G^3-LQ: Marrying Hyperbolic Alignment with Explicit Semantic-Geometric Modeling for 3D Visual Grounding

Yuan Wang[1,2] Yali Li[1,2] Shengjin Wang[1,2‡]
[1]Department of Electronic Engineering, Tsinghua University, China
[2]Beijing National Research Center for Information Science and Technology (BNRist), China
wy23@mails.tsinghua.edu.cn {liyali13, wgsgj}@tsinghua.edu.cn ‡Corresponding Author

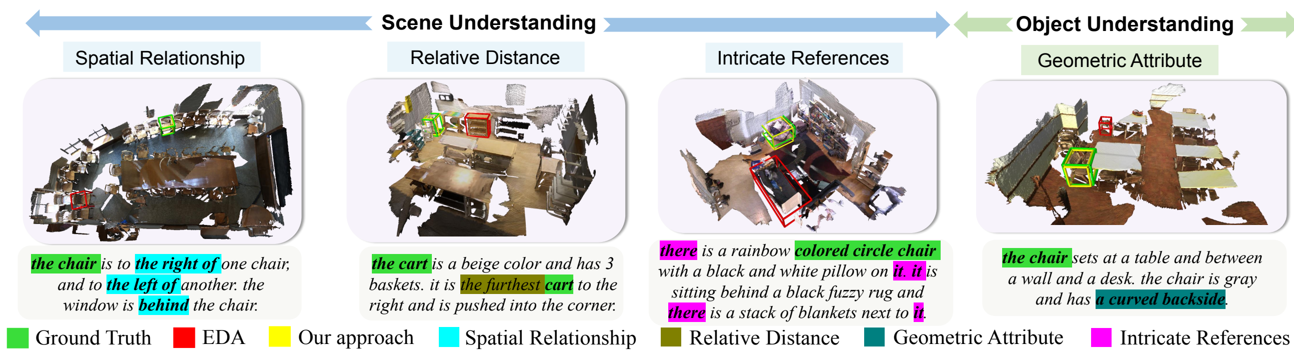

Figure 1. Illustration of the proposed G^3-LQ method on the *ScanRefer* benchmark. The versatile *scene-level and object-level* understanding properties empower G^3-LQ to comprehend **spatial relationship**, **relative distance**, **geometric attribute** as well as **intricate references**.

Abstract

*Grounding referred objects in 3D scenes is a burgeoning vision-language task pivotal for propelling Embodied AI, as it endeavors to connect the 3D physical world with free-form descriptions. Compared to the 2D counterparts, challenges posed by the variability of 3D visual grounding remain relatively unsolved in existing studies: 1) the underlying geometric and complex spatial relationships in 3D scene. 2) the inherent complexity of 3D grounded language. 3) the inconsistencies between text and geometric features. To tackle these issues, we propose G^3-LQ, a DEtection TRansformer-based model tailored for 3D visual grounding task. G^3-LQ explicitly models **G**eometric-aware visual representations and **G**enerates fine-**G**rained **L**anguage-guided object **Q**ueries in an overarching framework, which comprises two dedicated modules. Specifically, the Position Adaptive Geometric Exploring (PAGE) unearths underlying information of 3D objects in the geometric details and spatial relationships perspectives. The Fine-grained Language-guided Query Selection (Flan-QS) delves into syntactic structure of texts and generates object queries that exhibit higher relevance towards fine-grained text features. Finally, a pioneering Poincaré Semantic Alignment (PSA) loss establishes semantic-geometry consistencies by modeling non-linear vision-text feature mappings and aligning them on a hyperbolic prototype—Poincaré ball. Extensive experiments verify the superiority of our G^3-LQ method, trumping the state-of-the-arts by a considerable margin.*

1. Introduction

Grounding 3D physical realm with natural languages has propelled to the forefront in advancing Embodied AI [10, 18, 35], an ability that empowers agents to comprehend human instructions in real-world contexts. 3D Visual Grounding (3D VG) [6, 31, 43, 48] has garnered substantial attention as a crucial cross-modal 3D perception task in *defacto* applications, *e.g.*, assistive robots, AR/VR, and metaverse.

Compared with extensive explorations of 2D counterpart [20, 28, 29, 49], 3D VG poses prominent challenge known as "**Semantic-Geometric Coupling**". As shown in Fig. 1, we offer distinctive glimpse into it: 1) *Intricate geometric details and spatial relations*. 3D scenes encounter multiple objects with intricate layout, multi-level occlusion, and multifaceted spatial relation. 2) *Heightened flexibility and complexity of descriptions*. Free-form utterances with complex syntax convey precise spatial information, intricate references, and detailed relational terms. 3) *Inconsistency of the semantic-geometric features*. A single semantic conception yields multiple potential correspondences within point cloud or exhibit diverse geometry in varied contexts.

To ameliorate these issues, some pioneering methods have been proposed. They have largely focused on enhancing the representation capacity of 3D point clouds with 2D image priors [44]; capturing spatial relationships of 3D objects with powerful transformer [5, 15, 48] or graph convolution network [3, 12]; generating discriminative object proposals through instance segmentation [16] or linguistic guidance [31]. Despite commendable advancements be achieved, an equally important yet underexplored problem is **explicit semantic-geometric modeling**. Prior studies exhibit a significant dearth in explicitly modeling the intrinsic geometric attributes and spatial layouts of 3D objects, accentuating the dilemma to associate text descriptions with grounding objects. Moreover, object queries contain rich instance characteristics. But the *object proposal* [15, 48] and *query generation* [19, 31, 43] in current DETR-like models overlook explicit fine-grained interaction between text and object features, which in turn limit the model's capacity to refer potential language clues, *e.g.*, appearances, shapes, or textures. Finally, major 3D VG systems [19, 21, 43, 51] strive to establish the alignment between point cloud and semantic concepts in canonical Euclidean space. However, their efficacy falters in unraveling the complex non-linear relationships between multi-modal features and upholding the intrinsic hierarchy natures in 3D scenes and texts.

In this paper, our primary goal endeavors to explicitly delve into fine-grained representations of geometric and semantic concepts, then aligns two modalities in hyperbolic space. We thus devise an overarching G^3-LQ framework:

1) **For geometric feature modeling**, we firstly propose a **P**roxy **A**daptive **G**eometric **E**xploring (PAGE) module. PAGE is composed of the Point-Proxy Geometric Extraction and Proxy Adaptive Geometric Refinement component. The former explicitly capture the local geometric attributes (*e.g.*, distance and orientation) in a 3D local proxy, while the latter adaptively refine the geometric structure according to the position of point proxy. The geometric features are fed into the geometric encoder for high-level embedding and further achieve the interaction with text features for cross-modality enhancement. In this fashion, the PAGE module facilitates a seamless alignment of text descriptions with intrinsic geometric properties of 3D objects.

2) **For complex utterance understanding**, inspired by the dependency parsing, we aim to explore context dependency and syntactic structure of texts for guiding query selection. To this end, we propose a **F**ine-grained **Lan**guage-guided **Q**uery **S**election (Flan-QS) module, comprising the Language Scene Graph and Fine-grained Query Selection components. We decouple the free-form utterances to tree-like semantic components with directed dependency relationships as edges. Then, we establish and update the language scene graph to represent the context-aware entry features via node/edge embedding and graph context learning. Finally, we employ the constructed language scene graph to steer the generation of subsequent object candidates that are densely-aligned with fine-grained linguistic attributes.

3) **For semantic-geometric consistency**, we embed vision and text features to a hyperbolic prototype—Poincaré ball. The emerging appreciation [11, 13, 34] suggest that the hyperbolic space offers overwhelming superiority to handle data correlations. 3D scene (free-form text) exhibits an inherent hierarchical nature ranging from `scene`, `object`, `part` (from `sentence`, `phrase`, `word`). The hyperbolic space with negative curvature is **the only space** that can successfully embed the tree-shaped hierarchy [33]. Moreover, the compact Poincaré ball [13, 34] holds suitable geometric properties, enhancing its capacity to model non-linear mappings of vision and text features. Therefore, we design a **P**oincaré **S**emantic **A**lignment loss (PSA) to probe the complex mappings between vision and text features, facilitating the semantics-geometry consistency.

Our chief contributions are threefold:

- We roundly analysis the **Semantic-Geometric Coupling** of 3D VG task. Towards this issue, we propose a unified G^3-LQ framework to model intrinsic geometric features (PAGE module) and explicitly generate fine-grained object queries for 3D visual grounding (Flan-QS module).
- We propose a novel PSA loss to capture the non-linear relationship and underlying hierarchy of the vision-text features in the hyperbolic space, which further paves the way for semantics-geometry consistency learning.
- A battery of experiments on ScanRefer and Nr3d/Sr3d datasets demonstrate the effectiveness of our G^3-LQ method. G^3-LQ has proven its mettle, outperforming all existing SOTA methods by a considerable margin.

2. Related Works

2.1. 3D Visual Grounding

3D Visual Grounding has been an area of intense investigation that dedicates to pinpointing the target object in a 3D scene based on the text description. Prevailing methods are typically two-stage, which conforms to the *detection-then-matching* paradigm. Firstly, the 3D object detectors [30, 37, 48] are harnessed to yield candidate 3D object proposals, while the text features are encoded by language models [9, 27]. Secondly, these advanced methods perform a pivotal endeavor to align the vision and text features for target objects grounding. Among these, TGNN [16] utilizes the graph neural network to deduce spatial relationships among 3D object proposals, which are further enriched by text features. FFL-3DOG [12] captures the intra-modal and cross-modal relationships of text descriptions and 3D point cloud via scene graphs interaction. Recent endeavors embrace the formidable transformers [40] to model the relationship of proposals and text features. 3DVG [48] and TransRefer3D [15] achieve exceptional performance by establishing self-attention and cross-attention protocols for context perception. LanguageRefer [39] seamlessly converts the cross-modal task into a unified language modeling challenge, underpinned by predicted object labels.

However, two-stage methods encounter a noteworthy impediment: the detection stage neglects to leverage language contexts to prioritize the objects that are essential to the referring task. An alternative approach to the problem is one-stage methods [19, 31, 43], where extracted vision features are directly fused with the language features for object grounding. For instance, 3D-SPS [31] leverages text features to guide visual keypoints selection, enabling the progressive object grounding. BUTD-DETR [19] pioneers a DETR-like model, which fuses the vision-text features via a co-attention paradigm and then decodes objects in the utterance from the contextualized features. Building upon this, EDA [43] proposes a text-decoupled strategy and performs dense alignment of 3D objects and associated texts.

Despite one-stage methods unfold prowess performance, inherent limitations have been yet underexplored: 1) *They overlook the intrinsic geometric attributes* (distance, orientation, and shape) in 3D point clouds, leading to ambiguous visual representations. Conversely, our proposed PAGE module explicitly captures geometric features to avoid ambiguity. 2) *The queries generation lack the fine-grained language guidance*, limiting the model's capacity to infer potential language clues. Our tailored Flan-QS module explores language prior and context dependency to generate precise proposals. 3) *They establish the vision-text alignment in Euclidean space*, incapable of modeling the non-linear relation of vision and text features. Our method embeds multimodal features on the Poincaré ball then leverage the PSA loss to foster the semantic-geometric alignment.

2.2. Geometric Extraction in 3D Point Clouds

The explicit modeling of geometric structures has recently garnered considerable attention in the realms of 3D point cloud, *e.g.*, point cloud understanding [8, 26, 32, 41], point cloud completion [46], point cloud registration [38], and 3D object detection [45]. DGCNN [41] proposes an EdgeConv operator, by which the point features are aggregated through graph convolution to capture intricate local geometric structures. RSCNN [26] explicitly encodes geometric relationship of points with relation-aware convolution. PointMixer [8] embeds geometric relationships of 3D point in a MLP-Mixer's streamline. PoinTr [46] devises a geometry-aware transformer to learn structural knowledge and local geometric relationships for point cloud completion. HGNet [45] delves into localized shape information to generate the proposals with a hierarchical graph model for 3D object detection. Our cutting-edge G^3-LQ incorporates the geometric attributes of point cloud into the 3D VG task, thereby enhancing the discriminability of vision features.

3. Method

3.1. Overview

Provided a point cloud $\mathcal{P} = \{\boldsymbol{p}_i\}_{i=1}^{N} \in \mathbb{R}^{N\times(3+F)}$ with F-dim auxiliary features (*e.g.*, RGB, normals, or multi-view features) of N points, and a free-form text $\mathcal{T}$ with L words, the primary goal of 3D VG is to establish a mapping $\mathcal{M}$ that links $\mathcal{P}$ and $\mathcal{T}$ to the target object $\boldsymbol{o}$, *i.e.*, $\mathcal{M}(\mathcal{P}, \mathcal{T}) \to \boldsymbol{o}$.

Fig. 2 gives the outline of our proposed G^3-LQ method. Firstly, we leverage the well-accepted PointNet++ [36] to tokenize the point cloud into vision token $\mathbf{V} \in \mathbb{R}^{n\times d}$. The text is encoded by the pre-trained RoBERTa [27] and yields the vanilla text token $\mathbf{T} \in \mathbb{R}^{t\times d}$. The vision token $\mathbf{V}$ is fed into the PAGE module to capture underlying geometric features $\mathbf{V}_g \in \mathbb{R}^{n\times d}$, which adjusts according to the point positions, thereby furnishing precise geometric clues. Further, the geometric visual encoder embeds $\mathbf{V}_g$ via self-attention and interact with text features $\mathbf{T}$ via cross-attention.

Finally, we build the language scene graph $\mathcal{G}$ and obtain the updated fine-grained text features, which will explicitly guide the queries selection. Like the object queries in most DETR-like models [22, 25, 43], the selected queries $\mathbf{Q} \in \mathbb{R}^{K\times d}$ are fed into a cross-modal decoder to probe desired features and update themselves. The decoded queries $\mathbf{Q}' \in \mathbb{R}^{K\times l}$ are fed into an MLP to predict object boxes.

3.2. Geometric Exploring and Embedding

In this section, we propose a PAGE module to explicitly capture geometric properties of 3D scenes, which serves to encode holistic visual representations for 3D VG tasks.

Point-Proxy Geometric Extraction. To facilitate local geometric features modeling, we firstly instantiate point proxies predicated on the proximity of the points, thereby serving as the local regions of the point cloud. As illustrated on Fig 2, we search k neighbors of $\boldsymbol{p}_i \in \mathcal{P}$ as a point proxy with the KNN algorithm, denoted as: $\mathcal{X}_i = k\,\mathrm{NN}\,(\mathcal{P}, \boldsymbol{p}_i)$.

The point proxy $\mathcal{X}_i$ denotes index set with k closest points of $\boldsymbol{p}_i$ and $\boldsymbol{v}_{m_i} : \{\boldsymbol{v}_{m_{ij}} | j \in \mathcal{X}_i\} \in \mathbb{R}^{k\times d}$ is the corresponding features. With the Point-Proxy Geometric Extraction component, we capture the local structure by learning the geometric topology features within the point proxy $\mathcal{X}_i$. In this fashion, it is effective to achieve an inductive local representations that integrate explicit reasoning pertaining to geometric structures and objects spatial layout. We primarily compute pre-defined geometric prior $\widetilde{\boldsymbol{p}}_{ij} \in \mathbb{R}^7$:

$$\widetilde{\boldsymbol{p}}_{ij} = \mathrm{concat}\left(\boldsymbol{p}_i - \boldsymbol{p}_{ij}, \boldsymbol{p}_i, \left\|\boldsymbol{p}_i - \boldsymbol{p}_{ij}\right\|\right) \tag{1}$$

$\widetilde{\boldsymbol{p}}_{ij}$ embeds relative **orientation** and **distance** of different objects. Further, the low-level geometric features are encoded to high-level features $\boldsymbol{h}_{ij} \in \mathbb{R}^d$ with a linear layer $\phi[\cdot]$ to obtain abstract geometric conception and excavate

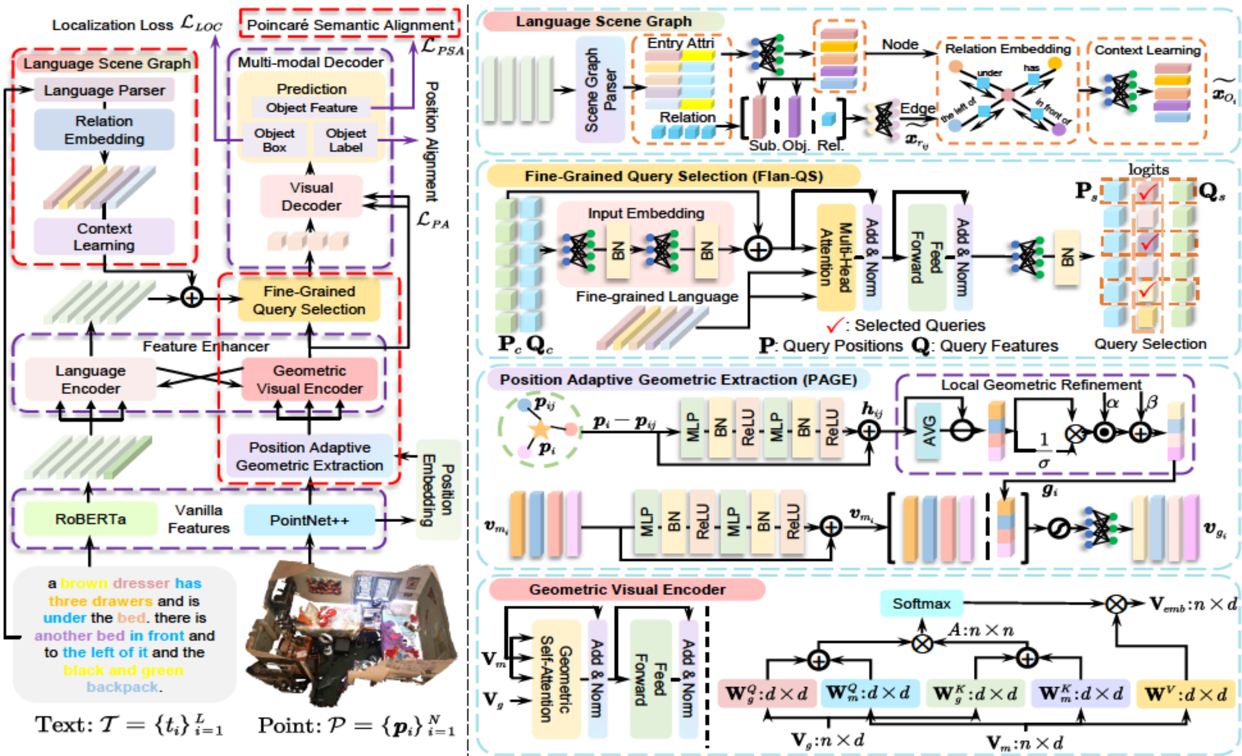

Figure 2. Illustration of (a) our proposed G^3-LQ framework. The red dashed frames denote our novel modules. (b) PAGE moudle to model the underlying geometric structure of 3D point cloud and (c) the Geometric Visual Encoder further to embed the geometric topology features. (d) Flan-QS module to capture fine-grained language context with (e) Language Scene Graph for guiding query selection.

discriminative geometric structures, which is described as:

$$h_{ij} = \widetilde{p}_{ij} + \phi\left[\widetilde{p}_{ij}\right], \forall j \in \mathcal{X}_i \tag{2}$$

Proxy-Adaptive Geometric Refinement. The presence of diverse geometric structures across distinct point proxies necessitate anisotropic extractors. Further, we flesh out this intuition and engineer a Proxy-Adaptive Geometric Refinement component, poised to model refined geometric features effectively of diverse point proxies. Given the geometric prior $\{h_{ij} | j \in \mathcal{X}_i\} \in \mathbb{R}^{k \times d}$ in $\mathcal{X}_i$, we establish local geometric refinement by the following streamline:

$$\{g_{ij}\} = \alpha \odot \frac{\{h_{ij}\} - h_i}{\Omega + \varepsilon} + \beta \tag{3}$$

where ε is a small constant for numerical stability and α, β are learnable parameters for features refinement. Ω is the standard deviation across all point proxies and channels. h_i is the center of $\mathcal{X}_i$. $g_i = \{g_{ij} | j \in \mathcal{X}_i\} \in \mathbb{R}^{k \times d}$ represents the anisotropic geometric structures in point proxy $\mathcal{X}_i$.

Finally, we opt for a softmax aggregation strategy to densely relate point features interaction in each $\mathcal{X}_i$:

$$\begin{aligned} \widetilde{v}_{g_i} &= \text{Linear}\left(\text{concat}\left(v_{m_i}, g_i\right)\right) \in \mathbb{R}^{k \times d} \\ v_{g_i} &= \sum_{j \in \mathcal{X}_i} \text{softmax}\left(\widetilde{v}_{g_{ij}}\right) \cdot \widetilde{v}_{g_{ij}} \in \mathbb{R}^{d}, \forall j \in \mathcal{X}_i, \end{aligned} \tag{4}$$

where Linear$(\cdot)$ denotes the linear projection, $\mathbf{V}_g = \{v_{g_i}\}_{i=1}^{n}$ is the output token with geometric representations.

Geometric Visual Encoder. Geometric perception enables to strengthen shape features of 3D objects and demonstrate enhanced viewpoint sensitivity, which bestows flexibility in capturing spatial relation and improves precise 3D objects grounding. Hence, we embed geometric topology features $\mathbf{V}_g$ via a tailored Geometric Encoder. As shown in Fig. 2, the core of Geometric Encoder is *Geometric self-attention layer*. Given the high-level vision features $\mathbf{V}_m \in \mathbb{R}^{n \times d}$ and the explicit geometric features $\mathbf{V}_g$, the output $\mathbf{V}_{emb} \in \mathbb{R}^{n \times d}$ is a weighted sum of projected features:

$$v_i^{emb} = \sum_{j=1}^{n} a_{ij} \left(v_{m_j} \mathbf{W}^V\right) \in \mathbb{R}^d \tag{5}$$

The attention weight a_{ij} is computed by a row-wise softmax, which can be formulated as:

$$a_{ij} = \sigma\left[\frac{\left(v_{g_i}\mathbf{W}_g^Q + v_{m_i}\mathbf{W}_m^Q\right)\left(v_{g_j}\mathbf{W}_g^K + v_{m_j}\mathbf{W}_m^K\right)^T}{\sqrt{d}}\right] \tag{6}$$

$\mathbf{W}_g^Q$, $\mathbf{W}_g^K$ are *geometric* projection matrices of query and key, $\mathbf{W}_m^Q$, $\mathbf{W}_m^K$ and $\mathbf{W}_m^V$ are *semantic* projection matrices of query, key and value. σ is the row-wise softmax operator.

Further, the geometric-enhanced features $\mathbf{V}_{emb}$ pass through the cross-attention layer to interact with the text features and box features (used in two stage), for obtaining the cross-modal features $\mathbf{V}' \in \mathbb{R}^{n\times d}$ and $\mathbf{T}' \in \mathbb{R}^{t\times d}$. Finally, $\mathbf{V}'$ is linearly projected as the candidate query $\mathbf{Q}_c$.

3.3. Query Selection

In the DETR-like model [24, 42, 50], object queries play essential roles to determine potential regions of targets, which are equally prominent for 3G VG systems, decoded into a box center and size that conform to vision tokens. Assisted by global text features, GroundingDINO [25] employs a language-guided query selection protocol to generate query proposals. However, such text features are *relatively coarse-grained*, leading to inaccurate query generation. Actually, free-form texts harbor intricate syntactic structure and possess a context-dependency nature. It further serves as inspiration for us to generate precise vision queries via improving *fine-grained* semantic understanding.

Language Scene Graph. We intend to delve into *fine-grained* text information, which is defined as *decoupled semantic components* (*e.g.*, object, attributes, and relations) and *syntactic structures*. To this end, we perform a syntactic dependency parsing to construct language graph $\mathcal{G} = \{\mathcal{O}, \mathcal{R}\}$ by the off-the-shelf Scene Graph Parser [1], which serves as a reasoning inductive bias of 3D VG tasks. $\mathcal{O} = \{\boldsymbol{o}_i\}$ and $\mathcal{R} = \{\boldsymbol{r}_{ij}\}$ are nodes and edges set. Each object $\boldsymbol{o}_i$ is denoted as a phrase with a set of attributes (appearance, shape, texture). Harnessing the final layer of the pre-trained RoBERTa model, we then acquire the object phrase embedding $\boldsymbol{x}_{o_i} \in \mathbb{R}^d$ as the graph node $\boldsymbol{o}_i \in \mathcal{O}$. In a similar vein, the relation phrases are also encoded into the context-aware embeddings $\boldsymbol{x}_{r_{ij}} \in \mathbb{R}^d$ akin to the graph edges $\boldsymbol{r}_{ij} \in \mathcal{R}$.

Inspired by the message passing mechanism [14, 23], we proceed to enhance relation and object embeddings via language graph convolution. Firstly, to enrich the contextual representations of relation embedding in connection with its affiliated nodes, we aggregate messages and update its features with the subject $\boldsymbol{x}_{o_i}$ and object node $\boldsymbol{x}_{o_j}$:

$$\widetilde{\boldsymbol{x}}_{r_{ij}} = \boldsymbol{x}_{r_{ij}} + F_r\left(\left[\boldsymbol{x}_{o_i}; \boldsymbol{x}_{r_{ij}}; \boldsymbol{x}_{o_j}\right]\right) \tag{7}$$

where F_r is a linear projection layer and $\widetilde{\boldsymbol{x}}_{r_{ij}}$ is the context-aware relation embedding. We then update the phrase embedding $\boldsymbol{x}_{o_i}$ by passing messages from all connected nodes $\mathit{\Gamma}(i)$ and the edges with a tailored attention mechanism:

$$\widetilde{\boldsymbol{x}}_{o_i} = \boldsymbol{x}_{o_j} + \sum_{j\in\mathit{\Gamma}(i)} \boldsymbol{w}_{o_{ij}} F_o\left(\left[\boldsymbol{x}_{o_j}; \boldsymbol{x}_{r_{ij}}\right]\right) \tag{8}$$

where F_o is another linear layer and $\widetilde{\boldsymbol{x}}_{o_i}$ denotes context-aware object features. $\boldsymbol{w}_{o_{ij}}$ is the tailored attention weight:

$$\boldsymbol{w}_{o_{ij}} = \mathrm{softmax}\left[F_o\left(\left[\boldsymbol{x}_{o_i}; \widetilde{\boldsymbol{x}}_{r_{ij}}\right]\right)^T F_o\left(\left[\boldsymbol{x}_{o_j}; \widetilde{\boldsymbol{x}}_{r_{ij}}\right]\right)\right] \tag{9}$$

The attention protocol manifests the capability to discern crucial contextual features from language graph, serving as potent language priors for query selection.

Fine-grained Query Selection. We input the candidate query features $\mathbf{Q}_c \in \mathbb{R}^{n\times d}$ from the cross-modal encoding layer into the Query Embedding component, which consists of several stacked Fully-connected (FC) layers and BN layers. Further, we leverage the fine-grained text features T_f to guide the query selection through a cross-attention fusion layer. The resulting query tokens are fed into an MLP layer for confidence score prediction. Through selecting the top-K highest scoring tokens for the following decoder, we derive the updated object queries $\mathbf{Q}_s \in \mathbb{R}^{K\times d}$, which are utilized to predict a box center, height and width.

3.4. Poincaré Semantic Alignment loss

We linearly embed the queries $\mathbf{Q}_s$ to the predicted object features $\boldsymbol{q} \in \mathbb{R}^{K\times 64}$, which aligns with the embedded text features $\boldsymbol{t} \in \mathbb{R}^{t\times 64}$. Noteworthy, point clouds (spanning from scenes, objects to parts) and texts (ranging from sentences, phrases to words) encompass inherent hierarchy. Further, granularity disparities of geometric and semantic representations complicate the intricate correlations. However, it is the Achilles' heel to model hierarchical structures and complex relationships in canonical Euclidean space.

To tackle this issue, we provide the insight to map vision-text features to a special hyperbolic space—Poincaré ball, for its powerful natural aptitude in modeling tree-shaped features. Based on this, we carefully devise a PSA loss to pave the way for **semantic-geometric consistency**. Benefiting from exponential volume expansion of hyperbolic embedding, PSA explicitly models complex semantic-geometric relationship for capturing hierarchical structures.

Since the Poincaré ball is a Riemannian manifold, we project the vision $\boldsymbol{q}$ and text features $\boldsymbol{t}$ to the Poincaré ball $\mathbb{D}_c^d$ with the curvature c via an exponential mapping function $\exp_{\boldsymbol{x}}^c : \mathbb{R}^d \to \mathbb{D}_c^d$, which can be defined as follows:

$$\boldsymbol{q}^c = \exp_{\boldsymbol{x}}^c(\boldsymbol{q}) = \boldsymbol{x} \oplus_c \left(\tanh\left(\sqrt{c}\frac{\lambda_{\boldsymbol{x}}^c \|\boldsymbol{q}\|}{2}\right)\frac{\boldsymbol{q}}{\sqrt{c}\|\boldsymbol{q}\|}\right) \tag{10}$$

where $\boldsymbol{x}$ is a fixed base point to make formulas less cumbersome and $\lambda_{\boldsymbol{x}}^c = 2/\left(1 - c\|\boldsymbol{x}\|^2\right)$. $\boldsymbol{t}^c$ is defined in a similar approach. $\oplus_c$ denotes the addition operator in the Mobius gyrovector space. The *geodesic distance* between the a vision-text feature pair $\left(\boldsymbol{q}_i^c, \boldsymbol{t}_j^c\right)$ on the Poincaré ball $\mathbb{D}_c^d$ is:

$$d_c\left(\boldsymbol{q}_i^c, \boldsymbol{t}_j^c\right) = \frac{2}{\sqrt{c}}\operatorname{arctanh}\left(\sqrt{c}\|{-\boldsymbol{q}_i^c} \oplus_c \boldsymbol{t}_j^c\|\right) \tag{11}$$

Inspired by [43], we term the PSA loss to densely align the fine-grained vision-text features via contrastive learning on Poincaré ball. The query loss is defined as:

$$\mathcal{L}_{\text{psa}}^q = \sum_{i=1}^{K} \frac{-1}{\left|\mathcal{T}_i^+\right|} \sum_{\boldsymbol{t}_i^c \in \mathcal{T}_i^+} \log\left[\frac{\exp\left[\boldsymbol{w}_+(-d_c(\boldsymbol{q}_i^c, \boldsymbol{t}_i^c)/\tau)\right]}{\sum\limits_{j=1}^{l} \exp\left[\boldsymbol{w}_-(-d_c(\boldsymbol{q}_i^c, \boldsymbol{t}_j^c)/\tau)\right]}\right] \tag{12}$$

where k and l are the number of object tokens and text to-

Method	Venue	Input	Unique (∼19%) 0.25	0.5	Multiple (∼81%) 0.25	0.5	Overall 0.25	0.5
ScanRefer [6]	ECCV′20	3D+2D	76.33	53.51	32.73	21.11	41.19	27.40
TGNN [16]	AAAI′21	3D	68.61	56.80	29.84	23.18	37.37	29.70
InstanceRefer [47]	ICCV′21	3D	77.45	66.83	31.27	24.77	40.23	32.93
SAT [44]	ICCV′21	3D+2D	73.21	50.83	37.64	25.16	44.54	30.14
FFL-3DOG [12]	ICCV′21	3D	78.80	67.94	35.19	25.7	41.33	34.01
3DVG [48]	ICCV′21	3D+2D	81.93	60.64	39.3	28.42	47.57	34.67
3D-SPS [31]	CVPR′22	3D+2D	84.12	66.72	40.32	29.82	48.82	36.98
BUTD-DETR [19]	ECCV′22	3D	82.88	64.98	44.73	33.97	50.42	38.60
ViL3DRel [7]	NeurIPS′22	3D	81.58	68.62	40.30	30.71	47.94	37.73
EDA [43]	CVPR′23	3D	85.76	68.57	49.13	37.64	54.59	42.26
3DJCG [5]	CVPR′22	3D+2D	83.37	64.34	41.39	30.82	49.56	37.33
3D-VLP [21]	CVPR′23	3D+2D	84.23	64.61	43.51	33.41	51.41	39.46
3D-VisTA(Scratch) [51]	ICCV′23	3D	77.40	70.90	38.70	34.80	45.90	41.50
G^3-LQ(Two-Stage)	——	3D	**88.09**	**72.73**	**51.48**	**40.80**	**56.90**	**45.58**
3D-SPS(One-stage) [31]	CVPR′22	3D	81.63	64.77	39.48	29.61	47.65	36.43
BUTD-DETR(One-stage) [19]	ECCV′22	3D	81.47	61.24	44.20	32.81	49.76	37.05
EDA(One-stage) [43]	CVPR′23	3D	86.40	69.42	48.11	36.82	53.83	41.70
G^3-LQ(One-stage)	——	3D	**88.59**	**73.28**	**50.23**	**39.72**	**55.95**	**44.72**

Table 1. Comparison results on the *ScanRefer* dataset, in terms of the accuracy evaluated by IoU 0.25 and IoU 0.5. The unique denotes samples devoid of distracting objects, while multiple applies to remaining samples. In the one-stage setting, there is no reliance on the additional 3D object detection step. Our proposed G^3-LQ performs favorably over current state-of-the-art approach EDA [43].

kens, τ is the temperature coefficient. $\boldsymbol{w}^+$, $\boldsymbol{w}^-$ are weights of positive and negative terms. Similar to EDA, $\boldsymbol{t}_i$ is the positive text features of the i-th object query and $\mathcal{T}_i^+$ are the positive features set of $\boldsymbol{t}_i$, which primarily includes the object entries, attributes, relations, and pronouns. Moreover, the text loss can be formulated similarly:

$$\mathcal{L}_{\text{psa}}^t = \sum_{i=1}^{l} \frac{-1}{\left|\mathcal{Q}_i^+\right|} \sum_{\boldsymbol{q}_i^c \in \mathcal{Q}_i^+} \log\left[\frac{\exp\left[\boldsymbol{w}_+(-d_c(\boldsymbol{t}_i^c, \boldsymbol{q}_i^c)/\tau)\right]}{\sum_{j=1}^{K} \exp\left[\boldsymbol{w}_-(-d_c(\boldsymbol{t}_i^c, \boldsymbol{q}_j^c)/\tau)\right]}\right] \quad (13)$$

where $\boldsymbol{q}_i^c \in \mathcal{Q}_i^+$ is the positive object features of $\boldsymbol{t}_i^c$. The final loss $\mathcal{L}_{\text{psa}}$ is the average of the above two terms.

Following [43], the total loss of our method also includes the box regression loss $\mathcal{L}_{\text{loc}}$, the position alignment loss $\mathcal{L}_{\text{pa}}$, which details in the **Supplementary Material**.

4. Experiment

Firstly, we carry out comparisons with SOTA methods on ScanRefer and Nr3D/Sr3D dataset in Sec. 4.1. Further, we delve into ablation studies in Sec. 4.2. Finally, we visualize and analyze the 3D VG results in Sec. 4.3.

4.1. Quantitative Comparisons

Performance on the ScanRefer. As illustrated in Table. 1, our method consistently exhibits admirably performance, trumping all competitors by a substantial margin on all test subsets. Our method achieves 56.90% and 45.58% performance in terms of the *overall* accuracy, with a remarkable improvement compared with EDA [43], 2.31% and 2.34%. 1) Compared with 2D assistance methods [17, 44] that furnish 2D image priors imbued with elaborate semantics and textures, the proposed G^3-LQ method showcases its superiority in performance. This serves as further evidence that our method explicitly captures geometric features, thus facilitating the shape understanding of target object and the vision-text alignment. 2) Our approach, featuring fine-grained language-guided query selection, surpasses all recently published DETR-like models [19, 43]. To explain, the proposed Flan-QS module thoroughly exploits fine-grained language prior and captures global context dependency, which effectively mitigates the issues of grounding ambiguity. 3) Compared with the 3D pre-trained

Method	Nr3D Overall	Hard	Sr3D Overall	Hard
TGNN [16]	37.3	30.6	45.0	36.9
InstanceRefer [47]	38.8	31.8	48.0	40.5
3DVG [48]	40.8	34.8	51.4	44.9
LanguageRefer [39]	43.9	36.6	56.0	49.3
TransRefer3D [15]	48.0	39.6	57.4	50.2
SAT [44]	49.2	42.4	57.9	50.0
LAR [4]	48.9	42.3	59.4	51.2
3DRef [2]	47.0	38.3	39.0	32.0
3D-SPS [31]	51.5	45.1	62.6	65.4
MVT [17]	55.1	49.1	64.5	58.8
BUTD-DETR [19]	54.6	48.4	67.0	63.2
EDA [43]	52.1	46.1	68.1	62.9
3D-VisTA [51]	57.5	49.4	69.6	63.6
G^3-LQ	**58.4**	**50.7**	**73.1**	**66.3**

Table 2. Quantitative comparisons on *Nr3D* and *Sr3D* dataset. We have highlighted the top-performing three methods in purple.

ID	PAGE	Flan-QS	PSA	Unique 0.25	Unique 0.5	Multiple 0.25	Multiple 0.5
(a)	—	—	—	85.67	68.57	49.13	37.64
(b)	✓	—	—	86.89	70.26	50.46	38.61
(c)	—	✓	—	86.75	69.76	50.11	38.58
(d)	—	—	✓	86.32	69.20	49.58	38.18
(e)	✓	—	✓	87.45	71.95	51.08	40.49
(f)	—	✓	✓	87.31	70.75	50.90	40.19
(g)	✓	✓	—	87.66	72.16	50.87	40.05
(h)	✓	✓	✓	**88.09**	**72.73**	**51.48**	**40.80**

Table 3. Ablation study on the effectiveness of the proposed PAGE, Flan-QS and PSA. Evaluated on the *ScanRefer* dataset.

models [21, 51], G^3-LQ achieves SOTA accuracy, because of the semantic-geometric alignment in Poincaré space with the tailored PSA loss, which aids in modeling the complex relationships of vision-text features. 4) The "multiple" setting entails a 3D scene where the description distinguishes the target object amidst numerous distractors. Under this setting, we reach a promising performance of 51.14% and 40.08%, providing further validation for the efficacy of G^3-LQ in addressing the Semantic-Geometric Coupling issue.

Performance on the Nr3D/Sr3D. The Nr3D/Sr3D task is geared towards locating the target object among all provided ground truth candidate boxes, representing a departure from the ScanRefer dataset. The overall accuracy of the proposed G^3-LQ along with other exceptional methods are reported in Table. 2. We attain peak performance of 58.4% and 73.1% on Nr3D and Sr3D. In Nr3D, descriptions exhibit noteworthy intricacy and detail, inducing additional challenges to 3D VG task. However, our method demonstrates a noteworthy superiority over alternative methods that rely on 2D images priors [44] or 3D vision-language pre-training models [51]. In the more challenging "*Hard*" subset, our method notably improves the accuracy by 1.3% in Nr3D and 2.7% in Sr3D, which underscores our method is beneficial for distinguishing ambiguous objects by effectively modeling the underlying geometric shape and unraveling the complex relation between vision-text features.

4.2. Ablation Study and Analysis

Effectiveness of Proposed Components. We develop several alternative designs and undertake experiments to highlight the advantages yielded by various components. We herein train our G^3-LQ model on the ScanRefer dataset. (a) outlines the performance demonstrated by EDA [43], while (h) is our proposed method. 1) Comparing (a) and (b), we witness that the PAGE module delivers a remarkable gain over the baseline model. The comparison further accentuates the paramount role of explicit 3D geometric features in modeling object shape and spatial relationship, conferring benefits to 3D visual grounding. 2) The comparative result of (b) and (g) furnishes compelling evidence that the Flan-

$\boldsymbol{p}_i$	$\|\boldsymbol{p}_i - \boldsymbol{p}_{ij}\|$	$\boldsymbol{p}_i - \boldsymbol{p}_{ij}$	Unique(0.25)	Multiple(0.25)
✓	—	—	85.67	49.13
✓	✓	—	85.83	49.19
✓	—	maxpool	85.98	49.32
✓	—	softmax	86.04	49.37
✓	—	Afine	86.57	50.06
✓	✓	Afine	**86.89**	**50.46**

Table 4. Ablation study of the geometric priors selection in the proposed PAGE module on the *ScanRefer* dataset.

Query Selection	Unique(∼19%) 0.25	Unique(∼19%) 0.5	Multiple(∼81%) 0.25	Multiple(∼81%) 0.5
Parametric	81.40	60.13	38.02	28.05
Top-K(EDA)	85.67	68.57	49.13	37.64
Language	87.10	71.59	50.63	39.31
Language Graph	**87.66**	**72.16**	**50.87**	**40.05**

Table 5. Ablation study of the query selection. "Parametric" represents the selection strategy with *random* parametric queries. "Language" denotes the guidance of *sentence-level* language.

QS module exploiting language prior and context dependency, helps to alleviate the problem of grounding ambiguity and generate precise query proposals. 3) Comparing (h) and (g), it showcases the effectiveness of the proposed PSA loss to capture complex non-linear mappings of point cloud and text embeddings on Poincaré ball. When we integrate all proposed modules, a discernible surge is observed, yielding an impressive overall accuracy of 56.90% and 45.58%.

Selection of Geometric Priors. The fundamental objective of the PAGE module lies in embedding low-level geometric priors into high-level geometric features, thus the definition of $\tilde{\boldsymbol{p}}$ emerges as a topic deserving exploration. We undertake ablations with $\boldsymbol{p}_i$, $\boldsymbol{p}_i - \boldsymbol{p}_{ij}$, $\|\boldsymbol{p}_i - \boldsymbol{p}_{ij}\|$ as illustrative examples, which unveils the inherent **relative position**, **shapes**, and **distance** (see Fig. 3) among 3D objects. As shown in Table. 4, using $\|\boldsymbol{p}_i - \boldsymbol{p}_{ij}\|$ or $\boldsymbol{p}_i - \boldsymbol{p}_{ij}$ alone yields inferior performance compared with their gradual integration, the accuracy of which reach 86.89% (Uniqu) and 50.46% (Multiple) respectively. Further, when compared with max pooling and softmax operators for features aggregation in $\mathcal{X}_i$, our geometric refinement strategy attests to an improved performance by 0.59% and 0.59% (Unique). The noteworthy results accentuate the significance of the *geometric refinement* in the PAGE module, given its potential to capture position-adaptive geometric structures.

Methods of Query Selection. We carry out several experiments to validate the proposed query selection method. We delineate the *overall* performance on the ScanRefer dataset in Table. 5. 1) We adopt the *parametric queries* used in MDETR [22]. However, we observe that the random parametric queries in performance decline for its incapability to capture contextual information of the text and 3D scene. 2)

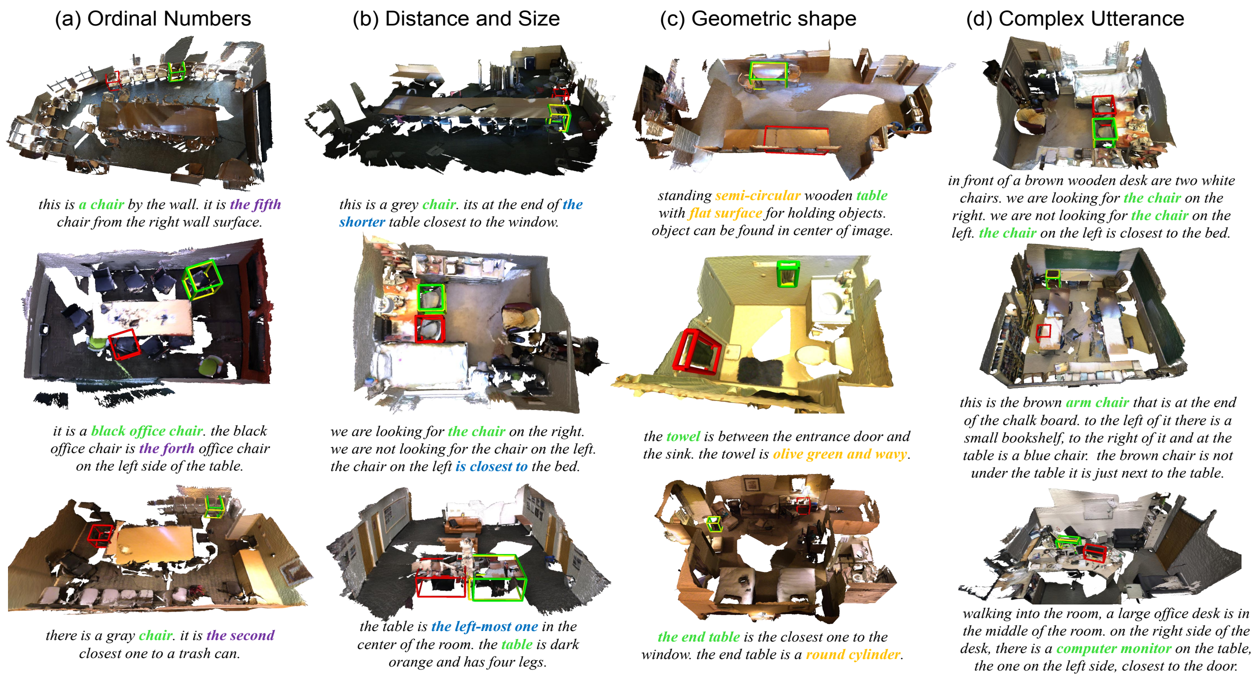

Figure 3. Qualitative results of the EDA [43] and our proposed G^3-LQ on the *ScanRefer* dataset. Our G^3-LQ manifests excellent grounding performance in terms of **ordinal numbers**, **spatial distance or object size**, **geometric attribute** and complex utterances.

We generate the query proposals guided by the *sentence-level* features. The performance of the language-guided query selection mechanism stands out prominently, underscoring the crucial role of text features in the query generation. 3) Our Flan-QS module showcases superior performance over the prevailing top-K selection and sentence-level methods. The results give explanation of the efficacy of explicit fine-grained text features. They precisely model the attributes and positions of target objects, thereby facilitating a more accurate generation of query proposals.

4.3. Qualitative Comparisons

To offer profound insight into the effectiveness of the proposed G^3-LQ, we visualize the grounding results of our method alongside the SOTA EDA [43]. As shown in Fig. 3, the predicted boxes of our G^3-LQ method and EDA are visually highlighted in yellow and red, with ground-truth boxes marked in green. The localization ability of EDA is found wanting, encountering challenges in handling the 3D objects described by **ordinal numbers**, **relative distance**, and **geometric shapes** (see Fig. 3(a)-(c)). Conversely, our G^3-LQ incorporates intrinsic geometric features and spatial relation, fostering an enriched understanding of both the scene and objects, manifests commendable grounding prowess. Secondly, Fig. 3(d) exemplifies the commendable capabilities of our method in grounding 3D objects narrated by complex utterances. To elucidate, the G^3-LQ model with the tailored Flan-QS, excels in harnessing language priors and capturing global context dependencies, generating high-quality semantic-aware query representations.

5. Conclusion

In this paper, we analysis the Semantic-Geometric Coupling issue of 3D VG tasks in three aspects: *intricate geometric details, complex text descriptions*, and *semantic-geometric inconsistency*. To this end, we propose a G^3-LQ framework to explicitly capture the fine-grained geometric and semantic concepts, which includes three crucial novelties. Firstly, a Position Adaptive Geometric Exploring module is designed to capture explicit geometric features. Secondly, with decoupled multiple semantic components, a Fine-grained Language-guided Query Selection module generates object queries densely aligned by fine-grained text features. Finally, a Poincaré Semantic Alignment loss capitalizes on vision-text hierarchy natures and achieves alignment in Poincaré space, encouraging semantic-geometric consistency. Experimental findings showcase the superior performance of our G^3-LQ across the ScanRefer and Nr3D/Sr3D benchmarks, establishing a new *state-of-the-art* standard.

6. Acknowledgement

This work is supported by the state key development program in 14th Five-Year under Grant No. 2021YFF0602103, 2021YFF0602102, 2021QY1702. Thanks for the research fund under Grant No. 2019GQG0001 from the Institute for Guo Qiang, Tsinghua University. This study is also supported by Tsinghua UniversityToyota Joint Research Center for AI Technology of Automated Vehicle(TTAD 2023-07).

References

[1] Scene graph parser. https://github.com/vacancy/SceneGraphParser, 2019. 5

[2] Ahmed Abdelreheem, Ujjwal Upadhyay, Ivan Skorokhodov, Rawan Al Yahya, Jun Chen, and Mohamed Elhoseiny. 3dref-transformer: Fine-grained object identification in real-world scenes using natural language. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 3941–3950, 2022. 6

[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. 2

[4] Eslam Bakr, Yasmeen Alsaedy, and Mohamed Elhoseiny. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. *Advances in Neural Information Processing Systems*, 35:37146–37158, 2022. 6

[5] 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. 2, 6

[6] 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, 6

[7] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in Neural Information Processing Systems*, 35:20522–20535, 2022. 6

[8] Jaesung Choe, Chunghyun Park, Francois Rameau, and Jaesik Park. Pointmixer: Mlp-mixer for point cloud understanding. In *Proceedings of the European Conference on Computer Vision*, pages 620–640. Springer, 2022. 3

[9] 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. 2

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

[11] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7409–7419, 2022. 2

[12] 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 International Conference on Computer Vision*, pages 3722–3731, 2021. 2, 6

[13] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in Neural Information Processing Systems*, 31, 2018. 2

[14] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, pages 1025–1035, 2017. 5

[15] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2344–2352, 2021. 2, 3, 6

[16] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1610–1618, 2021. 2, 6

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

[18] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022. 1

[19] 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

[20] Haojun Jiang, Yuanze Lin, Dongchen Han, Shiji Song, and Gao Huang. Pseudo-q: Generating pseudo language queries for visual grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15513–15523, 2022. 1

[21] Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, and Yinjie Lei. Context-aware alignment and mutual masking for 3d-language pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10984–10994, 2023. 2, 6, 7

[22] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1780–1790, 2021. 3, 7

[23] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. 5

[24] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 5

[25] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3, 5

[26] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8895–8904, 2019. 3

[27] 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. 2, 3

[28] 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*, pages 11645–11652, 2020. 1

[29] Yongfei Liu, Bo Wan, Lin Ma, and Xuming He. Relation-aware instance refinement for weakly supervised visual grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5612–5621, 2021. 1

[30] 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. 2

[31] 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. 1, 2, 3, 6

[32] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In *International Conference on Learning Representations*, 2021. 3

[33] Antonio Montanaro, Diego Valsesia, and Enrico Magli. Rethinking the compositionality of point clouds through regularization in the hyperbolic space. *Advances in Neural Information Processing Systems*, 35:33741–33753, 2022. 2

[34] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[35] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2017–2025, 2022. 1

[36] 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. 3

[37] 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 International Conference on Computer Vision*, pages 9277–9286, 2019. 2

[38] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11143–11152, 2022. 3

[39] 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. 3, 6

[40] 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

[41] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):1–12, 2019. 3

[42] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2567–2575, 2022. 5

[43] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual and language learning. *arXiv preprint arXiv:2209.14941*, 2022. 1, 2, 3, 5, 6, 7, 8

[44] Zhengyuan Yang, Songyang Zhang, 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. 2, 6, 7

[45] Ting Yao, Yehao Li, and Tao Mei. Hgnet: Learning hierarchical geometry from points, edges, and surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21846–21855, 2023. 3

[46] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 12498–12507, 2021. 3

[47] 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. 6

[48] 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. 1, 2, 3, 6

[49] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, and Xiaoshuai Sun. Seqtr: A simple yet universal network for visual grounding. In *Proceedings of the European Conference on Computer Vision*, pages 598–615. Springer, 2022. 1

[50] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 5

[51] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. *arXiv preprint arXiv:2308.04352*, 2023. 2, 6, 7