge within the text modality, we propose to utilize the abundant linguistic knowledge in LLMs, e.g., GPT-3 [4], to expand the original simple grounding texts. As discussed in [1], the 3D-text data in 3D visual grounding task can be divided into view-dependent and view-independent pairs, depending on whether the expression is constrained to the speaker's perspective. For both of them, we construct general command templates fed into the GPT-3 to generate M LLM-expanded texts. Further, considering their different semantic compositions, we adopt specific expanding strategies for the two categories.

View-Dependent Text. The description in such texts is strongly related to speaker's viewpoint. Therefore, for an input view-dependent text, we first generate its synonymous sentence that refers to the same target but under different speaker's views. Specifically, we create a view-related phrase dictionary, the keys of which include phrases for orientation like "looking at", and phrases describing positional relationships like "left". The corresponding values to the keys are the opposite phrases of them, e.g., "looking at - with back to", and "left - right". Then, we construct a command template as the input for GPT-3, which includes the input text and aforementioned view-related phrases in the dictionary. Such general template is designed as

$$\begin{aligned}&\text{“ Rephrase the sentence of ‘[TEXT]’}\\ &\quad\text{to the opposite perspective,}\\ &\text{which contains phrases ‘[PHRASEs]’: ”}\end{aligned} \tag{2}$$

where [TEXT] and [PHRASEs] are replaced by the input text and view-related phrases, respectively. For example, given the initial text of "*Facing the front of the couch, pick the table that is to the right of the couch*", we complete the command as

" Rephrase the sentence of '*Facing the front of the couch, pick the table that is to the right of the couch*' to the opposite perspective, which contains phrase '*with back to*', '*left*': " .

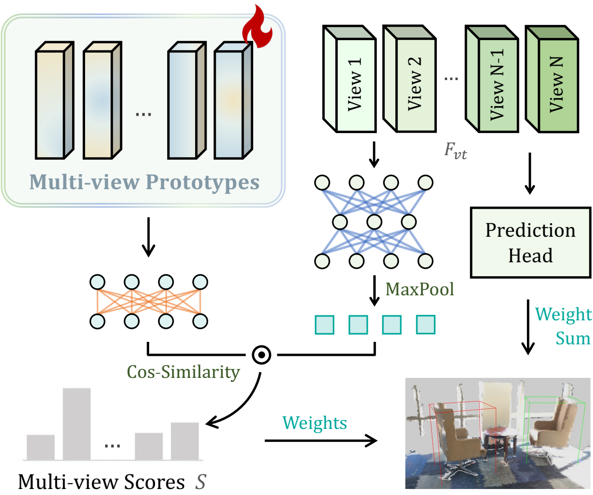


Figure 3: **View-guided Scoring Strategy.** Guided by the learnable multi-view prototypes, we assess the importance of each view to predict the final grounding logits.

By this command, we obtain GPT-3's response as *"With back to the front of the couch, pick the table that is on the left side of the couch"*. In this way, the rephrased description contains opposite-view information. Together with the original single text, we enrich the text modality with more view-dependent semantics. To further unleash LLMs' linguistic knowledge, we respectively rephrase the input sentence and its opposite-view synonym for $(M-2)/2$ times with a simple command template of

$$\text{“ Rephrase the sentence of ‘[TEXT]’: ”} \tag{3}$$

where [TEXT] is replaced by the input text or its opposite-view synonym. Through this expanding strategy, we enrich the initial single text into a set of M LLM-expanded texts, which contain abundant multi-view cues and effectively alleviate the discrepancy issue.

View-Independent Text. The description in view-independent texts is irrelevant to speaker's viewpoint. Thus, we adopt no opposite-view enrichment on these view-independent inputs, and directly rephrase them for $M-1$ times with Template 3. The generated M LLM-expanded texts also inherit the rich linguistic knowledge from LLMs, and diversify the text embedding space for visual grounding performance.

Method	Sr3D					Nr3D				
	Overall	Easy	Hard	View Dep.	View Indep.	Overall	Easy	Hard	View Dep.	View Indep.
ReferIt3D [1]	40.8%	44.7%	31.5%	39.2%	40.8%	35.6%	43.6%	27.9%	32.5%	37.1%
TGNN [19]	45.0%	48.5%	36.9%	45.8%	45.0%	37.3%	44.2%	30.6%	35.8%	38.0%
InstanceRefer [45]	48.0%	51.1%	40.5%	45.4%	48.1%	38.8%	46.0%	31.8%	34.5%	41.9%
3DVG-Transformer [57]	51.4%	54.2%	44.9%	44.6%	51.7%	40.8%	48.5%	34.8%	34.8%	43.7%
LanguageRefer [33]	56.0%	58.9%	49.3%	49.2%	56.3%	43.9%	51.0%	36.6%	41.7%	45.0%
TransRefer3D [15]	57.4%	60.5%	50.2%	49.9%	57.7%	42.1%	48.5%	36.0%	36.5%	44.9%
SAT [43]	57.9%	61.2%	50.0%	49.2%	58.3%	49.2%	56.3%	42.4%	46.9%	50.4%
MVT [20]	64.5%	66.9%	58.8%	58.4%	64.7%	55.1%	61.3%	49.1%	54.3%	55.4%
MVT*[20]	64.4%	67.3%	57.1%	55.7%	64.6%	54.9%	61.2%	48.9%	54.4%	55.4%
ViewRefer	**67.2%**	**69.7%**	**61.7%**	**56.9%**	**67.8%**	**56.1%**	**63.1%**	**50.0%**	**56.1%**	**56.9%**
Gain	+2.8%	+2.4%	+4.6%	+1.2%	+3.2%	+1.2%	+1.9%	+1.1%	+1.7%	+1.5%

Table 1: **Performance of ViewRefer on Sr3D and Nr3D [1].** We report the performance on the overall dataset and all its splits. '*' denotes our implementation results[†].

3.3. Fusion Transformer

To integrate 3D multi-view and textual features for grounding, we introduce a multi-modal transformer composed of cascaded fusion blocks. Each block includes three types of attention mechanisms with residual connections. Firstly, an intra-view attention layer is adopted for the N-view features F_v, which independently interacts K object features within each view. Secondly, we utilize a multi-modal cross-attention layer referring to [20] to infuse textual information from F_t' into the multi-view features. We formulate them as

$$\begin{aligned} F_v &= F_v \ + \text{Intra-Attn}(F_v), \\ F_{vt} &= F_v + \text{Cross-Attn}(F_v,\ F_t'), \end{aligned} \tag{4}$$

where $F_{vt} \in \mathbb{R}^{N \times K \times D}$ denotes the multi-modal features.

After that, to boost the cross-view communication, we subsequently introduce an inter-view attention layer. Orthogonal to intra-view attention, this layer calculates attention weights across different views for the same object, formulated as

$$F_{vt} = \left(F_{vt}^T + \text{Inter-Attn}(F_{vt}^T)\right)^T. \tag{5}$$

With such interaction, the network can better capture the inter-view differences from one object, and focus on more informative views to encode multi-modal features F_{vt}.

3.4. Multi-view Prototypes

Besides our LLM-expanded texts and fusion transformer, we further introduce a set of multi-view prototypes to conduct view-guided 3D visual grounding in ViewRefer. The prototypes are randomly initialized embeddings and represented as $Proto \in \mathbb{R}^{N \times D}$, where N, D denote the view number and feature dimension. They aim to memorize the 3D-text view-consistent knowledge from a common multi-modal space, which serve as high-level guidance for the grounding process from the following two aspects.

View-guided Textual Context. We first utilize the prototypes to incorporate view-guided contexts into the textual features F_t. As shown in the bottom right of Figure 2, we first randomly initialize a learnable context $Q \in \mathbb{R}^{M \times D}$, which serves as 'query' to extract 3D-text multi-view information from the prototypes $Proto$. Then, we regard the prototypes as 'key' and 'value', and propose a view-guided attention module with residual connections for feature interaction. We utilize one cross-attention layer in the module, formulated as

$$F_q = Q + \text{VG-Attn}(Proto,\ Q), \tag{6}$$

where VG-Attn denotes the view-guided attention module. After this, the query features F_q have adaptively aggregate informative multi-view cues and are element-wisely added to the input text features F_t as

$$F_t' = F_t + \alpha \cdot F_q, \tag{7}$$

where α denotes a frozen balance factor weighing the importance of view-guided context, which is initialized before training. From the interaction with prototypes, the text features F_t' become view-aware and perform better multi-modal fusion in the subsequent transformer.

[†] We reproduce the results based on the official open-source code [20] on a single NVIDIA A100 GPU.

Method	Unique		Multiple		Overall	
	Acc.@0.25	Acc.@0.50	Acc.@0.25	Acc.@0.50	Acc.@0.25	Acc.@0.50
ReferIt3D [1]	53.75%	37.47%	21.03%	12.83%	26.44%	16.90%
ScanRefer [6]	65.00%	43.31%	30.63%	19.75%	37.30%	24.32%
TGNN [19]	64.50%	53.01%	27.01%	21.88%	34.29%	27.92%
InstanceRefer [45]	77.45%	66.83%	31.27%	24.77%	40.23%	32.93%
MVT [20]	77.67%	66.45%	31.92%	25.26%	40.80%	33.26%
MVT* [20]	75.18%	63.84%	31.46%	24.95%	40.65%	32.96%
ViewRefer	**76.35%**	**64.27%**	**33.08%**	**26.50%**	**41.35%**	**33.69%**
Gain	+1.17%	+0.43%	+1.62%	+1.55%	+0.70%	+0.73%

Table 2: **Performance of ViewRefer on ScanRefer [6]**. We report the Acc@0.25 and Acc@0.50 on the overall dataset and its two splits, "Unique" and "Multiple". '*' denotes our implementation results†.

View-guided Scoring Strategy. Furthermore, we leverage the multi-view prototypes to conduct a scoring for the predicted logits after the fusion transformer. As shown in Figure 3, we calculate N-view scores for the multi-modal features F_{vt}, which assess the relative importance of each view in the final prediction process. We first apply a max pooling operation to F_{vt} along the dimension of intra-view objects, which aggregates the global features of different views. We then adopt linear layers to project the prototypes and multi-modal features into a unified embedding space. Upon that, we calculate the multi-view scores by cosine similarities of each N-view pair as

$$S = \text{Sim}\left(Proto,\ \text{MaxPool}(F_{vt})\right), \tag{8}$$

where $S \in \mathbb{R}^{N\times 1}$ denotes the score for each N view. After this, the multi-view scores serve as the weights to aggregate predicted logits among different views. Such scoring can adaptively constrain the prediction of those views inconsistent with the input text, and emphasize the informative views for accurate 3D visual grounding.

4. Experiments

In this section, we evaluate the performance of ViewRefer on three commonly used benchmarks, i.e., Sr3D [1], Nr3D [1], and ScanRefer [6].

4.1. Datasets

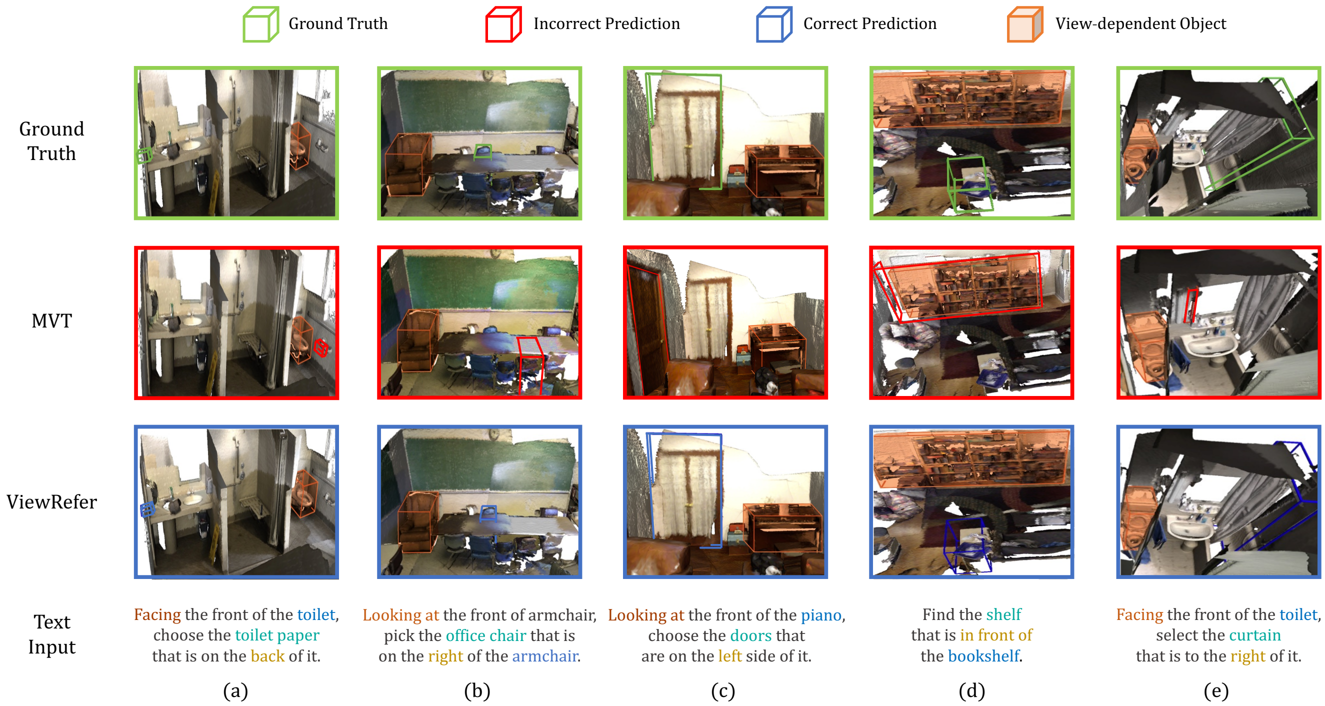
Sr3D and Nr3D. The Sr3D dataset [1] contains 83,572 template-based utterances leveraging spatial relationships among objects to localize a referred object in a 3D scene. The scenes are constrained to have no more than six distractors, i.e., objects belonging to the same category as the target. The Nr3D dataset [1] provides annotations for the ScanNet [7], an indoor 3D scene dataset, comprising 45,503 natural and free-form utterances. The dataset includes 707 distinct indoor scenes with target objects belonging to 76 fine-grained categories. Two distinct data splits are employed in Sr3D and Nr3D, namely the "Easy" and "Hard" splits that differ based on the number of distractors in the scene, and the "View-dependent" and "View-independent" splits that vary based on whether the referring expression relies on the speaker's viewpoint.

ScanRefer. The ScanRefer dataset [6] annotates 800 indoor scenes in ScanNet [7] and contains 51,583 utterances, composed of 36,665 samples for the training set, 9,508 for val set, and 5,410 for the test set. Depending on whether there are objects of the same target class in the scene, the dataset can be split into two parts, i.e., the "Unique" part and the "Multiple" one.

4.2. Experimental Settings

Evaluation Metrics. Following existing works [20, 45], we utilize the ground truth object proposals for Nr3D and Sr3D [1], while adopting detector-generated object proposals via pre-trained detector, PointGroup [21], for ScanRefer dataset [6]. As for the evaluation metrics, we measure the networks via the accuracy of the target predictions for Nr3D and Sr3D, and the Acc@mIoU metric for ScanRefer where, $m \in \{0.25, 0.50\}$. The Acc@mIoU measures the proportion of text input with predicted bounding box overlapping the ground truth box by intersection over the union (IoU) higher than m.

Implementation Details. For Sr3D and Nr3D datasets [1], we train the network for 100 epochs with a batch size of 24. We utilize AdamW [23] as the optimizer with an initial learning rate of 5×10^{-5} for the fusion transformer, 5×10^{-4} for other modules, and weight decay


Ground Truth
Incorrect Prediction
Correct Prediction
View-dependent Object
Ground Truth
MVT
ViewRefer
Text Input
Facing the front of the toilet, choose the toilet paper that is on the back of it.
Looking at the front of armchair, pick the office chair that is on the right of the armchair.
Looking at the front of the piano, choose the doors that are on the left side of it.
Find the shelf that is in front of the bookshelf.
Facing the front of the toilet, select the curtain that is to the right of it.

(a) (b) (c) (d) (e)

Figure 4: **Visualization of the 3D Visual Grounding Results.** For the presented 3D scenes, we utilize green, red, blue, and orange boxes to represent the ground truth, incorrect predictions, correct predictions, and view-dependent objects, respectively.

Decoder	Multi-view Input	Inter-view Attention	LLM-expanded Text	Multi-view Prototype		Overall	Easy	Hard	View Dep.	View Indep.
				VG. Score.	VG. Cont.					
-	-	-	-	-	-	22.4%	23.9%	18.7%	26.9%	22.2%
✓	-	-	-	-	-	61.5%	64.2%	55.2%	51.6%	61.9%
✓	✓	-	-	-	-	64.4%	67.0%	57.6%	54.5%	64.6%
✓	✓	✓	-	-	-	65.1%	67.2%	59.9%	54.7%	65.7%
✓	✓	✓	✓	-	-	66.0%	68.4%	60.4%	55.2%	66.4%
✓	✓	✓	✓	✓	-	66.5%	68.7%	60.8%	55.3%	66.9%
✓	✓	✓	✓	✓	✓	**67.2%**	**69.7%**	**61.7%**	**56.9%**	**67.8%**

Table 3: **Ablation Study on Different Components for Grasping Multi-view Knowledge on Sr3D [1].** 'VG. Score.' and 'VG. Cont.' denote the view-guided scoring strategy and view-guided textual context, respectively.

as 1×10^{-3}. After 40 epochs, we decline the learning rate by multiplying 0.65 per 10 epochs. For ScanRefer [6], we set the max epoch number as 30, and batch size as 32. We adopt AdamW [23] as the optimizer with an initial learning rate of 5×10^{-4} with no weight decay. As for the structure of the network, we set both the view number N and the generated text number M as 4. We utilize 3 layers for the text encoder and 4 transformer blocks for the fusion transformer with 8 attention heads. For data augmentation, we conduct translation on each 3D object and random rotation on 3D scenes during training. Please refer to the Supplementary Material for more implementation details.

4.3. Quantitative Analysis

Performance on Sr3D. In Table 1, we report the performance of ViewRefer on Sr3D dataset [1] for 3D visual grounding. Our ViewRefer outperforms prior works in all splits, surpassing the *state-of-the-art* method [20] on the overall accuracy by +2.8%, on the "view-dependent" split by +1.2%, which demonstrates the effectiveness of the view knowledge grasped by ViewRefer. Also, the significant amplification of +4.6% on the challenging "Hard" split shows the effectiveness of our ViewRefer in 3D scenes with complex spatial relations between objects, which further indicates our superior understanding of spatial relations.

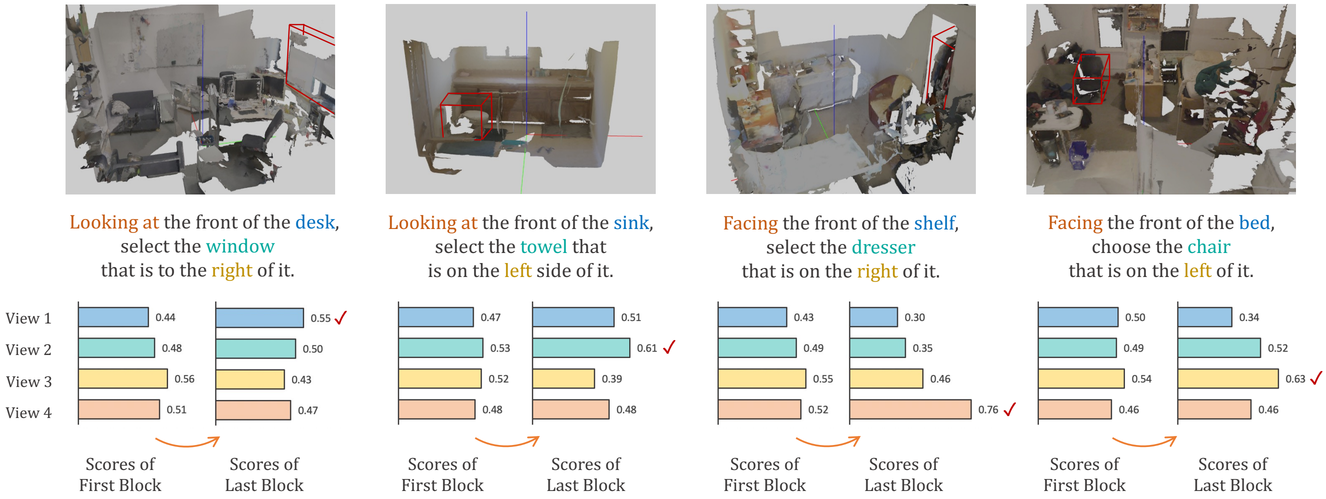


Figure 5: **Visualization of View-guided Scoring Strategy.** We present four view-dependent samples, and illustrate the multi-view scores calculated from the first to the last fusion transformer blocks. We mark the correct view with '✓'.

Performance on Nr3D. In Table 1, we evaluate ViewRefer on Nr3D dataset, where ViewRefer exceeds best competitor [20] by an overall improvement of +1.2%, and +1.7% on "view-dependent" split, indicating the superiority of ViewRefer on data with intricate view-dependent natural referential utterances, which further demonstrates the effectiveness of our view-guided designs.

Results on ScanRefer. We also evaluate ViewRefer on ScanRefer [6] and report the Acc@mIoU performance of ViewRefer in Table 2. Clearly, compared with prior works, our ViewRefer achieves the highest scores on all metrics. This well illustrates the superior view-understanding capability of our multi-view training framework.

4.4. Qualitative Analysis

3D Visual Grounding Results. In Figure 4, we select some view-dependent cases from Sr3D [1] and visualize the grounding truth boxes, predictions of MVT [20], and predictions of ViewRefer in each column from top to bottom. We emphasize the view-dependent objects in orange boxes with highlights for view alignment. As shown in (a), MVT fails to ensure the correct view as the text input describes, and grounds to wrong object with the same category, while ViewRefer successfully predicts the target under the correct perspective. From the predictions in (b) and (c), we find although ViewRefer and MVT both understand the correct views, only ViewRefer grounds the right target, which shows that ViewRefer achieves a comprehensive understanding of the spatial relations between the objects. Additionally, in (d) and (e), we observe that ViewRefer shows better object recognition ability, while MVT is relatively easy to predict objects of the wrong categories.

Intra-view Aggregation			Context		Overall	View Dep.
Avg	Max	Max+Avg	Each	Glo.		
✓	-	-	✓	-	64.9%	52.7%
✓	-	-	-	✓	65.4%	54.0%
-	-	✓	✓	-	66.1%	53.8%
-	-	✓	-	✓	66.2%	54.1%
-	✓	-	✓	-	66.6%	55.0%
-	✓	-	-	✓	**67.2%**	**56.9%**

Table 4: **Ablation Study on Multi-view Prototypes,** including View-guided Scoring Strategy and Textual Query Context. 'Avg', 'Max', and 'Max+Avg' denote average pooling, max pooling, and the summation of both to obtain the global view features. 'Context' denotes the combination of textual query features.

Interpretability of View-guided Scoring Strategy. To demonstrate the interpretability of the proposed view-guided scoring strategy, we utilize the multi-view prototypes and output features from the fusion transformer blocks, to calculate and analyze the intermediary multi-view scores. In Figure 5, we present the scores calculated from the first and last blocks of view-dependent cases, where we mark the correct view with "✓". As shown from the last blocks' scores, with the proposed scoring strategy, ViewRefer can effectively assess the relative importance of each view for exact visual grounding. Also, we observe the upward trend of the scores of correct view, which shows that with the view-guided designs, our network gradually captures the view knowledge as the network deepens.

5. Ablation Study

To explore the effectiveness of each design in ViewRefer, we conduct extensive ablation study on Sr3D [1] dataset and evaluate the overall accuracy on 3D visual grounding. In Table 3, we report the results of different combinations of proposed designs. In Table 4, we explore the implementation of enhancements from multi-view prototypes.

Effectiveness of Each Component. Based on the baseline network, we conduct ablation studies by adding the designed modules one by one until the final structure of ViewRefer. In Table 3, the first two rows in light gray are baselines with basic fusion modules as described in [20]. As reported, the import of each major component benefits the object location separately, which shows the effectiveness of each design in ViewRefer.

Implementation of Multi-view Prototypes. Additionally, we conduct ablation study on the implementation regarding the two enhancements of multi-view prototypes, i.e., aggregation in view-guided scoring, and combination of textual query features F_q. For the aggregation in view-guided scoring, we investigate different pooling operations to integrate features of all objects for a certain view. Meanwhile, we also evaluate two combination types of textual query features: only add the textual query features into the first global token that contains the whole sentence's feature ('Glo.'), or broadcast and add into each token simultaneously ('Each'). As reported in Table 4, 'max pooling' with 'global' combination performs the best, which is our default in all experiments.

6. Conclusion

We propose **ViewRefer**, a multi-view framework for 3D visual grounding that grasps view knowledge from both text and 3D modalities to alleviate the challenging view discrepancy issue. In the text modality, we introduce LLM-expanded grounding texts to utilize the diverse linguistic knowledge of large-scale language models for view understanding. For 3D modality, we propose a fusion transformer with inter-view attention to grasp rich multi-view knowledge. Furthermore, with a set of learnable multi-view prototypes, we enhance ViewRefer from two perspectives, view-guided textual context and a view-guided scoring strategy, which further enhances view-guided multi-modal fusion for exact grounding. Extensive experiments are conducted to demonstrate the superiority of ViewRefer on 3D visual grounding. We expect ViewRefer can inspire more works to further explore the possibility of grasping view knowledge to benefit view understanding in 3D visual grounding.

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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020. 1, 2, 4, 5, 6, 7, 8, 9

[2] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9902–9912, 2022. 3

[3] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9287–9296, 2019. 2

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 4

[5] Anthony Chen, Kevin Zhang, Renrui Zhang, Zihan Wang, Yuheng Lu, Yandong Guo, and Shanghang Zhang. Pimae: Point cloud and image interactive masked autoencoders for 3d object detection. *arXiv preprint arXiv:2303.08129*, 2023. 3

[6] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, pages 202–221. Springer, 2020. 1, 2, 6, 7, 8

[7] 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. 2, 6

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[9] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022. 1

[10] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 3

[11] Ankit Goyal, Hei Law, Bowei Liu, Alejandro Newell, and Jia Deng. Revisiting point cloud shape classification with a simple and effective baseline. In *International Conference on Machine Learning*, pages 3809–3820. PMLR, 2021. 2

[12] Ziyu Guo, Xianzhi Li, and Pheng Ann Heng. Joint-mae: 2d-3d joint masked autoencoders for 3d point cloud pre-training. *arXiv preprint arXiv:2302.14007*, 2023. 3

[13] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. *arXiv preprint arXiv:2209.14169*, 2022. 3

[14] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021. 1

[15] 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 *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2344–2352, 2021. 2, 5

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 3

[17] Yu Hong, Hang Dai, and Yong Ding. Cross-modality knowledge distillation network for monocular 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*, pages 87–104. Springer, 2022. 2

[18] Peixiang Huang, Li Liu, Renrui Zhang, Song Zhang, Xinli Xu, Baichao Wang, and Guoyi Liu. Tig-bev: Multi-view bev 3d object detection via target inner-geometry learning. *arXiv preprint arXiv:2212.13979*, 2022. 2

[19] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1610–1618, 2021. 5, 6

[20] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15524–15533, 2022. 1, 2, 3, 4, 5, 6, 7, 8, 9

[21] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020. 2, 6

[22] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 1

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6, 7

[24] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 2

[25] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng

Li. Frozen clip models are efficient video learners. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 388–404. Springer, 2022. 3

[26] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019. 1

[27] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 531–548. Springer, 2022. 2

[28] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16454–16463, 2022. 2

[29] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2

[30] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[31] Longtian Qiu, Renrui Zhang, Ziyu Guo, Ziyao Zeng, Yafeng Li, and Guangnan Zhang. Vt-clip: Enhancing vision-language models with visual-guided texts. *arXiv preprint arXiv:2112.02399*, 2021. 3

[32] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[33] 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. 1, 2, 4, 5

[34] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 1

[35] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 2

[36] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 2

[37] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[38] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6629–6638, 2019. 1

[39] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2

[40] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual and language learning. *arXiv preprint arXiv:2209.14941*, 2022. 2

[41] Chenfeng Xu, Shijia Yang, Tomer Galanti, Bichen Wu, Xiangyu Yue, Bohan Zhai, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Image2point: 3d point-cloud understanding with 2d image pretrained models. *arXiv preprint arXiv:2106.04180*, 2021. 3

[42] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 1

[43] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1856–1866, 2021. 2, 5

[44] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 3

[45] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1791–1800, 2021. 2, 5, 6

[46] Renrui Zhang, Hanqiu Deng, Bohao Li, Wei Zhang, Hao Dong, Hongsheng Li, Peng Gao, and Yu Qiao. Collaboration of pre-trained models makes better few-shot learner. *arXiv preprint arXiv:2209.12255*, 2022. 3

[47] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *arXiv preprint arXiv:2205.14401*, 2022. 2

[48] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8552–8562, 2022. 2, 3

[49] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *arXiv preprint arXiv:2303.02151*, 2023. 3

[50] Renrui Zhang, Liuhui Wang, Ziyu Guo, and Jianbo Shi. Nearest neighbors meet deep neural networks for point cloud analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1246–1255, 2023. 2

[51] Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. *arXiv preprint arXiv:2212.06785*, 2022. 2, 3

[52] Renrui Zhang, Liuhui Wang, Yali Wang, Peng Gao, Hongsheng Li, and Jianbo Shi. Parameter is not all you need: Starting from non-parametric networks for 3d point cloud analysis. *arXiv preprint arXiv:2303.08134*, 2023. 2

[53] Renrui Zhang, Ziyao Zeng, Ziyu Guo, Xinben Gao, Kexue Fu, and Jianbo Shi. Dspoint: Dual-scale point cloud recognition with high-frequency fusion. *arXiv preprint arXiv:2111.10332*, 2021. 3

[54] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6868–6874, 2022. 3

[55] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 493–510. Springer, 2022. 3

[56] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 3

[57] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021. 2, 5

[58] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 3

[59] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3

[60] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10012–10022, 2020. 1

[61] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022. 3