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, wsgsj\}@tsinghua.edu.cn  \(^\dagger\)Corresponding Author

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 geometric-aware visual representations and generates fine-grained Language-guided object Queries in an overarching framework, which comprises two dedicated modules. Specifically, the Position Adaptive Geometric Exploring (PAGE) uneartths 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 concep-
ection 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 being 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] struggle 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 Proxy Adaptive Geometric Exploring (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 Fine-grained Language-guided Query Selection (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 hyperboloid 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 Poincaré Semantic Alignment loss (PSA) to probe the complex mappings between vision and text features, facilitating the semantics-geometry consistency.

Our chief contributions are threefold:

- We roundly analyze 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 intramodal and cross-modal relationships of text descriptions and 3D point cloud via scene graphs interaction. Recent endeavors embrace the formidable transformers [40] to model

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 queries selection. Like the object queries in most DETR-like models [22, 25, 43], the selected queries Q ∈ RK×d are fed into a cross-modal decoder to probe desired features and update themselves. The decoded queries Q′ ∈ RK×l are fed into an MLP to predict object boxes.

3. Method

3.1. Overview

Provided a point cloud P = {pi}N i=1 ∈ RN×(3+F) with F-dim auxiliary features (e.g., RGB, normals, or multi-view features) of N points, and a free-form text T with L words, the primary goal of 3D VG is to establish a mapping M that links P and T to the target object o, i.e., M(P, T) → o.

Fig. 2 gives the outline of our proposed G3-LQ method. Firstly, we leverage the well-accepted PointNet++ [36] to tokenize the point cloud into vision token V ∈ Rn×d. The text is encoded by the pre-trained RoBERTa [27] and yields the vanilla text token T ∈ Rd×d. The vision token V is fed into the PAGE module to capture underlying geometric features Vg ∈ Rn×d, which adjusts according to the point positions, thereby furnishing precise geometric clues. Further, the geometric visual encoder embeds Vg via self-attention and interact with text features T via cross-attention.

Finally, we build the language scene graph 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 Q ∈ RK×d are fed into a cross-modal decoder to probe desired features and update themselves. The decoded queries Q′ ∈ RK×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 pi ∈ P as a point proxy with the KNN algorithm, denoted as: X′ i = kNN (P, pi).

The point proxy X′ i denotes index set with k closest points of pi and vi,m : {vi,m | j ∈ X′ i} ∈ RK×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 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 p̃ij ∈ R7:

\[ p̃ij = \text{concat}(pi - p̃ij, pi, ∥pi - p̃ij∥) \]  (1)

p̃ij embeds relative orientation and distance of different objects. Further, the low-level geometric features are encoded to high-level features h̃ij ∈ R d with a linear layer Φ(·) to obtain abstract geometric conception and excavate
discriminative geometric structures, which is described as:

$$h_{ij} = \tilde{p}_{ij} + \phi \left[ p_{ij} \right], \forall j \in \mathcal{X}_i$$  \hspace{1cm} (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:

$$\{g_{ij}\} = \alpha \odot \frac{[h_{ij} - h_i]}{\Omega + \varepsilon} + \beta$$  \hspace{1cm} (3)

where \(\varepsilon\) is a small constant for numerical stability and \(\alpha, \beta\) are learnable parameters for features refinement. \(\Omega\) is the standard deviation across all point proxies and channels. \(h_i\) is the center of \(\mathcal{X}_i\). \(g_{ij} = \{g_{ij}|j \in \mathcal{X}_i\} \in \mathbb{R}^{k \times d}\) 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 \(\mathcal{X}_i\):

$$\tilde{v}_{gi} = \text{Linear} \left( \text{concat} \left( v_{m_i}, g_{ij} \right) \right) \in \mathbb{R}^{k \times d}$$

$$v_{gi} = \sum_{j \in \mathcal{X}_i} \text{softmax} \left( \tilde{v}_{gij} \right) \cdot \tilde{v}_{gij} \in \mathbb{R}^{d}, \forall j \in \mathcal{X}_i,$$  \hspace{1cm} (4)

where Linear \((\cdot)\) denotes the linear projection, \(V_g = \{g_{ij}\}_{i=1}^n\) 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 \(V_g\) 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 \(V_m \in \mathbb{R}^{n \times d}\) and the explicit geometric features \(V_g\), the output \(V_{emb} \in \mathbb{R}^{n \times d}\) is a weighted sum of projected features:

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

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

$$a_{ij} = \sigma \left( \frac{(v_{g_j}W^Q_g + v_{m_j}W^Q_m)(v_{g_j}W^K_g + v_{m_j}W^K_m)^T}{\sqrt{d}} \right)$$  \hspace{1cm} (6)

\(W^Q_g\) and \(W^K_g\) are geometric projection matrices of query and key, \(W^Q_m\) and \(W^K_m\) are semantic projection matrices of query, key and value. \(\sigma\) is the row-wise softmax operator.

Further, the geometric-enhanced features \(V_{emb}\) 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 \( V' \in \mathbb{R}^{n \times d} \) and \( T' \in \mathbb{R}^{t \times d} \). Finally, \( V' \) is linearly projected as the candidate query \( Q_c \).

### 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 \( G = \{O, R\} \) by the off-the-shelf Scene Graph Parser \([1]\), which serves as a reasoning inductive bias of 3D VG tasks. \( O = \{o_i\} \) and \( R = \{r_{ij}\} \) are nodes and edges set. Each object \( o_i \) 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 \( x_{o_i} \in \mathbb{R}^d \) as the graph node \( o_i \in O \). In a similar vein, the relation phrases are also encoded into the context-aware embeddings \( x_{r_{ij}} \in \mathbb{R}^d \) akin to the graph edges \( r_{ij} \in R \).

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 \( x_{o_i} \) and object node \( x_{o_j} \):

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

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

\[
\tilde{x}_{o_i} = x_{o_i} + \sum_{j \in I(i)} w_{o_{ij}} F_o \left( [x_{o_j}; x_{r_{ij}}] \right) \tag{8}
\]

where \( F_o \) is another linear layer and \( \tilde{x}_{o_i} \) denotes context-aware object features. \( w_{o_{ij}} \) is the tailored attention weight:

\[
w_{o_{ij}} = \text{softmax} \left( F_o \left( [x_{o_j}; \tilde{x}_{r_{ij}}] \right)^T T F_o \left( [x_{o_j}; \tilde{x}_{r_{ij}}] \right) \right) \tag{9}
\]

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 \( Q_c \in \mathbb{R}^{K \times d} \) 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 \( T_j \) 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 \( Q_s \in \mathbb{R}^{K \times d} \), which are utilized to predict a box center, height and width.

### 3.4. Poincaré Semantic Alignment loss

We linearly embed the queries \( Q_s \) to the predicted object features \( q \in \mathbb{R}^{K \times 64} \), which aligns with the embedded text features \( t \in \mathbb{R}^{t \times 64} \). 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 \( q \) and text features \( t \) to the Poincaré ball \( \mathbb{D}_c^d \) with the curvature \( c \) via an exponential mapping function \( \exp^c_x : \mathbb{D}_c^d \rightarrow \mathbb{D}_c^d \), which can be defined as follows:

\[
q^c = \exp^c_x(q) = x \oplus_c \left( \tanh \left( \frac{c \lambda^c_q \|q\|}{2} \right) \frac{q}{\sqrt{c} \|q\|} \right) \tag{10}
\]

where \( x \) is a fixed base point to make formulas less cumbersome and \( \lambda^c_q = 2 / (1 - c \|x\|^2) \). \( t^c \) 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 \( (q^c_i, t^c_j) \) on the Poincaré ball \( \mathbb{D}_c^d \) is:

\[
d^c_c(q^c_i, t^c_j) = \frac{2}{\sqrt{c}} \arctanh \left( \sqrt{c} \| -q^c_i \oplus_c t^c_j \| \right) \tag{11}
\]

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:

\[
L^q_{\text{psa}} = \frac{1}{K} \sum_{i=1}^{K} \log \left[ \frac{\exp[w_c(-d^c_c(q^c_i, t^c_j)) / \tau]}{\sum_{j=1}^{K} \exp[w_c(-d^c_c(q^c_i, t^c_j)) / \tau]} \right] \tag{12}
\]

where \( k \) and \( l \) are the number of object tokens and text to-
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].

<table>
<thead>
<tr>
<th>Method</th>
<th>Venue</th>
<th>Input</th>
<th>Unique (~19%)</th>
<th>Multiple (~81%)</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScanRefer [6]</td>
<td>ECCV’20</td>
<td>3D+2D</td>
<td>76.33</td>
<td>53.51</td>
<td>21.11</td>
</tr>
<tr>
<td>TGNN [16]</td>
<td>AAAI’21</td>
<td>3D</td>
<td>68.61</td>
<td>56.80</td>
<td>29.84</td>
</tr>
<tr>
<td>InstanceRefer [47]</td>
<td>ICCV’21</td>
<td>3D</td>
<td>77.45</td>
<td>66.83</td>
<td>31.27</td>
</tr>
<tr>
<td>SAT [44]</td>
<td>ICCV’21</td>
<td>3D+2D</td>
<td>73.21</td>
<td>50.83</td>
<td>37.64</td>
</tr>
<tr>
<td>FFL-3DOG [12]</td>
<td>ICCV’21</td>
<td>3D</td>
<td>78.80</td>
<td>67.94</td>
<td>35.19</td>
</tr>
<tr>
<td>3DVG [48]</td>
<td>ICCV’21</td>
<td>3D+2D</td>
<td>81.93</td>
<td>60.64</td>
<td>39.3</td>
</tr>
<tr>
<td>3D-SPS [31]</td>
<td>CVPR’22</td>
<td>3D+2D</td>
<td>84.12</td>
<td>66.72</td>
<td>40.32</td>
</tr>
<tr>
<td>BUTD-DETR [19]</td>
<td>ECCV’22</td>
<td>3D</td>
<td>82.88</td>
<td>64.98</td>
<td>44.73</td>
</tr>
<tr>
<td>ViL3DRel [7]</td>
<td>NeurIPS’22</td>
<td>3D</td>
<td>81.58</td>
<td>68.62</td>
<td>40.30</td>
</tr>
<tr>
<td>EDA [43]</td>
<td>CVPR’23</td>
<td>3D</td>
<td>85.76</td>
<td>68.57</td>
<td>49.13</td>
</tr>
<tr>
<td>3D-JCG [5]</td>
<td>CVPR’22</td>
<td>3D+2D</td>
<td>83.37</td>
<td>64.34</td>
<td>41.39</td>
</tr>
<tr>
<td>3D-VLP [21]</td>
<td>CVPR’23</td>
<td>3D+2D</td>
<td>84.23</td>
<td>64.61</td>
<td>43.51</td>
</tr>
<tr>
<td>3D-VisTa(Scratch)</td>
<td>ICCV’23</td>
<td>3D</td>
<td>77.40</td>
<td>70.90</td>
<td>38.70</td>
</tr>
<tr>
<td>$G^3$-LQ(Two-Stage)</td>
<td>——</td>
<td>3D</td>
<td>88.09</td>
<td>72.73</td>
<td>51.48</td>
</tr>
<tr>
<td>3D-SPS(One-stage)</td>
<td>CVPR’22</td>
<td>3D</td>
<td>81.63</td>
<td>64.77</td>
<td>39.48</td>
</tr>
<tr>
<td>BUTD-DETR(One-stage)</td>
<td>ECCV’22</td>
<td>3D</td>
<td>81.47</td>
<td>61.24</td>
<td>44.20</td>
</tr>
<tr>
<td>EDA(One-stage)</td>
<td>CVPR’23</td>
<td>3D</td>
<td>86.40</td>
<td>69.42</td>
<td>48.11</td>
</tr>
<tr>
<td>$G^3$-LQ(One-stage)</td>
<td>——</td>
<td>3D</td>
<td>88.59</td>
<td>73.28</td>
<td>50.23</td>
</tr>
</tbody>
</table>

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

$$L_{pa}^t = \sum_{i=1}^{l} \frac{-1}{|Q_i^+|} \sum_{q^+_i \in Q_i^+} \log \left[ \frac{\exp \left[ w^+ \cdot (d_c(t^+_i, q^+_i)) / \tau \right]}{\sum_{j=1}^{K} \exp \left[ w^- \cdot (d_c(t^+_i, q^-_j)) / \tau \right]} \right]$$

where $q^+_i \in Q_i^+$ is the positive object features of $t_i$. The final loss $L_{pa}$ is the average of the above two terms.

Following [43], the total loss of our method also includes the box regression loss $L_{loc}$, the position alignment loss $L_{pa}$, 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]

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

Table 2. Quantitative comparisons on Nr3D and Sr3D dataset. We have highlighted the top-performing three methods in purple. 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 detects global context dependency, which effectively mitigates the issues of grounding ambiguity. 3) Compared with the 3D pre-trained
models [21, 51], G$^3$-LQ achieves SOTA accuracy, because of the semantic-geometric alignment in Poincare 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-QS module exploits 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 $\hat{p}$ emerges as a topic deserving exploration. We undertake ablations with $p_i$, $p_i - p_{ij}$, $\|p_i - p_{ij}\|$ as illustrative examples, which unveils the inherent relative position, shapes, and distance (see Fig. 3) among 3D objects. As shown in Table. 4, using $\|p_i - p_{ij}\|$ or $p_i - p_{ij}$ alone yields inferior performance compared with their gradual integration, the accuracy of which reach 86.89% (Unique) and 50.46% (Multiple) respectively. Further, when compared with max pooling and softmax operators for features aggregation in $X_i$, 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)
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. 2019GQG001 from the Institute for Guo Qiang, Tsinghua University. This study is also supported by Tsinghua University-Toyota Joint Research Center for AI Technology of Automated Vehicle(TTAD 2023-07).
References


