LIBA: Language Instructed Multi-granularity Bridge Assistant for 3D Visual Grounding

Yuan Wang[1,2], **Ya-Li Li**[1,2], **W U Eastman Z Y**[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

Abstract

3D Vision Grounding (3D-VG) seeks to unravel referential language and identify targets in 3D physical world. Prevailing methods align with the 2D-VG's pipeline to pinpoint the referred object in a categorical multi-modal reasoning manner. However, the geometric complexities of 3D scenes and the nuanced syntactic structures of language, exacerbates the **granularity inconsistency** of point cloud and text features, hindering the development of 3D-VG systems in complex scenarios. Towards this issue, we propose LIBA, a Language-Instructed multi-granularity Bridge Assistant tailored for 3D-VG task. LIBA tackles this issue as follows. (1) *How to establish a multi-granularity 3D vision-text feature alignment in a unified model*? We advance a bilateral Dynamic Bridge Adapter (DBA) build multi-granularity interaction of 3D vision and language backnones during feature extraction. We further develop the Language-aware Cross-scale Object Modulation (LCOM) module to integrate multi-scale point cloud features modulated by language information. (2) After aligning multi-modal features, *how to fully harness language model's knowledge to bolster vision concepts understanding*? A LLM-guided Hierarchical Query Selection (LLM-HQS) module incorporates world knowledge of Large Language Model (LLM) to ground the target referral via an Attribute-then-Relation reasoning process. In this manner, our LIBA inherits reasoning prowess and world knowledge of LLM to bridge point clouds and texts at multiple granularities. Experiments on ScanRefer and Nr3D/Sr3D benchmarks substantiate the superiority of our LIBA, trumping state-of-the-arts by a considerable margin.

Introduction

Pinpointing object properties and relations in 3D physical realm with natural utterances indicates a significant leap in advancing Embodied AI (Duan et al. 2022; Huang et al. 2022b), a capability that empowers agents to comprehend human directives in real-world contexts. Notably, 3D Vision Grounding (3D-VG) (Chen, Chang, and Nießner 2020; Wu et al. 2022; Luo et al. 2022) has garnered substantial attention as crucial cross-modal 3D Vision-Language (3D-VL) perception task. The 3D-VG task aims to discern a designated object in a 3D scene that corresponds to a provided text, facilitating informed decision-making of agents based primarily on the 3D world understandings.

Existing studies direct their efforts towards enhanced multi-modal representations with multi-view images (Yang, Zhang, and Luo 2021; Huang et al. 2022a), spatial relationships of 3D objects (He et al. 2021; Bakr 2022; Cai et al. 2022), fine-grained textual information (Wu et al. 2022; Feng et al. 2021; Bakr et al. 2023), and unified representations of 3D point cloud and textual features (Zhu et al. 2023; Fu et al. 2024). Despite advancements, the granularity inconsistency issue remains poorly addressed. As shown in Fig. 1(b), prior efforts (Wu et al. 2022; Wang, Li, and Wang 2024) construct language-irrelevant point cloud visual maps via **post-feature extraction**, worsening the granularity inconsistency of point clouds and texts. Further, point clouds exhibit inherent multi-scale nature, providing holistic geometric details at different levels. Existing methods (Zhao et al. 2021; Roh et al. 2022) rely on complex multi-modal feature interaction. As outlined in Fig. 1(a), they fuse point clouds and texts at a specific scale and neglect multi-scale vision-text feature alignment—a prerequisite for enriching scene representations and contextual semantics. Finally, current 3D-VG solutions fail to mimic human reasoning system nor be well-interpretable, struggling to understand ambiguities in referential language.

Toward granularity inconsistency, we propose a Language-Instructed multi-granularity Bridge Assistant (LIBA), which bridges 3D vision-text multi-granularity interaction and unleashes the abilities of Large Language Models (LLMs) for ground scene understanding with world knowledge. LIBA tackles this issue from the following two viewpoints:

(1) *How to establish a multi-granularity 3D vision-text feature alignment in a unified model*? Extensive researches (Devlin et al. 2018; Liu et al. 2019; Lin et al. 2023) confirm that distinct layers of 3D vision and language models feature hierarchical information. BERT-family models (Devlin et al. 2018; Liu et al. 2019) embed a rich hierarchy of linguistic information, while 3D vision models (Qi et al. 2017) attend hierarchical part-to-whole relationships. We thus advance Dynamic Bridge Adapter (DBA) to perceive multi-modal features **actively during extraction** and mitigate domain discrepancy of 3D vision-text features via multi-granularity interaction. DBA supports the top-down cross-modal alignment of 3D vision-text representations across different semantic hierarchies within pre-trained uni-modal encoders.

*Corresponding author.

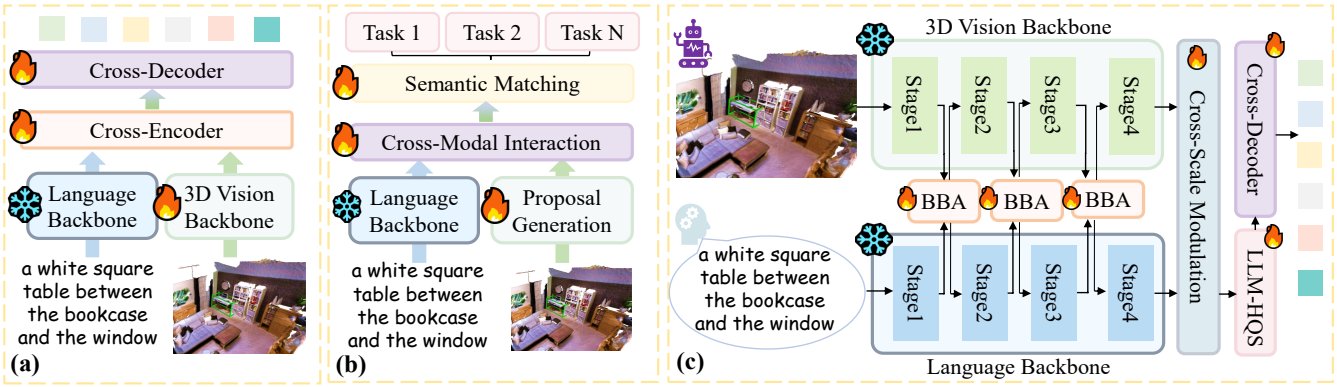

Figure 1: (a) The one-stage 3D-VG passively perceives 3D vision-text features, which are fused via single-granularity encoder. (b) The two-stage 3D-VG follows the *detection-then-matching* protocol with language-irrelevant object proposals. (c) Our **LIBA** bridge feature interaction at **multiple granularities** and generate query proposals assisted by the world knowledge of the LLMs.

Further, multi-scale feature fusion exhibits promising potential for grounding objects at different granularities (Cai et al. 2016). Existing studies depend on either single-scale features or fuse multi-scale features in a language-agnostic way, worsening the granularity inconsistency of point clouds and texts. We introduce a Language-aware Cross-scale Object Modulation (LCOM) to capture global context and local details modulated by textual representations. Unlike single-scale methods, LCOM comprehensively captures visual details and semantic attributes by aligning geometric and abstract concepts across multiple levels.

(2) *With multi-modal features aligned, how to fully harness LLM's knowledge to strengthen vision concepts understanding*? Our intuition is that language models gain knowledge about informative entry relationships from massive corpus during general sentence pre-training. Despite the rich knowledge embodied in pre-trained LLMs, it is *non-trivial* to directly transfer this knowledge to 3D object grounding. First, LLMs pale in reasoning and locating based on the 3D understandings since the remarkable modality gap in migrating linguistic knowledge to the 3D-VG task. An alternative method is to develop 3D Multi-modal Large Language Models. However, training such a scene-level 3D-MLLM requires a substantial collection of 3D scene-text pairs with the computational overhead remaining considerable. Second, existing pre-trained LLMs often exhibit subpar performance in tasks necessitating structured representations, a pivotal capacity in explicit 3D object grounding. To this end, we develop a LLM-guided Hierarchical Query Selection (LLM-HQS) module. It exploits linguistic knowledge from the Visually-Prompted Large Language Model (ViP-LLM) and Language Scene Graph Knowledge (LSGK) to guide query selection via step-wise visual clues. ViP-LLM as a fine-tuned 3D cross-modal LLM, integrates object queries as visual prompts, along with textual instructions and linguistic descriptions. It grants the LLMs to interpret the attribute of objects while maintaining its reasoning capabilities. Inspired by LISA (Lai et al. 2023), we incorporate an additional learnable token, $<$ GROUND $>$, into the vocabulary to furnish the LLM with 3D grounding prowess. Moreover, the LSGK constructs Language Scene Graph (LSG) to explore structural knowledge and context-aware text priors via Scene Graph Parser (Sce 2019). The constructed LSG provides decoupled semantic components that are densely-aligned with fine-grained spatial relationship. Such a step-wise **Attribute-then-Relation** protocol is highly interpretable and leverages the world knowledge of the LLM for understanding referential objects. Our LIBA method surpasses all state-of-the-arts (even those with pre-trained 3D-VL models) by a substantial margin on the ScanRefer and Nr3D/Sr3D benchmarks. Our contributions are as follows:

- We roundly investigate the essential issue of granularity inconsistency in the 3D-VG task. To tackle this issue, we accordingly develop a LIBA framework from the perspectives of *multi-granularity cross-modal alignment* and *knowledge-guided vision concepts understanding*.
- We propose a multi-granularity DBA module to establish holistic interaction of 3D vision and language backnones via bridging tower. A LCOM module is introduced to integrate cross-scale visual clues guided by textual features.
- We devise a LLM-HQS module which exploits LLM-based world knowledge and structural text representations to understand visual concepts in 3D physical world. It guides the query selection in a hierarchical Attribute-then-Relation protocol that mimics human reasoning.

Related Works

3D Visual Grounding

Pioneering efforts are unfolding within the nascent 3D-VG field, which aims to pinpoint the referent in a 3D scene predicated upon utterances. ScanRefer (Chen, Chang, and Nießner 2020) and ReferIt3D (Achlioptas et al. 2020) emerge as pioneers for the 3D-VG task. Prevailing methods are primarily *two-stage*, following a *detection-then-matching* paradigm. They first utilize 3D object detectors (Qi et al. 2019; Liu et al. 2021) and language models (Devlin et al. 2018; Liu et al. 2019) to generate independent 3D proposals and textual features in a passive fashion. Further, they endeavor to align

the 3D-VL features for semantic matching. (Zhao et al. 2021; Roh et al. 2022; Yang, Zhang, and Luo 2021; He et al. 2021) adopt transformer-based networks to model complex relationship for generating context-aware object proposals. Other approaches (Huang et al. 2021; Feng et al. 2021; Yuan et al. 2021) regard the object proposals as nodes and employ graph neural networks guided by textual features to aggregate object information. Besides this, some researches have ventured into the single-stage paradigms. In this contextual landscape, 3D-SPS (Luo et al. 2022) puts forth to progressively select text-relevant visual keypoints. EDA (Wu et al. 2022), BUTD-DETR (Jain et al. 2022) and G^3-LQ (Wang, Li, and Wang 2024) pioneer DETR-style models with language and objectness guidance in a bottom-up top-down fashion. However, they are limited in addressing 3D vision-language granularity inconsistency issue. Recent endeavors (Zhu et al. 2023; Jia et al. 2024; Wang et al. 2023) embark on the 3D-VL pre-training model for grounded scene understanding. (Zhu et al. 2023; Jia et al. 2024) develop substantial datasets and learn 3D joint multi-modal representations with tailored proxy scheme. (Chen et al. 2023b; Cai et al. 2022) connects the 3D-VG and 3D captioning task via unified multi-modal transformers. While excelling in 3D-VL concept comprehension, they prioritize generalized patterns and overlook fine-grained alignment associated with object referral. In departure from these, our LIBA bridges 3D vision-text alignment and harmonizes modality inconsistency at varied granularity.

3D Point Cloud Multi-modal Models

During 3D-VL models carnival in generalized point cloud understanding, CLIP-family works (Zhang et al. 2022; Huang et al. 2023) achieve text alignment with depth image projections of point clouds, seamlessly interfacing with 2D CLIP models. Tri-modality methods (Dong et al. 2022; Chen et al. 2023a) integrate object triplets (point clouds, texts, images) to learn a shared visual and textual space via building 3D pre-training models. GPT4Point (Qi et al. 2023) presents a 3D multi-modal understanding and generation framework which executes manifold point-text reference tasks. In parallel, some works have emerged in achieving 3D point cloud understanding and reasoning using LLMs. Drawing upon large-scale curated dateset, Point-LLM (Xu et al. 2023) and Scene-LLM (Fu et al. 2024) retrain the 3D-MLLM with an alignment-then-tuning strategy. 3D-LLM (Hong et al. 2023) injects 3D world into LLMs, presenting mighty capacities a machine could understand various 3D scenes following human instructions. In contrast, our work unleashes LLMs' world knowledge with a visually-prompted tuning for ground scene understanding and object relation reasoning.

Method

Dynamic Bridge Adapter

For a point cloud $\mathbf{P} = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \cdots \boldsymbol{p}_N\}$, we evenly partition the 3D point embedding layers of PointNet++ (Qi et al. 2017) into L_b blocks. Further, we bridge multi-granularity interaction of the last L_b-1 blocks, the inputs of which are visual tokens $\mathbf{F}_V^i \in \mathcal{R}^{n_i \times d_i}, i \in \{2, \cdots, L_b\}$. d_i represents the feature dimension in the i-th block and n_i is the number of points. The BERT-family DeBERTa (He et al. 2020) model is also split into L_b blocks. DeBERTa extracts word-level textual feature $\mathbf{F}_T^i \in \mathcal{R}^{L \times d}$, where L is the text length. $\mathbf{T}_g \in \mathcal{R}^d$ denotes the sentence-level textual feature.

As shown in Fig. 2(b), to capture manifold geometric patterns inherent in real-world 3D point cloud, we commence with the EdgeConv (Wang et al. 2019) layer on multi-scale visual tokens to perceive local geometric shapes and spatial relationship of 3D objects. EdgeConv **dynamically** models graph structures at varying granularities of the DBA module, thereby enriching the geometric information of point cloud. Further, we utilize the geometric-informed visual features $\mathbf{F}_V^i$ and textual features $\mathbf{F}_T^i$ to derive the language-related visual features $\hat{\mathbf{H}}_V^i$ and visual-aware textual features $\hat{\mathbf{H}}_T^i$:

$$\hat{\mathbf{H}}_V^i, \hat{\mathbf{H}}_T^i = \mathrm{BridgeAdapter}^i(\mathbf{F}_V^i + \hat{\mathbf{H}}_V^{i-1}, \mathbf{F}_T^i + \hat{\mathbf{H}}_T^{i-1}), \quad (1)$$

As depicted in Fig. 2(b), the Bridge Adapter employs a self-attention mechanism followed by a feed-forward network (FFN) to capture global relation and refine the 3D visual and textual features within their respective modalities. Further, 3D vision-text tokens cross-attend to another and update via standard key-value attention. The cross-attention weights $\mathbf{A}_T^i$, $\mathbf{A}_V^i$ are computed by a row-wise softmax normalization. We derive the cross-attended 3D vision-text features in parallel:

$$\begin{aligned} \hat{\mathbf{H}}_T^i &= \mathrm{FFN}\left[\mathbf{A}_T^{i\top} \cdot \left(\hat{\mathbf{H}}_V^i \mathbf{W}_3^{V,i}\right)\right] \\ \hat{\mathbf{H}}_V^i &= \mathrm{FFN}\left[\mathbf{A}_V^{i\top} \cdot \left(\hat{\mathbf{H}}_T^i \mathbf{W}_3^{T,i}\right)\right]. \end{aligned} \quad (2)$$

To harmonize the domain discrepancy, we propose a bilateral weighting approach to update the visual and textual features, obtaining learned bilateral residual features $\mathbf{F}_V^{i+1}$ and $\mathbf{F}_T^{i+1}$:

$$\mathbf{F}_V^{i+1} = \tau \mathbf{H}_V^i + \mathbf{F}_V^{i+1}, \mathbf{F}_T^{i+1} = \gamma \mathbf{H}_T^i + \mathbf{F}_T^{i+1} \quad (3)$$

τ and γ control the proportion of the updated features.

Language-aware Cross-scale Object Modulation

Modeling multi-scale features is a pre-requisite for enriching scene representations and contextual semantics, thus enhancing the multi-granularity alignment of vision-text features. In Fig. 3, we propose Language-aware Cross-scale Object Modulation (LCOM) to capture inter-scale dependency and integrate cross-scale visual clues guided by textual features.

After obtaining visual tokens $\mathbf{F}_V^i$ from the hierarchical 3D visual encoder, we feed them into the LCOM module for multi-scale context modeling. We set $L_b = 4$ as default and $\mathbf{F}_V^2$, $\mathbf{F}_V^3$, $\mathbf{F}_V^4$ denote the low-, middle-, high-level features. As illustrated in Fig. 3, the tri-directional feature flow conveys diverse information across various scales. Compared to $\mathbf{F}_V^3$, $\mathbf{F}_V^4$ exhibits *heightened contextual semantic*, and $\mathbf{F}_V^2$ excels in *finer details*. Bearing in mind the details and semantic features of point cloud are crucial to the 3D-VG task, we endeavor to establish the cross-scale fusion by preserving the geometric details from $\mathbf{F}_V^2$ while keeping the semantic features from $\mathbf{F}_V^4$ in a neighbor-scale interaction protocol.

To propel cross-scale visual features interaction, we add the middle-level token $\mathbf{F}_V^3$ alongside high-level token $\mathbf{F}_V^4$ and low-level token $\mathbf{F}_V^2$, yielding $\mathbf{V}_3'' \in \mathcal{R}^{n \times d}$ and $\mathbf{V}_3' \in \mathcal{R}^{n \times d}$.

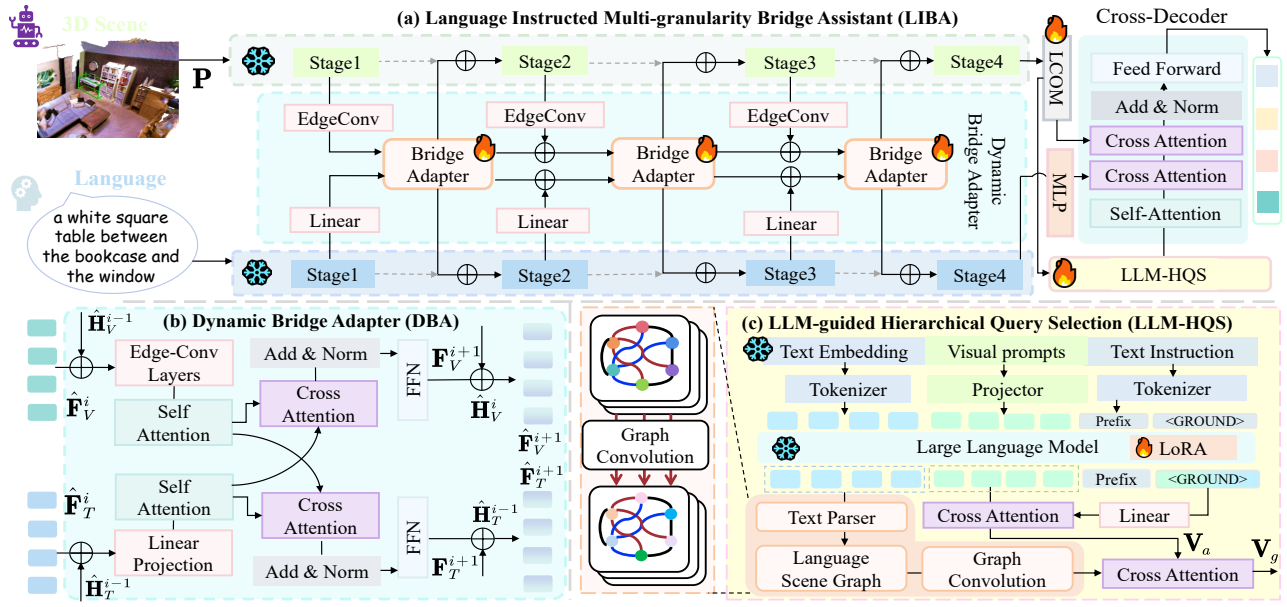

Figure 2: Illustration of our (a) LIBA. The DBA module composed of (b) bilateral Bridge Adapter blocks to bridge multi-granularity alignment of 3D vision-text features. LCOM captures multi-scale context modulated by textual features. (c) LLM-HQS exploits LLM's world knowledge and reasoning capacity for step-wise object grounding.

$\mathbf{V}_3''$ captures high-level semantics, with a broader context of the 3D scene, while $\mathbf{V}_3'$ furnishes a high-resolution representation, encapsulating intricate object details. As previously outlined, $\mathbf{T}_g$ offers a holistic scene description, while $\mathbf{F}_T^{L_b}$ provides granular object attributes. Thus, we calculate the element-wise production of $\mathbf{T}_g$ and $\mathbf{V}_3'$, $\mathbf{F}_T^{L_b}$ and $\mathbf{V}_3''$, obtaining *semantic* attention $\mathbf{A}_{\text{sem}}$ and the *detail* attention $\mathbf{A}_{\text{det}}$.

Modulated by textual features, we dynamically aggregate language-aligned visual attributes spanning different scales, thereby enriching multi-level semantic modeling of 3D point cloud and improving consistency in granularity. As shown in Fig. 3, we apply $\mathbf{A}_{\text{sem}}$ on $\mathbf{F}_V^4$ and $\mathbf{A}_{\text{det}}$ on $\mathbf{F}_V^2$ to acquire modulated cross-scale 3D features $\mathbf{V}_{\text{sem}}$ and $\mathbf{V}_{\text{det}}$.

LLM-guided Hierarchical Query Selection

Object queries in the DETR-based 3D-VG framework (Wu et al. 2022; Zhao et al. 2021; He et al. 2021; Jain et al. 2022) capture wealthy positional information, decoded into box centers and sizes for vision tokens. To obtain object queries anchored in the free-form utterance, GroundingDINO (Liu et al. 2023) generates query proposals based on global textual features, but struggles with explicit localization in complex scenarios and vague referrals. To mimic human cognitive reasoning in a step-wise fashion, we propose a LLM-HQS module inspired by the Chain-of-Thought (CoT) (Wei et al. 2022; Chowdhery et al. 2023). We model query selection into the **Attribute-then-Relation** chains: (1) the ViP-LLM excavates world knowledge for selecting all objects of the same attribute as candidates. (2) the rule-based LSGK explores structural knowledge to reason relation among objects.

Visually-Prompted Large Language Model. Acquiring $\mathbf{V}_m$, language-modulated visual features aligned with texts are designated as decoder queries. We utilize the world knowledge of LLMs to promote query generation. However, the formidable modality gap poses challenges in perceiving 3D world. *Owing to DBA, the association of visual and textual features has been bridged across multiple granular levels.* On this premise, we devise a ViP-LLM for explicit 3D object understanding in a low-resource way and equip LLMs with an awareness of 3D scenes via visually-prompted tuning.

We formulate the visual-aware prompt template with 3D visual feature and text embedding, *i.e.*, $\mathcal{I}$="[*Text embedding*][*Visually-prefixed prompts*][*Textual instruction*]". Among these, [*Visually-prefixed prompts*] refers to the 3D scene-aware token crafted by an MLP layer from multi-scale features $\mathbf{V}_m \in \mathcal{R}^{n\times d}$. Similar to (Lai et al. 2023; Wei et al. 2023), we expand LLM's vocabulary with an additional token $<$ GROUND $>$, empowering LLM with 3D grounding ability. [*Textual instruction*] represents the textual prompt for instruction learning, *e.g.*, "The object to be referred is $<$ GROUND $>$". During the training regimen, we provide the visual-aware template $\mathcal{I}$ to the LLM $\mathcal{F}$ which is fine-tuned by the LoRA (Hu et al. 2021) technique to anticipate grounding tokens positioned at the $<$ GROUND $>$. It is filled with the semantic information of the referent, projected to g_a via MLP layers for feature spaces alignment between BERT-based model and the fine-tuned LLM. Finally, the visual feature $\mathbf{V}_m$ and the token g_a are integrated into a cross-attention layer for obtaining attribute-aware queries $\mathbf{V}_a$.

Language Scene Graph Knowledge. Inspired by the graph reasoning (Kipf and Welling 2017; Hamilton, Ying, and Leskovec 2017), we decouple the textual narrative $\mathbf{T}$ with the Scene Graph Parser (Sce 2019), while rendering the object phrase $\mathcal{X} = \{\boldsymbol{x}_i\}$ as nodes and object relation

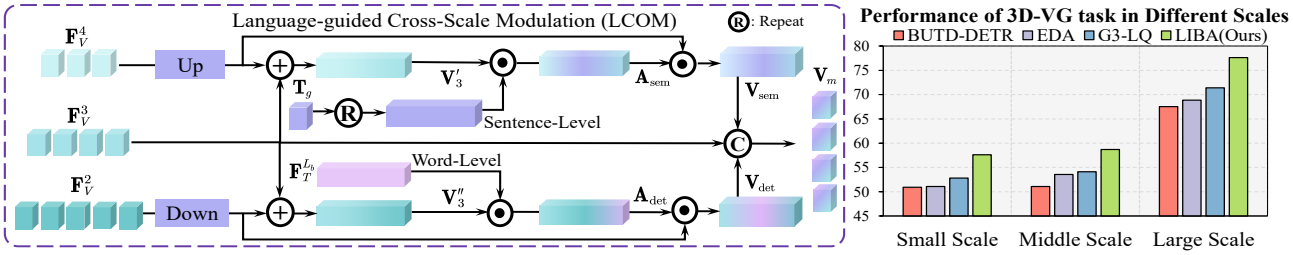

Figure 3: (*left*) Overview of proposed LCOM module. (*right*) Experiments of our LIBA and other DETR-like methods on locating 3D objects at different scales, measured by the accuracy of IoU@0.25 on ScanRefer (Chen, Chang, and Nießner 2020).

$\mathcal{E} = \{\boldsymbol{r}_{ij}\}$ as edges. To capture enriched contextual representations of relationships between affiliated nodes, we bolster the relation embedding with the subject $\boldsymbol{x}_i$ and object node $\boldsymbol{x}_j$ via graph convolution in Language Scene Graph (LSG): $\boldsymbol{r}_{ij} = \boldsymbol{r}_{ij} + \Theta_r([\boldsymbol{x}_i; \boldsymbol{r}_{ij}; \boldsymbol{x}_j])$. Θ_r is a projection layer and $\boldsymbol{r}_{ij}$ embeds context-aware relationship. We update object embedding $\boldsymbol{x}_i$ by passing messages from neighbor nodes $\mathcal{N}(i)$ with a graph attention mechanism:

$$\widetilde{\boldsymbol{x}}_i = \boldsymbol{x}_i + \sum\nolimits_{j \in \mathcal{N}(i)} \boldsymbol{w}_{ij} \Theta_o([\boldsymbol{x}_j; \boldsymbol{r}_{ij}]) \tag{4}$$

where Θ_o is a linear layer to embed context-aware object features. $\widetilde{\boldsymbol{x}}_i$ is the refined noun phrase feature and $\boldsymbol{w}_{ij}$ is the attention weight. LSGK manifests the capability to discern object relationship clues from LSG, serving as structured text priors. Finally, we utilize the updated object feature $g_r = \{\widetilde{\boldsymbol{x}}_i\}$ and $\mathbf{V}_a$ to guide query selection via another cross-attention layer, resulting in $\mathbf{V}_g$. Relation-aware object queries $\mathbf{V}_g$ are fed into the cross-decoder to predict a box center and size.

Training Objectives

Language Modeling Loss. We utilize $\mathcal{L}_{\text{LLM}}$ to supervise the LoRA (Hu et al. 2021) adapter and align predicted $\boldsymbol{y}_j$ and GT tokens, conditioning on the previous token $\boldsymbol{y}_{j'}$, visual feature $\mathbf{V}_m$ and reference text $\mathbf{T}$.

$$\mathcal{L}_{\text{LLM}} = -\sum \log p\left[\boldsymbol{y}_j \mid \boldsymbol{y}_{j'}, \mathbf{V}_m, \mathbf{T}\right], \quad j' < j \tag{5}$$

Multi-level Scene-Text Alignment Loss. For scene-text alignment, we feed the object-level 3D visual feature $\mathbf{F}_V^i$ and word-level textual feature $\mathbf{F}_T^i$ into linear layers, yielding the scene and caption features via max-pooling operations. The scene-text alignment protocol is carried out through:

$$\begin{aligned}
\mathcal{L}_{\text{MST}}^{i,vt} &= -\frac{1}{B}\sum_{k=1}^{B}\left[\log \frac{\exp\left(\boldsymbol{v}^{i,k} \cdot \boldsymbol{t}^{i,k}/\tau\right)}{\sum_{j=1}^{B}\exp\left(\boldsymbol{v}^{i,k} \cdot \boldsymbol{t}^{i,j}/\tau\right)}\right], \\
\mathcal{L}_{\text{MST}}^{i,tv} &= -\frac{1}{B}\sum_{k=1}^{B}\left[\log \frac{\exp\left(\boldsymbol{t}^{i,k} \cdot \boldsymbol{v}^{i,k}/\tau\right)}{\sum_{j=1}^{B}\exp\left(\boldsymbol{t}^{i,k} \cdot \boldsymbol{v}^{i,j}/\tau\right)}\right]
\end{aligned} \tag{6}$$

The overall multi-level scene-text alignment loss is defined as $\mathcal{L}_{\text{MST}} = \sum_{i=2}^{L_b-1}\left(\mathcal{L}_{\text{MST}}^{i,vt} + \mathcal{L}_{\text{MST}}^{i,tv}\right)$.

Experiments

Quantitative Comparisons

Performance on the ScanRefer dataset. As shown in Table 1, our LIBA method outperforms all competitors by a significant margin across all test subsets. Compared to 3D-VL pre-training methods (Zhu et al. 2023; Jia et al. 2024; Jin et al. 2023; Chen et al. 2023b) (¶) that align 3D vision-text features at a single granularity, our method excels in 3D scenes with multiple objects, where objects of the same category vary in location, scale, or environment. To explain, DBA aligns visual and textual features across various granularities. LCOM captures multi-scale visual details and semantic attributes related to textual descriptions. By leveraging the step-wise LLM-HQS module, our approach surpasses all recent DETR-like models (Wu et al. 2022; Jain et al. 2022) (♣). The superiority is blessed with twofold advantages: (1) ViP-LLM guarantees the LLMs' world knowledge grounded in the 3D scene towards explicit query selection and relation reasoning. (2) LSGK leverages structural knowledge and contextual text priors for fine-grained query selection, with exceptional performance highlighting LIBA's potential in resolving granularity inconsistency.

Performance on the Nr3D/Sr3D dataset. In Table 2, the LIBA's performance on the Sr3D/Nr3D dataset showcases an exceptional overall evaluation of 64.5% and 75.8%, the best one in addition to the 3D-VL pre-training-based methods, *i.e.*, GPS(F) and 3D-VisTA(F). However, these methods leverage proliferative 3D scene-level datasets for grounded 3D understanding and improve performance via targeted fine-tuning on downstream datasets. When compared with the *training from scratch* GPS(S) and 3D-VisTA(S) method, our method with DBA and LCOM modules to establish multi-granularity 3D vision-text alignment manifests unparalleled performance. Amid nuanced texts in the Nr3D dataset, EDA and G^3-LQ show reduced performance due to their reliance on rule-based LSGs for text parsing. In contrast, LIBA leverages LLM knowledge and enhanced spatial relations through structured LSGs, using an Attribute-then-Relation protocol, and outperforms in handling complex textual descriptions.

Grounding without Object Name. To affirm the reasoning capacity of our LIBA model, we evaluate the "*grounding without object names*" setting following EDA (Wu et al. 2022). In Table 3, our method showcases remarkable grounding capabilities, with an overall performance of 5.51% (0.25) and 4.2% (0.5). To explain, the DBA and LCOM bridges cross-modal alignment of bottom-up 3D vision-text features at different semantic granularities, which achieves precise grounding rooted in object color, geometric shape and spatial relationship, with lessened bias of object name. Further, the

Method	Venue	Input	Unique (∼19%)		Multiple(∼81%)		Overall	
			0.25	0.5	0.25	0.5	0.25	0.5
FFL-3DOG (Feng et al. 2021)	ICCV′21	3D	78.80	67.94	35.19	25.70	41.33	34.01
3DVG (Zhao et al. 2021)	ICCV′21	3D+2D	81.93	60.64	39.30	28.42	47.57	34.67
3D-SPS (Luo et al. 2022)	CVPR′22	3D+2D	84.12	66.72	40.32	29.82	48.82	36.98
BUTD-DETR (Jain et al. 2022)♣	ECCV′22	3D	82.88	64.98	44.73	33.97	50.42	38.60
D3Net (Chen et al. 2022a)	ECCV′22	3D+2D	—	70.35	—	30.50	—	37.87
ViL3DRel (Chen et al. 2022b)	NeurIPS′22	3D	81.58	68.62	40.30	30.71	47.94	37.73
3DJCG (Cai et al. 2022)	CVPR′22	3D+2D	83.37	64.34	41.39	30.82	49.56	37.33
EDA (Wu et al. 2022)♣	CVPR′23	3D	85.76	68.57	49.13	37.64	54.59	42.26
3DLP (Jin et al. 2023)¶	CVPR′23	3D+2D	84.23	64.61	43.51	33.41	51.41	39.46
M3DRef-CLIP (Zhang 2023)	ICCV′23	3D	—	77.20	—	36.80	—	44.70
3D-VisTA (Zhu et al. 2023)¶	ICCV′23	3D	81.60	75.10	43.70	39.10	50.60	45.80
UniT3D (Chen et al. 2023b)¶	ICCV′23	3D	82.75	73.14	36.36	31.05	45.27	39.14
3D-VLP (Zhang et al. 2024)¶	AAAI′24	3D	85.18	70.04	43.65	33.40	51.70	40.51
GPS (Jia et al. 2024)¶	CVPR′24	3D	—	**77.90**	—	42.70	—	48.10
G^3-LQ (Wang, Li, and Wang 2024)♣	CVPR′24	3D	88.09	72.73	51.48	40.80	56.90	45.58
MCLN (Qian et al. 2024)	ECCV′24	3D	86.89	72.73	51.96	40.76	57.17	45.53
LIBA(Ours)	—	3D	**88.81**	74.27	**54.42**	**44.41**	**59.57**	**48.96**

Table 1: Comparison results on the ScanRefer (Chen, Chang, and Nießner 2020) 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.

Method	Nr3D			Sr3D		
	All	Hard	VD	All	Hard	VD
TGNN	37.3	30.6	35.8	45.0	36.9	45.8
InstanceRefer	38.8	31.8	34.5	48.0	40.5	45.4
3DVG	40.8	34.8	34.8	51.4	44.9	44.6
LanguageRefer	43.9	36.6	41.7	56.0	49.3	49.2
TransRefer3D	48.0	39.6	42.5	57.4	50.2	49.9
SAT	49.2	42.4	46.9	57.9	50.0	49.2
LAR	48.9	42.3	47.4	59.4	51.2	50.0
ViL3DRel	64.4	57.4	62.0	72.8	67.9	63.8
3DRef	47.0	38.3	44.3	39.0	32.0	34.7
3D-SPS	51.5	45.1	48.0	62.6	65.4	49.2
MVT	55.1	49.1	54.3	64.5	58.8	58.4
BUTD-DETR	54.6	48.4	46.0	67.0	63.2	53.0
EDA	52.1	46.1	50.2	68.1	62.9	54.1
M3DRef-CLIP	49.4	43.4	42.3	—	—	—
3D-VisTA(F)	64.2	56.7	61.5	76.4	71.3	58.9
3D-VisTA(S)	57.5	49.4	53.7	69.6	63.6	57.9
GPS(F)	64.7	57.8	56.9	77.5	71.6	62.8
GPS(S)	58.7	50.9	55.8	68.4	63.4	53.1
G^3-LQ	57.0	50.7	53.8	73.1	66.3	57.2
LIBA(Ours)	64.5	57.2	60.3	75.8	70.2	61.7

Table 2: Evaluations on the Nr3D/Sr3D benchmark. (F) denotes the *fine-tuning* and (S) means training from *scratch*.

LLM-HQS utilizes the reasoning ability of LLM to guide query proposals via *descriptive features and contextual semantics*, moving beyond the reliance on object names.

Ablation Study and Analysis

The strategy of Query Selection. Table 4 reports the grounding performance on ScanRefer (Chen, Chang, and Nießner 2020) dataset with different query selection methods. (1) The heuristic rule, *e.g.*, Topk-QS (Wu et al. 2022), reveals limited generalizability when handling complex texts and delivers mundane outcomes. (2) The language-driven query selection strategy stands comparably impressive performance, which generates contextually-informed queries and unearths consistent semantic cues of vision-text features. (3) The standalone LLM- or LSGK-based protocols produce queries with object attribute and spatial position related to fine-grained textual priors, bolstering the grounding performance. (4) Leveraging explicit reasoning and structured knowledge, our LLM-HQS module refines query proposals with hierarchical *Attribute-then-Relation* protocol demonstrates its superior capabilities.

Method	Subsets			Overall	
	Att only	Rel only	Att+Rel	0.25	0.5
ScanRefer	11.17	10.53	10.29	10.51	6.20
TGNN	10.52	13.32	11.35	11.64	9.51
Instance	14.74	13.71	13.81	13.92	11.47
BUTD	12.30	12.11	11.86	11.99	8.95
EDA	25.40	25.82	26.96	26.50	21.20
G^3-LQ	26.61	26.92	27.88	27.55	21.89
Ours	**31.03**	**32.68**	**33.79**	**33.06**	**26.09**

Table 3: Results on *grounding without object name*. The accuracy of attribute / relation subsets is measured by IoU@0.25.

Effectiveness of proposed modules. In Table 5, we develop alternative designs over ScanRefer (Chen, Chang, and Nießner 2020) dataset to verify the advantage of proposed modules. Comparisons of (a)(b) reveals a discernible gain of **2.18**% (0.25) and **2.21**% (0.5) for *overall* performance, underscoring the necessity of multi-granularity alignment in dealing with 3D clustering scenes. Comparing (b)(e), the improvement is due to the LCOM module, which extracts

Figure 4: Qualitative results with *ScanRefer* texts. Our LIBA method delivers superior performance over G^3-LQ (Wang, Li, and Wang 2024) on the (a) View-Dependent. (b) Complex Spatial Relationship. (c) Grounding without Object Name cases.

Query Selection	Unique		Multiple	
	0.25	0.5	0.25	0.5
Random	82.15	67.08	41.67	33.84
Topk-QS	87.85	72.34	52.71	42.28
Language	87.98	72.70	53.29	42.79
Only LSG	88.09	72.98	53.92	43.99
Only LLM	88.42	73.66	53.46	43.49
LLM-HQS	**88.81**	**74.27**	**54.42**	**44.11**

Table 4: Ablations of the query selection. "Random" represents the random selection. "Topk-QS" is the top-k query selection protocol. "LSG" denotes the Language Scene Graph.

ID	DBA	LCOM	LLM-HQS	@0.25	@0.50
(a)	—	—	—	54.59	42.25
(b)	√	—	—	56.37	45.46
(c)	—	√	—	55.68	44.81
(d)	—	—	√	55.10	43.05
(e)	√	√	—	57.95	46.77
(f)	—	√	√	57.23	46.06
(g)	√	—	√	58.82	47.33
(h)	√	√	√	59.55	48.61

Table 5: Ablation study on the effectiveness of the proposed components on the ScanRefer dataset.

multi-scale contextual information modulated by textual features, enhancing multi-granularity consistency of 3D vision and text. Further, (e)(h) highlights the effectiveness of the LLM-HQS module, which uses hierarchical query selection to interpret visual concepts with LLM reasoning. The suboptimal performance of LLM-HQS alone underscores the need for multi-granularity feature alignment via the DBA module. Integrating all modules significantly boosts performance.

Qualitative Comparisons

To provide profound insight into our LIBA, we offer a visually compelling results in Fig. 4 (Green denotes the GT Box). Compared with G^3-LQ, LIBA performs admirably in the "*view-dependent*" setting in Fig. 4(a). Our multi-granularity feature interaction mechanism captures generalized object details and global context across views. The ViP-LLM enhances visual concept understanding, discerning 3D referent attributes without viewpoint-specific labels. Using LSGK, our method comprehends spatial relationships and fine-grained contextual information among 3D objects. Fig.4(b) confirms its superior spatial reasoning, while Fig.4(c) highlights LIBA's exceptional grounding in the "*Grounding without Object Name*" setting, using world knowledge and additional descriptive clues to pinpoint 3D objects.

Conclusion

In this work, we aim to tackle 3D vision-language granularity inconsistency issue in 3D-VG task, focusing on multi-granularity 3D vision-text feature alignment and knowledge-guided vision concept understanding. We introduce LIBA, a Language Instructed multi-granularity Bridge Assistant for 3D visual grounding. LIBA features a Dynamic Bridge Adapter for cross-modal alignment through multi-granularity interaction, a Language-aware Cross-scale Object Modulation module to capture cross-scale contextual visual cue around 3D objects, and an LLM-guided Hierarchical Query Selection for explicit query selection via an Attribute-then-Relation approach. Experiments deliver compelling proof of the LIBA's exceptional performance, showcasing its superiority over existing methods on prevailing benchmarks.

References

2019. Scene Graph Parser. https://github.com/vacancy/SceneGraphParser.

Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Proceedings of the European Conference on Computer Vision*, 422–440. Springer.

Bakr, E. 2022. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. *Advances in Neural Information Processing Systems*, 35: 37146–37158.

Bakr, E. M.; Ayman, M.; Ahmed, M.; Slim, H.; and Elhoseiny, M. 2023. CoT3DRef: Chain-of-Thoughts Data-Efficient 3D Visual Grounding. *arXiv preprint arXiv:2310.06214*.

Cai, D.; Zhao, L.; Zhang, J.; Sheng, L.; and Xu, D. 2022. 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*, 16464–16473.

Cai, Z.; Fan, Q.; Feris, R. S.; and Vasconcelos, N. 2016. A unified multi-scale deep convolutional neural network for fast object detection. In *Proceedings of the European Conference on Computer Vision*, 354–370. Springer.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Proceedings of the European Conference on Computer Vision*, 202–221. Springer.

Chen, D. Z.; Wu, Q.; Nießner, M.; and Chang, A. X. 2022a. D 3 Net: A Unified Speaker-Listener Architecture for 3D Dense Captioning and Visual Grounding. In *Proceedings of the European Conference on Computer Vision*, 487–505. Springer.

Chen, R.; Liu, Y.; Kong, L.; Zhu, X.; Ma, Y.; Li, Y.; Hou, Y.; Qiao, Y.; and Wang, W. 2023a. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7020–7030.

Chen, S.; Guhur, P.-L.; Tapaswi, M.; Schmid, C.; and Laptev, I. 2022b. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in Neural Information Processing Systems*, 35: 20522–20535.

Chen, Z.; Hu, R.; Chen, X.; Nießner, M.; and Chang, A. X. 2023b. Unit3d: A unified transformer for 3d dense captioning and visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, 18109–18119.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, R.; Qi, Z.; Zhang, L.; Zhang, J.; Sun, J.; Ge, Z.; Yi, L.; and Ma, K. 2022. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*.

Duan, J.; Yu, S.; Tan, H. L.; Zhu, H.; and Tan, C. 2022. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2): 230–244.

Feng, M.; Li, Z.; Li, Q.; Zhang, L.; Zhang, X.; Zhu, G.; Zhang, H.; Wang, Y.; and Mian, A. 2021. Free-form description guided 3d visual graph network for object grounding in point cloud. In *Proceedings of the IEEE International Conference on Computer Vision*, 3722–3731.

Fu, R.; Liu, J.; Chen, X.; Nie, Y.; and Xiong, W. 2024. Scene-LLM: Extending Language Model for 3D Visual Understanding and Reasoning. *arXiv preprint arXiv:2403.11401*.

Hamilton, W. L.; Ying, R.; and Leskovec, J. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, 1025–1035.

He, D.; Zhao, Y.; Luo, J.; Hui, T.; Huang, S.; Zhang, A.; and Liu, S. 2021. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2344–2352.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hong, Y.; Zhen, H.; Chen, P.; Zheng, S.; Du, Y.; Chen, Z.; and Gan, C. 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36: 20482–20494.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, P.-H.; Lee, H.-H.; Chen, H.-T.; and Liu, T.-L. 2021. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1610–1618.

Huang, S.; Chen, Y.; Jia, J.; and Wang, L. 2022a. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 15524–15533.

Huang, T.; Dong, B.; Yang, Y.; Huang, X.; Lau, R. W.; Ouyang, W.; and Zuo, W. 2023. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE International Conference on Computer Vision*, 22157–22167.

Huang, W.; Abbeel, P.; Pathak, D.; and Mordatch, I. 2022b. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, 9118–9147. PMLR.

Jain, A.; Gkanatsios, N.; Mediratta, I.; and Fragkiadaki, K. 2022. Bottom up top down detection transformers for language grounding in images and point clouds. In *Proceedings of the European Conference on Computer Vision*, 417–433. Springer.

Jia, B.; Chen, Y.; Yu, H.; Wang, Y.; Niu, X.; Liu, T.; Li, Q.; and Huang, S. 2024. SceneVerse: Scaling 3D Vision-Language Learning for Grounded Scene Understanding. *arXiv preprint arXiv:2401.09340*.

Jin, Z.; Hayat, M.; Yang, Y.; Guo, Y.; and Lei, Y. 2023. Context-aware Alignment and Mutual Masking for 3D-Language Pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10984–10994.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Lin, F.; Yue, Y.; Hou, S.; Yu, X.; Xu, Y.; Yamada, K. D.; and Zhang, Z. 2023. Hyperbolic chamfer distance for point cloud completion. In *Proceedings of the IEEE International Conference on Computer Vision*, 14595–14606.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Zhang, Z.; Cao, Y.; Hu, H.; and Tong, X. 2021. Group-free 3d object detection via transformers. In *Proceedings of the IEEE International Conference on Computer Vision*, 2949–2958.

Luo, J.; Fu, J.; Kong, X.; Gao, C.; Ren, H.; Shen, H.; Xia, H.; and Liu, S. 2022. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16454–16463.

Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 9277–9286.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in Neural Information Processing Systems*, 30.

Qi, Z.; Fang, Y.; Sun, Z.; Wu, X.; Wu, T.; Wang, J.; Lin, D.; and Zhao, H. 2023. GPT4Point: A Unified Framework for Point-Language Understanding and Generation. *arXiv preprint arXiv:2312.02980*.

Qian, Z.; Ma, Y.; Lin, Z.; Ji, J.; Zheng, X.; Sun, X.; and Ji, R. 2024. Multi-branch Collaborative Learning Network for 3D Visual Grounding. *arXiv preprint arXiv:2407.05363*.

Roh, J.; Desingh, K.; Farhadi, A.; and Fox, D. 2022. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, 1046–1056. PMLR.

Wang, T.; Mao, X.; Zhu, C.; Xu, R.; Lyu, R.; Li, P.; Chen, X.; Zhang, W.; Chen, K.; Xue, T.; et al. 2023. EmbodiedScan: A Holistic Multi-Modal 3D Perception Suite Towards Embodied AI. *arXiv preprint arXiv:2312.16170*.

Wang, Y.; Li, Y.; and Wang, S. 2024. Gˆ 3-LQ: Marrying Hyperbolic Alignment with Explicit Semantic-Geometric Modeling for 3D Visual Grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 13917–13926.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5): 1–12.

Wei, F.; Zhang, X.; Zhang, A.; Zhang, B.; and Chu, X. 2023. Lenna: Language enhanced reasoning detection assistant. *arXiv preprint arXiv:2312.02433*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, Y.; Cheng, X.; Zhang, R.; Cheng, Z.; and Zhang, J. 2022. Eda: Explicit text-decoupling and dense alignment for 3d visual and language learning. *arXiv preprint arXiv:2209.14941*.

Xu, R.; Wang, X.; Wang, T.; Chen, Y.; Pang, J.; and Lin, D. 2023. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*.

Yang, Z.; Zhang, S.; and Luo, J. 2021. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, 1856–1866.

Yuan, Z.; Yan, X.; Liao, Y.; Zhang, R.; Wang, S.; Li, Z.; and Cui, S. 2021. 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*, 1791–1800.

Zhang, R.; Guo, Z.; Zhang, W.; Li, K.; Miao, X.; Cui, B.; Qiao, Y.; Gao, P.; and Li, H. 2022. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8552–8562.

Zhang, T.; He, S.; Dai, T.; Wang, Z.; Chen, B.; and Xia, S.-T. 2024. Vision-Language Pre-training with Object Contrastive Learning for 3D Scene Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7296–7304.

Zhang, Y. 2023. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE International Conference on Computer Vision*, 15225–15236.

Zhao, L.; Cai, D.; Sheng, L.; and Xu, D. 2021. 3DVG-Transformer:Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2928–2937.

Zhu, Z.; Ma, X.; Chen, Y.; Deng, Z.; Huang, S.; and Li, Q. 2023. 3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment. *arXiv preprint arXiv:2308.04352*.