Rethinking 3D Dense Caption and Visual Grounding in A Unified Framework through Prompt-based Localization

Yongdong Luo[1,*], **Haojia Lin**[1,*], **Xiawu Zheng**[1], **Yigeng Jiang**[1], **Fei Chao**[1], **Jie Hu**, **Guannan Jiang**, **Songan Zhang**, **Rongrong Ji**[1],

[1]Xiamen University.

Abstract

3D Visual Grounding (3DVG) and 3D Dense Captioning (3DDC) are two crucial tasks in various 3D applications, which require both shared and complementary information in localization and visual-language relationships. Therefore, existing approaches adopt the two-stage "detect-then-describe/discriminate" pipeline, which relies heavily on the performance of the detector, resulting in suboptimal performance. Inspired by DETR, we propose a unified framework, 3DGCTR, to jointly solve these two distinct but closely related tasks in an end-to-end fashion. The key idea is to reconsider the prompt-based localization ability of the 3DVG model. In this way, the 3DVG model with a well-designed prompt as input can assist the 3DDC task by extracting localization information from the prompt. In terms of implementation, we integrate a **Lightweight Caption Head** into the existing 3DVG network with a **Caption Text Prompt** as a connection, effectively harnessing the existing 3DVG model's inherent localization capacity, thereby boosting 3DDC capability. This integration facilitates simultaneous multi-task training on both tasks, mutually enhancing their performance. Extensive experimental results demonstrate the effectiveness of this approach. Specifically, on the ScanRefer dataset, 3DGCTR surpasses the state-of-the-art 3DDC method by 4.30% in CIDEr@0.5IoU in MLE training and improves upon the SOTA 3DVG method by 3.16% in Acc@0.25IoU.

1 Introduction

The intersection of 3D scene understanding and natural language processing has emerged as a focal point in research, evident in tasks like 3D visual grounding (3DVG) [Guo *et al.*, 2023; Zhang *et al.*, 2023; Hsu *et al.*, 2023; Chen *et al.*, 2022; Jain *et al.*, 2022; Wu *et al.*, 2023; Zhao *et al.*, 2021; Yang *et al.*, 2021; Yuan *et al.*, 2021; Roh *et al.*, 2022; Luo *et al.*, 2022; Yin *et al.*, 2023; Huang *et al.*, 2021] and 3D dense captioning (3DDC) [Yuan *et al.*, 2022; Chen *et al.*, 2021b; Jiao *et al.*, 2022; Zhong *et al.*, 2022a; Yu *et al.*, 2023]. 3DVG involves processing a text paired with a point cloud to identify the specified object with a bounding box. Its counterpart, 3DDC, takes a point cloud and produces detailed bounding boxes and descriptions for each object. These tasks are pivotal for developing innovative applications, including assistive robotics and intuitive language-based interaction in AR/VR environments.

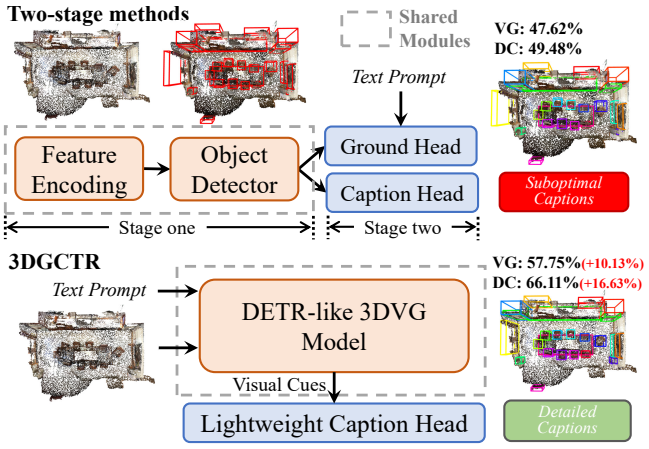

Figure 1: Illustration of existing method (upper) with two-stage pipeline and our single-stage 3DGCTR (bottom). These two-stage methods heavily depend on a detector's output and also suffer from low reuse of task-agnostic modules. Therefore, we propose a transformer-based single-stage model that simply builds upon a mature 3DVG model, thus giving 3DVG model 3DDC capability. Compared to the SOTA method 3DJCG [Cai *et al.*, 2022] that jointly trains the two tasks, our method achieves a significant improvement.

Considering that 3DVG and 3DDC contain both shared and complementary information in nature, previous works [Chen *et al.*, 2021a; Cai *et al.*, 2022; Chen *et al.*, 2023b] attempt to integrate these two tasks into a unified work. D3Net [Chen *et al.*, 2021a] proposes a speaker-listener pipeline to improve the performance of these two tasks in a self-critical manner. Based on a pre-trained detector, 3DJCG [Cai *et al.*, 2022] utilizes an attribute and relation-aware module to enhance the task-agnostic features. Unit3D [Chen *et al.*, 2023b] propose a transformer-based fusion module that is based on a PointGroup [Jiang *et al.*, 2020] detection backbone and a BERT [Devlin *et al.*, 2018] encoder, to learn the multi-model repre-

sentations between objects and text inputs. Among existing methods, they all adopt a detector-based architecture [Chen *et al.*, 2021a; Cai *et al.*, 2022; Chen *et al.*, 2023b], in which two task-specific modules were stacked on a shared detector [Ding *et al.*, 2019; Jiang *et al.*, 2020] to unify these two tasks in a single framework. Though these *two-stage* methods have achieved remarkable performance, the detector-based architecture yields suboptimal performance due to the following issues: **1) For 3DVG**, the language-irrelevant proposals generated by the detector easily cause incorrect localization, which is demonstrated by 3D-SPS [Luo *et al.*, 2022]. **2) For 3DDC**, the serial and explicit reasoning highly limits the mutual promotion of localization and captioning.

Recently, DETR-like [Carion *et al.*, 2020] models developed based on Transformers [Vaswani *et al.*, 2017] have achieved inspiring progress on both 3DVG [Luo *et al.*, 2022; Jain *et al.*, 2022; Wu *et al.*, 2023] and 3DDC [Chen *et al.*, 2023a] tasks in a *single-stage* way. It is a natural idea to integrate these two tasks into a single DETR-like architecture, thus maximizing the use of task-agnostic modules to achieve end-to-end training. However, this idea is not easy to implement due to the following reasons: **1) Variations in Input Types.** In terms of inputs, 3DDC exclusively requires a point cloud from a scene, whereas 3DVG also demands additional text references. **2) Divergent Optimization Objectives for Query Embeddings.** Query embeddings are an important component of the DETR-like architecture, which represents the object proposals to be located in a scene. In 3DVG methods [Luo *et al.*, 2022; Jain *et al.*, 2022; Wu *et al.*, 2023], the query embeddings are learned to align the target object described in the input referencing text. In contrast, 3DDC [Chen *et al.*, 2023a] focuses on aligning query embeddings with all predetermined ground-truth objects in the scene.

To address the above issues and naturally merge two different tasks, we rethink that a well-crafted grounding referencing text can act as a prompt to boost the detection capability of the 3DVG model, which is also supported by the previous work [Jain *et al.*, 2022]. By utilizing such a prompt as a link, it is possible to cast the detection part of the dense caption model as referential grounding. Specifically, as described above, the query embeddings are trained to align the input text in the 3DVG task. When a prompt such as "*cabinet . bed . chair . sofa .* " composed of multiple object class names is fed into the 3DVG model, the 3DVG model can locate the objects mentioned in the prompt, thus turning it into a detector. This allows us to naturally merge the two tasks into the DETR-like architecture by harmonizing the input form across both tasks (point cloud together with prompt) and the optimization objectives for query embeddings (align the objects mentioned in the prompt). Combined with our rethinking, we proposed a unified Transformer framework termed 3DGCTR, which integrates the Dual-Clued Captioner (DCC), a lightweight caption head outlined in [Chen *et al.*, 2023a], into the 3DVG model. Together with different text prompts for the two tasks, we enable the model to train both tasks end-to-end.

Experiments show that both the 3DDC and 3DVG performance of 3DGCTR have achieved state-of-the-art on the ScanRefer[Chen and Chang, 2020] benchmark. To be specific, 3DGCTR surpasses the 3DDC method by 4.30% in CIDEr@0.5IoU in MLE training and improves upon the 3DVG method by 3.16% in Acc@0.25IoU. Meanwhile, through joint training, 3DGCTR can achieve the mutual promotion of the two tasks. Specifically, the 3DDC and 3DVG tasks increased CIDEr@0.5IoU by 1.27% and Acc@0.25IoU by 0.3%, respectively.

To sum up, the main contributions are as follows:

- **Rethinking the role of prompts in the 3DVG model.** We provide a new perspective on the prompt-based localization ability of the 3DVG model, transforming it into a dual-purpose tool that effectively facilitates both 3DVG and 3DDC tasks.
- **Advanced Unified Framework for 3D Visual Grounding and 3D Dense Captioning.** By adding a lightweight caption head to the 3DVG model and using a well-designed prompt for the 3DDC task, our 3DGCTR framework integrates 3DVG and 3DDC tasks within a DETR-like architecture, enabling efficient end-to-end training.
- **State-of-the-Art Performance.** Not only does our integration approach produce outstanding results, but in our end-to-end training, both DC and VG tasks were mutually enhanced, which set new benchmarks for both 3D captioning and grounding tasks, further establishing our framework's dominance in the domain.

2 Related Work

2.1 3D Visual Grounding

3D Visual Grounding (3DVG) aims to identify the targeted object in a 3D scene based on linguistic cues. Broadly, 3DVG encompasses two pivotal tasks: 3DREC and 3DRES.

Two-Stage Method. Within the 3DREC realm, most recent works [Zhao *et al.*, 2021; Yang *et al.*, 2021] employ a two-stage pipeline. Initially, they utilize either ground truth or a 3D object detector [Jiang *et al.*, 2020; Liu *et al.*, 2021] for generating object proposals. Subsequently, these methods use text and 3D encoder [Liu *et al.*, 2019; Qi *et al.*, 2017] to extract features and then ground the target one after the feature fusion. For instance, InstanceRefer [Yuan *et al.*, 2021] ingeniously construes the task as instance-matching, while LanguageRefer [Roh *et al.*, 2022] reinterprets it as a linguistic modeling task by replacing the 3D features with predicted object labels.

Single-Stage Method. 3D-SPS [Luo *et al.*, 2022] is the first to introduce a single-stage network for 3DREC. Meanwhile, recent state-of-the-art 3DREC approaches, such as BUTD-DETR [Jain *et al.*, 2022] and EDA [Wu *et al.*, 2023], have incorporated the query selection modules from Groupfree [Liu *et al.*, 2021], indicating a converging trend of these technologies. It is worth noting that the two-stage setting introduced in [Wu *et al.*, 2023] is different from that described in our paper. Specifically, the outputs of the detector participate in the cross-attention of the query embedding as auxiliary information, which in essence still belongs to the single-stage training method. In comparison, the 3DRES field remains largely unexplored, with the only notable single-stage method TGNN [Huang *et al.*, 2021]. TGNN builds on

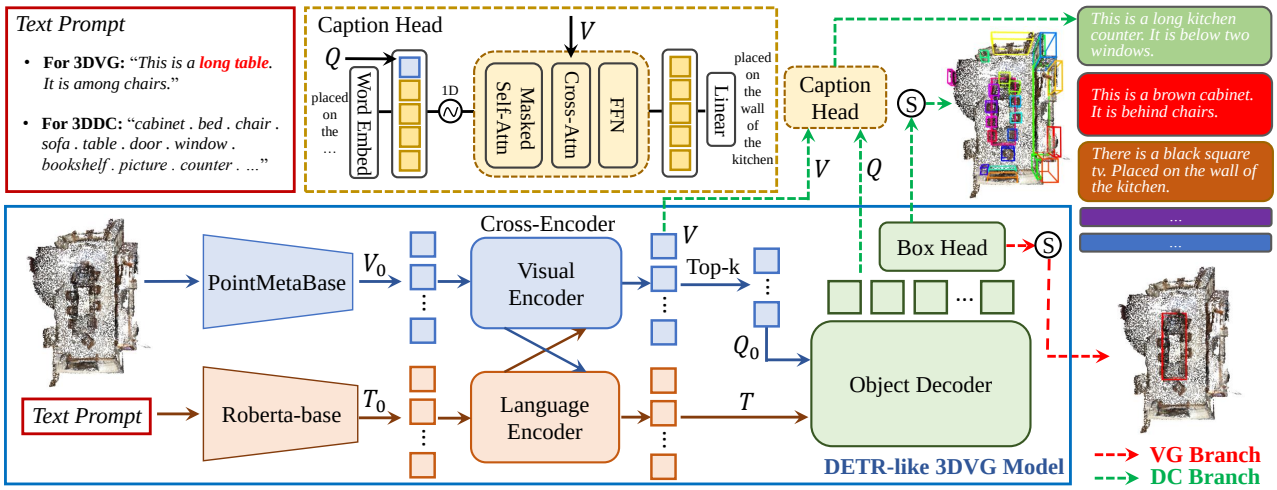

Figure 2: The framework of 3DGCTR builds upon a mature DETR-like 3DVG model (bottom in the figure) with a lightweight caption head. After obtaining the fused visual tokens V and decoder output query embeddings Q of each scene, the caption head uses Q as caption prefix to identify the described region, and contextual features V surrounding the vote query to complement with more surrounding information for more descriptive caption generation. Finally, the referring/detection boxes are selected from the candidate boxes via the referring scores.

an instance segmentation framework, employing text-guided Graph Neural Networks to pinpoint the center of the instance referred to in the text. In this paper, we will focus our research on 3DREC because it has been fully explored.

2.2 3D Dense Captioning

3D dense captioning is a task that requires translating 3D scene information to a set of bounding boxes and natural language descriptions, is challenging and has raised great interest among scholars in recent years.

Two-Stage Method. Scan2Cap [Chen *et al.*, 2021b] and MORE [Jiao *et al.*, 2022] create graphs based on a detector's [Qi *et al.*, 2019; Jiang *et al.*, 2020] box predictions, using predefined rules to decipher complex object relations in 3D scenes. SpaCap3D [Wang *et al.*, 2022a] develops a spatially-informed transformer to understand spatial relationships among detector outputs. 3DJCG [Cai *et al.*, 2022] and D3Net [Chen *et al.*, 2021a] explore the mutual enhancement of 3D dense captioning and 3D visual grounding. $\mathcal{X}$-Trans2Cap [Yuan *et al.*, 2022] incorporates 2D priors to enhance 3D dense captioning through knowledge transfer. Recently, [Zhong *et al.*, 2022b] has focused on contextual data for recognizing non-object details. These methods have significantly advanced the 3D dense captioning challenge. Yet, they are heavily reliant on the accuracy of the detector.

Single-Stage Method. Vote2Cap-DETR [Chen *et al.*, 2023a] is a unified, single-stage approach that concurrently detects objects and generates captions, addressing 3D dense captioning as a set prediction issue. Mirroring Vote2Cap-DETR [Chen *et al.*, 2023a], our 3DGCTR model seamlessly integrates 3DDC and 3DVG tasks within a DETR-like framework. Furthermore, 3DGCTR's visual grounding aspect enhances the dense captioning task more effectively than joint training of detection and captioning tasks, owing to its superior text-visual alignment.

3 Method

The key idea of 3DGCTR is to harness the relation-oriented capability of the 3DVG task and the object-oriented ability of the 3DDC task to complement each other. The architecture of the adopted 3DVG model is illustrated in Sec. 3.1. To enhance the feature extraction ability, we replace the visual backbone in EDA [Wu *et al.*, 2023] with PointMetaBased [Lin *et al.*, 2023] as our 3DVG model, termed EDA-PMB, which is introduced in Sec. 3.2. To develop the 3DDC capacity, a lightweight caption head is appended to the 3DVG model for caption generation, which is detailed in Sec. 3.3. To naturally unify 3DDC and 3DVG tasks in DETR-like architecture, we propose a text prompt for 3DDC as an input to the 3DVG backbone, as described in Sec. 3.4. Last, the training scheme and the inference process are illustrated in Sec. 3.5 and Sec. 3.6, respectively.

3.1 Preliminaries: 3DVG Model

To ensure the end-to-end training characteristics of our model, we introduce EDA [Wu *et al.*, 2023] with the setting of single-stage as the basic 3DVG model. EDA is a typical DETR-like [Carion *et al.*, 2020] model composed of a backbone, a cross-encoder, and an object decoder.

Visual and Textual Backbone. As shown in Figure. 2, PointMetaBased [Lin *et al.*, 2023] produces the visual tokens $V_0 \in \mathbb{R}^{n\times d}$ from the input point cloud $P \in \mathbb{R}^{N\times 3}$, while Roberta-base [Liu *et al.*, 2019] processes referential text for text tokens $T_0 \in \mathbb{R}^{l\times d}$, where n denotes the number of visual token, l denotes the number of text token, N denotes the point cloud size, and d is the feature dimension.

Cross-Encoder. The visual tokens V_0 and text tokens T_0 enter a dual-pathway cross-encoder alternatively stacking with self-attention, cross-attention, and FFN layers, producing thoroughly fused features $V \in \mathbb{R}^{n\times d}$ and $T \in \mathbb{R}^{l\times d}$. The output visual tokens $V \in \mathbb{R}^{n\times d}$ is input to the KPS module proposed by [Liu *et al.*, 2021]. The top-k object queries

$Q_0 \in \mathbb{R}^{k\times d}$ are chosen and then fed into the object decoder, caption head and the query scoring branch.

Object Decoder. The top-k object queries Q_0 and the text tokens T enter the object decoder which consists of stacked Transformer decoder layers, producing the final query embeddings $Q \in \mathbb{R}^{k\times d}$. On top of these object query embeddings Q, the decoder has two branches respectively for box regression and box-text alignment. The box branch dynamically updates boxes $B \in \mathbb{R}^{k\times 6}$ and provides position embeddings for query embeddings in each Transformer decoder layer. The alignment branch outputs the referring scores $s \in \mathbb{R}^k$ for the query embeddings to determine which query best matches the referent. Given the existing box and alignment branches, the query embeddings Q in the 3DVG model exhibit inherent localization abilities. In the caption branch, referring scores are also used to filter scene objects, thus describing the scene as a set of multiple objects.

3.2 Integrating PointMetaBase into EDA

A more powerful visual backbone can better extract the attributes and relative position relationships of objects in 3D scenes, which is helpful for more fine-grained interaction with language. Hence, we replaced the PointNet++ [Qi *et al.*, 2017] in EDA [Wu *et al.*, 2023] with our proposed PointMetaBase [Lin *et al.*, 2023]. The main contribution of PointMetaBase is to abstract the point cloud feature extraction network into a *Meta Architecture* consisting of four key modules and explore the optimal design of each module. However, as the K-Nearest query is used in the module *Neighbor Update*, the number of downsampled point clouds output by the backbone is not fixed, and thus can't be used as the visual tokens V_0 directly. To solve this problem, we design a parameter-free visual token query module. We first use the farthest point sampling (FPS) algorithm to sample a fixed number of n points from the original point cloud P as the candidates C:

$$C = FPS(P \in \mathbb{R}^{N\times 3}) \in \mathbb{R}^{n\times 3}$$

Then, given the output tokens $V' \in \mathbb{R}^{n'\times 3}$ from the PointMetaBase, we employ the ball query [Qi *et al.*, 2017] with a ball radius of r to find k_q nearest neighbor tokens in the coordinate space of V' for each candidate, then assign the features of V' to them using max-pooling. Finally, we can obtain the visual tokens $V_0 \in \mathbb{R}^{n\times 3}$:

$$V_0 = MaxPool(BallQuery(V', C, k_q, r) \in \mathbb{R}^{n\times 3\times k_q})$$

3.3 Lightweight Caption Head

We argue that the key to unambiguous detailed caption generation is to obtain the relationship between the target object and its close surroundings. As described in Vote2Cap-DETR [Chen *et al.*, 2023a], the vote queries fail to provide adequate attribute and spatial relations. To address this issue, they proposed a caption head named Dual-Clued Captioner (DCC), which introduced the vote queries' k nearest local context token features as their local surroundings and keys for cross attention. In our method, the top-k object queries Q_0 generated in the KPS module also have the same problem as the vote queries generated in Vote2Cap-DETR. However, different from Vote2Cap-DETR, where vote queries are constructed based on spatial bias and content information for better caption generation, our KPS-based method focuses on the precise localization of key points of objects to tackle both 3DVG and 3DDC tasks simultaneously. Therefore, we use DCC as our caption head with some modifications to fit this difference. Specifically, instead of using vote queries' k nearest for cross attention, we introduce the visual tokens V that integrate the broader scene context as the object's interaction and relation information with its surroundings, leading to richer and more contextually nuanced captions.

As shown in Figure 2, DCC is a lightweight transformer decoder-based model comprising two identical transformer decoder blocks, sinusoidal position embeddings $PE(\cdot)$, and a linear classification head. For effective caption generation, DCC processes dual visual cues $V_c = (Q, V)$. Firstly, following [Wang *et al.*, 2022b], in captioning a proposal, the standard 'Start Of Sequences' (SOS) prefix is replaced with the query Q from the described query, identifying the object in focus. Then, for each query embedding $q \in Q$, DDC gives its corresponding dense caption. The masked self-attention and cross-attention mechanism of the transformer decodes the contextual relationship between the query and the visual scene, which can be formulated as:

$$Atten(V_c) = CrossAtten(SelfAtten(PE(Q), mask), V)$$

Finally, the output of the cross-attention is then processed through a Feed-Forward Network (FFN) together with the linear classification head to map the processed features to the caption vocabulary:

$$Captions = Linear(FFN(Atten(V_c)))$$

3.4 Caption Text Prompt

BUTD-DETR [Jain *et al.*, 2022] suggests that the task of object detection is essentially a form of referential language grounding, where the utterance is simply the object's category label. Essentially, object detection can be viewed as the grounding process of detection prompts.

Concretely, utilizing the detector's library of object categories, we form prompts by stringing together the labels of objects to be identified, like "*cabinet . bed . chair . sofa . *". These concatenated labels are treated as reference text for grounding: the goal is to locate all instances of the mentioned categories in the scene, if present. Additionally, incorporating negative category labels (for which no instances exist in the scene) serves as negative training, teaching the model to avoid matching any boxes to these negative labels.

Through this approach, we can perform visual grounding and dense captioning in a single-stage manner. This integration enables end-to-end training within a single framework, fostering a synergistic relationship where each task mutually reinforces the other.

3.5 Multi-task End-to-end Training

For the 3DVG task, we follow the training scheme of EDA [Wu *et al.*, 2023], which adopts five loss functions for each layer of the object decoder: box center coordinate prediction with a smooth-L_1 loss $\mathcal{L}_{\text{coord}}$, box size prediction with a smooth-L_1 loss $\mathcal{L}_{\text{size}}$, GIoU loss [Rezatofighi *et al.*, 2019]

$\mathcal{L}_{\text{giou}}$, the semantic alignment loss $\mathcal{L}_{\text{sem}}$, and the position alignment loss $\mathcal{L}_{\text{pos}}$. The loss of l-th decoder layer is the combination of these 5 loss terms by weighted summation:

$$\mathcal{L}_{\text{dec}}^{(l)} = \beta_1 \mathcal{L}_{\text{coord}}^{(l)} + \beta_2 \mathcal{L}_{\text{size}}^{(l)} + \beta_3 \mathcal{L}_{\text{giou}}^{(l)} + \beta_4 \mathcal{L}_{\text{sem}}^{(l)} + \beta_5 \mathcal{L}_{\text{pos}}^{(l)} .$$

The losses on all decoder layers are averaged to form the total 3DVG loss:

$$\mathcal{L}_{\text{vg}} = \frac{1}{L} \sum_{l=1}^{L} \mathcal{L}_{\text{dec}}^{(l)} .$$

For the 3DDC task, we follow Vote2Cap-DETR [Chen *et al.*, 2023a] and apply the standard cross-entropy loss (MLE training), then fine-tune it with Self-Critical Sequence Training (SCST). We adopt the standard SCST in our model, whose reward function is CIDEr [Vedantam *et al.*, 2015] score. We termed the caption loss as $\mathcal{L}_{\text{cap}}$. During MLE training, together with the KPS loss [Liu *et al.*, 2021] $\mathcal{L}_{\text{kps}}$ for the query selection, the final loss for 3DGCTR in end-to-end manner is as follows:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{\text{vg}} + \alpha_2 \mathcal{L}_{\text{cap}} + \alpha_3 \mathcal{L}_{\text{kps}} .$$

Note that when the batch of training data is 3DVG, $\mathcal{L}_{\text{cap}} = 0$. While in SCST training, the VG model is frozen and only the caption head is trained with the caption loss $\mathcal{L}_{\text{cap}}$:

$$\mathcal{L} = \mathcal{L}_{\text{cap}}.$$

We also use the set-to-set training strategy proposed by Vote2Cap-DERT [Chen *et al.*, 2023a] for better performance.

3.6 Inference

During inference, given an input point cloud and a referring text, 3DGCTR outputs the object boxes $B \in \mathbb{R}^{k \times 6}$. Subsequently, the referent box will be selected according to the referring scores $s \in \mathbb{R}^k$. Specifically, in 3DVG, the target box described in the referring text will be selected, while in 3DDC, all label boxes described in the prompt are selected, and the caption head provides a dense caption for each box.

4 Experiments

4.1 DataSets and Metrics

We assess our approach on 3D referring datasets ScanRefer [Chen and Chang, 2020]. ScanRefer derives from ScanNet [Dai *et al.*, 2017], featuring 1,513 detailed 3D in-door scene reconstructions. Specifically, ScanNet [Dai *et al.*, 2017] comprises 1,201 training and 312 validation indoor 3D scenes. ScanRefer adheres to the official ScanNet splits, it offers 51,583 descriptions, averaging 13.81 objects and 64.48 descriptions per scene.

In terms of the evaluation metric, for 3DVG, we use Acc@IoU to measure the proportion of descriptions where the predicted box and ground truth overlap with an IoU greater than 0.25 and 0.5. Descriptions are categorized into "unique" if the object is the sole representative of its class in the scene, or "multiple" otherwise. As for 3DDC, following [Chen *et al.*, 2023a], the evaluation metric is $m@k$IoU that m could be any metric for natural language generation, such as CIDEr [Vedantam *et al.*, 2015], METEOR [Banerjee and Lavie, 2005], BLEU-4 [Papineni *et al.*, 2002], and ROUGE-L [Lin, 2004]. In practice, k is set to 0.25 and 0.5.

4.2 Implementation Details

We first pre-train the 3DVG model without the caption head on ScanRefer datasets [Chen and Chang, 2020] following the training settings in [Wu *et al.*, 2023]. During MLE training, we load the pre-trained 3DVG weight and jointly train the whole network for another 30 epochs in both VG and DC tasks. To prevent overfitting, the initial learning rates of the 3DVG model and caption head are empirically set to 2e-6 and 2e-4, respectively. We apply learning rate decay at epochs 10 and 20 with a rate of 0.1. As for the input type, we use XYZ coordinates and RGB values as the input, and the number of visual tokens V and query embeddings Q are empirically set to 1024 and 256, respectively. The number of decoder layers L is set to 6. The balancing factors are set default as α_1 = 1.0 / (L + 1), α_2 = 5.0, α_3 = 8.0, β_1 = 5.0, β_2 = 1.0, β_3 = 1.0, β_4 = 0.5 and β_5 = 0.5 for the ScanRefer dataset. During SCST training, due to the high memory cost, we tune the caption head with a batch size of 8 and freeze the rest of the modules for 400 epochs with a fixed learning rate of 5e-6 and apply learning rate decay at epochs 100 and 200 with a rate of 0.1.

4.3 Quantitative Comparison

In this subsection, we perform a qualitative comparison on ScanRefer [Chen and Chang, 2020] datasets.

3DVG. To evaluate the 3DVG performance, we compare 3DGCTR with several existing 3DVG works in the 3D Referring Expression Comprehension (3DREC) task on ScanRefer, involving the state-of-the-art DETR-like methods EDA [Wu *et al.*, 2023] and BUTD-DETR [Jain *et al.*, 2022], among others. As evidenced in Table. 1, 3DGCTR consistently exhibits superior performance over these prevailing methods in both two-stage and single-stage settings. This evidence strongly suggests that the newly incorporated lightweight caption head has indeed augmented the performance of the 3DVG task.

3DDC. In 3DDC performance, among the evaluated methods, most employ the standard VoteNet [Ding *et al.*, 2019] detector, except D3Net [Chen *et al.*, 2021a] and 3DJCG [Cai *et al.*, 2022]. The state-of-the-art method Vote2Cap-DETR [Chen *et al.*, 2023a] is the first attempt to jointly train 3D Detection and 3D Dense Caption tasks in a single-stage manner, which gains state-of-the-art performance. Table. 2 reports comparisons on ScanRefer [Chen and Chang, 2020] dataset. Our 3DGCTR surpasses current state-of-the-art methods. For example, our 3DGCTR achieves 66.11% C@0.5 while Vote2Cap-DETR. [Chen *et al.*, 2023a] achieves 61.81% (4.30% C@0.5↑). Additionally, under SCST, our 3DGCTR outperforms the state-of-the-art method C@0.5 (0.18% C@0.5↑).

4.4 Ablation Study

The experiments in this subsection are conducted on Scanrefer using MLE training, and metrics adopted for 3DVG and 3DDC performance are Acc@0.25IoU and CIDer@0.5IoU, respectively.

Influence of 3DVG model. We replace the pre-trained 3DVG model with EDA [Wu *et al.*, 2023] in both single- and two-stage settings. As illustrated in Table. 3, under the joint training of 3DDC and 3DVG, various 3DVG backbones achieve enhanced performance in 3DVG tasks

Method	Unique(19%)		Multiple(81%)		Overall	
	0.25	0.5	0.25	0.5	0.25	0.5
ScanRefer [Chen and Chang, 2020]	67.64	49.19	32.06	21.26	38.97	26.10
TGNN [Huang *et al.*, 2021]	68.61	56.80	29.84	23.18	37.37	29.70
3DJCG [Cai *et al.*, 2022]	78.75	61.30	40.13	30.08	47.62	36.14
D3Net [Chen *et al.*, 2021a]	-	70.35	-	30.50	-	37.87
UniT3D [Chen *et al.*, 2023b]	82.75	73.14	36.36	31.05	45.27	39.14
3DSPS [Luo *et al.*, 2022]	84.12	66.72	40.32	29.82	48.82	36.98
BUTD-DETR [Jain *et al.*, 2022]	82.88	64.98	44.73	33.97	50.42	38.60
EDA [Wu *et al.*, 2023]	85.76	68.57	49.13	37.64	54.59	42.26
EDA-PMB (ours backbone)	**88.22**	**73.40**	51.76	40.82	57.45	45.91
3DGCTR (ours)	88.01	73.13	**52.15**	**41.31**	**57.75**	**46.28**

Table 1: The 3D referring expression comprehension results on ScanRefer in terms of Acc@0.25IoU and Acc@0.5IoU. Note that what is shown in the table is the optimal accuracy of each method as reported in its paper. Our 3DGCTR surpasses existing state-of-the-art works by a significant margin. EDA-PMB refers to the replacement of single-stage EDA's [Wu *et al.*, 2023] visual backbone from Pointnet++ [Qi *et al.*, 2017] to PointMetaBase [Lin *et al.*, 2023].

Method	IoU=0.25				IoU=0.5			
	C	B-4	M	R	C	B-4	M	R
MLE training								
Scan3Cap [Chen *et al.*, 2021b]	56.82	34.18	26.29	55.27	39.08	23.32	21.97	44.78
MORE [Jiao *et al.*, 2022]	62.91	36.25	26.75	56.33	40.94	22.93	21.66	44.42
SpaCap3d [Wang *et al.*, 2022a]	63.30	36.46	26.71	55.71	44.02	25.26	22.33	45.36
3DJCG [Cai *et al.*, 2022]	64.70	40.17	27.66	59.23	49.48	31.03	24.22	50.80
D3Net [Chen *et al.*, 2021a]	-	-	-	-	46.07	30.29	24.35	51.67
UniT3D [Chen *et al.*, 2023b]	-	-	-	-	46.69	27.22	21.91	45.98
Vote2Cap-DETR [Chen *et al.*, 2023a]	71.45	39.34	28.25	59.33	61.81	34.46	**26.22**	**54.40**
3DGCTR (ours)	**84.87**	**44.58**	**29.68**	**63.24**	**66.11**	**35.85**	26.12	54.29
SCST training								
$\mathcal{X}$-Trans2Cap [Yuan *et al.*, 2022]	61.83	35.65	26.61	54.70	43.87	25.05	22.46	45.28
Scan3Cap [Chen *et al.*, 2021b]	-	-	-	-	48.38	26.09	22.15	44.74
D3Net [Chen *et al.*, 2021a]	-	-	-	-	62.64	35.68	25.72	53.90
Vote2Cap-DETR [Chen *et al.*, 2023a]	84.15	42.51	28.47	59.26	73.77	38.21	**26.64**	54.71
3DGCTR (ours)	**95.96**	**48.13**	**29.91**	**64.23**	**73.95**	**38.71**	26.29	**56.27**

Table 2: The 3D referring expression comprehension results on ScanRefer in terms of CIDEr (**C**), METEOR (**M**), BLEU-4 (**B-4**) and ROUGE-L (**R**), with the setting of m@0.25IoU and m@0.5IoU in both four metrics. We compare 3DGCTR with all published 3D dense caption methods on the ScanRefer dataset, which demonstrates that 3DGCRE achieves new state-of-the-art under both MLE and SCST training.

(Acc@0.25IoU). Meanwhile, when using the state-of-the-art 3DVG model [Wu *et al.*, 2023] as the component of our method, we obtain the state-of-the-art 3DDC performance compared to Vote2Cap-DETR [Chen *et al.*, 2023a] (3.29% C@0.5↑), which demonstrate that our "3DDC builds upon 3DVG model" rethinking pipeline offers greater advantages.

Joint Training Scheme. We explore the impacts of different joint training schemes on task performance. After pre-training the 3DVG model without the caption head, if only dense caption data is used for training, the training scheme can be divided into the following three types: 1) only 3DDC: training all components with the same learning rate, 2) only 3DDC (two lr): training all components with different learning rates ($2e-4$ and $2e-6$), 3) only 3DDC (frozen vg): freezing 3DVG components and training 3DDC components. If both data to jointly train 3DDC and 3DVG tasks, the training scheme can be divided into the following two types: 4) joint training: jointly training 3DDC and 3DVG with the same learning rate, 5) joint training (two lr): jointly training 3DDC and 3DVG with different learning rates ($2e-4$ and $2e-6$), As displayed in Table. 4, The first three schemes bring about significant 3DDC performance gain but cause a suboptimal

Method	3DDC	3DVG
Vote2Cap [Chen *et al.*, 2023a]	61.81	-
EDA^s [Wu *et al.*, 2023]	-	53.83
3DGCTR-EDA^s	58.85	54.37**(+0.54%)**
EDA^t [Wu *et al.*, 2023]	-	54.59
3DGCTR-EDA^t	65.10	55.09**(+0.50%)**
EDA-PMB	-	57.45
3DGCTR	66.11	57.75**(+0.30%)**

Table 3: Performance on Scanrefer affected by different 3DVG backbone during MLE training. The expressions EDA^s and EDA^t denote the application of EDA [Wu *et al.*, 2023] in single- and two-stage settings, respectively. The acronyms 3DGCTR-EDA^s and 3DGCTR-EDA^t are used to indicate the utilization of EDA^s and EDA^t as the 3DVG backbone in the corresponding settings. The red font represents the performance improvement after joint training.

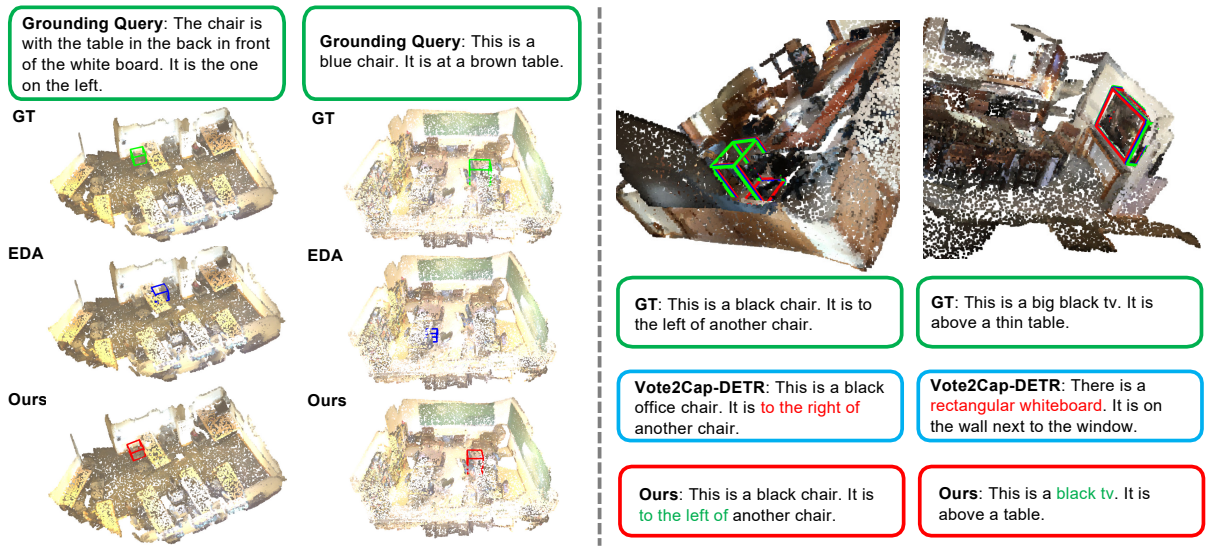

Figure 3: Qualitative Comparisons. We compare qualitative results with two state-of-the-art methods in 3DVG (left part in the figure) and 3DDC (right part in the figure) tasks, EDA [Wu *et al.*, 2023] and Vote2Cap-DETR [Chen *et al.*, 2023a]. We mark correct attribute words in green and wrong descriptions in red. Our method produces right bounding boxes close to ground truth annotations and produces accurate descriptions of object attributes, classes and spatial relationships.

3DVG performance. Using the same learning rate in the first scheme significantly decreases 3DVG performance, as solely employing 3DDC training data biases the model's statistical parameters towards DC data (e.g., mean as well as variation in the batch normalization layer), negatively impacting 3DVG. This is mitigated in the second and third schemes by either freezing or reducing the learning rate for 3DVG components, which preserves 3DVG localization and improves 3DDC performance. However, applying the same learning rate for both 3DDC and 3DVG data in the fourth scheme leads to overfitting in 3DVG training, skewing the model's statistical parameters towards VG data and impairing 3DDC performance. In contrast, the fifth scheme, which uses different learning rates, achieves optimal 3DDC accuracy and slightly better 3DVG performance. It can also be seen from the table that after joint training, 3DGCTR's 3DVG performance increased by 0.3% compared with EDA-PMB, and 3DDC performance increased by 1.27% compared with only 3DDC training, suggesting that a well-designed training approach allows for effective end-to-end training of both 3DDC and 3DVG tasks within the same framework, benefiting each other.

4.5 Qualitative Comparison

As for the 3DVG task, we compare qualitative results to the state-of-the-art model EDA [Wu *et al.*, 2023] in the left part of Figure 3. Our method produces the right bounding boxes in some difficult samples, which indicates that by jointly training the 3DDC and 3DVG tasks, the 3DVG task benefits from the 3DDC task's comprehensive characterization of object attributes. As for the 3DDC task, we compare qualitative results to the state-of-the-art model Vote2Cap-DETR [Chen *et al.*, 2023a] in the right part of Figure 3. Our method can produce accurate descriptions of object attributes, classes, and spatial relationships. This suggests the 3DDC task benefits from the 3DVG task's grasp of the relationships between objects in the scene to generate more accurate caption information. Through qualitative comparison, we can more intuitively recognize the advantages of an end-to-end training scheme compared to the two-stage approach.

Training Scheme	3DDC	3DVG
EDA-PMB	-	57.45
Vote2Cap [Chen *et al.*, 2023a]	61.81	-
1) only 3DDC	62.89	44.84
2) only 3DDC (two lr)	64.84	57.46
3) only 3DDC (frozen vg)	63.83	57.45
4) joint training	60.17	55.85
5) joint training (two lr)	**66.11**(+1.27%)	**57.75**(+0.30%)

Table 4: Performance on Scanrefer affected by different training schemes during MLE training. 'two lr' means jointly training 3DDC and 3DVG components with different learning rates, and 'frozen vg' means only training 3DDC components while freezing the 3DVG backbone. The red font represents the performance improvement after joint training, while underline represents the performance to be compared with our optimal precision.

5 Conclusions

In conclusion, we introduce a significant shift in 3DVG and 3DDC integration using the proposed 3DGCTR. By rethinking prompt-based localization in 3DVG to enhance 3DDC, this approach moves away from traditional two-stage, detector-dependent methods towards a more effective, unified strategy. The DETR-like framework enables end-to-end training, utilizing shared modules for multitasking. 3DGCTR's exceptional performance on ScanRefer and its ability to mutually enhance 3DVG and 3DDC tasks underline its transformative impact on 3D scene understanding.

References

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

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 Conference on Computer Vision and Pattern Recognition*, pages 16464–16473, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

Dave Zhenyu Chen and Angel X Chang. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X Chang. D3net: a speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans. 2021.

Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021.

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.

Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11124–11133, 2023.

Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18109–18119, 2023.

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

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

Zhipeng Ding, Xu Han, and Marc Niethammer. Votenet: A deep learning label fusion method for multi-atlas segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part III 22*, pages 202–210. Springer, 2019.

Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15372–15383, 2023.

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*, pages 2614–2623, 2023.

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

Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pages 417–433. Springer, 2022.

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 conference on computer vision and Pattern recognition*, pages 4867–4876, 2020.

Yang Jiao, Shaoxiang Chen, Zequn Jie, Jingjing Chen, Lin Ma, and Yu-Gang Jiang. More: Multi-order relation mining for dense captioning in 3d scenes. In *European Conference on Computer Vision*, pages 528–545. Springer, 2022.

Haojia Lin, Xiawu Zheng, Lijiang Li, Fei Chao, Shanshan Wang, Yan Wang, Yonghong Tian, and Rongrong Ji. Meta architecture for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17682–17691, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

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.

Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2949–2958, 2021.

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 Conference on Computer Vision and Pattern Recognition*, pages 16454–16463, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

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

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*, pages 9277–9286, 2019.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 658–666, 2019.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*, 2022.

Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong Cai. Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*, 2022.

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 Conference on Computer Vision and Pattern Recognition*, pages 19231–19242, 2023.

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

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multimodal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*, 2023.

Ting Yu, Xiaojun Lin, Shuhui Wang, Weiguo Sheng, Qingming Huang, and Jun Yu. A comprehensive survey of 3d dense captioning: Localizing and describing objects in 3d scenes. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2023.

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 International Conference on Computer Vision*, pages 1791–1800, 2021.

Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8563–8573, 2022.

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*, pages 15225–15236, 2023.

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

Yufeng Zhong, Long Xu, Jiebo Luo, and Lin Ma. Contextual modeling for 3d dense captioning on point clouds. *arXiv preprint arXiv:2210.03925*, 2022.

Yufeng Zhong, Long Xu, Jiebo Luo, and Lin Ma. Contextual modeling for 3d dense captioning on point clouds. *arXiv preprint arXiv:2210.03925*, 2022.