Data-Efficient 3D Visual Grounding via Order-Aware Referring

Tung-Yu Wu[1*] Sheng-Yu Huang[1*] Yu-Chiang Frank Wang[1,2]
[1]Department of Electrical Engineering, National Taiwan University
[2]NVIDIA
{b08901133, f08942095}@ntu.edu.tw, frankwang@nvidia.com

Abstract

*3D visual grounding aims to identify the target object within a 3D point cloud scene referred to by a natural language description. Previous works usually require significant data relating to point color and their descriptions to exploit the corresponding complicated verbo-visual relations. In our work, we introduce Vigor, a novel Data-Efficient 3D **Vi**sual **G**rounding framework via **O**rder-aware **R**eferring. Vigor leverages LLM to produce a desirable referential order from the input description for 3D visual grounding. With the proposed stacked object-referring blocks, the predicted anchor objects in the above order allow one to locate the target object progressively without supervision on the identities of anchor objects or exact relations between anchor/target objects. We also present an order-aware warm-up training strategy, which augments referential orders for pre-training the visual grounding framework, allowing us to better capture the complex verbo-visual relations and benefit the desirable data-efficient learning scheme. Experimental results on the NR3D and ScanRefer datasets demonstrate our superiority in low-resource scenarios. In particular, Vigor surpasses current state-of-the-art frameworks by 9.3% and 7.6% grounding accuracy under 1% data and 10% data settings on the NR3D dataset, respectively.*

1. Introduction

Visual grounding is an emerging task that aims to ground a target object in a given 2D/3D scene from a natural description, where the description contains information to identify the target object (e.g., color, shape, or relations to other anchor objects). This task is potentially related to industrial applications to AR/VR and robotics [3, 24, 25]. Compared to object detection, the main challenge of visual grounding lies in the requirement to find *the only one* object described in the given natural description, while there might be multiple objects with the same class of the target object appearing in the scene. Therefore, the model is expected to identify the relations between all objects in the scene to find the ideal target object according to the given description. In recent years, significant progress has been made in image-based 2D visual grounding [22,27,37,43,49,52]. However, comparatively fewer efforts are directed towards addressing the more intricate challenge of 3D visual grounding, raised from joint consideration of the unstructuredness of natural language descriptions and scattered object arrangements in the 3D scene. The complications of the two modalities make it challenging to directly refer to the target object with plain cross-modal interaction between the features of the scene point cloud and the referring description, showing the need for additional research in the field of 3D visual grounding.

As pioneers of 3D visual grounding, Referit3D [2] and ScanRefer [8] are two benchmark datasets that build upon the point cloud scene provided by the ScanNet [12] dataset. The former presents a graph neural network (GNN) [35]-based framework to explicitly learn the object spatial relations as the baseline. The latter designs a verbo-visual cross-modal feature extraction and fusion pipeline as the baseline. Following the settings of the aforementioned benchmarks, several subsequent methods are presented [4, 11, 14, 18–20, 29, 40, 45]. However, these approaches use a referring head to localize the target object directly. Without explicitly considering any additional information about the anchor objects mentioned in the description, models must implicitly discover the relation between the anchor objects and the target object. They may be misled by other similar objects presented, as pointed out in [5]. To overcome this issue, some approaches [1, 41, 47] propose to incorporate anchor objects during training by including their label annotations [5, 17]. Nonetheless, human annotators are usually required to obtain this additional linguistic information [1, 47], causing potential difficulty in scaling up to larger datasets for real-world applications.

To eliminate the need of human annotation for training grounding models, recent approaches [5,17,39,53] leverage

[*]Equal Contribution

Figure 1. **Referential orders for 3D grounding.** The order manifests an anchor-to-target referring process that helps the grounding model identify the target object described in the input.

pre-trained 2D priors (e.g., SAM [23], LDM [34]) or large language models (LLMs) for automatic dissection of the descriptions and generation of prior linguistic knowledge. For example, Diff2Scene [53] conducts the use of LDM to obtain text-conditioned 2D semantic maps as pseudo labels to guide a 3D segmentation model to achieve scene understanding and produce zero-shot 3D visual grounding. Unfortunately, limited by the spatial understanding ability of 2D diffusion models between multiple objects as described in [30], the ability of Diffi2Scene to achieve effective visual grounding for complex description with multiple anchor objects is still unclear. On the other hand, NS3D [17] utilizes Codex [10] to parse descriptions into nested expressions and designs a neuro-symbolic framework to find the target object step-by-step. However, it only considers fixed-template relations between objects (e.g., below/above, near/far, etc.) and cannot be easily extended to arbitrary natural descriptions. Inspired by the mechanism of human perception system [9, 31], CoT3DRef [5] generates the referential order of a description that points from anchor objects to the final target object using LLM. For example, for a description "*Find the water bottle on the table nearest to the door.*", the referential order is generated as { "*door*" (anchor), "*table*" (anchor), "*water bottle*" (target) }. Additionally, it utilizes a rule-based algorithm to localize the identities of the above anchor/target objects, which guides a transformer-based module to predict the final target object. However, as noted in [5], such rule-based identity prediction might not be applicable for scenarios with complex language descriptions.

In this paper, we propose a data-efficient 3D **Vi**sual **G**rounding framework via **O**rder-aware **R**eferring (Vigor). Leveraging the LLM-parsed referential order, Vigor exploits the awareness of anchor objects from the textual description, as depicted in Fig. 1. With such ordered anchor objects as guidance, a series of Object Referring blocks are deployed to process the corresponding objects, each performing feature enhancement to update the visual features of corresponding objects. Since only the ground-truth grounding information of the target object is available during training (no ground-truth referential order observed), we additionally introduce a unique warm-up learning strategy to Vigor. This pre-training scheme can be viewed as augmenting object labels and referential orders to initialize Vigor so that it can be realized in data-efficient training schemes. Our experiments on real-world benchmark datasets confirm that Vigor performs favorably against recent 3D visual grounding methods, especially when the size of training data is limited.

We now summarize our contribution as follows:

- We present a Data-Efficient 3D **Vi**sual **G**rounding Framework via **O**rder-Aware **R**eferring (Vigor), which performs 3D visual grounding from natural description inputs.
- By utilizing sequential yet consecutive Object Referring blocks, Vigor is able to locate anchor/target objects mentioned in the description by considering plausible referential orders established by LLM.
- We introduce a warm-up strategy that introduces the model with the ability to locate anchor/target object identities by synthesizing training examples of reliable labels and referential orders.
- Through comprehensive experiments, we show that Vigor achieves satisfactory performances in various low-source settings, surpassing current grounding approaches significantly.

2. Related Work

2.1. 2D Visual Grounding

2D visual grounding aims to locate the target object in an image referred to by a natural language description, with various approaches being proposed in recent years [22, 26, 27, 37, 38, 43, 46, 49]. Among them, verbo-visual feature alignment frameworks [22, 26, 27, 49] have proven themselves to be an effective way to equip models with abilities to tackle description contexts and image semantics simultaneously. Particularly, as one of the pioneers, MDETR [22] extends DETR [7], an end-to-end object detection framework, to incorporate text modalities with the proposed text-image alignment contrastive losses. GLIP [49] takes a step forward to improve the performance of visual grounding by proposing unified multi-task learning that includes object localization and scene understanding tasks, showing that these tasks could gain mutual benefits from each other. Grounding DINO [27] further designs the large-scale grounding pretraining for DINO [15], reaching the capability of open-set grounding. Although great progress is achieved, extending these 2D visual grounding methods to 3D scenarios is not easy due to the additional depth information in 3D data that triggers more complicated object arrangements and more complex descriptions to describe the relations between objects, leaving 3D visual grounding as an unsolved research area.

2.2. 3D Visual Grounding

In 3D visual grounding, models are designed to jointly handle complicated natural language descriptions and scattered objects within a point cloud scene. Previous approaches attempt to solve this task by either constructing text-point-cloud feature alignment frameworks [20], designing pipelines to better exploit the 3D spatial relations of objects [2, 11, 14, 18, 44, 48], or bringing in auxiliary visual features [4, 19, 29, 45]. Specifically, BUTD-DETR [20] extends MDETR [22] to 3D visual grounding by adapting a text-point-cloud alignment loss to pull the features of point cloud and text together. To exploit the 3D spatial relations, graph-based methods [2, 14, 18, 48] utilize GNNs to model the 3D scene, with nodes and edges being the objects and object-to-object relations, to learn their correlations explicitly. Also, some studies craft specialized modules [11, 16, 40, 44], such as the spatial self-attention presented in ViL3DRel [11] and relation matching network in CORE-3DVG [44], to capture spatial relations among objects.

To better identify the target object, [4, 19, 29, 45] aim to produce richer input semantic information for learning the grounding model. For example, [4, 29, 45] introduce image features by acquiring 2D images of the scene to obtain more color/shape information. MVT [19] projects the point cloud into multiple views for more position information. Although these approaches have achieved great progress in dealing with scattered object arrangements, such methods typically extract a global, sentence-level feature [13,28] from the given natural description. As a result, detailed information such as the target object, anchor objects, and their relations may not be preserved and leveraged properly, potentially reducing training efficiency and prediction accuracy as discussed in [5].

To address the above issue, some works have put their efforts into mining the natural descriptions to acquire additional prior knowledge for improved learning [1, 5, 17, 41, 47]. Specifically, ScanEnts3d [1] and 3DPAG [47] recruit human annotators to establish one-to-one matching between each anchor object mentioned in the description and the corresponding object entity in the 3D scene. With such additional information, they design dense word-object alignment losses to improve the training. However, annotating one-to-one text-3D relations requires considerable labor effort. For example, it takes more than 3600 hours of workforce commitment in 3DPAG to annotate anchor objects for 88k descriptions.

2.3. Data-Efficient 3D Visual Grounding

To eliminate the need for human annotators for learning grounding models, NS3D [17] makes the first attempt to use the LLM. It leverages Codex [10] to parse fixed-template descriptions into nested logical expressions, followed by a neural-symbolic framework to execute the logical expressions implemented as programmatic functions. By doing so, NS3D correctly locates the anchor/target objects mentioned in the given description and achieves impressive performance with only 0.5% training data on synthetic datasets. Unfortunately, NS3D is not designed to handle arbitrary natural descriptions, resulting in complicated expressions and unforeseen functions and hindering the framework from successful execution [17]. Recently, CoT3DRef [5] proposes to utilize LLMs to acquire the referential order of the description, listing from anchor objects to the final target object. It deploys a rule-based searching method using a traditional sentence parser [36] to construct the one-to-one matching between class names in the order and potential anchor/target objects in the scene at once. The matching information is encoded by positional encoding as pseudo labels of anchor/target objects and as additional inputs for the proposed transformer-based CoT module to learn the referring process implicitly and predict the final target object. By the above design, CoT3DRef is able to reduce the training data to 10% while preserving competitive performance on real-world datasets. Nevertheless, since the parsed pseudo labels may not be accurate, taking them as additional inputs might result in noisy information and degrade the grounding process, impeding correct target prediction and affecting training stability.

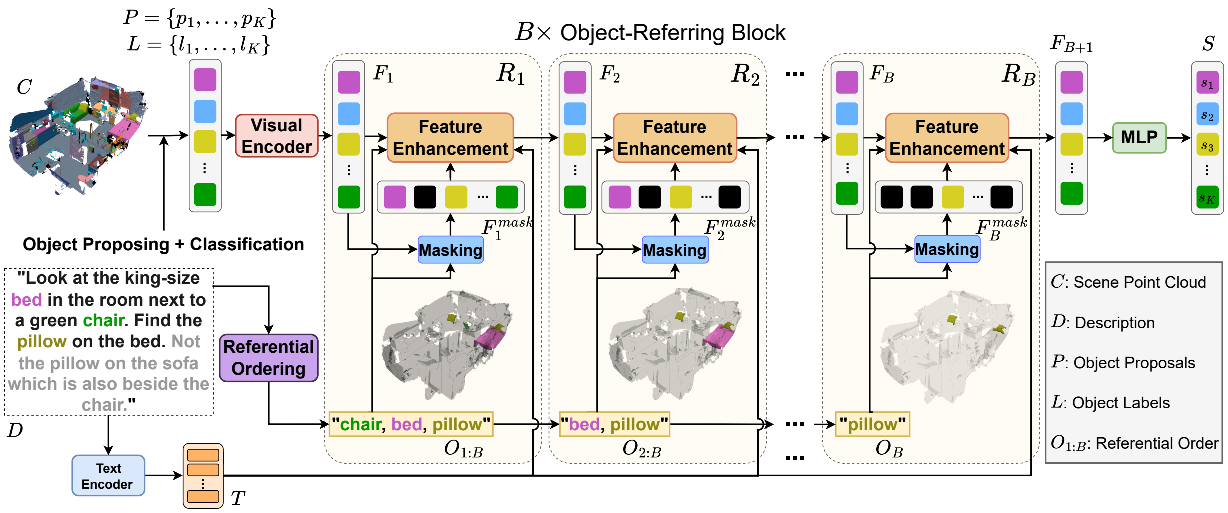


Figure 2. **Architecture of our 3D Visual Grounding Framework with Order-Aware Referring (Vigor).** By taking a point cloud scene C and a natural description D as inputs, our Vigor produces a referential order of anchor/target objects $O_{1:B}$ and conduct Object-Referring blocks $R_{1:B}$ to locate the target object progressively.

3. Methodology

3.1. Problem Formulation and Model Overview

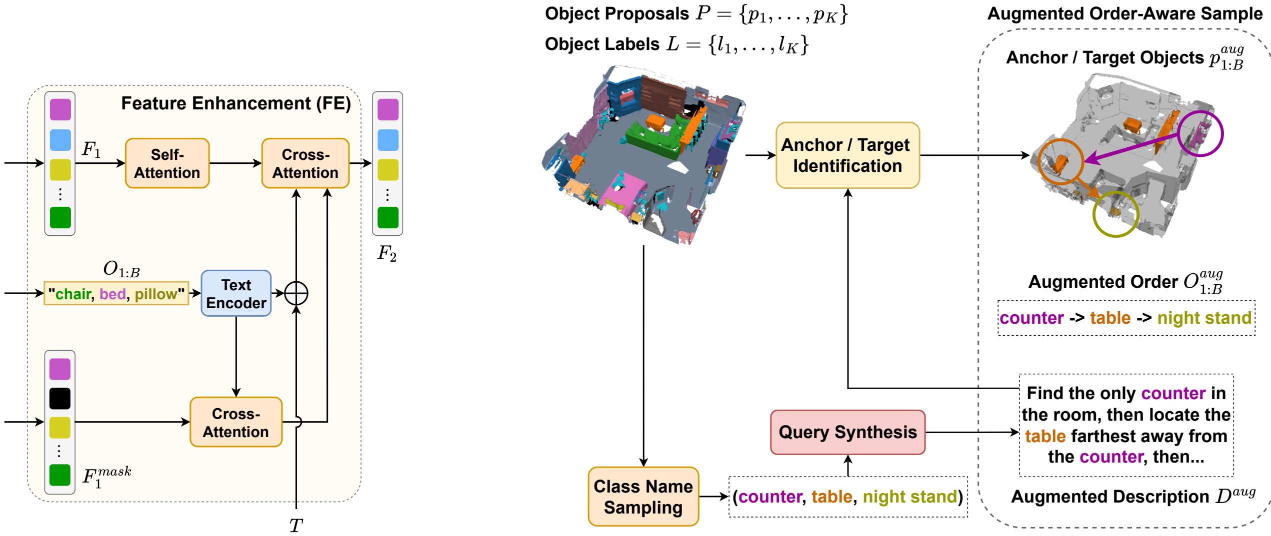
3.1.1 Problem formulation

We first define the setting and notations used in this paper. For each indoor scene, we have a set of colored point cloud $C \in \mathbb{R}^{N\times 6}$, where N denotes the number of the points in the scene, with each point represented in terms of its three-dimensional coordinate and RGB spaces. C is processed to acquire K object proposals $P = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_K\}$ that represent possible objects in the scene, with each proposal containing I points (i.e., $\boldsymbol{p}_n \in \mathbb{R}^{I\times 6}$, $n \in [1, \cdots, M]$). P is obtained either by pre-trained object segmentation networks [21, 32, 42] or directly from the dataset. Along with P, the class labels $L = \{l_1, \cdots, l_K\}$ for all proposals are additionally predicted by a Pointnet++ [33] classifier. For 3D grounding, a text description D is given, illustrating the target object in C by describing its color, shape, or relations to other anchor objects. Given the above inputs, our goal is to identify the exact target object that matches D among all objects in the scene by predicting a K-dimensional confidence score $S = \{s_1, \cdots, s_K\}$ for classifying the target object.

3.1.2 Model overview

As shown in Fig. 2, Vigor is composed of B consecutive Object-Referring blocks $\{R_1, \cdots, R_B\}$ to progressively locate the target object. In particular, by taking both P and D as the inputs, Vigor utilizes the Object-Referring blocks $R_{1:B}$ in Fig. 2 to sequentially produce anchor objects to guide the grounding process. Each R_i takes the object feature $F_i \in \mathbb{R}^{K\times d_i}$ and the text feature $T \in \mathbb{R}^{(|D|+1)\times 768}$ as the inputs. Note that T contains a 1×768-dimensional sentence-level feature and $|D| \times 768$-dimensional word-level feature, where $|D|$ denotes the length of D after tokenization. By observing F_i and T, the Object-Referring block R_i aims to produce the refined feature F_{i+1} along with the updated anchor/target objects and their relations for grounding purposes.

To enable our Object-Referring blocks to capture proper information about the anchor/target objects, we apply a Large Language Model (LLM) to D to generate a Referential Order $O_{1:B} = \{O_1, \cdots, O_B\}$ that mimics human perception system of searching target object [9, 31] by extracting and arranging the class names of the anchor and target objects, similar to [5]. Specifically, $\{O_1, \cdots, O_{B-1}\}$ represent the class names of the anchor objects, and O_B is the class name of the target object (please refer to Supp. F for details of Referential Order generation). Note that $O_{i:B}$ is observed by the i-th Object-Referring block R_i as guidance, which updates the features of anchor/target objects with the proposed Feature Enhancing (FE) module. Since the ground truth referential order is *not* available during training, we introduce a unique warm-up strategy for training Vigor. This is achieved by synthesizing *accurate* referential order and anchor/target object labels. It is worth noting that, with the above design, Vigor can be applied for 3D grounding tasks and achieve satisfactory performances with a respectively limited amount of training data. We now detail the design of our Vigor in the following subsections.



(a) **Feature Enhancement (FE).** Taking R_1 as an example, FE processes the objects described mentioned in $O_{1:B}$ and the relations between them by attending the masked feature F_1^{mask}. Thus, only object features related to $O_{1:B}$ would be refined as F_2.

(b) **Synthesizing a referential order and description for order-aware learning warm-up.** Given P and L, several class names are sampled to construct D^{aug}. The identities of anchor/target objects $p_{1:B}^{aug}$ described in D^{aug} are then located by considering the center coordinates and class name of each proposal. By the above design, the augmented referential order $O_{1:B}^{aug}$ is uniquely determined (i.e., the appearance order of each sampled class name in D^{aug}).

Figure 3. Illustration of feature enhancement and synthesizing warmup data in Vigor.

3.2. 3D Visual Grounding with Order-Aware Object Referring

3.2.1 Object-referring blocks.

Given the object proposals P, the corresponding labels $L = \{l_1, \cdots, l_K\}$, the encoded text features T, and the derived referential order $O_{1:B}$ the as inputs, our Vigor deploys a series of Object-Referring blocks $\{R_1, \cdots, R_B\}$ to perform the grounding task. As depicted in Fig. 2, this referring process is conducted by leaving out an anchor object and updating the visual features in each step until only the final step locates the target object of interest. Thus, the deployment of Object-Referring blocks allows one to focus on the anchor/target objects so that their visual features and spatial relations between them can be exploited while those of irrelevant objects are disregarded.

Take the i-th referring block observing $O_{i:B}$ for example, a *masked feature* $F_i^{mask} = F_i \odot M_i$ is derived by applying a Hadamard product between F_i and a K-dimensional binary mask M_i to replace features of object proposals in F_i not belonging to any of the object classes in $O_{i:B}$. Such a masking strategy ensures that F_i^{mask} contains objects described in $O_{i:B}$ and hence explicitly suppresses the effects of irrelevant objects that are not in our interests. Thus, the j-th entry of M_i (denoted as m_{ij}) is defined as:

$$m_{ij} = \begin{cases} 1 & \text{if class name of } l_j \text{ is in } O_{i:B}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

F_i and F_i^{mask} are refined into F_{i+1} for the next referring block via the feature enhancement module (as discussed later). At the final stage, the output F_{B+1} of R_B is utilized to predict the confidence score S that represents the identity of the target object, supervised by a cross-entropy loss $\mathcal{L}_{ref}$. Additionally, to ensure the text feature T properly describing the anchor/target objects, we follow CoT3DRef [5] and apply the language classification loss $\mathcal{L}_{text}$ to T for matching the associated class labels.

3.2.2 Object feature enhancement

With the above masking process, each object-referring block is expected to update the object features related to the anchor and target objects. This is realized by our attention-based Feature Enhancement (FE) module. To be more precise, in order to update the features associated with the anchor and target objects in R_i according to F_i, F_i^{mask}, T and $O_{i:B}$, our FE module aims to exploit their visual features and spatial relations through attention mechanisms.

Take the FE module in R_1 as an example, as depicted in Fig. 3a, we start with the lower branch which locally emphasizes the features of the potential anchor/target objects via a cross-attention layer by treating F_i^{mask} as value/key and encoded text feature of $O_{1:B}$ (denoted as $T_{O_{1:B}}$) as query. On the other hand, the upper branch of Fig. 3a explores the spatial relations between all objects by treating the self-attended F_1 as the key/value of another cross-attention, with the concatenation of $T_{O_{1:B}}$ and T being query. Finally, an additional cross-attention layer is applied to the previous output features of both branches to obtain the enhanced proposal feature F_2, which enriches not only the information of anchor/target objects but also the relations between them.

On the other hand, to prevent the information extracted from the anchor/target objects from vanishing (i.e., F_2 becomes identical to F_1) during FE, we introduce an additional masking loss $\mathcal{L}_{mask}$ by projecting F_2 from $K \times d_2$-dimensional to $K \times 1$-dimensional digits with MLPs to classify if each proposal in F_2 is previously masked in F_1^{mask}. The masking loss $\mathcal{L}_{mask}$ is defined as:

$$\mathcal{L}_{mask} = \mathcal{L}_{BCE}(MLP(F_2), M_1), \tag{2}$$

where M_1 represents the K-dimensional binary mask as defined in Eqn. 1. It is worth noting that, $\mathcal{L}_{mask}$ is applied to output features of each referring block with a similar formulation to ensure each output feature contains the information of the current anchor/target objects correspondingly.

3.3. Order-Aware Warm-up with Synthetic Referential Order

Although Vigor is designed to produce a referential order of anchor objects for localizing the target object, only the ground truth point cloud information of the target object is given during training. Thus, the above framework is viewed as a weakly-supervised learning scheme since there is *no* ground truth referential order available during training. To provide better training supervision, we warm-up Vigor with a simple yet proper synthetic 3D visual grounding task, where the ground-truth labels of anchor/target objects and descriptions with accurate referential orders can be obtained. This warm-up strategy is presented below.

3.3.1 Augmenting plausible referential order and description

To provide better training supervision and to ensure the reliability of our synthesized data, the constructed description and the corresponding referential order need to be easily and uniquely determined based on anchor/target objects. In our work, we choose to consider spatial relations between objects that are independent of viewpoint (e.g., *"nearest"* or *"farthest"*) as the constructing descriptions, suggesting the referential order of anchor/target objects during this data augmentation stage. As highlighted in Fig. 3b and Algorithm A1 of our supplementary material, given P, L, and B, we construct an augmented referential order $O_{1:B}^{aug}$ by choosing B different class labels $\{l_1^{aug}, \cdots, l_B^{aug}\}$ from L and extracting their class names. The augmented description D^{aug} is then derived as:

"There is a $\{O_1^{aug}\}$ in the room, find the $\{O_2^{aug}\}$ farthest to it, and then find the $\{O_3^{aug}\}$ farthest to that $\{O_2^{aug}\}$, $\{\dots\}$, finally you can see the $\{O_B^{aug}\}$ farthest to that $\{O_{B-1}^{aug}\}$."

Since D^{aug} is constructed following the appearing sequence of object names in $O_{1:B}^{aug}$, it is guaranteed that $O_{1:B}^{aug}$ is a correct referential order w.r.t. D^{aug} and thus can be served as ground truth supervision for pre-training Vigor.

It is worth noting that, to have each object in D^{aug} uniquely defined, we only keep one proposal p_1^{aug} with the class name of $\{O_1^{aug}\}$ in P and remove all the other proposals with that class name. As a result, all the anchor and target objects in P (denoted as $p_{1:B}^{aug} = \{p_1^{aug}, \cdots, p_B^{aug}\}$) according to D^{aug} and their corresponding identities are uniquely determined (i.e., p_2^{aug} is assigned by finding the farthest proposal against p_1^{aug} with label l_2^{aug}, and the rest of the anchor/target objects are determined consecutively with the same strategy).

3.3.2 Warm-up objectives

To have Vigor follow $O_{1:B}^{aug}$ to refer p_i^{aug} in the i-th referring block R_i, we design a coordinate loss $\mathcal{L}_{crd}$ to encourage the output feature F_{i+1} of R_i to identify the coordinate of all proposals in P w.r.t. p_i^{aug}. Thus, we calculate $\mathcal{L}_{crd}$ as:

$$\mathcal{L}_{crd} = \frac{1}{B}\sum_{i=1}^{B} \mathcal{L}_{MSE}(MLP(F_{i+1}), V - \mathbb{I}\cdot v_i), \tag{3}$$

where $MLP(\cdot)$ represents MLP layers applied to F_{i+1}, V is a $K \times 3$ matrix representing center coordinates of calculated bounding-boxes of all K proposals in P, $\mathbb{I}$ stands for a $K \times 1$-dimensional identity vector, and v_i is the 1×3-dimensional center coordinate of p_i^{aug}.

We note that, the referential loss $\mathcal{L}_{ref}$ mentioned in Sec. 3.2 is also extended to classify both the identity of anchor objects using $F_{2:B}$ and the identity of the target object for F_{B+1} during the warm-up process. To this end, we can define the objectives used during our warm-up process by summing up the referential loss $\mathcal{L}_{ref}$ (for both anchor and target objects), the masking loss $\mathcal{L}_{mask}$, the language classification loss $\mathcal{L}_{text}$ and the coordinate loss $\mathcal{L}_{crd}$. With this warm-up strategy, Vigor is initialized to observe relations between anchor/target objects before the subsequent fully-supervised training stage. Later we will verify that, with this pre-training scheme, our Vigor produces satisfactory grounding performances especially when the amount of supervised training data is limited.

Table 1. **Data Efficient Grounding accuracy (%) on NR3D.** Note that each column shows the results trained with a specific amount of training data.

Method	Labeled Training Data			
	1%	2.5%	5%	10%
Referit3D [2]	4.4	13.6	20.3	23.3
TransRefer3D [16]	11.0	16.1	21.9	25.7
SAT [45]	11.6	16.0	21.4	25.0
BUTD-DETR [20]	<u>24.2</u>	<u>28.6</u>	31.2	33.3
MVT [19]	9.9	16.1	21.6	26.5
MVT + CoT3DRef [5]	9.4	17.3	26.5	38.2
ViL3DRel + CoT3DRef [5]	22.4	27.3	<u>33.8</u>	<u>38.4</u>
Vigor (Ours)	**33.5**	**36.1**	**41.5**	**46.0**

Table 2. **Grounding accuracy (%) on the ScanRefer validation set**. In this table, different amounts of training data are considered.

Method	Labeled Training Data			
	5%		10%	
	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer [8]	23.0	12.0	27.4	15.0
3DVG-Trans. [51]	35.3	23.3	39.0	29.0
3D-SPS [29]	28.4	16.9	32.9	22.9
BUTD-DETR [20]	<u>38.2</u>	26.2	<u>40.3</u>	28.3
M3DRef-CLIP [50]	37.2	<u>29.9</u>	40.0	<u>32.4</u>
Vigor (Ours)	**39.8**	**31.5**	**43.6**	**34.9**

3.4. Overall Training Pipeline

We now summarize the training of Vigor. With the warm-up stage noted in Sec. 3.3, we take point cloud data with real-world natural descriptions to continue the training process. Since the identity of anchor objects is unknown, we only apply $\mathcal{L}_{ref}$ (for the target object only), $\mathcal{L}_{mask}$, and $\mathcal{L}_{text}$ as supervision. The overall training pipeline is summarized in Algorithm A2 in supplementary.

4. Experiments

4.1. Dataset

NR3D NR3D [2] dataset consists of 707 indoor scenes in ScanNet [12] with 28715/7485 description-target pairs in the training/testing set, where the descriptions are collected from human annotators. There are 524 different object classes in the scenes in total. NR3D provides ground-truth class-agnostic object proposals, where each point in the scene is properly assigned to its corresponding proposal. As a result, models are only required to classify the target object that uniquely matches the description among all proposals in the scene, with classification accuracy (Acc in %) being the metric.

ScanRefer ScanRefer [8] contains 36665/9508 description-target pairs across a total of 800 indoor scenes in its training/validation set, where the descriptions are also collected from human annotators. Also derived from ScanNet [12] but different from NR3D, perfect object proposals are not available in ScanRefer, and therefore, additional object proposers are required for all methods. Nevertheless, Acc in % under 0.25 and 0.5 intersection over union (IoU) are used as the metrics for ScanRefer. In ScanRefer, since ground-truth object proposals are unavailable, we adopt the visual encoder of M3DRef-CLIP [50] that applies PointGroup [21] to perform object segmentation as proposals P and object classification as labels L. We do not perform object-aware pre-training for ScanRefer since imperfect object proposals may lead to noisy synthetic samples and hinder Vigor's training stability.

4.2. Quantitative Results for Data Efficiency

We present the quantitative results on NR3D and ScanRefer, with consideration of different amounts of available training samples against several baselines by reproducing from their official implementation. Results of NR3D are shown in Table 1. Vigor possesses a considerably superior performance when training with 1% (287), 2.5% (717), 5% (1435), and 10% (2871) NR3D training set samples. Specifically, using only 1% data, Vigor achieves 33.5 overall Acc, which even surpasses SOTA methods with 10% of data. This suggests that Vigor is preferable for 3D grounding, especially when paired training data is limited. Table 2 shows the results on ScanRefer datasets with 5% (1833) and 10% (3666) training samples. Vigor still surpasses all baselines under the conditions where imperfect object proposals are used, indicating the robustness of our methods. The detailed performance on different official subsets of NR3D are in the Supp. B.1.

4.3. Ablation Studies

We ablate different components of our Vigor in Table 3, with performances on NR3D under **1%**, **10%**, and **100%** training samples available to explore the effectiveness of each component with different amounts of data. Baseline model A extracts class names of anchor/target objects from D using a language parser [6] and constructs the referential order according to the appearance of the names in D directly. Also, we employ each of our Referring Blocks R_i in model A without the FE module, i.e., directly using F_i to calculate attention with text features of $O_{i:B}$ and D. Model B enhances the accuracy on 1% and 10% training data by conducting our two-stage object ordering with LLM. When further applying the order-aware pre-training in model C, significant improvements in all settings, especially for 1% available data, are observed. By conducting our pre-training strategy, our Vigor is able to learn foundational concepts of ordering and relations to locate the target objects progressively. Finally, our full model in the last row, incorporating

Table 3. **Ablation studies of proposed components in Vigor.** For methods without LLM Object Ordering, we form $O_{1:B}$ according to the appearance of anchor/target object names in D. Cases of 1%, 10%, and 100% training data are considered.

	FE Module	Order-Aware Pre-training	LLM Object Ordering	1%	10%	100%
A				8.9	35.2	53.9
B			✓	10.3	38.8	53.8
C		✓	✓	25.8	42.5	58.0
Ours	✓	✓	✓	33.5	46.0	59.7

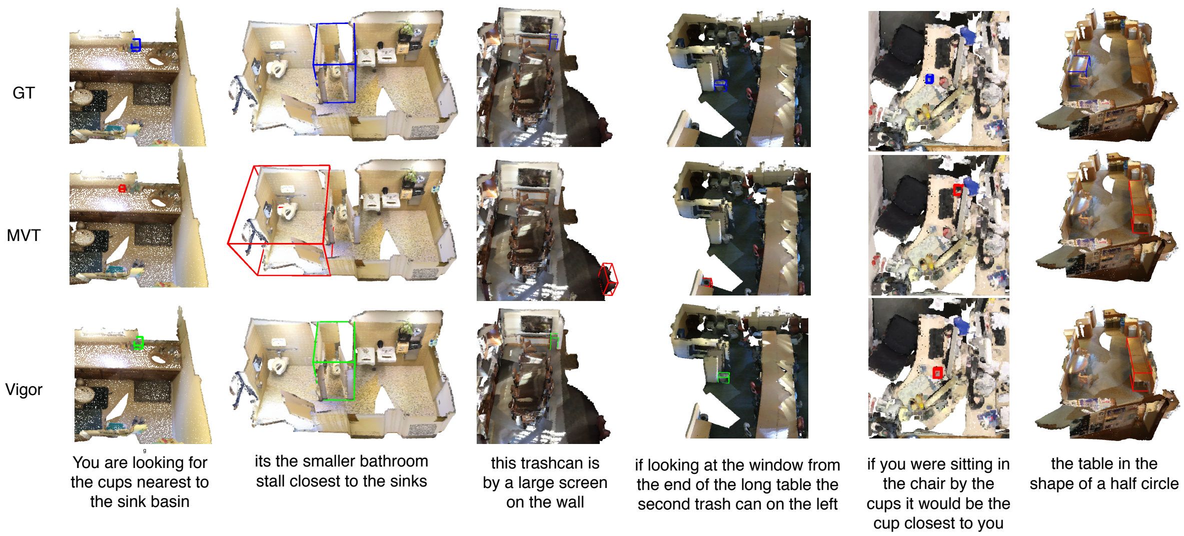


Figure 4. **3D grounding examples of NR3D.** Note that blue/green/red boxes denote ground truth/correct/incorrect predictions. While both MVT and Vigor fail on the last two cases, it is due to the fact that the size of the target object is extremely small (e.g., cup) and the description does *not* describe any anchor objects.

FE modules each R_i to enhance features of anchor/target objects with F_i^{mask}, achieves optimal results on both settings. This verifies the success of our proposed modules and warm-up strategy, especially when the available training data is very limited.

4.4. Qualitative Results

Fig. 4 demonstrates the qualitative results of Vigor on NR3D, with MVT being the baseline. We display four successful cases and two failed cases. It is shown that Vigor can successfully identify the target object referred by one to multiple anchor objects, even in lengthy descriptions. Failed cases include those that refer by shapes or have a very small target object that is hard for Pointnet++ to capture visual information.

5. Conclusions

We presented a data-efficient 3D **Vi**sual **G**rounding Framework with **O**rder-Aware **R**eferring (Vigor) in this paper. Vigor identifies anchor/target objects from the LLM-parsed referential order of input description and guides the updates of the associated object features for grounding purposes. The above process is realized by stacked Object-Referring blocks in Vigor, which progressively process the features of the objects of interest in the above referential order. In addition, a unique warm-up scheme to pre-train Vigor was presented, that augments a pseudo yet desirable series of anchor/target objects and enables Vigor to realize relations between objects before formal training. Experiments on benchmark datasets demonstrate that our Vigor performed favorably against SOTA 3D grounding works in a data-efficient manner.

Acknowledgment. This work is supported in part by the National Science and Technology Council via grant NSTC 112-2634-F-002-007 and NSTC 113-2640-E-002-003, and the Center of Data Intelligence: Technologies, Applications, and Systems, National Taiwan University (grant nos.113L900902, from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) of Taiwan). We also thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

[1] Ahmed Abdelreheem, Kyle Olszewski, Hsin-Ying Lee, Peter Wonka, and Panos Achlioptas. Scanents3d: Exploiting phrase-to-3d-object correspondences for improved visio-linguistic models in 3d scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3524–3534, 2024. 1, 3

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 422–440. Springer, 2020. 1, 3, 7

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3674–3683, 2018. 1

[4] Eslam Bakr, Yasmeen Alsaedy, and Mohamed Elhoseiny. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. 1, 3

[5] Eslam Mohamed Bakr, Mohamed Ayman, Mahmoud Ahmed, Habib Slim, and Mohamed Elhoseiny. Cot3dref: Chain-of-thoughts data-efficient 3d visual grounding. *arXiv preprint arXiv:2310.06214*, 2023. 1, 2, 3, 4, 5, 7

[6] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 7

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the European conference on computer vision (ECCV)*, pages 213–229. Springer, 2020. 3

[8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Proceedings of the European conference on computer vision (ECCV)*, pages 202–221. Springer, 2020. 1, 7

[9] Lang Chen, Matthew A Lambon Ralph, and Timothy T Rogers. A unified model of human semantic knowledge and its disorders. *Nature human behaviour*, page 0039, 2017. 2, 4

[10] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 2, 3

[11] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. 1, 3

[12] 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 (CVPR)*, pages 5828–5839, 2017. 1, 7

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019. 3

[14] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3722–3731, 2021. 1, 3

[15] Zhang Hao, Li Feng, Liu Shilong, Zhang Lei, Su Hang, Zhu Jun, Lionel M. Ni, and Shum Heung-Yeung. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 3

[16] 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 (MM)*, pages 2344–2352, 2021. 3, 7

[17] Joy Hsu, Jiayuan Mao, and Jiajun Wu. Ns3d: Neuro-symbolic grounding of 3d objects and relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2614–2623, 2023. 1, 2, 3

[18] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1610–1618, 2021. 1, 3

[19] 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 (CVPR)*, pages 15524–15533, 2022. 1, 3, 7

[20] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 417–433. Springer, 2022. 1, 3, 7

[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 (CVPR)*, pages 4867–4876, 2020. 4, 7

[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 (ICCV)*, pages 1780–1790, 2021. 1, 3

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-

head, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[24] Ryan B Kochanski, Joseph M Lombardi, Joseph L Laratta, Ronald A Lehman, and John E O'Toole. Image-guided navigation and robotics in spine surgery. *Neurosurgery*, pages 1179–1189, 2019. 1

[25] Lik-Hang Lee, Tristan Braud, Pengyuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, and Pan Hui. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *arXiv preprint arXiv:2110.05352*, 2021. 1

[26] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10975, 2022. 3

[27] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 3

[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 3

[29] 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 (CVPR)*, pages 16454–16463, 2022. 1, 3, 7

[30] Wan-Duo Kurt Ma, Avisek Lahiri, John P Lewis, Thomas Leung, and W Bastiaan Kleijn. Directed diffusion: Direct control of object placement through attention guidance. In *AAAI*, 2024. 2

[31] Jennifer C McVay and Michael J Kane. Conducting the train of thought: working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, page 196, 2009. 2, 4

[32] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9277–9286, 2019. 4

[33] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Proceedings of Advances in neural information processing systems (NeurIPS)*, volume 30, 2017. 4

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[35] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, pages 61–80, 2008. 1

[36] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, pages 70–80, 2015. 3

[37] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11987–11997, 2023. 1, 3

[38] Sanjay Subramanian, Will Merrill, Trevor Darrell, Matt Gardner, Sameer Singh, and Anna Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. 3

[39] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *NeurIPS*, 2023. 1

[40] Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. 3DRP-net: 3D relative position-aware network for 3D visual grounding. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. 1, 3

[41] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19231–19242, 2023. 1, 3

[42] Yizheng Wu, Min Shi, Shuaiyuan Du, Hao Lu, Zhiguo Cao, and Weicai Zhong. 3d instances as 1d kernels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 235–252. Springer, 2022. 4

[43] Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. In *Proceedings of Advances in Neural Information Processing Systems*, 2024. 1, 3

[44] Li Yang, Ziqi Zhang, Zhongang Qi, Yan Xu, Wei Liu, Ying Shan, Bing Li, Weiping Yang, Peng Li, Yan Wang, et al. Exploiting contextual objects and relations for 3d visual grounding. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. 3

[45] 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 (ICCV)*, pages 1856–1866, 2021. 1, 3, 7

[46] Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021. 3

[47] Zhihao Yuan, Xu Yan, Zhuo Li, Xuhao Li, Yao Guo, Shuguang Cui, and Zhen Li. Toward explainable and fine-grained 3d grounding through referring textual phrases. *arXiv preprint arXiv:2207.01821*, 2022. 1, 3

[48] 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 (ICCV)*, pages 1791–1800, 2021. 3

[49] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. 1, 3

[50] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 15225–15236, 2023. 7

[51] 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 (ICCV)*, pages 2928–2937, 2021. 7

[52] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 1

[53] Xiaoyu Zhu, Hao Zhou, Pengfei Xing, Long Zhao, Hao Xu, Junwei Liang, Alexander Hauptmann, Ting Liu, and Andrew Gallagher. Open-vocabulary 3d semantic segmentation with text-to-image diffusion models. *ECCV*, 2024. 1, 2