DenseG: Alleviating Vision-Language Feature Sparsity in Multi-View 3D Visual Grounding

Henry Zheng[1,*], Hao Shi[1,*], Yong Xien Chng[1], Rui Huang[1], Zanlin Ni[1], Tianyi Tan[1], Qihang Peng[1], Yepeng Weng[2], Zhongchao Shi[2], Gao Huang[1,†]
[1]Tsinghua University, [2]Lenovo
jh-zheng22@mails.tsinghua.edu.cn

Abstract

*Multi-view 3D visual grounding aims to locate target objects in 3D space based on natural language. Existing methods often suffer from sparse feature spaces due to limited contextual information in language descriptions and the sparse fusion of point clouds with multi-view images. In this work, we propose a novel method, DenseG, to address these challenges. For text modality, we propose an LLM Assisted Augmentation Pipeline that leverages LLMs to enhance input languages with more anchors and diverse viewpoints and construct a scene information database to provide better context, thereby enriching the language feature space. For the visual modality, we introduce a Bidirectional Text-View Images Interaction Module that retains multi-view semantic information by facilitating interaction between textual and global multi-view visual features. Our approach significantly outperforms the baseline and other teams, achieving **first place** in the Multi-View 3D Visual Grounding Track of CVPR 2024 Autonomous Grand Challenge with an AP of 59.59% at IoU 0.25 and an AP of 34.72% at IoU 0.50.*

1. Introduction

Multi-view 3D visual grounding, which aims to locate target objects in a 3D environment based on natural language descriptions, is a fundamental task for embodied agents. In recent years, the increasing attention in the field of embodied AI has garnered several works in this field [5, 7, 8, 10–13].

The task of multi-view 3D visual grounding requires a deep understanding of space, semantics, and language. Current methods integrate information from 2D, 3D, and language modalities to enrich the feature space with contextual information. Some methods [8, 11–13] reconstruct 3D from RGB-D for better spatial feature representation to locate the corresponding object described in the language. However, due to the inherent sparsity of 3D point clouds, a significant amount of multi-view semantic information is lost at the input stage. While some approaches [5, 7] attempt to enrich multi-view semantic features by applying various rotations to the point clouds, the sparse nature of point clouds still

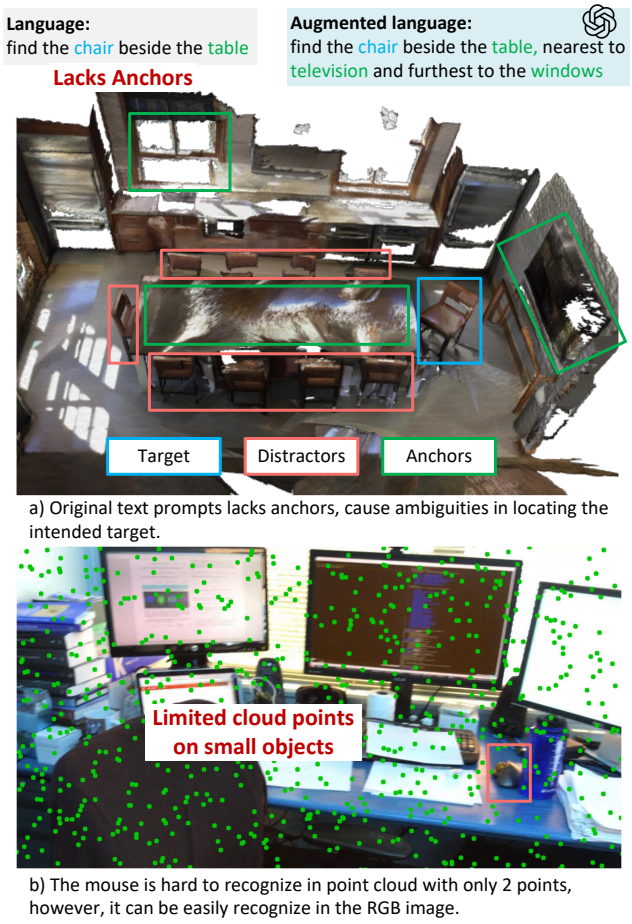

Figure 1. The sparsity problems in the language descriptions and visual features

[*]Equal Contributions
[†]Corresponding Author

limits their performance.

Instead of using painted point clouds or multiview point clouds, EmbodiedScan [10] encodes the point cloud and RGB images separately aiming to unleash the advantage of both modalities. EmbodiedScan first extracts features from 3D point clouds and multi-view 2D images. They then project the point cloud features onto the images and perform a sparse fusion before integrating language features. This approach allows for some interaction between features of adjacent regions during feature extraction, which partially alleviates the loss of multi-view semantic information. However, some crucial semantic pixels in the 2D images are often lost during the sparse fusion process, leading to a sparse visual semantic space that compromises the model's ability to retain object semantics, especially for smaller objects.

Furthermore, the input language descriptions often lack necessary anchors and are confined to single-view perspectives, leading to a sparse textual feature or lack of anchors. This absence of essential contextual details can cause the model to be easily disturbed by similar objects. Previous methods[5] have attempted to enrich these input descriptions with large language models (LLMs). However, this approach only enhances the view without improving the sparse anchor problem and simply replacing the words in the text limits the diversity in generated text.

In this work, we propose DenseG, a novel method for multi-view 3D visual grounding that alleviates the sparsity in both visual and textual features. Specifically, for the text modality, we construct a scene information database based on the existing dataset and leverage LLMs to enrich input descriptions with more anchors and diverse viewpoints, reducing confusion with similar objects and providing more robust representations. The database provides external knowledge to improve the diversity in the generated text while concurrently mitigating hallucinations in LLM. We further utilize LLMs to verify the enriched descriptions to ensure accuracy and reliability, further reducing hallucinations.

For the visual modality, we introduce a Bidirectional Text-View Images Interaction Module (Bi-TVI) to retain more multi-view semantic information. Before fusing point cloud and image features, we append a learnable token to the flattened feature map of each view image to obtain the scene's global features by self-attention mechanism. These visual global features then interact with the textual features through cross-attention layers, ensuring the visual features are enriched with multi-view semantics.

We conducted extensive experiments on the EmbodiedScan benchmark dataset [10]. Our method outperforms other teams and achieved **first place** in the CVPR 2024 Autonomous Grand Challenge Track on Multi-View 3D Visual Grounding, with an AP@IoU$_{0.25}$ of 59.59% and an AP@IoU$_{0.5}$ of 34.72%, significantly surpassing the baseline.

2. Method

This section is structured into three parts. Sec 2.1 describes the overview of network architecture. Sec 2.2 explains the LLMs-assisted text augmentation to alleviate the anchor sparsity in the language feature space. Sec 2.3 introduces the Bidirectional Text-View Images Interaction Module (Bi-TVI), which alleviates sparsity in the visual feature space.

2.1. Overview

Given the excellent performance of EmbodiedScan in multi-view 3D visual grounding, we adopt their framework and incorporate our proposed modules. We show the overview of our framework in Fig. 2.

2.2. LLM Assisted Augmentation Pipeline

Due to the lack of anchors and limited viewpoints in input language, which result in sparse textual feature space that causes description ambiguities, we propose the LLM Assisted Augmentation Pipeline. This pipeline enriches input descriptions by leveraging LLMs grounded with a scene information database, providing more contextual details of the corresponding scene. The pipeline involves three key steps, as illustrated in Fig. 3.

Step 1: Construct Scene Information Database. We first collect the text descriptions in the EmbodiedScan[10] dataset by their scene. For each text to be rephrased, we select the K language descriptions that are most relevant to the target object from the same scene. These descriptions, which encapsulate rich spatial and semantic context, constitute the scene information database. This database is leveraged to provide external knowledge to the LLM, thereby enriching the input text with reliable context information.

Step 2: Enrich Text with LLM and Database. The text to be rephrased, along with its corresponding scene information database constructed in Step 1, is fed into the LLM. The LLM leverages the database to draw on spatial and semantic details, enriching the text with additional contextual anchors to prevent confusion with similar-looking objects. Moreover, the LLM addresses the single-view limitation by generating from opposite viewpoints, similar to [5].

To ensure reliability and reduce hallucinations, our prompt is designed with several key principles in mind:

1. Sufficient and Reliable Anchors: Utilize multiple positional relationships of objects in the scene to describe the target object, ensuring enough reliable anchor points to minimize distractions.

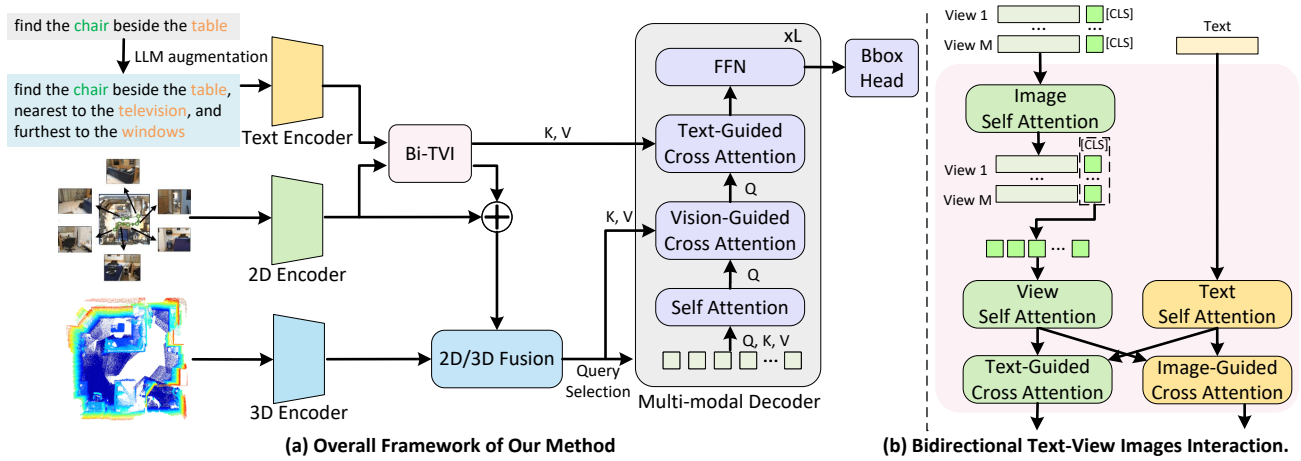

Figure 2. The overall framework is shown on the left and the detailed Bidirectional Text-View Images Interaction module (Bi-TVI) is shown on the right.

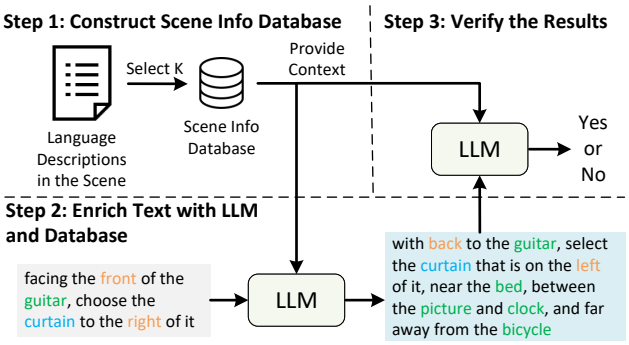

Figure 3. LLM Augmentation Pipeline.

2. Opposite View only for View-Dependent Text: For view-dependent text, rephrase it from the opposite perspective and logically adjust the spatial relationships, e.g., *"facing the front of xx, select xx to the right of"* becomes *"with back to xx, choose xx on the left side of"*. For view-independent text, simply add more anchor points to enhance its context.

3. High Confidence in Modifications: Ensure that all modifications, including the addition of new positional relationships and perspectives, are made with a high degree of confidence in their accuracy. The target object and its positional relationships must remain consistent.

By adhering to these principles, the rephrased text becomes more detailed and reliable, enriching the textual feature space. For example:

> Original: *"facing the front of the guitar, choose the curtain to the right of it"*
>
> Rephrased: *"with back to the guitar, select the curtain that is on the left of it, near the bed, between the picture and clock, and far away from the bicycle"*

Step 3: Verify the Generated Results. The final step is to verify the accuracy of the descriptions generated in Step 2. The LLM compares the rephrased sentences with the original sentences and scene descriptions from the database to identify possible inaccuracies or hallucinations, focusing on spatial relationships and perspective adjustments. This step ensures the enriched descriptions are accurate and reliable, further reducing hallucinations.

2.3. Bidirectional Text-View Images Interaction

To preserve more information in the visual feature representation, we propose a Bidirectional Text-View Images Interaction Module (Bi-TVI) to perform text and view-images feature interaction before the point cloud and image fusion. Specifically, we first obtain the global features of each view by appending a learnable token g_n to the flattened feature maps of each view from the output of the last image backbone layer. We followed this by performing a self-attention mechanism on the features of each of the scenes and using the output corresponding to g'_n to represent the global feature of view n of the scene.

$$g'_n = \phi(\text{SelfAttn}(g_n, F_{\text{view\_}n})) \tag{1}$$

where $1 \leq n \leq N$ views, $g \in \mathbb{R}^{1\times C'}$ is the learnable embedding, $F_{\text{view\_}n} \in \mathbb{R}^{\frac{H}{32}\frac{W}{32}\times C'}$, and ϕ denotes a function that outputs the first token in the token list.

To allow feature interaction between image views, we perform another self-attention among the g'_n denoted by:

$$G' = \text{SelfAttn}(g'_1, ..., g'_n) \tag{2}$$

Lastly, we use a bidirectional cross-attention mechanism that performs interaction between text features and view

Method	AP@IoU$_{0.25}$	AP@IoU$_{0.5}$
EmbodiedScan*	41.87	15.61
DenseG-8e	47.23	21.12
Mi-Robot (Top3)	46.91	20.38
chanc101 (Top2)	58.58	34.63
DenseG-E (Top1)	**59.59**	**34.72**

* Our implementation.

-E denotes Ensembled Predictions

Table 1. Test Results: Leaderboard Results. Note that DenseG-E is the ensembled prediction of several full 12-epoch models, that are not tested on the test set individually due to submission limits.

Method	AP@IoU$_{0.25}$	AP@IoU$_{0.5}$
EmbodiedScan	33.59	14.40
EmbodiedScan*	34.66	14.39
+ CLIP encoder	35.71	14.77
+ CBGS	37.35	16.06
+ LLMAug(10K)	38.93	17.43
+ Bi-TVI	39.70	18.31

* Our implementation.

Table 2. Mini Val Result: Performance of the models on Official Mini Validation Set. The + denotes adding to the model in the previous row.

features.

$$\mathcal{T}, \mathcal{G}' = \text{BiCrossAttn}(T, G') \tag{3}$$

We then add the feature vectors in $\mathcal{G}'$ to the corresponding image views and replace the text features T with $\mathcal{T}$ for succeeding network parts.

3. Experiments

3.1. Implementation Details

Architecture details. We followed EmbodiedScan [10] by using feature encoders, sparse fusion modules, and a DETR-based [2] decoder. Specifically, we use ResNet50 [6], MinkNet34 [4], and CLIP [9] text encoder as our image, point cloud, and language feature encoders respectively. We keep the vision-language sparse fusion module and decoder layer unchanged as in EmbodiedScan while incorporating our Bi-TVI before the vision-language fusion module.

Training details. We first pre-train our image and point cloud encoders on the 3D object detection task. We follow the training settings in EmbodiedScan while incorporating CBGS[14] during model training to alleviate the long-tail distribution in dataset classes.

Our multi-view visual grounding model, DenseG, is trained using AdamW optimizer with a learning rate of 5×10^{-4}, weight decay of 5×10^{-4}, and batch size of 24 with gradient accumulation for 2 iterations in training on mini dataset and batch size of 48 for full dataset training. We train the model with 12 epochs and decay the learning rate by 0.1 at epochs 8 and 11. We also incorporate exponential moving average weights updating with a momentum of 0.9998 and a gamma of 2000. Other training specifications are kept the same as those of EmbodiedScan. We note that we only use the LLM augmented text during training.

Training data. We train multiple variants of DenseGon different sets of data. Specifically, for the full setting, we train various variants of models with different datasets. Aside from the official training and validation set, LLM-augmented dataset of 308k samples, we also adopted the publicly available dataset such as Nr3D [1], Sr3D [1] and ScanRefer [3] for model training.

Ensembling details. We ensemble 5 variants of our models, including our implementation of EmbodiedScan baseline, and different variants of our method trained on different datasets. The models are ensembled with an IoU threshold of 0.4 using NMS.

3.2. Main Results

Table 1 presents the top-performing teams on the leaderboard. Our ensembled model predictions excel in AP@IoU$_{0.25}$ and AP@IoU$_{0.5}$, showing the superior grounding performance of our method. For comparison, we also provide the performance of our model, at epoch 8 (the full model is trained on 12 epochs), trained using official data and our augmented data. A simple ablation study on the proposed components is displayed in Table 2. This illustrates the effectiveness of our proposed language augmentation and modules.

4. Conclusion

In conclusion, in this technical report, we show our innovative approach, DenseG, effectively addresses feature sparsity in multi-view 3D visual grounding by enhancing both visual and textual modalities and finally achieving the best performance among the submitted entries. By combining a scene information database with large language models, we enrich textual descriptions, reduce confusion, and minimize hallucinations. Additionally, our Bidirectional Text-View Images Interaction Module significantly improves the retention of multi-view semantics. Our leading results on the EmbodiedScan benchmark highlight the effectiveness of our method, setting a new record in the leaderboard and paving the way for future advancements in embodied multi-view grounding.

References

[1] 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 *ECCV*, pages 422–440. Springer, 2020. 4

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 4

[3] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, pages 202–221. Springer, 2020. 4

[4] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, pages 3075–3084, 2019. 4

[5] Ziyu Guo, Yiwen Tang, Renrui Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding with gpt and prototype guidance. *arXiv preprint arXiv:2303.16894*, 2023. 1, 2

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[7] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *CVPR*, pages 15524–15533, 2022. 1

[8] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *ECCV*, pages 417–433. Springer, 2022. 1

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 4

[10] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024. 1, 2, 4

[11] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *CVPR*, pages 19231–19242, 2023. 1

[12] Li Yang, Ziqi Zhang, Zhongang Qi, Yan Xu, Wei Liu, Ying Shan, Bing Li, Weiping Yang, Peng Li, Yan Wang, et al. Exploiting contextual objects and relations for 3d visual grounding. *NeurIPS*, 36, 2024.

[13] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, pages 2928–2937, 2021. 1

[14] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 4