EDA: Explicit Text-Decoupling and Dense Alignment for 3D Visual Grounding

Yanmin Wu[1] Xinhua Cheng[1] Renrui Zhang[2,3] Zesen Cheng[1] Jian Zhang[1]*

[1] Shenzhen Graduate School, Peking University, China
[2] The Chinese University of Hong Kong, China [3] Shanghai AI Laboratory, China

wuyanmin@stu.pku.edu.cn zhangjian.sz@pku.edu.cn

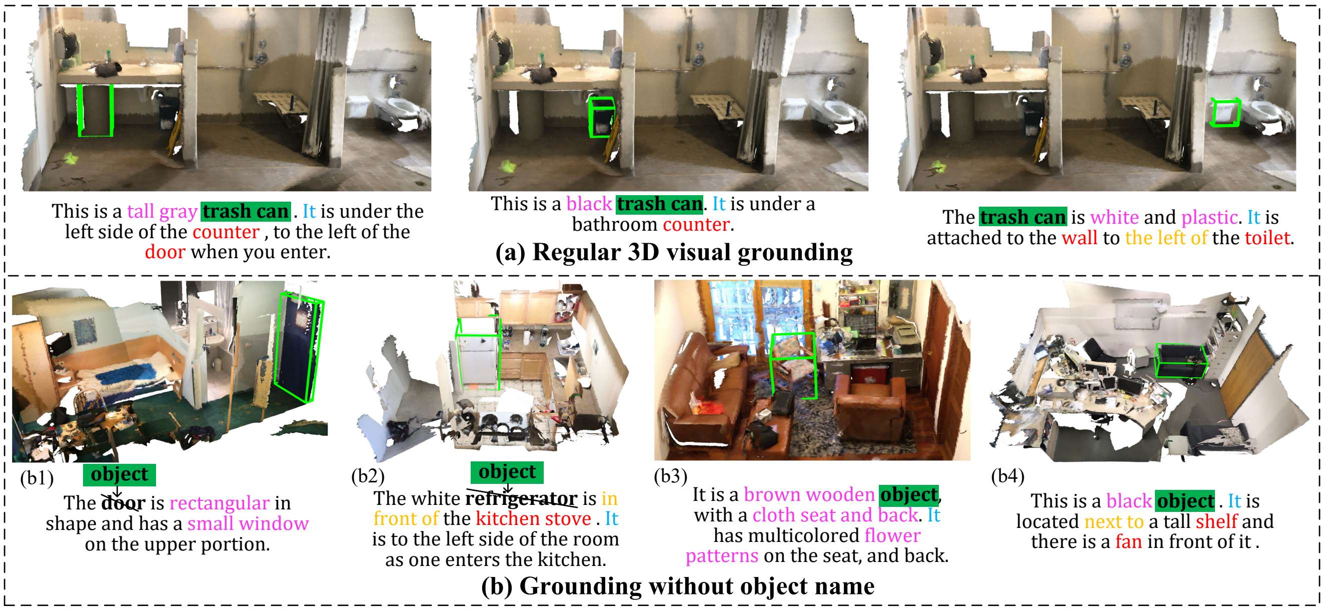


Figure 1. Text-decoupled, dense aligned 3D visual grounding. Different colours in the text correspond to different decoupled components. (a) Regular 3D visual grounding: locating objects requires comprehensively considering multiple semantic cues such as appearance attributes, object names, and spatial relationships. (b) Grounding without object name: not mentioning object names, avoiding short-cuts and forcing the model to predict the target based on other attributes.

Abstract

*3D visual grounding aims to find the object within point clouds mentioned by free-form natural language descriptions with rich semantic cues. However, existing methods either extract the sentence-level features coupling all words or focus more on object names, which would lose the word-level information or neglect other attributes. To alleviate these issues, we present **EDA** that **E**xplicitly **D**ecouples the textual attributes in a sentence and conducts **D**ense **A**lignment between such fine-grained language and point cloud objects. Specifically, we first propose a text decoupling module to produce textual features for every semantic component. Then, we design two losses to supervise the dense matching between two modalities: position alignment loss and semantic alignment loss. On top of that, we further introduce a new visual grounding task, locating objects without object names, which can thoroughly evaluate the model's dense alignment capacity. Through experiments, we achieve state-of-the-art performance on two widely-adopted 3D visual grounding datasets, ScanRefer and SR3D/NR3D, and obtain absolute leadership on our newly-proposed task. The source code is available at* https://github.com/yanmin-wu/EDA.

1. Introduction

Multi-modal cues can highly benefit the 3D environment perception of an agent, including 2D images, 3D point clouds, and language. Recently, **3D visual grounding** (3D VG) [8,50], also known as 3D object referencing [3], has at-

*Corresponding author. This work was supported in part by Shenzhen Research Project under Grant JCYJ20220531093215035.

tached much attention as an important 3D cross-modal task. Its objective is to find the target object in point cloud scenes by analyzing the descriptive query language, which requires understanding both 3D visual and linguistic context.

Language utterances typically involve words describing appearance attributes, object categories, spatial relationships and other characteristics, as shown by different colours in Fig. 1(a), requiring that the model integrate multiple cues to locate the mentioned object. Compared with 2D Visual Grounding [18, 19, 67, 71], the sparseness and incompleteness of point clouds, and the diversity of language descriptions produced by 3D multi-view, make 3D VG more challenging. Existing works made significant progress from the following perspectives: improving point cloud features extraction by sparse convolution [70] or 2D images assistance [68]; generating more discriminative object candidates through instance segmentation [33] or language modulation [46]; identifying complex spatial relationships between entities via graph convolution [23] or attention [6].

However, we observe two issues that remain unexplored. **1) Imbalance**: The object name can exclude most candidates, and even in some cases, there is only one name-matched object, as the *"door"* and *"refrigerator"* in Fig. 1(b1, b2). This shortcut may lead to an inductive bias in the model that pays more attention to object names while weakening other properties such as appearance and relationships, resulting in imbalanced learning. **2) Ambiguity**: Utterances frequently refer to multiple objects and attributes (such as *"black object, tall shelf, fan"* in Fig. 1(b4)), while the model's objective is to identify only the main object, leading to an ambiguous understanding of language descriptions. These insufficiencies of existing works stem from their characteristic of feature coupling and fusing implicitly. They input a sentence with different attribute words but output only one globally **coupled** sentence-level feature that subsequently matches the visual features of candidate objects. The coupled feature is ambiguous because some words may not describe the main object (green text in Fig. 1) but other auxiliary objects (red text in Fig. 1). Alternatively, using the cross-modal attention of the Transformer [21, 63] automatically and **implicitly** to fuse visual and text features. However, this may encourage the model to take shortcuts, such as focusing on object categories and ignoring other attributes, as previously discussed.

Instead, we propose a more intuitive decoupled and explicit strategy. First, we parse the input text to **decouple** different semantic components, including the main object word, pronoun, attributes, relations, and auxiliary object words. Then, performing **dense alignment** between point cloud objects and multiple related decoupled components achieves fine-grained feature matching, which avoids the inductive bias resulting from imbalanced learning of different textual components. As the final grounding result, we **explicitly** select the object with the highest similarity to the decoupled text components (instead of the entire sentence), avoiding ambiguity caused by irrelevant components. Additionally, to explore the limits of VG and examine the comprehensiveness and fine-graininess of visual-language perception of the model, we suggest a challenging new task: **Grounding without object name (VG-w/o-ON)**, where the name is replaced by *"object"* (see Fig. 1(b)), forcing the model to locate objects based on other attributes and relationships. This setting makes sense because utterances that do not mention object names are common expressions in daily life, and in addition to testing whether the model takes shortcuts. Benefiting from our text decoupling operation and the supervision of dense aligned losses, all text components are aligned with visual features, making it possible to locate objects independent of object names.

To sum up, the main contributions of this paper are as follows: **1)** We propose a text decoupling module to parse linguistic descriptions into multiple semantic components, followed by suggesting two well-designed dense aligned losses for supervising fine-grained visual-language feature fusion and preventing imbalance and ambiguity learning. **2)** The challenging new 3D VG task of grounding without object names is proposed to comprehensively examine the model's robust performance. **3)** We achieve state-of-the-art performance on two datasets (ScanRefer and SR3D/NR3D) on the regular 3D VG task and absolute leadership on the new task evaluated by the same model without retraining.

2. Related Work

2.1. 3D Vision and Language

3D vision [52, 53, 75] and language are vital manners for humans to understand the environment, and they are also important research topics for the evolution of machines to be like humans. Previously, the two fields evolved independently. Due to the advance of multimodality [25, 26, 54, 56, 57, 72–74, 76], many promising works across 3D vision and language have been introduced recently. In 3D visual grounding [1, 8–10, 12], the speaker (like a human) describes an object in language. The listener (such a robot) needs to understand the language description and the 3D visual scene to grounding the target object. On the contrary, the 3D dense caption [9, 13, 14, 36, 65, 69] is analogous to an inverse process in which the input is a 3D scene, and the output is textual descriptions of each object. Language-modulated 3D detection or segmentation [7, 32, 34, 59, 78] enriches the diversity of text queries by matching visual-linguistic feature spaces rather than predicting the probability of a set number of categories. Furthermore, some studies explore the application of 3D visual language in agents such as robot perception [27, 61], vision-and-language navigation (VLN) [15, 31, 55], and embodied question answering

(EQA) [4,22,30,45,47,62]. In this paper, we focus on point clouds-based 3D visual grounding, which is the fundamental technology for many embodied AI [24,38,48] tasks.

2.2. 3D Visual Grounding

The majority of current mainstream techniques are two-stage. In the first stage, obtain the features of the query language and candidate point cloud objects independently by a pre-trained language model [16, 20, 49] and a pre-trained 3D detector [44, 51] or segmenter [11, 35, 64]. In the second stage, the researchers focus on fusing the two modal features and then selecting the best-matched object. **1)** The most straightforward solution is to concatenate the two modal features and then consider it a binary classification problem [8, 39], which provides limited performance because the two features are not sufficiently fused. **2)** Taking advantage of the Transformer's attention mechanism, which is naturally suitable for multi-module feature fusion, He *et al.* [28] and Zhao *et al.* [77] achieve remarkable performance by performing self-attention and cross-attention to features. **3)** In contrast, other studies view feature fusion as a matching problem rather than a classification. Yuan *et al.* [70] and Abdelreheem *et al.* [2], supervised by the contrastive loss [29], compute the cosine similarity of visual features and textual features. Inspired by [43], Feng *et al.* [23] parses the text to generate a text scene graph, simultaneously builds a visual scene graph, and then performs graph node matching. **4)** Point clouds' sparse, noisy, incomplete, and lack of detail make learning objects' semantic information challenging. Yang *et al.* [68] and Cai *et al.* [6] use 2D images to aid visual-textual feature fusion, but at the cost of additional 2D-3D alignment and 2D feature extraction.

However, the two-stage method has a substantial detection bottleneck: objects overlooked in the first stage cannot be matched in the second. In contrast, object detection and feature extraction in the single-stage method is modulated by the query text, making it easier to identify the text-concerned object. Liu *et al.* [41] suggest fusing visual and linguistic features at the bottom level and producing text-related visual heatmaps. Similarly, Luo *et al.* [46] present a single-stage approach that employs textual features to guide visual keypoint selection and progressively localizes objects. BUTD-DETR [34] is also a single-stage capable framework. More importantly, inspired by the 2D image-language pre-train model (such MDETR [37], GLIP [40]), BUTD-DETR measures the similarity between each word and object and then selects the features of the word that correspond to the object's name to match the candidate object. However, there are two limitations: 1) Since multiple object names may be mentioned in a sentence, the ground truth annotation is needed to retrieve the target name, which limits its generalizability. Our text decoupling module separates

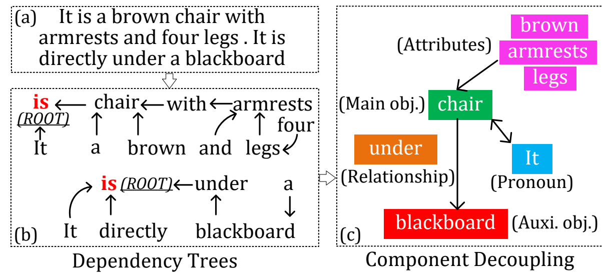


Figure 2. Text component decoupling: (a) The query text. (b) Dependency tree analysis. (c) Decoupled into five components.

text components and determines the target object name by grammatical analysis to avoid this restriction. 2) BUTD-DETR (and MDETR and GLIP in the 2D task) only consider the **sparse alignment** of main object words or noun phrases to visual features. Conversely, we align all object-related decoupled textual semantic components with visual features, which we refer to **dense alignment**, significantly enhancing the discriminability of multimodal features.

3. Proposed Method

The framework is illustrated in Fig. 3. First, the input text description is decoupled into multiple semantic components, and its affiliated text positions and features are obtained (Sec. 3.1). Concurrently, the Transformer-based encoder extracts and modulates features from point clouds and text, then decodes the visual features of candidate objects (Sec. 3.2). Finally, the dense aligned losses are derived between the decoupled text features and the decoded visual features (Sec. 3.3). The grounding result is the object with visual features most similar to text features (Sec. 3.4).

3.1. Text Decoupling

The text features of the coupled strategy are ambiguous, where features from multiple objects and attributes are coupled, such as *"a brown wooden chair next to the black table."* Among them, easy-to-learn clues (such as the category *"chair "* or the colour *"brown"*) may predominate, weakening other attributes (such as material *"wooden"*); words of other objects (such as the *"black table"*) may cause interference. To produce more discriminative text features and fine-grained cross-modal feature fusion, we decouple the query text into different semantic components, each independently aligned with visual features, avoiding the ambiguity caused by feature coupling.

Text Component Decoupling. Analyzing grammatical dependencies between words is a fundamental task in NLP. We first use the off-the-shelf tool [60, 66] to parse the language description grammatically to generate the grammatical dependency trees, as shown in Fig. 2(b). Each sentence contains only one ROOT node, and each remaining word has a corresponding parent node. Then according to

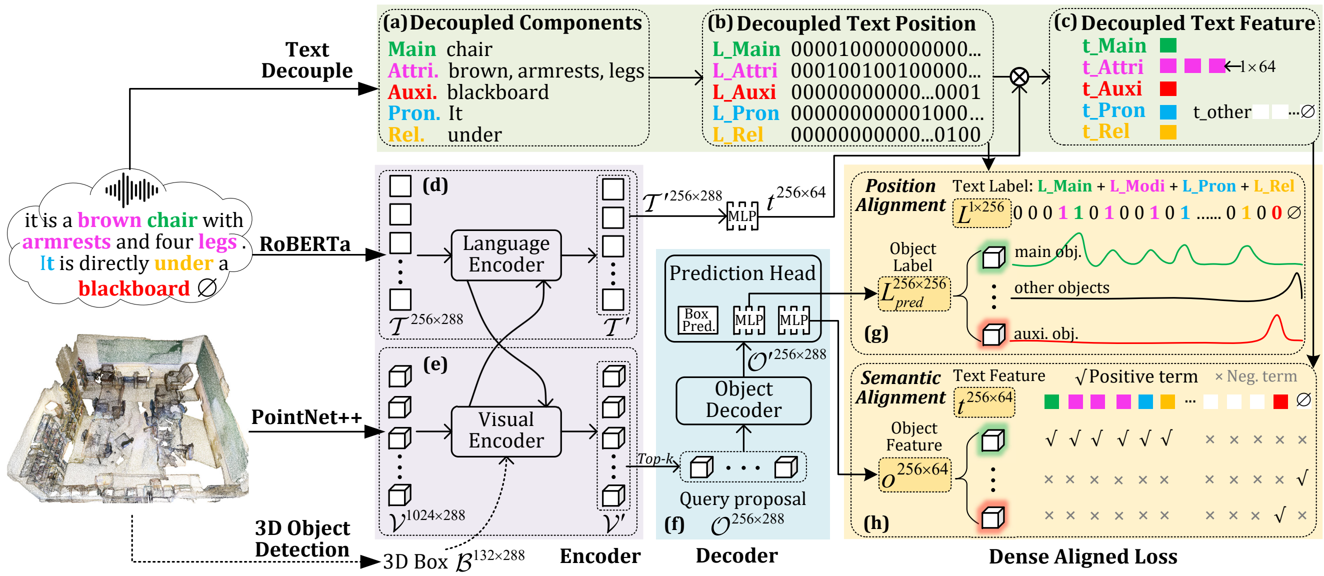


Figure 3. The system framework. (a-c): Decouple the input text into several components to acquire the position label L and features $\boldsymbol{t}$ of the decoupled text. (d-e): Transformer-based encoders for cross-modal visual-text feature extraction. (f): Decode proposal features $\mathcal{O}'$ and linearly project them as object position labels L_{pred} and object features $\boldsymbol{o}$, in addition to a box prediction head for regression of the bounding box. (g-h): Visual-text feature dense alignment. Note that the additional 3D object detection procedure is optional.

the words' part-of-speech and dependencies, we decouple the long text into ***five semantic components*** (see Fig. 2(c)): `Main object` - the target object mentioned in the utterance; `Auxiliary object` - the one used to assist in locating the main object; `Attributes` - objects' appearance, shape, *etc.*; `Pronoun` - the word instead of the main object; `Relationship` - the spatial relation between the main object and the auxiliary object. Note that attributes affiliated with pronouns are equivalent to attached with the main object, thus connecting two sentences in an utterance.

Text Position Decoupling. After decoupling each text component (Fig. 3(a)), we generate the position label (similar to a mask) $L_{main}, L_{attri}, L_{auxi}, L_{pron}, L_{rel} \in \mathbb{R}^{1\times l}$ for the component's associated word (Fig. 3(b)). Where l=256 is the maximum length of the text, each component's word position is set to 1 and the rest to 0. The label will be used to construct the position alignment loss and supervise the classification of objects. The classification result, is not one of a predetermined number of object categories but the position of the text with the highest semantic similarity.

Text Feature Decoupling. The feature of each word (token) is produced in the backbone of multimodal feature extraction (Fig. 3(d)). The text feature of the decoupled component can be derived by dot-multiplying all words' features $\boldsymbol{t}$ with its position label L, as shown in Fig. 3(c). The decoupled text features and visual features will be independently aligned under the supervision of the semantic alignment loss. Note that in the decoupled text features, the semantics of the corresponding components absolutely predominate, but as a result of the Transformer's attention mechanism, it also implicitly contains the global sentence's information. In other words, feature decoupling produces individual features while keeping the global context.

3.2. Multimodal Feature Extraction

We employ BUTD-DETR's encoder-decoder module for feature extraction and intermodulation of cross-modal features. We strongly recommend the reader to refer to Fig. 3.

Input Modal Tokenlization. The input text and 3D point clouds are encoded by the pre-trained RoBERTa [42] and PointNet++ [53] and produce text tokens $\mathcal{T} \in \mathbb{R}^{l\times d}$ and visual tokens $\mathcal{V} \in \mathbb{R}^{n\times d}$. Additionally, the GroupFree [44] detector is used to detect 3D boxes, which are subsequently encoded as box tokens $\mathcal{B} \in \mathbb{R}^{b\times d}$. Note that the GroupFree is ***optional***, the final predicted object of the network is from the prediction head (see below), and the box token is just to assist in better regression of the target object.

Encoder-Decoder. Self-attention and cross-attention are performed in the encoder to update both visual and text features, obtaining cross-modal features $\mathcal{V}', \mathcal{T}'$ while keeping the dimensions. The top-k (k=256) visual features are selected, linearly projected as query proposal features $\mathcal{O} \in \mathbb{R}^{k\times d}$, and updated as $\mathcal{O}'$ in the decoder.

Prediction Head. 1) The decoded proposal features $\mathcal{O}' \in \mathbb{R}^{k\times d}$ are fed into an MLP and output the predicted position labels $L_{pred} \in \mathbb{R}^{k\times l}$, which are then utilized to calculate the position alignment loss with decoupled text position labels $L \in \mathbb{R}^{1\times l}$. 2) Additionally, the proposal features are linearly projected as object features $\boldsymbol{o} \in \mathbb{R}^{k\times 64}$ by another MLP, which are then utilized to compute the semantic alignment loss with the similarly linearly projected text feature $\boldsymbol{t} \in \mathbb{R}^{l\times 64}$. 3) Lastly, a box prediction head [44] regresses the bounding box of the object.

3.3. Dense Aligned Loss

3.3.1 Dense Position Aligned Loss

The objective of position alignment is to ensure that the distribution of language-modulated visual features closely matches that of the query text description, as shown in Fig. 3(g). This process is similar to standard object detection's one-hot label prediction. However, rather than being limited by the number of categories, we predict the position of text that is similar to objects.

The constructed ground truth *text distribution* of the mentioned main object is obtained by element-wise summing the position labels of the associated decoupled text components:

$$P_{main} = \lambda_1 L_{main} + \lambda_2 L_{attri} + \lambda_3 L_{pron} + \lambda_4 L_{rel}, \quad (1)$$

where λ is the weight of different parts (refer to the parametric search in Supplementary Material.). $P_{auxi} = L_{auxi}$ represents the text distribution of the auxiliary object. The remaining candidate objects' text distribution is P_{oth}, with the final bit set to 1 (see $\varnothing$ in Fig. 3(g)). Therefore, all k candidate objects' ground truth text distribution is $P_{text} = \{P_{main}, P_{auxi}, P_{oth}\} \in \mathbb{R}^{k\times l}$.

The predicted *visual distribution* of k objects is produced by applying softmax to the output $L_{pred} \in \mathbb{R}^{k\times l}$ of the prediction head:

$$P_{obj} = Softmax(L_{pred}). \quad (2)$$

Their KL divergence is defined as the position-aligned loss:

$$\mathcal{L}_{pos} = \sum_{i=1}^{k} [P^i_{text} \log(P^i_{text}) - P^i_{text} \log(P^i_{obj})]. \quad (3)$$

We **highlight** that "dense alignment" indicates that the target object is aligned with the positions of multiple components (Eq. (1)), significantly different from BUTD-DETR (and MDETR for 2D tasks), which only sparsely aligns with the object name's position L_{main}.

3.3.2 Dense Semantic Aligned Loss

Semantic alignment aims to learn the similarity of visual-text multimodal features through contrastive learning. The object loss of semantic alignment is defined as follows:

$$\mathcal{L}_{sem\_o} = \sum_{i=1}^{k} \frac{1}{|\mathbf{T}_i^+|} \sum_{\boldsymbol{t}_i \in \mathbf{T}_i^+} -\log \left(\frac{\exp\left(w_+ * (\boldsymbol{o}_i^\top \boldsymbol{t}_i/\tau)\right)}{\sum_{j=1}^{l} \exp\left(w_- * (\boldsymbol{o}_i^\top \boldsymbol{t}_j/\tau)\right)} \right), \quad (4)$$

where $\boldsymbol{o}$ and $\boldsymbol{t}$ are the object and text features after linear projection, and $\boldsymbol{o}^\top \boldsymbol{t}/\tau$ is their similarity, as shown in Fig. 3 (h). k and l are the number of objects and words. $\boldsymbol{t}_i$ is the positive text feature of the i_{th} candidate object. Taking the main object as an example, the positive text feature $\mathbf{T}_i^+$ corresponding to it is:

$$\boldsymbol{t}_i \in \mathbf{T}_i^+ = \{\boldsymbol{t}_{main}, \boldsymbol{t}_{attri}, \boldsymbol{t}_{pron}, \boldsymbol{t}_{rel}\}, \quad (5)$$

and w_+ is the weight of each positive term. $\boldsymbol{t}_j$ is the feature of the i_{th} text, but note that the negative similarity weight w_- for auxiliary object term $\boldsymbol{t}_{auxi}$ is 2, while the rest weight 1. The text loss of semantic alignment defined similarly:

$$\mathcal{L}_{sem\_t} = \sum_{i=1}^{l} \frac{w_+}{|\mathbf{O}_i^+|} \sum_{\boldsymbol{o}_i \in \mathbf{O}_i^+} -\log \left(\frac{\exp\left(\boldsymbol{t}_i^\top \boldsymbol{o}_i/\tau\right)}{\sum_{j=1}^{k} \exp\left(\boldsymbol{t}_i^\top \boldsymbol{o}_j/\tau\right)} \right), \quad (6)$$

where $\boldsymbol{o}_i \in \mathbf{O}_i^+$ is the positive object feature of the i_{th} text, and $\boldsymbol{o}_j$ is the feature of the j_{th} object. The final semantic alignment loss is the mean of the two: $\mathcal{L}_{sem} = (\mathcal{L}_{sem\_o} + \mathcal{L}_{sem\_t})/2$.

Similarly, the semantic alignment of multiple text components (Eq. (5)) with visual features also illustrates our insight of "dense." This is intuitive, such as *"It is a brown chair with legs under a blackboard,"* where the main object's visual features should be not only similar to *"chair"* but also similar to *"brown, legs"* and distinct to *"blackboard"* as possible.

The total loss for training also includes the box regression loss. Refer to Supplementary Material for details.

3.4. Explicit Inference

Because of our text decoupling and dense alignment operations, object features fused multiple related text component features, allowing for the computation of the similarity between individual text components and candidate objects. For instance, $S_{main} = Softmax(\boldsymbol{o}^\top \boldsymbol{t}_{main}/\tau)$ indicates the similarity between objects $\boldsymbol{o}$ and the main text component $\boldsymbol{t}_{main}$. Therefore, the similarity of objects and related components can be explicitly combined to obtain the total score and select the candidate with the highest score:

$$S_{all} = S_{main} + S_{attri} + S_{pron} + S_{rel} - S_{auxi}, \quad (7)$$

where the definition of $S_{attri}, S_{pron}, S_{rel}$, and S_{auxi} is similar to S_{main}. If providing supervision of auxiliary objects during training, and the auxiliary object can be identified by solely computing the similarity between the object features and the auxiliary component's text features: $S_{all} = S_{attri}$. Being able to infer the object based on the part of the text is a significant sign that the network has learned well-aligned and fine-grained visual-text feature space.

4. Experiments

First, we conduct comprehensive and fair comparisons with SOTA methods in the Regular 3D Visual Grounding setting in Sec. 4.1. Then, in Sec. 4.2, we introduce our proposed new task, Grounding without Object Name, and perform comparison and analysis. Implementation details, additional experiments and more qualitative results are detailed in the supplementary material.

4.1. Regular 3D Visual Grounding

4.1.1 Experiment settings

We keep the same settings as existing works, with ScanRefer [8] and SR3D/NR3D [3] as datasets and Acc@0.25IoU and Acc@0.5IoU as metrics. Based on the visual data of ScanNet [17], **ScanRefer** adds 51,583 manually annotated text descriptions about objects. These complex and free-form descriptions involve object categories and attributes such as colour, shape, size, and spatial relationships. **SR3D/NR3D** is also proposed based on ScanNet, with SR3D including 83,572 simple machine-generated descriptions and NR3D containing 41,503 descriptions similar to ScanRefer's human annotation. The ***difference*** is that in the ScanRefer configuration, detecting and matching objects are required, while SR3D/NR3D is simpler. It supplies GT boxes for all candidate objects and only needs to classify the classes of the boxes and choose the target object.

4.1.2 Comparison to the state of the art

ScanRefer. Table 1 reports the results on the ScanRefer dataset. **i)** Our method achieves state-of-the-art performance by a substantial margin, with an overall improvement of 4.2% and 3.7% to 54.59% and 42.26%. **ii)** Some studies [6, 8, 9, 46, 68, 77] proved that supplemented 2D images with detailed and dense semantics could learn better point cloud features. Surprisingly, we only use sparse 3D point cloud features and even outperformed 2D assistance methods. This superiority illustrates that our decoupling and dense alignment strategies mine more efficient and meaningful visual-text co-representations. **iii)** Another finding is that the accuracy of most existing techniques is less than 40% and 30% in the **"multiple"** setting because multiple means that the category of the target object mentioned in the language is not unique, with more interference candidates with the same category. However, we reached a remarkable 49.13% and 37.64%. To identify similar objects, a finer-grained understanding of the text and vision is required in this complex setting. **iv)** The last three rows in Table 1 compare single-stage methods, where our method's single-stage implementation is without the object detection step ($\mathcal{B}$ in Fig. 3) in training and inference. The result illustrates that while not requiring an additional pre-trained 3D object detector, our approach also can achieve SOTA performance. **v)** The qualitative results are depicted in Fig. 4(a-c), which reveals that our method with an excellent perception of appearance attributes, spatial relationships, and even ordinal numbers.

SR3D/NR3D. Table 2 shows the accuracy on the SR3D/NR3D dataset, where we achieve the best performance of 68.1% and 52.1%. In SR3D, since the language descriptions are concise and the object is easy to identify, our method and [34, 46] reach an accuracy of over 60%. Conversely, in NR3D, descriptions are too detailed and complex, causing additional challenges for text decoupling. However, we still achieve SOTA accuracy with the 3D-only data, while other comparable methods [46, 68] rely on additional 2D images for training. Some methods [6, 8, 9] compared in Table 1 are not discussed here because they are not evaluated on the SR3D/NR3D dataset. In addition, because GT boxes of candidate objects are provided in this setting, the single-stage methods are not applicable and discussed.

4.1.3 Ablation studies

Loss ablation. The ablation of the position-aligned loss and the semantic-aligned loss is shown in Table 3. The performance of the semantic-aligned loss is marginally better because its contrastive loss not only shortens the distance between similar text-visual features but also enlarges the distance between dis-matched features (such as $\boldsymbol{t}_{auxi}$ is a negative term in Eq. (4)). Whereas position-aligned loss only considers object-related components (as in Eq. (1)). When both losses supervise together, the best accuracy is achieved, demonstrating that they can produce complementary performance.

Dense components ablation. To demonstrate our insight into dense alignment, we perform ablation analysis on the different decoupled text components, and the results are displayed in the **"Regular VG"** column in Table 4. Analysis: **i)** (a) is our baseline implementation of the sparse concept, using only the "Main object" component decoupled from the text. In contrast to BUTD-DETR, the text decoupling module (Sec. 3.1) is used to obtain text labels and features during training and inference instead of employing ground truth labels. **ii)** Dense-aligned sub-methods (b)-(h) outperform the sparse alignment (a) because of the finer-grained visual-linguistic feature fusion. **iii)** (b)-(e) indicate that adding any other component on top of the "Main object" improves performance, demonstrating the validity of each text component. The "Attribute" component aids in identifying characteristics such as colour, and shape, frequently mentioned in language descriptions. Unexpectedly, the "Pronoun" component such as *"it, that, and which"* have little meaning when used alone but also function in our method, indicating that the pronoun learned contextual information from the sentence. The "Relationship" component facilitates comprehension of spatial relationships be-

Method	Venue	Modality	Unique (∼19%) 0.25	Unique (∼19%) 0.5	Multiple (∼81%) 0.25	Multiple (∼81%) 0.5	**Overall** **0.25**	**Overall** **0.5**
ScanRefer [8]	ECCV2020	3D	67.64	46.19	32.06	21.26	38.97	26.10
		3D+2D	76.33	53.51	32.73	21.11	41.19	27.40
ReferIt3D [3]	ECCV2020	3D	53.8	37.5	21.0	12.8	26.4	16.9
TGNN [33]	AAAI2021	3D	68.61	56.80	29.84	23.18	37.37	29.70
InstanceRefer [70]	ICCV2021	3D	77.45	66.83	31.27	24.77	40.23	32.93
SAT [68]	ICCV2021	3D+2D	73.21	50.83	37.64	25.16	44.54	30.14
FFL-3DOG [23]	ICCV2021	3D	78.80	67.94	35.19	25.70	41.33	34.01
3DVG-Transformer [77]	ICCV2021	3D	77.16	58.47	38.38	28.70	45.90	34.47
		3D+2D	81.93	60.64	39.30	28.42	47.57	34.67
3D-SPS [46]	CVPR2022	3D+2D	84.12	66.72	40.32	29.82	48.82	36.98
3DJCG [6]	CVPR2022	3D	78.75	61.30	40.13	30.08	47.62	36.14
		3D+2D	83.47	64.34	41.39	30.82	49.56	37.33
BUTD-DETR [34] †	ECCV2022	3D	82.88	64.98	44.73	33.97	50.42	38.60
D3Net [9]	ECCV2022	3D+2D	-	**70.35**	-	30.50	-	37.87
EDA	-	3D	**85.76**	68.57	**49.13**	**37.64**	**54.59** (+4.2%)	**42.26** (+3.7%)
3D-SPS(single-stage) [46]	CVPR2022	3D	81.63	64.77	39.48	29.61	47.65	36.43
BUTD-DETR (single-stage) [34]‡	ECCV2022	3D	81.47	61.24	44.20	32.81	49.76	37.05
EDA (single-stage) §	-	3D	**86.40**	**69.42**	**48.11**	**36.82**	**53.83**	**41.70**

Table 1. The 3D visual grounding results on ScanRefer, accuracy evaluated by IoU 0.25 and IoU 0.5. † The accuracy is reevaluated using our parsed text labels because the performance reported by BUTD-DETR used ground truth text labels and ignored some challenging samples (see supplementary materials for more details). § Our single-stage implementation without the assistance of the additional 3D object detection step (dotted arrows in Fig. 3). ‡ BUTD-DETR did not provide single-stage results and we retrained the model.

Method	Venue	Modality	SR3D	NR3D
ReferIt3D [3]	ECCV20	3D	39.8	35.6
TGNN [33]	AAAI21	3D	45.0	37.3
TransRefer3D [28]	MM21	3D	57.4	42.1
InstanceRefer [70]	ICCV21	3D	48.0	38.8
3DVG-Transfor. [77]	ICCV21	3D	51.4	40.8
FFL-3DOG [23]	ICCV21	3D	-	41.7
SAT [68]	ICCV21	3D+2D	57.9	49.2
3DReferTrans. [2]	WACV22	3D	47.0	39.0
LanguageRefer [58]	CoRL22	3D	56.0	43.9
3D-SPS [46]	CVPR22	3D+2D	62.6	51.5
BUTD-DETR † [34]	ECCV22	3D	65.6	49.1
LAR [5]	NeurIPS22	3D	59.6	48.9
EDA (Ours)	-	3D	**68.1**	**52.1**

Table 2. Performance on SR3D/NR3D datasets by Acc@0.25IoU as the metric. The detailed results of EDA in four subsets are provided in the Supplementary Material. † Reevaluated by parsed text labels (see supp. for more details).

	$+\mathcal{L}_{pos}$	$+\mathcal{L}_{sem}$	Acc@0.25	Acc@0.5
(a)	✓		51.2	39.6
(b)		✓	52.2	39.9
(c)	✓	✓	**54.6**	**42.3**

Table 3. Loss ablation on the ScanRefer dataset.

	Dense components Main	Attri.	Pron.	Auxi.	Rel.	Regular VG @0.25	Regular VG @0.5	VG-w/o-ON @0.25	VG-w/o-ON @0.5
(a)	✓					51.5	38.8	21.8	16.7
(b)	✓	✓				53.1	41.3	23.2	17.8
(c)	✓		✓			52.9	40.8	23.7	18.6
(d)	✓			✓		52.9	41.0	23.2	18.4
(e)	✓				✓	52.8	40.7	25.7	19.6
(f)	✓	✓	✓			53.8	41.8	24.2	18.5
(g)	✓	✓	✓	✓		54.2	42.3	24.0	18.8
(h)	✓	✓	✓	✓	✓	**54.6**	**42.3**	**26.5**	**21.2**

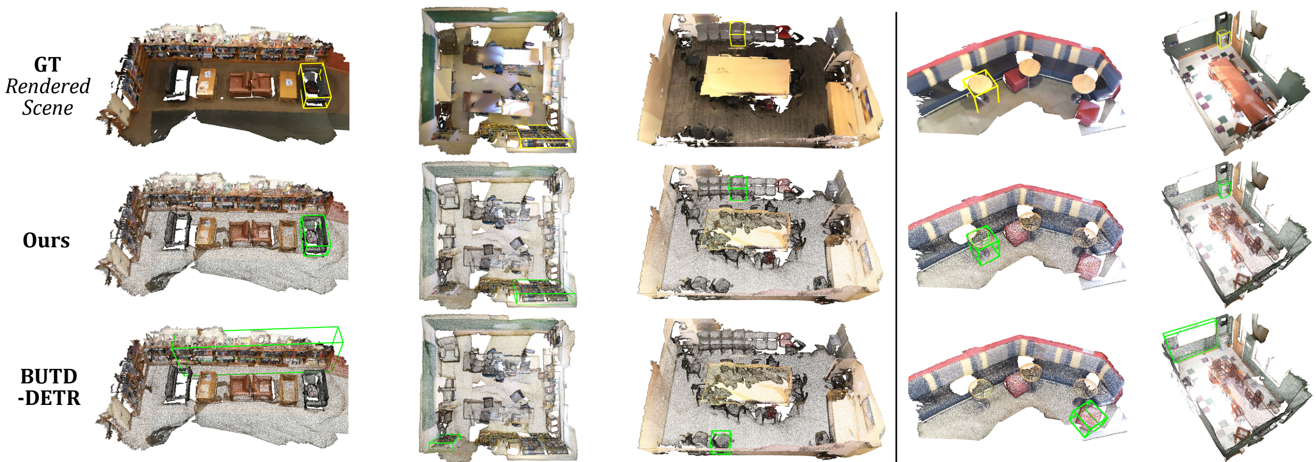
Table 4. Ablation studies of different text components on Regular VG and VG-w/o-ON tasks. Evaluated on the ScanRefer dataset.

tween objects. The component "Auxiliary object" is a negative term in the loss (Eq. (4)). During inference (Eq. (7)), its similarity is subtracted in the hopes that the predicted main object is as dissimilar to it as possible. **iv)** (f)-(h) integrate different components to make performance gains and reach the peak when all are involved, demonstrating that the functions of each component can be complementary, and there may be no overlap between the features of each one. The result reveals that our method effectively decouples and matches fine-grained multimodal features.

4.2. Grounding without Object Name (VG-w/o-ON)

4.2.1 Experiment settings

To evaluate the comprehensive reasoning ability of the model and avoid inductive biases about object names, we propose a new and more challenging task: grounding objects without mentioning object names (VG-w/o-ON). Specifically, we manually replace the object's name with *"object"* in the ScanRefer validation set. For instance: *"This is a brown wooden chair"* becomes *"this is a brown*



Text

(a) this is a black leather **loveseat**. if you were sitting in it, the long, short bookshelf would be on the right.

(b) wooden double **bookcase** filled with books. walking into the room it is in the right hand most corner next to the window.

(c) there is a **chair** with it is back to the wall. it is the fourth chair from the left.

(d) choose the first small circular **object** with a metal stand, on the right side. it has no red stool.

(e) the **object** is in the corner next to the door, below the whiteboard. it is gray, tall, and narrow.

Figure 4. Qualitative results with ScanRefer texts. (a-c): Regular 3D visual grounding. (d-e): Grounding without object name.

wooden object". A total of 9253 samples were annotated and discarded another 255 ambiguous samples. We divide this language set into four subsets: only mentioning object attributes (∼15%), only mentioning spatial relationships (∼20%), mentioning both attributes and relationships (∼63%), and others (∼2%), as the first row in Table 5. Note that, **without retraining**, we perform comparisons using our best model and comparative approaches' best models trained for the regular VG task (Sec. 4.1).

4.2.2 Result and analysis

Table 5 reports the experimental results. **i)** The performance of all methods on this challenging task is significantly lower than that of regular VG (see Table. 1), indicating that object categories provide rich semantic information, which is also conducive to classifying features. It's accessible to produce the category's inductive bias, especially in the "unique" setting. **ii)** Our method achieves absolute lead with the overall performance of 26.5% and 21.2%, which is over 10% higher than other methods. This preponderance demonstrates that our proposed text decoupling and dense alignment enable fine-grained visual-linguistic feature matching, where the model identifies the visual features most similar to other text components (*e.g.* attributes, relations, *etc.*). **iii)** Notably, in subset "Attri+Rel", our method performs better than in the other subsets because additional cues can be exploited for fine-grained localization. However, the performance of the comparison approaches on this subset drops, revealing that more clues make them suffer from ambiguity. **iv)** Column "VG-w/o-ON" in Table 4 shows ablation studies of text components for this new task. The performance increase offered by the additional components is more remarkable than during the regular VG task. Among them, the "Relationship" component plays the most significant role because the spatial relationship with other objects may provide more obvious indications in this setting. **v)** Fig. 4(d-e) shows qualitative examples. Even without knowing the target's name, our method can infer it by other cues, while BUTD-DETR's performance drops catastrophically.

Method	Subsets			**Overall**	
	Attri only	Rel only	Attri+Rel	**@0.25**	**@0.5**
ScanRefer [8]	11.17	10.53	10.29	10.51	6.20
TGNN [33]	10.52	13.32	11.35	11.64	9.51
InstanceRefer [70]	14.74	13.71	13.81	13.92	11.47
BUTD-DETR [34]	12.30	12.11	11.86	11.99	8.95
EDA (Ours)	**25.40**	**25.82**	**26.96**	**26.50**	**21.20**

Table 5. Performance of grounding without object name. The accuracy of subsets is measured by acc@0.25IoU, where the "other" subset is not reported due to its small proportion.

5. Conclusion

We present EDA, an explicit, dense-aligned method for 3D Visual Grounding tasks. By decoupling text into multiple semantic components and densely aligning it with visual features under the supervision of position-aligned and semantic-aligned loss, we enable fine-grained visual-text feature fusion, avoiding the imbalance and ambiguity of existing methods. Extensive experiments and ablation demonstrate the superiority of our method. In addition, we propose a new challenging 3D VG subtask of grounding without the object name to evaluate model robustness comprehensively. However, our limitations are the performance bottlenecks of multiple modules, such as PointNet++, RoBERTa, and text parsing modules. Especially when the text is lengthy, text decoupling may fail, resulting in performance degradation.

References

[1] Ahmed Abdelreheem, Kyle Olszewski, Hsin-Ying Lee, Peter Wonka, and Panos Achlioptas. Scanents3d: Exploiting phrase-to-3d-object correspondences for improved visio-linguistic models in 3d scenes. *arXiv preprint arXiv:2212.06250*, 2022. 2

[2] Ahmed Abdelreheem, Ujjwal Upadhyay, Ivan Skorokhodov, Rawan Al Yahya, Jun Chen, and Mohamed Elhoseiny. 3dreftransformer: Fine-grained object identification in real-world scenes using natural language. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3941–3950, 2022. 3, 7

[3] 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*, pages 422–440. Springer, 2020. 1, 6, 7

[4] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19129–19139, 2022. 3

[5] Eslam Mohamed BAKR, Yasmeen Youssef Alsaedy, and Mohamed Elhoseiny. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Proceedings of the Advances in Neural Information Processing Systems*, 2022. 7

[6] 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*, pages 16464–16473, 2022. 2, 3, 6, 7

[7] Anthony Chen, Kevin Zhang, Renrui Zhang, Zihan Wang, Yuheng Lu, Yandong Guo, and Shanghang Zhang. Pimae: Point cloud and image interactive masked autoencoders for 3d object detection. *arXiv preprint arXiv:2303.08129*, 2023. 2

[8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Proceedings of the European Conference on Computer Vision*, pages 202–221. Springer, 2020. 1, 2, 3, 6, 7, 8

[9] 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. In *Proceedings of the European Conference on Computer Vision*. Springer, 2022. 2, 6, 7

[10] 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. 2

[11] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15467–15476, 2021. 3

[12] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *Proceedings of the Advances in Neural Information Processing Systems*, 2022. 2

[13] Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Tao Chen, and Gang YU. End-to-end 3d dense captioning with vote2cap-detr. *arXiv preprint arXiv:2301.02508*, 2023. 2

[14] 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. 2

[15] Wenhao Cheng, Xingping Dong, Salman Khan, and Jianbing Shen. Learning disentanglement with decoupled labels for vision-language navigation. In *Proceedings of the European Conference on Computer Vision*, pages 309–329. Springer, 2022. 2

[16] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proceedings of the NIPS 2014 Workshop on Deep Learning*, 2014. 3

[17] 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. 6

[18] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7746–7755, 2018. 2

[19] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779, 2021. 2

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

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021. 2

[22] Yasaman Etesam, Leon Kochiev, and Angel X Chang. 3dvqa: Visual question answering for 3d environments. In *2022 19th Conference on Robots and Vision (CRV)*, pages 233–240. IEEE, 2022. 3

[23] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-form description guided 3d visual graph network for object grounding in point cloud. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 3722–3731, 2021. 2, 3, 7

[24] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Clip on wheels: Zero-shot object navigation as object localization and exploration. *arXiv preprint arXiv:2203.10421*, 2022. 3

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

[26] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. *arXiv preprint arXiv:2209.14169*, 2022. 2

[27] Huy Ha and Shuran Song. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. In *Proceedings of the 6th Annual Conference on Robot Learning*, 2022. 2

[28] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2344–2352, 2021. 3, 7

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[30] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[31] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2021. 2

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

[33] 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. 2, 7, 8

[34] 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*, pages 417–433. Springer, 2022. 2, 3, 6, 7, 8

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

[36] 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. *arXiv preprint arXiv:2203.05203*, 2022. 2

[37] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 3

[38] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838, 2022. 3

[39] Sumit Kumar, Badri N Patro, and Vinay P Namboodiri. Auto qa: The question is not only what, but also where. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 272–281, 2022. 3

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

[41] Haolin Liu, Anran Lin, Xiaoguang Han, Lei Yang, Yizhou Yu, and Shuguang Cui. Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6032–6041, 2021. 3

[42] 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. 4

[43] Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. Learning cross-modal context graph for visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11645–11652, 2020. 3

[44] 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*, pages 2949–2958, 2021. 3, 4, 5

[45] Haonan Luo, Guosheng Lin, Yazhou Yao, Fayao Liu, Zichuan Liu, and Zhenming Tang. Depth and video segmentation based visual attention for embodied question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[46] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16454–16463, 2022. 2, 3, 6, 7

[47] 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. 3

[48] Sudipta Paul, Amit K Roy-Chowdhury, and Anoop Cherian. Avlen: Audio-visual-language embodied navigation in 3d environments. *arXiv preprint arXiv:2210.07940*, 2022. 3

[49] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3

[50] Mihir Prabhudesai, Hsiao-Yu Fish Tung, Syed Ashar Javed, Maximilian Sieb, Adam W. Harley, and Katerina Fragkiadaki. Embodied language grounding with 3d visual feature representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[51] 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. 3

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

[53] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2, 4

[54] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020. 2

[55] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. Hop: History-and-order aware pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15418–15427, 2022. 2

[56] 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 *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[58] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Proceedings of the Conference on Robot Learning*, pages 1046–1056. PMLR, 2022. 7

[59] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *Proceedings of the European Conference on Computer Vision*. Springer, 2022. 2

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

[61] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv preprint arXiv:2210.05663*, 2022. 2

[62] Sinan Tan, Mengmeng Ge, Di Guo, Huaping Liu, and Fuchun Sun. Knowledge-based embodied question answering. *arXiv preprint arXiv:2109.07872*, 2021. 3

[63] 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. 2

[64] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 3

[65] 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. 2

[66] Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6609–6618, 2019. 3

[67] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9499–9508, 2022. 2

[68] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1856–1866, 2021. 2, 3, 6, 7

[69] 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. 2

[70] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1791–1800, 2021. 2, 3, 7, 8

[71] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. GLIPv2: Unifying localization and vision-language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Proceedings of the Advances in Neural Information Processing Systems*, 2022. 2

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

[73] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Hongsheng Li, Yu Qiao, and Peng Gao. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. *arXiv preprint arXiv:2303.02151*, 2023. 2

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

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

[76] Renrui Zhang, Ziyao Zeng, Ziyu Guo, and Yafeng Li. Can language understand depth? In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6868–6874, 2022. 2

[77] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021. 3, 6, 7

[78] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv preprint arXiv:2211.11682*, 2022. 2