

Towards CLIP-driven Language-free 3D Visual Grounding via 2D-3D Relational Enhancement and Consistency

Yuqi Zhang* Han Luo* Yinjie Lei✉

College of Electronics and Information Engineering, Sichuan University
 yqcheung@stu.scu.edu.cn luohan@stu.scu.edu.cn yinjie@scu.edu.cn

Abstract

3D visual grounding plays a crucial role in scene understanding, with extensive applications in AR/VR. Despite the significant progress made in recent methods, the requirement of dense textual descriptions for each individual object, which is time-consuming and costly, hinders their scalability. To mitigate reliance on text annotations during training, researchers have explored language-free training paradigms in the 2D field via explicit text generation or implicit feature substitution. Nevertheless, unlike 2D images, the complexity of spatial relations in 3D, coupled with the absence of robust 3D visual language pre-trained models, makes it challenging to directly transfer previous strategies. To tackle the above issues, in this paper, we introduce a language-free training framework for 3D visual grounding. By utilizing the visual-language joint embedding in 2D large cross-modality model as a bridge, we can expediently produce the pseudo-language features by leveraging the features of 2D images which are equivalent to that of real textual descriptions. We further develop a relation injection scheme, with a Neighboring Relation-aware Modeling module and a Cross-modality Relation Consistency module, aiming to enhance and preserve the complex relationships between the 2D and 3D embedding space. Extensive experiments demonstrate that our proposed language-free 3D visual grounding approach can obtain promising performance across three widely used datasets – ScanRefer, Nr3D and Sr3D. Our codes are available at <https://github.com/xibi777/3DLFVG>

1. Introduction

3D Visual Grounding (3DVG) [1, 3–6, 25, 36, 42, 43], also known as referring 3D object localization, aims to accurately locate and identify specific objects within an input point cloud based on provided textual descriptions. The ad-

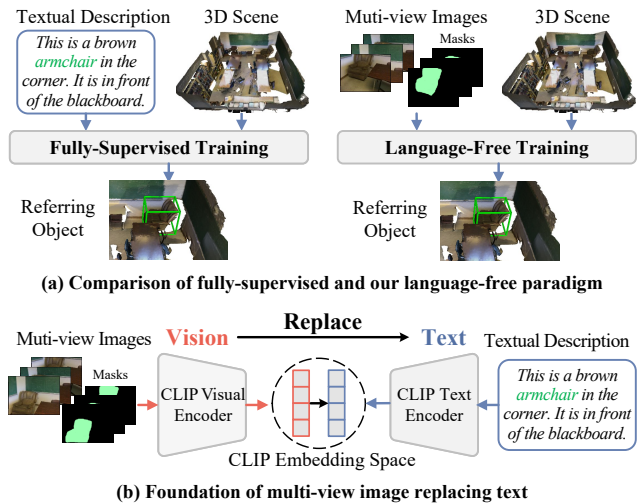


Figure 1. (a) Comparison of fully-supervised and our language-free training paradigm. (b) Based on CLIP embedding space, our language-free training method uses multi-view images corresponding to the scene instead of the textual descriptions.

vancements in visual grounding technology possess the potential to greatly enhance various real-world applications, including autonomous robots and augmented/virtual reality (AR/VR) systems [27, 38]. However, training current 3DVG models demands sufficient detailed text descriptions of each object, which are time-consuming and costly to acquire. As a case in point, the ScanRefer [4] collection, involving 1,929 AMT workers over a month, required approximately 4,984 man hours for both description collection and verification. In light of this, our research pivots towards a promising language-free training paradigm for 3DVG.

In recent years, language-free training methods have been widely explored for numerous 2D vision-language tasks. Early methods typically generate *explicit* pseudo-language to replace human-annotated texts during training. Notably, Nam *et al.* [28] and Jiang *et al.* [16] have respectively developed language-free video localization and image grounding methods. They both utilize off-the-shelf

* Equal contribution

✉ Corresponding Author: Yinjie Lei (yinjie@scu.edu.cn)

detectors [2, 34] and hand-crafted text templates to create simplified sentences as pseudo-language. Since the generated sentences are template-based, such kind of approaches tend to synthesize oversimplified and unnatural language, leading to model overfitting. With the advent of large image-language models [15, 20–22, 32, 33], researchers [19, 26, 45] begin to develop novel *implicit* language-free training methods. By leveraging the intrinsic image-text alignment capability embedded in the pre-trained model (*i.e.*, CLIP [32]), carefully crafted image features are typically used in place of text features for training. This approach circumvents the issues of direct text generation and yields significant performance improvement.

Although language-free training based on implicit feature substitution looks promising for various 2D vision-language tasks, it encounters several specific challenges when applied to 3D point clouds: **(1) Insufficient 3D-language alignment**: given the fact that 3D data (especially 3D-language pairs) is far less abundant than images, there is a lack of 3D pre-trained models that can provide vision-language alignment capability equivalent to CLIP; **(2) Complexity in 3D relation modeling**: different from objects in 2D images, where their positions can be described with simple texts, the descriptions of 3D objects in point clouds commonly include more intricate relational information. However, existing methods tend to neglect the relation modeling during pseudo-language feature synthesis.

To address the above issues, we propose a Language-Free training method for 3D Visual Grounding, named 3DLFVG. As shown in Fig. 1, our key idea is to use multi-view images which are readily available in existing datasets, *e.g.*, ScanNet [8], as input to generate pseudo-language features in place of manually annotated text samples. With the assistance of the image-text feature alignment facilitated by CLIP, our model is trained without language dependency yet able to ground objects described by texts during inference. Besides, in order to enhance the 2D-3D consistent relational expression ability of our model, we further propose a relation injection strategy, which consists of two modules: Neighboring Relation-aware Modeling (**NRM**) and Cross-modality Relation Consistency (**CRC**). NRM aims to inject richer relation information among neighboring 2D masks into the pseudo-language features. Building upon this, CRC is designed to constrain the alignment between 3D proposals and 2D mask relations, thereby endowing 3D models with relational reasoning capability. To summarize, we utilize multi-view images to construct relation-aware pseudo-language features, which serve as a bridge between 3D and language embedding to facilitate language-free 3DVG. Overall, our contributions can be summarized as follows:

- We introduce a CLIP-driven language-free 3DVG framework, which requires no manually annotated texts to effectively achieve 3D visual grounding on point clouds.

- We propose a **NRM** and **CRC** module, respectively to enrich the relational context of the pseudo-language features and enhance the consistency of the relation between 2D and 3D modality. These two modules collaborate to introduce multi-modal aligned relation features.
- Compared with several baselines across multiple datasets (*i.e.*, ScanRefer [4], Nr3D and Sr3D [1]), our approach achieves promising results for language-free 3D visual grounding, demonstrating its effectiveness.

2. Related Works

2.1. 3D Visual Grounding

3D visual grounding aims to locate objects within unstructured point clouds using linguistically formulated queries. Pioneering works such as “ScanRefer” [4] and “ReferIt3D” [1] have introduced 3D grounding datasets, wherein dense object-sentence connections are meticulously annotated on the point cloud dataset ScanNet [8]. These datasets pave the way for language-supervised training.

In general, methods for language-supervised training can be categorized into two main groups. The majority of 3D visual grounding methods [3, 4, 12, 13, 17, 39, 41, 43] adhere to a two-stage pipeline. These methods primarily focus on modeling object positions and relationships. For instance, 3DVG-Transformer [43] leverages transformer-based attention mechanism to achieve interactive fusion between point clouds and language. InstanceRefer [41] employs pre-segmented instances, thereby interacting with language, and comprehensively assessing proposals across three dimensions: attributes, locations, and relationships. Another type is single-stage methods [14, 36] exemplified by 3D-SPS [25]. It departs from the conventional two-stage framework and integrates language for progressive, point-by-point filtering to localize targets within a single-stage. Furthermore, EDA [36] explicitly decouples textual attributes in sentences and conducts a dense alignment between 3D point clouds and detailed linguistic descriptions, which is a feasible way to avoid confusion caused by too many mentioned objects in one sentence.

Nevertheless, all aforementioned methods rely on text supervision. Given the substantial expense associated with annotating language for 3D scenes, our proposed model is dedicated to learning the localization of target objects within 3D scenes without any reliance on text annotations.

2.2. Language-Free Paradigm

In the training for multi-modal tasks, the acquisition of high-quality visual-language training samples poses a significant challenge. To address this hurdle, certain research endeavors [28, 45] have introduced the concept of “language-free” training paradigm. This paradigm eliminates the necessity for language data in the training of

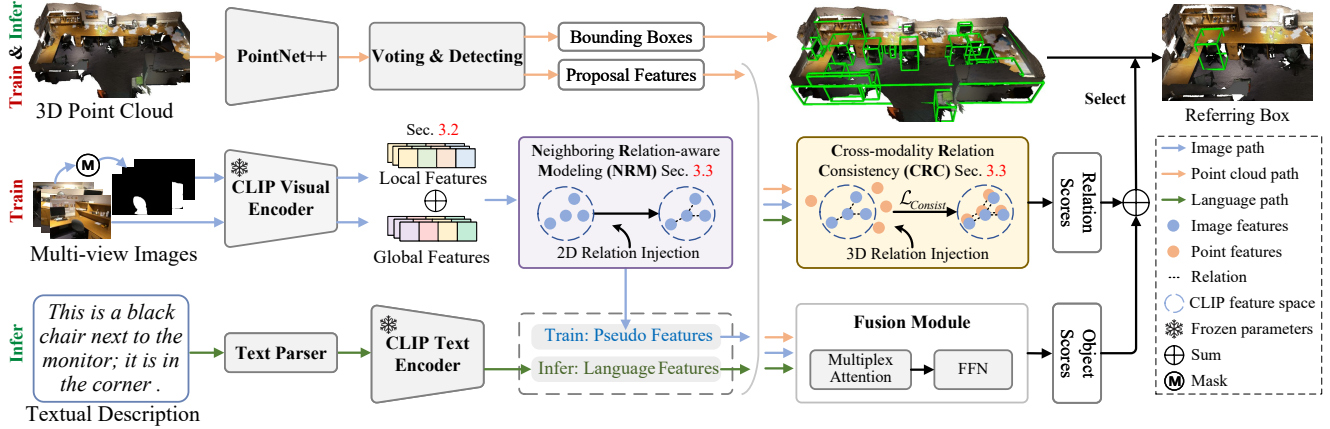


Figure 2. The overview of our method. During training, we first encode 3D point cloud and multi-view images with point cloud encoder and CLIP visual encoder separately. After that, we extract global and local features of the images (Sec. 3.2), and inject them with modeled neighboring relation by **NRM** (Sec. 3.3). Then we additionally introduce **CRC** module (Sec. 3.3) to model the relation of 3D proposals and enforce a consistency constraint between 2D and 3D. After training, we can perform inference of grounding through the textual description.

visual-language models, offering great convenience.

The most common language-free training approaches [10, 16, 28] are to generate pseudo texts paired with visual inputs. For example, Unsupervised Image Captioning [10] generates a pseudo caption for each image by leveraging the visual concept detector, and initializes the image captioning model using the pseudo image-sentence pairs. Pseudo-Q [16] uses a pre-trained detector to detect objects in the image and constructs pseudo language queries based on the relation among objects. However, it is noteworthy that the effectiveness of these methods heavily relies on the quality of designed text templates and pre-trained object detectors. This could be problematic due to the oversimplified language and domain gap between target datasets and object detector training datasets. Several other methods [19, 26, 45] leverage the pre-trained visual-language models, such as CLIP [32]. These methods take into account that CLIP has acquired the ability of aligning image and text features within the embedding space. As a result, they employ CLIP to implicitly generate pseudo text features directly from the image. Drawing inspiration from these approaches, we harness the CLIP embedding space for the analysis of 3D scenes, enabling the accomplishment of 3D visual grounding.

3. Methodology

In this section, we detail our proposed method for language-free training in 3D visual grounding. Sec. 3.1 introduces the language-free training paradigm, along with an overview of the proposed framework. Sec. 3.2 discusses how we extract local-global features from images as pseudo-language features. In Sec. 3.3, we describe the methods for augmenting

the pseudo-language features with more neighboring relation information and the construction of 2D and 3D relational consistency constraints. Finally, Sec. 3.4 describes the model’s training and inference process.

3.1. Overview

The overall pipeline of our language-free 3DVG is shown in Fig. 2. During training phase, the inputs consist of two parts: a point cloud $\mathbf{P} \in \mathbb{R}^{N \times (3+F)}$ (with 3D coordinates and F -dimensional auxiliary features) of N points, and corresponding multi-view images $\mathbf{M} = \{\mathbf{I}_i\}_{i=1}^{N_I}$, where N_I is the total number of the images. At inference stage, the inputs shift to include a point cloud $\mathbf{P} \in \mathbb{R}^{N \times (3+F)}$ and a sentence query $\mathbf{Q} \in \mathbb{R}^L$ designed to describe the target object. The objective of our method is to train a model to localize specified objects without using any language queries during training, yet capable of identifying targets described by texts in the inference phase. Similar to the regular 3DVG methods [3, 4, 12, 13, 17, 39, 41, 43], our model outputs a 3D bounding box $\mathbf{B} = \{\mathbf{c}, \mathbf{s}\}$ of the referring object, with center $\mathbf{c} = (c_x, c_y, c_z)$ and the size $\mathbf{s} = (s_x, s_y, s_z)$.

Our language-free 3DVG training framework comprises three key modules: Pseudo-Language Feature Generation (PFG), Neighboring Relation-aware Modeling (NRM), and Cross-modality Relation Consistency (CRC). The pseudo-language feature generation module utilizes the shared image-text embedding space to create visual features that can replace text features. These features are then fused with proposal features, produced by 3D Voting and Detecting part in [43], to generate object confidence scores. The neighboring relation-aware modeling module introduces dynamic 2D relations between adjacent masks, echoing the relational intricacies that are often present in textual

features. It produces mask relation features by facilitating interactions between the main object and its adjacent objects. Meanwhile, the cross-modality relation consistency module strengthens the alignment between 3D proposals and 2D mask relations, thus ensuring consistent relational understanding across modalities. It models the relation between each proposal in the scene and surrounding proposals to derive proposal relation features. By comparing these with the 2D mask relations, relation matching scores can be computed. Finally, we combine these scores with object confidence scores to calculate the final object-matching scores. Since our method capitalizes on the image-text feature alignment provided by CLIP, and incorporates extra modules that enhance the features with relation-aware capabilities. So, it ensures an efficient implementation of 3DVG given the texts during the inference phase. We will elaborate the each component in the following sections.

3.2. Pseudo Language Feature Generation

To train our language-free 3DVG model effectively, we need pseudo-language features that mirror the role of an actual text description by pinpointing the target object and its environmental context. For example, a typical description in ScanRefer dataset like “a brown table in the center of the room”. This inspires us to split our pseudo-language features into two parts: the local features that identify “a brown table” and the global features that describe its surroundings “in the center of the room”.

Local-Global Feature Generation. Firstly, we begin to generate $M_i = \{M_i^m\}_{m=1}^{N_M}$ class-agnostic masks across each multi-view image I_i with a pre-trained 2D mask generator [31, 35, 44], N_M is the mask number. We then isolate and crop the image around each mask, aiming to concentrate on the area delineated by the mask. Following this, the cropped image C_i^m is fed into the CLIP visual encoder [32] to distill the local features \mathbf{f}_{pseudo}^L of the object of interest.

Beyond the local features, we also extract the global features to include wider context information from the environment. Building on [40], we modify the CLIP visual encoder to obtain global features:

$$\mathbf{f}_{pseudo}^G = \mathcal{E}_{later}(\mathcal{E}_{former}(I_i) \odot M_i^m), \quad (1)$$

where \mathcal{E}_{former} represents the former l layers, and \mathcal{E}_{later} denotes the rest layers of CLIP visual encoder. Subsequently, we calculate the local-global visual features $\mathbf{f}_{pseudo}^V = \alpha \mathbf{f}_{pseudo}^L + (1 - \alpha) \mathbf{f}_{pseudo}^G$, α is a hyper-parameter that balances the proportion of local and global features.

Random Noise. So far, we have obtained the local-global visual features. However, it has been proved that merely utilizing the visual features might not adequately mimic actual language characteristics [45]. Consequently, we deliberately add random noise to perturb the local-global features from the pre-trained visual encoder to get the pseudo-

language features $\mathbf{f}_{pseudo} \in \mathbb{R}^{D_c}$, where D_c is the dimension of the CLIP encoded features.

Ultimately, we utilize multi-view images and mask proposals to create pseudo-language features for the target objects. However, considering the pre-trained mask generator will produce some extremely incomplete objects and prefer to notice larger entities, We further select the partial masks that best match the target boxes in the scene to ensure the representative of pseudo-language features. After that, we fuse these features with the proposal features derived from the 3D Voting and Detecting part to compute the final object confidence scores. At this point, we have established the baseline of language-free 3D visual grounding.

3.3. Relation Injection

With the pseudo-language features in hand, we can already initiate language-free 3DVG training. However, natural textual descriptions typically convey not just the target object’s local and global information but also its relation with neighboring objects — take an example from the ScanRefer [4] dataset, “The trash can sits along the wall in the kitchen next to the console table under the TV”. Such textual sample includes detailed descriptions of 3D spatial relationship among objects, which are often missing in 2D pre-training model like CLIP [32]. To bridge this gap and enhance the relation representation ability of our CLIP-driven pseudo-language features, we further introduce a neighboring relation-aware module and a cross-modality relation consistency module.

Neighboring Relation-aware Modeling. Since textual descriptions in 3DVG often detail the relation among the target object and its neighbors, we first identify the k nearest mask proposals to each generated mask from Section 3.2. Due to the limited number of objects and lack of complex 3D position information in a single multi-view image, we project the 2D mask proposals into 3D space to acquire more reliable spatial relation. Specifically, we use world-to-camera extrinsic and camera intrinsic parameters following [7] to map the center points of the 2D masks from all multi-view images in one scene to 3D space. Through this step, we locate the k adjacent instances for each mask proposal. Subsequently, we extract the local features of the mask proposal along with the features of the k nearest masks. From the interplay between these local features $\mathbf{f}_{pseudo}^L \in \mathbb{R}^{D_c}$ and neighboring features $\mathbf{f}_{neigh}^{2D} \in \mathbb{R}^{k \times D_c}$, we derive each mask’s 2D relational features as follows:

$$\mathbf{f}_{R2D} = \text{Linear}(\text{CrossAtt}(\mathbf{f}_{pseudo}^L, \mathbf{f}_{neigh}^{2D})). \quad (2)$$

Since there is no supervision of this relation during our training process, we introduce the proxy task of predicting the target object to achieve the optimization of neighboring relation awareness:

$$\mathcal{L}_{cls} = \text{CE}(\mathbf{f}_{R2D}, y_{cls}), \quad (3)$$

where y_{cls} represents the category of the target object.

Cross-modality Relation Consistency. Having integrated relational information into pseudo-language features, we then consider a similar enhancement for the 3D feature space. Through Voting and Detecting module [43], we obtain the proposal features $\mathbf{F}_{proposal} = \{\mathbf{f}_{proposal}^p\}_{p=1}^{N_p}$ of the point cloud, N_p denotes the number of 3D proposals. Initially, we assess the confidence score of each 3D proposal as a potential object center, selecting the most probable ones. For each chosen proposal, we identify k adjacent proposals to learn the 3D relational features:

$$\mathbf{f}_{R3D} = \text{Linear} \left(\text{CrossAtt} \left(\mathbf{f}_{proposal}^p, \mathbf{f}_{neigh}^{3D} \right) \right), \quad (4)$$

where $\mathbf{f}_{proposal}^p$ is the feature of one proposal, and \mathbf{f}_{neigh}^{3D} is its neighboring features. We now have the neighboring relation features for each 3D objectness proposal. The relational features of the 3D proposal corresponding to the target object indicated by a 2D mask should align with the 2D mask’s relational features:

$$\mathcal{L}_{\text{consist}} = -\frac{1}{N_p} \sum_{k=1}^{N_p} \left[\log \frac{\exp(\mathbf{f}_{R3D}^k \cdot \mathbf{f}_{R2D}^k / \tau)}{\sum_{n=1}^{N_p} \exp(\mathbf{f}_{R3D}^k \cdot \mathbf{f}_{R2D}^n / \tau)} \right]. \quad (5)$$

Ultimately, by enforcing a consistency constraint on relational features between modalities, we can more effectively disambiguate the grounding outcomes.

3.4. Training and Inference

Our language-free 3DVG method follows distinct training and inference processes as shown in Fig. 2. Here we first detail the network training objectives of learning with pseudo-language features, and then outline the inference process using point clouds with authentic language queries.

Training. We train the point cloud encoder with a detection loss \mathcal{L}_{Det} and a matching loss \mathcal{L}_{VG} as in [43]. For the proxy task in Neighboring Relation-aware Modeling, we predict the target object category using \mathcal{L}_{cls} . We also introduce the $\mathcal{L}_{\text{consist}}$ to constrain the consistency of the relationship between 2D and 3D modality.

Inference. Different from the training process, at the inference stage, an input is a point cloud and its corresponding complete sentences from the test set. To minimize the discrepancy between training and inference, we also utilize CLIP to extract both local and global text features. Specifically, we employ a pre-trained text parser [11] to identify the sentence’s target noun for local features encoding with the CLIP text encoder. Additionally, we process the entire sentence through CLIP to capture global features. And we combine the local and global features to derive the comprehensive language features. For relation injection during the inference stage, we diverge from the training approach by employing the text parser [11] to extract and encode neighboring entities from the text. Note that our training isn’t

tailored to any specific grounding dataset. Instead, we’ve produced a generalized embedding space, allowing us to infer directly on any grounding dataset without retraining.

4. Experiments

4.1. Datasets

Datasets. To validate the efficacy of our 3DLFVG method, we conducted evaluations using two widely recognized datasets: ScanRefer [4] and Nr3D/Sr3D [1] from the ReferIt3D. Our approach to point cloud-based visual grounding is unique in that it operates under a language-free paradigm; thus, textual descriptions were not employed during training but were utilized solely during testing.

ScanRefer. This dataset comprises 51,583 manually crafted descriptions for 11,046 objects across 800 scenes from the ScanNet [8]. On average, each scene features approximately 13.81 objects, each accompanied by 64.48 annotations. The correctness and distinctiveness of the data are preserved through the use of skilled annotators and trained verifiers. We follow the ScanRefer benchmark to divide our dataset into the train/val/test set with 36,655, 9,508, and 5,410 samples respectively, and utilize val set to evaluate our framework. The evaluation metrics of the dataset are Acc@0.25 IoU and Acc@0.5 IoU. These metrics are reported for both unique and multiple object categories.

Nr3D and Sr3D. Nr3D and Sr3D are subsets in ReferIt3D. Specifically, Nr3D is composed of 41,503 samples obtained through ReferItGame, while Sr3D encompasses 83,572 samples created using synthetic templates. We take the standard splits for Nr3D and Sr3D, using only the val sets for evaluation. Our method is assessed with metrics for “easy” and “hard” categories, along with “view-dep.” and “view-indep.” subsets based on the dependency of descriptions on the camera viewpoint.

4.2. Implementation Details

In our practice, we employ the PointNet++ [30] backbone along with the Object Proposal Generation Module module in 3DVG-Transformer [43] to generate 3D proposals and corresponding bounding boxes. For generating pseudo-language features, we utilize the CLIP [32] visual encoder from the ViT-B/16 model. During the training stage, we initially train our baseline model on the ScanNet [8] dataset for 200 epochs, followed by a further 50 epochs to train our NRM and CRC modules. The model is trained with the AdamW [24] optimizer and a batch size of 8. The learning rates for proposed *Neighboring Relation-aware Modeling* and *Cross-modality Relation Consistency* are empirically set at $2e-3$. For the voting & grouping module, detection head, and cross-modal fusion module, the learning rates are set as $2e-3$, $1e-4$, and $5e-4$, respectively, following the 3DVG-Transformer [43]. During inference, we use the

Table 1. Quantitative comparison of language-free (LF) 3DVG on ScanRefer [4] dataset. Results of relevant fully supervised (Fully) methods are also provided. Accuracy (Acc) under 0.25 and 0.5 IoU thresholds in “Unique”, “Multiple”, and “Overall” is reported respectively. Without language supervision, our method significantly outperforms previous methods. † indicates our re-implemented method on 3D.

Methods	Publication	Setting	Unique		Multiple		Overall	
			Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer [4]	ECCV2020	Fully	76.33	53.51	32.73	21.11	41.19	27.40
3DVG-Transformer [43]	ICCV2021	Fully	83.25	61.95	41.20	30.29	49.36	36.43
Random	-	LF	5.57	3.94	3.58	2.33	3.97	2.65
OpenScene [29]	CVPR2023	LF	27.60	-	8.60	-	13.00	5.10
LLM-Grounder [37]	arXiv2023	LF	33.60	-	12.10	-	17.10	5.30
Pseudo-Q† [16]	CVPR2022	LF	58.07	38.45	19.49	11.82	26.90	16.96
Zero-shot-RIS† [40]	CVPR2023	LF	61.12	45.90	18.34	13.85	26.67	20.02
Ours (3DLFVG)	-	LF	65.80	51.27	22.03	16.94	30.53	23.61

Table 2. Quantitative comparison of language-free 3DVG on Nr3D and Sr3D [1] datasets. We report accuracy (Acc) for the IoU@ m ($m \in \{0.25, 0.5\}$) metrics in several subsets and “Overall”. “Easy” and “Hard” mean whether there are more than 2 instances from the same category in the scene. “View-dep.” and “View-indep.” refer to whether the reference expressions are dependent of the camera view.

Method	Easy		Hard		View-dep.		View-indep.		Overall	
	$m=0.25$	$m=0.5$	$m=0.25$	$m=0.5$	$m=0.25$	$m=0.5$	$m=0.25$	$m=0.5$	$m=0.25$	$m=0.5$
Nr3D										
Random	6.70	2.40	6.34	2.75	6.59	2.91	6.47	2.41	6.51	2.59
Pseudo-Q† [16]	16.70	8.83	10.44	6.12	12.59	6.54	15.00	8.22	14.23	8.07
Zero-shot-RIS† [40]	23.56	16.52	11.57	7.99	14.40	10.19	18.65	13.56	16.99	11.89
Ours (3DLFVG)	22.35	16.53	13.15	8.96	15.10	11.13	19.01	14.02	18.28	13.32
Sr3D										
Random	8.81	5.66	7.57	4.97	7.28	4.80	8.65	5.61	8.17	5.30
Pseudo-Q† [16]	12.45	7.31	10.70	7.48	3.36	2.04	12.20	7.56	11.74	7.22
Zero-shot-RIS† [40]	20.13	15.62	12.27	9.09	12.18	10.18	17.51	13.79	17.43	13.47
Ours (3DLFVG)	21.00	16.63	15.16	11.51	11.17	9.92	19.07	14.82	19.25	14.99

same ViT-B/16 model to encode the local and global text features. The input point number N , the proposal number N_p , and the neighboring objects number k are set to 40000, 256 and 4, respectively. The balance weight α is set to 0.15. We train and evaluate our model using “xyz + normals + multiviews” inputs. All experiments are conducted using PyTorch on a single NVIDIA RTX 3090.

4.3. Compared Methods

Random. With access to all boxes produced by Object Proposal Generation Module, we randomly choose one box to represent the predicted result.

Pseudo-Q. Pseudo-Q [16] is currently a method that has achieved good performance in 2D language-free grounding. We replicated its main idea into a 3D scene as a comparative experiment. Due to the distinction between the 3D scene and the image, there are also some differences in reproduction details. The principal deviation lies in our utilization of

the Group-Free [23] model as the detector for 3D scene objects. Furthermore, we have incorporated global positional information into the generation of pseudo-language.

OpenScene. OpenScene [29] propose an open-vocabulary 3D scene understanding method. It aligns multi-view image features with point cloud features at a pixel-point level, facilitating text query grounding through cosine similarity between CLIP text embeddings and individual points. To produce bounding boxes by OpenScene, DBSCAN clustering [9] can be used on points with high cosine similarity and draw boxes around them. Given its ability to perform 3DVG without text-based training, akin to our proposed paradigm, OpenScene serves as a benchmark for comparison.

LLM-Grounder. LLM-Grounder [37] employs an open-vocabulary Large Language Model for 3D visual grounding tasks. It deconstructs complex natural language queries into their semantic parts using an LLM and then applies visual grounding techniques, including OpenScene [29] or LERF

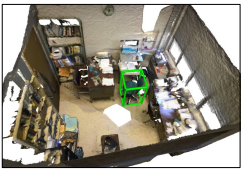
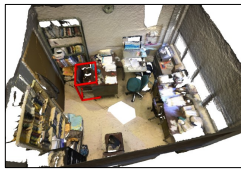
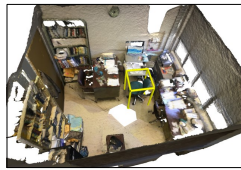
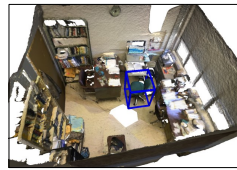








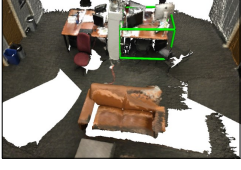
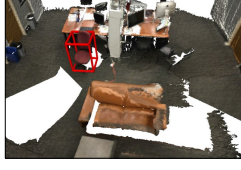
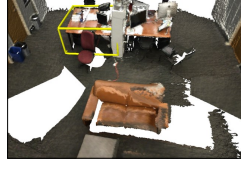

Description	Ground-truth	Pseudo-Q [16]	Zero-shot-RIS [37]	Ours
(a) <i>It is a blue desk chair with a black base and handles. It is positioned in the right corner of the room, in front of a gray desk.</i>				
(b) <i>The brown couch is in between the piano and the glass double doors. The brown couch is against the white wall.</i>				
(c) <i>A blue table. It is behind the blue couch.</i>				
(d) <i>This is a brown table with three computer monitors on it. It does not have a red chair near it.</i>				

Figure 3. Qualitative results from Pseudo-Q [16], Zero-shot-RIS [40] and our method. The GT boxes are marked in green. Boxes predicted by Pseudo-Q, Zero-shot-RIS, and ours are marked in red, yellow, and blue respectively.

[18], to locate objects within a 3D scene. Given its independence from textual training data, LLM-Grounder is utilized as a comparative benchmark to our method.

Zero-shot RIS. Zero-shot RIS [40] proposes a method for zero-shot referring image segmentation, using pre-trained cross-modal knowledge. This approach employs CLIP to encode local-global context features for images and text descriptions respectively. Then it segments fine-detailed instance-level groundings by calculating the similarity between these two features. Due to its capability to capture both the target object’s attributes and its environmental context, we adapt this method to 3D as a baseline for our study.

4.4. Quantitative Comparison

We show quantitative comparisons of our 3DLFVG and aforementioned methods on ScanRefer [4] and Nr3D/Sr3D [1] in Tab. 1 and Tab. 2 respectively. The results reveal significant improvements in overall accuracy (Acc@0.25 and Acc@0.5) over other baselines. The evaluation analysis highlights the following key observations: 1) Our method markedly outperforms the Random approach across all datasets, even nearly reaching the upper limit of ScanRefer. This indicates the efficacy of our framework and affirms the feasibility of language-free training in the 3D

visual grounding task. 2) The results show that our method outperforms other well-designed methods by a large margin under the language-free training setting. It’s worth mentioning that our method greatly surpasses other approaches on the “multiple” subset, which demonstrates our model’s capability to effectively perceive and model intricate relations within both 2D and 3D embedding spaces.

4.5. Qualitative Comparison

Fig. 3 visualizes the representative language-free visual grounding results of the Pseudo-Q [16], Zero-shot-RIS [40] and our method. These examples demonstrate that our method achieves more reliable 3D object localization results than explicit text generation or implicit feature substitution methods transferred from 2D fields. Taking the visualization results from the fourth row as an example, the Pseudo-Q method incorrectly grounds “red chair” instead of “brown table” mentioned in the text description and the Zero-shot-RIS wrongly predicts other entities belonging to the same category “table”. In contrast, our method correctly identifies the target object. The visualization results demonstrate that, in response to complex textual descriptions involving multiple objects, our model can accurately perceive the intricate relationships among objects interwoven in the

Table 3. Ablation study on main components of our method. We report the “overall” results in terms of Acc@0.25 and Acc@0.5.

PFG		Relation		Acc@0.25	Acc@0.5
LGFG	Noise	NRM	CRC		
✓	×	×	×	26.88	22.30
✓	✓	×	×	28.27	20.62
✓	×	✓	×	28.88	22.81
✓	✓	✓	×	29.60	23.02
✓	×	✓	✓	30.35	23.44
✓	✓	✓	✓	30.53	23.61

Table 4. Comparison on different 3D visual grounding baseline methods. We only report the “overall” results.

Method	Baseline	Acc@0.25	Acc@0.5
Pseudo-Q [16]	3D-SPS [25]	25.19	17.16
Ours	3D-SPS [25]	26.75	21.03
Random	3DVG-Transformer [43]	3.97	2.65
Pseudo-Q [16]	3DVG-Transformer [43]	26.91	16.94
Ours	3DVG-Transformer [43]	30.53	23.61

3D scene, thus promoting more effective inference.

4.6. Ablation Study

We conducted ablation studies on the ScanRefer dataset to further evaluate the effectiveness of our proposed method.

Components analysis. We conducted a detailed ablation study to dissect the impact of various components in our language-free 3DVG model, with Tab. 3 illustrating the performance of different module combinations. As shown in 3, Our module can be divided into two parts: PFG and Relation, where PFG represents the Pseudo-Language Feature Generation and Relation Injection. Upon further refinement, “LGFG” is the local-global features generation part, and “Noise” indicates the addition of random perturbations to the pseudo-language features generated by LGFG. The first row presents results achieved by merely substituting text with pseudo-language features during training. Note that we omit the validation of solely adding CRC module, as the consistency of neighboring relations between 3D and 2D modalities must be established on the basis of having already utilized the NRM module to construct 2D neighboring relations. The results reveal consistent performance improvements with the addition of each component, confirming the utility of our proposed modules.

Effects of different 3D baselines. In addition, to assess the flexibility of our proposed language-free training approach, we implemented it across various 3DVG baseline models. Our experimental setup involved the use of 3DVG-Transformer and 3D-SPS as the representative models for two-stage and single-stage methods, respectively. As shown

Table 5. Ablation study on different numbers (k) of neighboring objects in the NRM module. Here \mathcal{A} refers to Acc.

k	Unique		Multiple		Overall	
	$\mathcal{A}@0.25$	$\mathcal{A}@0.5$	$\mathcal{A}@0.25$	$\mathcal{A}@0.5$	$\mathcal{A}@0.25$	$\mathcal{A}@0.5$
0	64.38	50.54	20.76	15.93	29.22	22.62
2	63.90	49.65	21.32	16.46	29.58	22.90
4	65.80	51.57	22.03	16.94	30.53	23.61
6	65.50	51.06	21.60	16.78	30.12	23.43

in Tab. 4, our model performs better than the Pseudo-Q [16] on different baselines. This performance superiority not only demonstrates the robustness of our method but also its adaptability in enhancing different types of 3DVG models.

Numbers of neighboring objects in NRM. The influence of neighboring object numbers in NRM is analyzed by setting different $k \in \{0, 2, 4, 6\}$. As illustrated in Tab. 5, we find that selecting 4 neighbor objects yields the most effective results in relation modeling, which indicates the optimal number of neighbors to aggregate sufficient relational information. When too few neighboring objects, such as 0 or 2, are considered, the model may not capture enough contextual details to accurately understand the complex relationships present in the 3D point cloud. Conversely, since referring descriptions seldom mention too many (*e.g.*, 6) surrounding objects in one single sentence, a much higher value of k might introduce discrepancies between image and text relation modeling.

5. Conclusion

This paper presents a novel language-free training framework for 3DVG, eliminating the need for text annotations. This approach addresses the practical challenges of labor-intensive and time-consuming annotations. Our framework leverages the joint embedding capabilities of CLIP, using it as a bridge to generate pseudo-language features from multi-view images that closely mimic real text descriptions. To enhance the understanding of spatial relationships, we incorporate a Neighboring Relation-aware Modeling module and a Cross-modality Relation Consistency module. These modules are designed to effectively enhance and preserve relations between 2D and 3D modalities. Extensive experiments conducted on mainstream datasets demonstrate the robustness and efficiency of our approach.

Acknowledgement: This work was supported by the National Natural Science Foundation of China (No.U23B2013), was also partially supported by the National Natural Science Foundation of China (No.62276176). We would like to express our sincere gratitude to Professor Yifan Liu for her insightful guidance, may she rest in peace. We would also like to thank Zhao Jin and Yuwei Yang for their useful discussions and suggestions.

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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 422–440, 2020. 1, 2, 5, 6, 7
- [2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 6077–6086, 2018. 2
- [3] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16464–16473, 2022. 1, 2, 3
- [4] 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, 2020. 1, 2, 3, 4, 5, 6, 7
- [5] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D 3 net: A unified speaker-listener architecture for 3d dense captioning and visual grounding. In *European Conference on Computer Vision (ECCV)*, pages 487–505. Springer, 2022.
- [6] Jiaming Chen, Weixin Luo, Xiaolin Wei, Lin Ma, and Wei Zhang. Ham: Hierarchical attention model with high performance for 3d visual grounding. *arXiv preprint arXiv:2210.12513*, 2022. 1
- [7] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 4
- [8] 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. 2, 5
- [9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 6
- [10] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4125–4134, 2019. 3
- [11] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1373–1378, 2015. 5
- [12] 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 (AAAI)*, pages 1610–1618, 2021. 2, 3
- [13] 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. 2, 3
- [14] 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, 2022. 2
- [15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2
- [16] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15513–15523, 2022. 1, 3, 6, 7, 8
- [17] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10984–10994, 2023. 2, 3
- [18] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lrf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19729–19739, 2023. 7
- [19] Dahye Kim, Jungin Park, Jiyoung Lee, Seongheon Park, and Kwanghoon Sohn. Language-free training for zero-shot video grounding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2539–2548, 2023. 2, 3
- [20] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, pages 9694–9705, 2021. 2
- [21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR, 2022.
- [22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2
- [23] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2949–2958, 2021. 6

- [24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018. 5
- [25] 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, 2, 8
- [26] Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. Deltaedit: Exploring text-free training for text-driven image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6894–6903, 2023. 2, 3
- [27] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 1
- [28] Jinwoo Nam, Daechul Ahn, Dongyeop Kang, Seong Jong Ha, and Jonghyun Choi. Zero-shot natural language video localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1470–1479, 2021. 1, 2, 3
- [29] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023. 6
- [30] 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)*, pages 5099–5108, 2017. 5
- [31] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19446–19455, 2023. 4
- [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 (ICML)*, pages 8748–8763, 2021. 2, 3, 4, 5
- [33] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 2
- [34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2015. 2
- [35] Xinlong Wang, Zhiding Yu, Shalini De Mello, Jan Kautz, Anima Anandkumar, Chunhua Shen, and Jose M Alvarez. Freesolo: Learning to segment objects without annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14176–14186, 2022. 4
- [36] 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, 2
- [37] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. *arXiv preprint arXiv:2309.12311*, 2023. 6
- [38] Yuwei Yang, Munawar Hayat, Zhao Jin, Chao Ren, and Yinjie Lei. Geometry and uncertainty-aware 3d point cloud class-incremental semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21759–21768, 2023. 1
- [39] 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. 2, 3
- [40] Seonghoon Yu, Paul Hongsuck Seo, and Jeany Son. Zero-shot referring image segmentation with global-local context features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19456–19465, 2023. 4, 6, 7
- [41] 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. 2, 3
- [42] 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 (ICCV)*, pages 15225–15236, 2023. 1
- [43] 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. 1, 2, 3, 5, 6, 8
- [44] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision (ECCV)*, pages 350–368. Springer, 2022. 4
- [45] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17907–17917, 2022. 2, 3, 4