

3DGraphLLM: Combining Semantic Graphs and Large Language Models for 3D Scene Understanding

Tatiana Zemskova^{1,2} Dmitry Yudin^{1,2}
¹AIRI, ²MIPT

Abstract

A 3D scene graph represents a compact scene model by capturing both the objects present and the semantic relationships between them, making it a promising structure for robotic applications. To effectively interact with users, an embodied intelligent agent should be able to answer a wide range of natural language queries about the surrounding 3D environment. Large Language Models (LLMs) are beneficial solutions for user-robot interaction due to their natural language understanding and reasoning abilities. Recent methods for learning scene representations have shown that adapting these representations to the 3D world can significantly improve the quality of LLM responses. However, existing methods typically rely only on geometric information, such as object coordinates, and overlook the rich semantic relationships between objects. In this work, we propose 3DGraphLLM, a method for constructing a learnable representation of a 3D scene graph that explicitly incorporates semantic relationships. This representation is used as input to LLMs for performing 3D vision-language tasks. In our experiments on popular ScanRefer, Multi3DRefer, ScanQA, Sqa3D, and Scan2cap datasets, we demonstrate that our approach outperforms baselines that do not leverage semantic relationships between objects. The code is publicly available at <https://github.com/CognitiveAISystems/3DGraphLLM>.

1. Introduction

In this paper, we consider scene understanding in the context of 3D vision-language tasks: 3D referred object grounding task, 3D dense scene captioning and 3D visual question answering. The 3D referred object grounding task involves identifying a region within a 3D scene that corresponds to a natural language query. These queries often describe object properties (e.g., color, size) as well as spatial relationships (e.g., a mug on a table). A common setup of this problem assumes access to a 3D reconstruction of the scene, such as a point cloud, mesh, or NeRF. The objective is to predict the bounding box of the object or region referenced in the query.

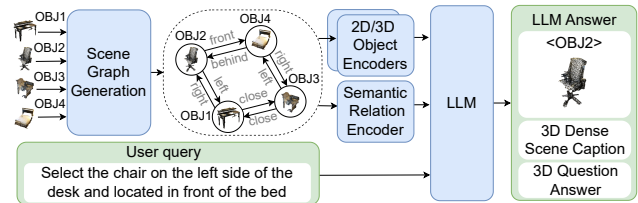


Figure 1. The proposed 3DGraphLLM approach leverages 3D semantic scene graph learnable representation supplied as input to an LLM to perform various 3D vision-language tasks.

The goal of 3D dense scene captioning is to generate a textual description of a selected object in the 3D scene, including its attributes or relationships. Finally, the goal of the 3D visual question answering task is to generate text answers to various questions about the properties of the scene. It seems promising to explicitly use a three-dimensional scene graph to solve these tasks.

A 3D scene graph provides a unified representation of a scene by storing multimodal information about individual objects, along with their semantic relationships [32, 52] and hierarchical organization [20, 53]. It also supports real-time updates in dynamic environments, making it suitable for interactive scenes [38, 45]. Furthermore, representing the scene as a graph enables the use of graph algorithms for tasks such as navigation [19, 20, 64] and object search based on textual queries [4, 14, 17, 53].

Solving 3D vision-language tasks is essential for embodied intelligent agents [3, 5, 9]. To interact effectively with users, such agents must be able to describe their environment and answer questions about its properties using natural language. Large Language Models (LLMs) are particularly well-suited for this, thanks to their strong capabilities in language understanding and commonsense reasoning. They can interpret user queries and match them to objects in a scene, even when the queries are vague or indirect [17, 22, 51]. By leveraging LLMs, it becomes easier to adapt the method to new object categories and relationships mentioned in referring expressions. LLMs can also handle complex queries that describe an object by its function rather than its name

(e.g., "somewhere to sit").

A 3D scene can be represented for input to an LLM either as text [17, 20, 34, 53, 55, 58] or as an implicit learnable representation [7, 8, 10, 22, 24]. Learnable representations encode objects and their relationships into embeddings, using significantly fewer tokens than textual descriptions. This compact form not only increases the speed of LLM inference but also enhances response quality by enabling better adaptation to 3D scenes. However, existing methods [7, 8, 22, 24] that use learnable 3D scene representations for vision-language tasks typically rely only on spatial coordinates and fail to incorporate semantic relationships between objects - limiting the expressiveness and reasoning capabilities of the model.

In this paper, we introduce 3DGraphLLM, a novel learnable representation of a 3D scene graph designed for use as input to an LLM (see Fig. 1). The representation consists of a list of learnable embeddings for scene objects, where each object is modeled as a local subgraph that includes the object itself and its nearest neighbors. These subgraphs are provided to the LLM as a sequence of triplets (*object1*, *relation*, *object2*). Semantic relations are encoded using features derived from the semantic edges of the scene graph, generated by state-of-the-art methods such as VL-SAT [52]. Our experiments show that incorporating semantic relationships between objects significantly improves the accuracy of LLM responses in 3D vision-language tasks, outperforming baseline methods that use learnable scene representations without semantic context.

To summarize, our contributions are as follows:

- We introduce 3DGraphLLM, the first method for creating a learnable 3D scene graph representation specifically designed for LLMs. It enables semantic relationships between objects in a scene to be mapped directly into the LLM’s token embedding space.
- We propose an algorithm that generates a flat sequence of graph embedding tokens by selecting object subgraphs using k-nearest neighbors with Non-Maximum Suppression (NMS) and a minimum distance filters between objects. This approach reduces the number of tokens needed to describe the scene, thereby improving inference speed.
- 3DGraphLLM outperforms the baseline method which does not use semantic relationships on the 3D referred object grounding task, achieving improvements of +7.5% F1@0.5 on the Multi3DRefer[60] and +6.4% Acc@0.5 on ScanRefer [5] benchmarks. It also improves performance on 3D scene captioning, with a +3.9% CIDEr@0.5 score on the Scan2Cap [9] dataset. 3DGraphLLM achieves state-of-the-art results in 3D referred object grounding while requiring up to five times less inference time compared to LVLM-based methods.

2. Related works

3D Language Scene Understanding. 3D scene understanding is a complex computer vision task that involves identifying the semantic, physical, and functional properties of objects, as well as their mutual relations. One of the goals of 3D scene understanding is to develop methods capable of responding to natural language queries about the scene. The queries may correspond to different visual-language tasks such as 3D referred object grounding [5, 36, 60], question answering [3], and dense scene captioning [9]. Recent approaches address these queries by reconstructing the scene as a 3D mesh [41] or point cloud [6, 61, 65], often enhanced with instance segmentation [65].

The emergence of transformer models [48] has enabled the development of neural network models that create a learnable representation of a scene for answering various language queries. MultiCLIP [12] proposes to align 3D scene representation with text queries and multi-view 2D CLIP [44] embeddings to improve the quality of question answering. 3DVG-Transformer [61] and Vi3DRef [6] methods introduce modules for modeling spatial relationships between objects to improve the quality of object grounding. 3D-VisTA [65] presents a transformer model for aligning 3D object and text representations, coupled with an unsupervised pre-training scheme to solve various 3D vision-text problems using specialized task-specific heads. However, these approaches face challenges in generalizing to new tasks and domains. In contrast, leveraging large language models (LLMs) for scene understanding enhances generalization capabilities and taps into the extensive knowledge LLMs contain about the physical world [22].

Scene Graphs. The concept of a scene graph was initially developed for 2D images, providing a structured representation of a scene’s semantics by incorporating relationships between the semantic elements [29]. In the context of images, scene graphs have proven effective for tasks such as content-based image retrieval [29, 40], 2D referring expression comprehension [18, 47, 56], image caption [42, 57], image generation [13, 30].

In 3D scenes, a scene graph is commonly used to address robotics challenges such as planning [20, 53], object grounding for navigation [17, 20, 34, 53] and manipulation [20], as well as scene generation [16, 59]. Our approach is part of a class of methods that utilize an implicit representation of the scene graph, such as OVSG [4], which frames the problem of 3D object grounding as subgraph retrieval. 3DGraphQA [54] proposes to use the bilinear graph neural network for feature fusion between scene and question graphs for question answering task. FFL-3DOG [14] builds a graph based on a text query, which is used to refine the visual graph to select from its vertices the one that best fits the description. However, the application scope of this method is limited to specific tasks such as 3D referred object grounding or question answering.

In contrast, we propose a more versatile method capable of solving various 3D vision-language tasks.

Large Language Models for Scene Understanding.

Large language models (LLMs) offer several advantages for scene understanding, notably enhancing the ability to address complex queries that require common knowledge. LLMs can serve as agents that decompose user queries into elementary tasks, which can then be addressed by other methods [55, 58]. Additionally, LLMs can act as an interface for reasoning by processing textual descriptions of the scene as input [17, 34]. BBQ [34] and ConceptGraphs [17] demonstrate that using a text-based graph representation with an LLM interface significantly improves the quality of object retrieval compared to using CLIP features of objects. HOV-SG [53] constructs a hierarchical graph consisting of objects, rooms, and floors, and demonstrates the effectiveness of such a representation for the task of object grounding given a query containing object location hints. The authors of the MOMA [20] method propose using a hierarchical scene graph together with a navigational Voronoi graph as input to LLM to predict a high-level policy for object search for navigation and manipulation. However, using text to describe an object in a scene graph inevitably leads to the loss of some of the information contained in its RGB point cloud. Additionally, in the case of using a text graph, several hundred tokens may be required to describe one object (its semantic class, pose), which will significantly slow down LLM inference in the case of a large number of objects in the scene.

Recent advancements have successfully integrated point cloud data into LLMs by employing pre-trained point cloud encoders and training adapters to align the resulting representations with the LLM embedding space. 3D-LLM [21] aggregates 3D point cloud features from a sequence of 2D images and then solves the grounding problem as a prediction of a sequence of location tokens added to the LLM dictionary. Chat-Scene [25] generates 2D and 3D features for each object in the scene and introduces learnable object identifier tokens to solve object grounding, dense scene captioning, and question answering problems. LLA3D [7] proposes to use a set of trainable fixed-length query tokens obtained by interacting potential visual cues, text cues, and object point cloud features in a transformer model. Grounded 3D-LLM [8] uses referent tokens to decode object masks in point clouds. Additionally, research has demonstrated that incorporating spatial information, such as object coordinates [24] or depth maps [10], enhances the accuracy of responses to user queries.

Despite recent advances, existing methods do not fully leverage the rich semantic information in object relationships. In this paper, we introduce 3DGraphLLM, a method that demonstrates the effectiveness of utilizing semantic relationships between objects to enhance performance across

various scene understanding tasks.

3. Method

Our approach uses a set of point clouds of scene objects as input. The objects’ point clouds can be obtained either from ground-truth annotations or through state-of-the-art point cloud instance segmentation methods. These point clouds are used to extract scene graph features (see Sec. 3.1). A scene graph consists of nodes representing the objects and edges corresponding to semantic relationships between them. To convert the scene graph into a token sequence, we represent each object by an identifier, its 2D object feature, and a subgraph comprising the object’s k nearest neighbors. The relationships between an object and its neighbors are encoded as triplets $(object_i, relation_{ij}, object_j)$. The scheme of the 3DGraphLLM approach is shown in Fig. 2. For more details on the scene graph representation, refer to Sec. 3.2. Our training process is two-stage. First, we pre-train the model on a dataset for various 3D scene understanding tasks using ground-truth instance segmentation. Next, we fine-tune 3DGraphLLM with predicted instance segmentation of scene point clouds, considering a scenario where ground-truth segmentation is unavailable (see Sec. 3.3).

3.1. Model Architecture

The model architecture includes pre-trained encoders for 2D images, 3D point clouds, and point clouds semantic relationships, alongside a pre-trained LLM. We train projection layers to map the extracted object features and their relationships into the LLM’s token embedding space. Following the approach of Chat-Scene [25], we introduce additional object identifier tokens $\{< OBJ_i >\}_{i=1}^n$ into the LLM’s vocabulary. Here and throughout, we use n to denote the number of objects in the scene. These learned identifiers, with the features from object subgraphs composed of nearest neighbors for each object, are used to create a flat representation of the scene graph, which is then fed into the LLM.

Object Proposals. We use point clouds of objects in the scene as vertices in the scene graph G . In our experiments, we evaluate 3DGraphLLM in various modes, including ground-truth scene segmentation and instance segmentation using state-of-the-art neural network methods like Mask3D [46] and OneFormer3D [33]. Thus, the set V of vertices of the graph consists of n point clouds $\{P_i\}_{i=1}^n$, where $P_i \in \mathbb{R}^{m_i \times 6}$. Here, m_i is the number of points in the i -th object proposal of instance segmentation of scene point cloud, and 6 dimensions of each point correspond to its 3D coordinates and RGB color.

Object Identifiers. Following the approach in Chat-Scene, we add a set of learnable identifier tokens $\{< OBJ_i >\}_{i=1}^n$ to the LLM’s vocabulary for object identification. These tokens allow the model to identify objects in the scene by simply predicting the corresponding object identifier to-

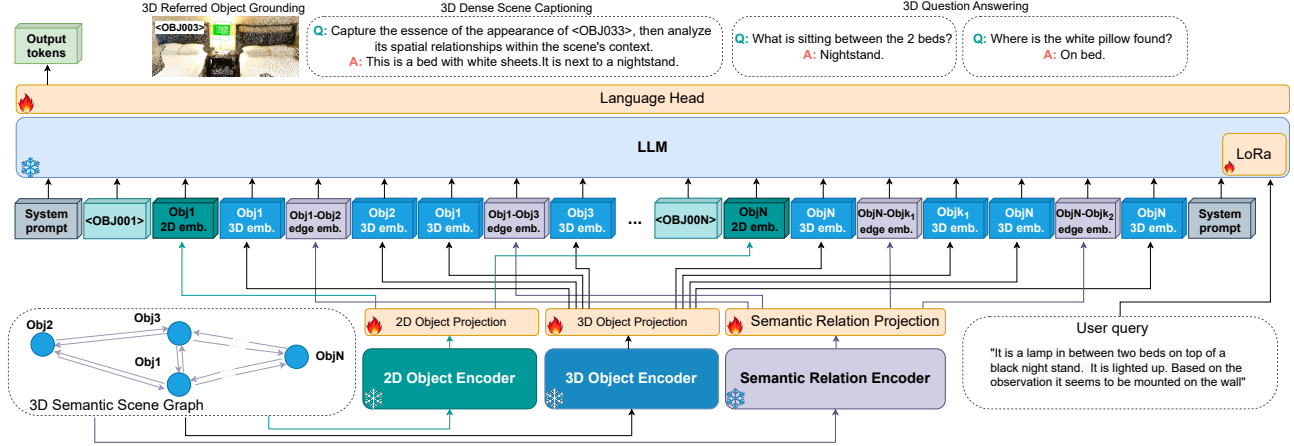


Figure 2. The overall architecture of our approach. We introduce trainable layers to map the extracted graph node and edge features into the token embedding space of a pre-trained LLM. The scene graph is flattened for input into the LLM, with each object represented by a subgraph of its k nearest neighbors. To further adapt the LLM to 3D vision-language tasks, we add new object tokens to the LLM’s vocabulary alongside with objects’ 2D features and fine-tune the LLM using LoRa.

ken. In our experiments, we assume a maximum of 200 objects per scene.

2D Object Encoder. The results of Chat-Scene demonstrate that adding aggregated 2D DINOv2[37] features increases the LLM performance on 3D vision-language tasks. Therefore, we add DINOv2 $Z_i^{2d} \in \mathbb{R}^{1 \times 1024}$ features as an additional token describing the object subgraph. DINOv2 object features are obtained by aggregating features from the masked multi-view images where masks come from the projection of the object’s 3D point cloud.

3D Object Encoder. We extract vertex features using a pre-trained Uni3D [63] encoder, which generates point cloud features aligned with their textual descriptions. Since this model is pre-trained on a large dataset, it enables us to produce high-quality graph vertex embeddings across various data domains. For each object point cloud P_i , we extract Uni3D feature $Z_i^{vp} \in \mathbb{R}^{1 \times 1024}$.

Edge Feature Encoder. One challenge in generating features for semantic relationships between objects is that most methods for 3D semantic scene graph generation are trained on 3RScan scenes [50], while visual grounding tasks are typically tested on ScanNet scenes [11]. Although both datasets belong to the indoor scene domain, existing methods struggle with performance in cross-domain testing, resulting in a drop in accuracy for the grounding task [36].

To extract semantic relationships between objects, we use VL-SAT [52], a method for generating 3D semantic scene graphs from point clouds. One of its key advantages is that it only requires 3D point cloud coordinates as input during prediction while leveraging knowledge transfer from the pre-trained CLIP model [44]. This allows the method to perform well when applied to new scene domains [52], as confirmed

by our experiments (see Sec. 4.2). For each pair of point clouds P_i and P_j , we generate a latent feature representing their relationship $Z_{ij}^e \in \mathbb{R}^{1 \times 512}$, which corresponds to VL-SAT graph neural network feature before the classification head assigning semantic categories to the graph edges. While VL-SAT predicts a fixed set of relationships between objects, these relationships are not mutually exclusive (e.g., ”larger” and ”close”). Therefore, we use latent features to capture possible combinations of these semantic relationships.

2D/3D object, and semantic relation projection. To adapt the extracted features for the language model, we use three trainable projection modules: the 2D Object Projection $f_{2d}(\cdot)$, which maps the 2D image features of objects, the 3D Object Projection $f_v(\cdot)$, which maps the point cloud features of objects, and the Semantic Relation Projection $f_e(\cdot)$, which maps the features of semantic relationships between objects. Therefore, for the i -th object, the 2D and 3D object features are projected to token embeddings F_i^v and F_i^{2d} , respectively. For the pair of i -th and j -th objects, the semantic relation feature is projected to token embedding F_{ij}^e :

$$F_i^{2d} = f_v(Z_i^{2d}), F_i^v = f_v(Z_i^v), F_{ij}^e = f_e(Z_{ij}^e). \quad (1)$$

3.2. Flat Graph Representation

The scene graph is a complete graph since we can generate connections between all pairs of objects. Such a graph contains $n \cdot (n - 1)$ edges between objects, and using the complete graph as a sequence for the LLM would significantly increase the sequence length. Intuitively, the most relevant relationships for answering user questions are those between an object and its nearest neighbors. Therefore, for each object, we consider a subgraph of its k nearest neigh-

System:	A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user’s questions. The conversation centers around an indoor scene: [<OBJ001> $F_1^{2d}, F_1^v, F_1^e, F_2^v F_1^v, F_{14}^e, F_4^v \dots$ <OBJN> $F_N^{2d}, F_N^v, F_N^e, F_{Nk_1}^v, F_{k_1}^v F_N^v, F_{Nk_2}^e, F_{k_2}^v$]
User:	According to the given description, <i>there are brown wooden cabinets, placed on the side of the kitchen</i> , please provide the ID of the object that closely matches this description.
Assistant:	<OBJ001>.

Table 1. Example of prompt for the language model containing scene graph.

bors. The relationships between objects are encoded using features extracted from point clouds $\{F_i^v\}_{i=1}^n$ and semantic relations features $\{F_{ij}^e, i \in \{1, \dots, n\}, j \in \{1, \dots, n\}\}$, represented as a triplet (F_i^v, F_{ij}^e, F_j^v) .

When using the complete scene graph the number of tokens required to describe the scene is $2 \cdot n + 3n \cdot (n - 1)$. For 100 objects, which matches the number of object proposals in the Mask3D [46] instance segmentation, this totals 29900 tokens. By using a k -nearest neighbor subgraph, we reduce the token count to $2 \cdot n + 3n \cdot k$. As shown in Sec. 4.2 (see Fig. 4) and Supplementary Materials, setting $k = 2$ improves accuracy in 3D visual-language tasks while reducing the number of tokens needed to describe a scene with 100 objects to 800. We analyze how the number of objects affects inference speed and GPU memory usage in Supplementary Materials.

Prompt template. We integrate the scene description as a sequence of object subgraphs into the prompt for LLM similar to the integration of the list of object embeddings in the Chat-Scene method [25]. An example of a prompt for LLM containing a system prompt, a scene description in the form of an object identifier, a 2D object feature and an object subgraph, a user request, and an LLM assistant response is given in Tab. 1. The sequence describing an object i starts with its identification token <OBJ*i*> and 2D object feature F_i^{2d} . Then there are k triplets $\{(F_i^v, F_{ijk}^e, F_{jk}^v)\}_{jk=1}^k$ describing the relationship between the object and its k nearest neighbors.

3.3. Training Strategy

Following the strategy used in Chat-Scene[25], we implement a training approach that involves simultaneously training the projection layers and the language model. We also conduct joint training for various tasks, including visual grounding (ScanRefer [5], Multi3DRefer [60], RioRefer [36]), 3D scene description (Scan2Cap [9], Nr3D [1], RioRefer [36]), and 3D visual question answering (ScanQA [3], SQA3D [35], 3RQA [26]). This adaptation of the tasks is designed for user-assistant interactions, as proposed by the authors of Chat-Scene. During training, we aim to optimize the trainable parameters θ of both the language model and the projection layers to minimize the negative log-likelihood

of the target response s^{res} compared to the response predicted by the model. We use the following loss function:

$$L(\theta) = - \sum_{i=1}^{\ell} \log P(s_i^{\text{res}} | s_{[1, \dots, i-1]}^{\text{res}}, s^{\text{prefix}}), \quad (2)$$

where ℓ is the length of the token sequence in the LLM response, $s_{[1, \dots, i-1]}^{\text{res}}$ is the sequence generated up to the i -th token, s^{prefix} is the input prefix sequence containing system and user prompts. The trainable parameters θ include the parameters of 2D/3D Object Projections and Semantic Relation Projection Layers, added object identifier token embeddings, and the language model.

We use the semantic relationships encoder [52] pre-trained using ground-truth (GT) point cloud scene segmentation data. Since the predicted point cloud segmentation typically contains more noise than the GT segmentation, we anticipate that the edge features derived from the GT segmentation will be of higher quality than those from the neural network instance segmentation. To address this problem, we employ a two-stage training strategy for 3DGraphLLM. First, we pre-train the projection layers and the language model on the GT instance segmentation data to achieve effective projections of the semantic embeddings of relations and objects into the language model’s embedding space. Then, we fine-tune 3DGraphLLM using the noisy data from the neural network segmentation. Sec. 4.2 presents the experimental results, demonstrating the effectiveness of two-stage training and comparing different pre-training datasets.

4. Experiments

Datasets. For pretraining 3DGraphLLM using GT instance segmentation, we employ a combined 3D Vision-Language dataset for ScanNet [11] and 3RScan [50] scenes. For ScanNet scenes, we utilize data from five 3D vision-language benchmarks: visual grounding tasks (ScanRefer [5], Multi3DRefer [60]), scene description (Scan2Cap [9]), and 3D visual question answering (ScanQA [3], SQA3D [35]). Each of these datasets follows a standard split into training and validation sets, corresponding to 1201 training scans and 312 validation scans from ScanNet. For 3RScan scenes, we use data from the RioRefer dataset [36] for object grounding, and the 3RQA dataset [26] for question answering. For 3RScan data, we follow the standard train/validation scan split and use the scans present in the RioRefer dataset for training, resulting in 1175 training scans and 157 validation scans. To augment the data for the scene description task, we use data from the RioRefer [36] and Nr3D [1] datasets, taking object grounding queries provided in these datasets as reference descriptions of objects in the scene. To assess 3DGraphLLM performance under realistic conditions, we perform fine-tuning on predicted instance segmentation using 3D vision-language benchmarks

Methods	2D features	3D features	LLM	ScanRefer		Multi3DRefer		Scan2Cap		ScanQA		Sqa3D	
				A@0.25↑	A@0.5↑	F1@0.25↑	F1@0.5↑	C@0.5↑	B-4@0.5↑	C↑	B-4↑	EM↑	
<i>Expert models</i>	ScanRefer [5]	✓	✓	✗	37.3	24.3	-	-	-	-	-	-	
	MVT [27]	✓	✓	✗	40.8	33.3	-	-	-	-	-	-	
	3DVG-Trans [61]	✓	✓	✗	45.9	34.5	-	-	-	-	-	-	
	ViL3DRel [6]	✗	✓	✗	47.9	37.7	-	-	-	-	-	-	
	M3DRef-CLIP [60]	✓	✓	✗	51.9	44.7	42.8	38.4	-	-	-	-	
	Scan2Cap [9]	✓	✓	✗	-	-	-	-	35.2	22.4	-	-	
	ScanQA [3]	✓	✓	✗	-	-	-	-	-	-	64.9	10.1	
	Sqa3D [35]	✗	✓	✗	-	-	-	-	-	-	-	47.2	
	3D-VisTA [65]	✗	✓	✗	50.6	45.8	-	-	66.9	34.0	72.9	13.1	48.5
	BUTD-DETR [28]	✗	✓	✗	52.2	39.8	-	-	-	-	-	-	-
PQ3D [66]	✓	✓	✗	-	51.2	-	50.1	80.3	36.0	87.8	-	47.1	
<i>LLM-based models</i>	ZSVG3D [58]	✓	✓	GPT4	36.4	32.7	-	-	-	-	-	-	
	3D-LLM [21]	✓	✓	Flamingo	21.2	-	-	-	-	-	59.2	7.2	
	3D-LLM [21]	✗	✓	BLIP2-flant5	30.3	-	-	-	-	-	69.4	12.0	
	Chat-3D v2 [24]	✗	✓	Vicuna-7B-v0	35.9	30.4	-	-	-	-	77.1	7.3	
	Scene-LLM [15]	✓	✓	Llama-2-7B	-	-	-	-	-	-	80.0	12.0	54.2
	LEO [26]	✗	✓	Vicuna-7B-v1.1	-	-	-	-	72.4	38.2	101.4	13.2	50.0
	LL3DA [7]	✗	✓	OPT-1.3B	-	-	-	-	65.2	36.8	76.8	13.5	-
	Grounded 3D-LLM [8]	✗	✓	Tiny-Vicuna-1B	47.9	44.1	45.2	40.6	70.6	35.5	72.7	13.4	-
	Robin3D [31]	✓	✓	Vicuna-7B-v1.5	60.8	55.1	64.9	59.7	87.2	38.4	-	-	56.0
	GPT4Scene-HD [43]	✓	✓	Qwen2-VL-7B	50.9	46.4	53.7	50.0	74.4	37.9	89.9	15.9	57.2
	GPT4Scene-HDM [43]	✓	✓	Qwen2-VL-7B	62.6	57.0	64.5	59.8	86.3	40.6	96.3	15.5	59.4
	Chat-Scene [25] (baseline)	✓	✓	Vicuna-7B-v1.5	55.5	50.2	57.1	52.4	77.1	36.3	87.7	14.3	54.6
	3DGraphLLM (ours)	✓	✓	Vicuna-7B-v1.5	58.6	53.0	61.9	57.3	79.2	34.7	91.2	13.7	55.1
3DGraphLLM (ours)	✓	✓	LLAMA3-8B-Instruct	62.4	56.6	64.7	59.9	81.0	36.5	88.8	15.9	55.9	

Table 2. Performance comparison of 3DGraphLLM with state-of-the-art approaches for 3D vision-language tasks. "Expert models" use specialized heads to deal with different 3D vision-language tasks. Our approach falls into the category of "LLM-based models" that consider different tasks as different user queries to a generative model. C denotes the CIDEr metric.

for ScanNet scenes: ScanRefer, Multi3DRefer, Scan2Cap, ScanQA, and SQA3D.

Implementation details. The projection layers for 2D/3D object features and their semantic relations are three-layer MLPs. In our experiments, we use LLAMA3-8B-Instruct [2], a state-of-the-art large language model, as well as Vicuna-1.5-7B [62] for ablation. For fine-tuning the language model, we apply LoRA [23] with a rank of 16. We use a batch size of 8 and train 3DGraphLLM for 3 epochs with an initial learning rate of $5 \cdot 10^{-6}$, following a cosine annealing schedule. Training is performed on a server equipped with 4 NVIDIA A100 GPUs, and the entire training process takes approximately 24 hours. In our experiments, we select $k = 2$ nearest neighbors to construct object subgraphs and, in the case of using Mask3D [46] instance scene point cloud segmentation, we use a NMS filter and a filter that ensures a minimum distance between nearest neighbors of 1 cm (see Sec. 4.2).

	Dataset	Method	
		3DGraphLLM	GPT4Scene
Input token number per scene		800	10400
	ScanRefer	0.4	1.9
Inference speed, sec	Multi3DRefer	0.5	2.0
	Scan2Cap	0.9	2.2
	ScanQA	0.4	1.9
	SQA3D	0.4	1.7

Table 3. Input tokens and inference speed comparison (Mask3D instance segmentation).

Evaluation metrics. For the visual grounding task on the ScanRefer [5] dataset, we use the standard metrics Acc@0.25 and Acc@0.5. A prediction is considered a true positive if the intersection-over-union (IoU) between the predicted object's 3D bounding box and the ground truth exceeds the thresholds of 0.25 and 0.5, respectively. The Multi3DRefer [60] dataset contains queries that may refer to multiple objects. Therefore, we use the benchmark-standard

F1 score at IoU thresholds of 0.25 and 0.5. We assess the quality of object descriptions using the Scan2Cap [9] benchmark metrics CIDEr@0.5 and BLEU-4@0.5. For the visual question answering task, we follow the validation strategy from Chat-Scene[25], applying CIDEr [49] and BLEU-4 [39] metrics for ScanQA [3], and exact match accuracy (EM) for SQA3D [35].

4.1. Experimental Results

Comparison with state-of-the-art approaches. As shown in Tab. 2, our method significantly outperforms the baseline approach Chat-Scene [25] on the two ScanNet 3D referred object grounding benchmarks, ScanRefer [5] and Multi3DRefer [60], as well as on the scene captioning benchmark Scan2Cap [9] and the question answering benchmarks ScanQA [3] and SQA3D [35]. These results highlight the effectiveness of a learnable graph-based scene representation 3D vision-language tasks. It's worth noting that the performance of our method surpasses state-of-the-art specialized models with separate heads for different language tasks, such as 3D-VisTA [65], PQ3D [66], and M3DRef-CLIP [60].

Notably, 3DGraphLLM demonstrates state-of-the-art quality for the 3D referred object grounding task for LLM-based methods. In particular, our 3DGraphLLM with LLAMA3-8B as the base LLM outperforms Robin3D [31] on ScanRefer benchmark showing comparable quality on Multi3DRefer and SQA3D benchmarks. Robin3D is trained on 1M instruction-following data that are not publicly available, while our approach uses only 370K instruction-following data. Our experiments in Tab. 4 highlight the importance of training data for 3DGraphLLM, suggesting that incorporating more data for fine-tuning could further improve its performance. 3DGraphLLM achieves results comparable to the state-of-the-art method GPT4Scene-

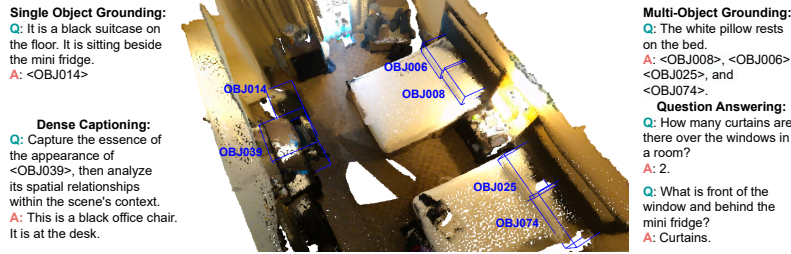


Figure 3. Qualitative examples of 3DGraphLLM performance on object grounding, dense captioning, and question answering tasks. We provide a visualization of the RGB point cloud along with blue objects bounding boxes.

Methods	LLM	Pre-train	Number of edges	Training scenes	ScanRefer	Multi3DRefer	Scan2Cap		ScanQA		Sqa3D
					Acc@0.5↑	F1@0.5↑	C@0.5↑	B-4@0.5↑	C↑	B-4↑	EM↑
3DGraphLLM-0	Vicuna1.5-7B	✗	0	ScanNet	50.2	52.4	77.1	36.3	87.7	14.3	54.6
3DGraphLLM-2	Vicuna1.5-7B	✗	2	ScanNet	50.1	52.7	80.4	36.9	92.2	15.5	54.7
3DGraphLLM-2	Vicuna1.5-7B	✓	2	ScanNet+3RScan	53.1	57.3	79.2	34.7	91.2	13.7	55.1
3DGraphLLM-0	LLAMA3-8B-Instruct	✗	0	ScanNet	52.0	55.1	80.0	37.5	84.0	15.8	53.8
3DGraphLLM-2	LLAMA3-8B-Instruct	✗	2	ScanNet	54.3	57.3	85.6	39.6	87.4	14.9	54.5
3DGraphLLM-2	LLAMA3-8B-Instruct	✓	2	ScanNet	56.2	58.7	82.9	37.3	85.4	15.1	55.6
3DGraphLLM-2	LLAMA3-8B-Instruct	✓	2	ScanNet+3RScan	56.6	59.9	81.0	36.5	88.8	15.9	55.9

Table 4. Ablation study on semantic edges role and training pipeline. C denotes the CIDEr metric.

HDM [43], showing the importance of semantic relations for this task. At the same time, 3DGraphLLM uses fewer tokens to describe the scene (see Tab. 3), allowing up to five times faster inference for object-grounding tasks.

Qualitative results. Fig. 3 shows the qualitative results of 3DGraphLLM using Mask3D [46] instance scene segmentation. 3DGraphLLM efficiently uses spatial cues for solving 3D Vision-Language tasks. For example, 3DGraphLLM distinguishes the black suitcase next to the refrigerator, despite there being another suitcase farther away from the refrigerator in the scene. In Supplementary Materials we provide more examples of 3DGraphLLM performance.

4.2. Ablation Studies

Role of Semantic Relations. To isolate the impact of using a scene graph representation, we conduct an experiment with different LLMs and training pipelines using Mask3D [46] instance segmentation. We train a version of 3DGraphLLM (3DGraphLLM-0) where the scene is represented as a sequence of object identifiers and features extracted by the 2D Object Encoder and the 3D Object Encoder, following the same training pipeline as 3DGraphLLM (3DGraphLLM-2) with two nearest neighbors. The 3DGraphLLM version with zero nearest neighbors serves as a baseline, equivalent to the Chat-Scene approach, which uses the same LLM as 3DGraphLLM-2. As shown in Tab. 4, incorporating a scene graph representation significantly improves the performance of the LLMs across all three 3D Vision-Language tasks: visual grounding, scene description, and question answering. However, the effect is more noticeable for the more recent LLAMA3-8B-Instruct.

Training pipeline. The pre-training on GT instance segmentation data improves the quality of the 3D referred ob-

ject grounding for LLAMA3-8B-Instruct and Vicuna-1.5-7B. For LLM Vicuna-1.5-7B, pre-training increases the scene captioning quality. For LLAMA3-8B-Instruct, pre-training improves the question answering on the SQA3D dataset. We compare two pre-training datasets for 3DGraphLLM using LLAMA3-8B-Instruct. The first contains only 3D Vision-Language data from ScanNet, while the second includes data from both ScanNet and 3RScan. Tab. 4 shows that incorporating 3RScan data further enhances object grounding and question answering performance. The most interpretable metrics for the role of semantic edges are the accuracy metrics in the 3D referred object grounding task, so we keep this pre-training as part of the 3DGraphLLM training pipeline.

It is worth noting that the n-gram-based evaluation metrics used in scene captioning and question answering benchmarks are not adequate for assessing the quality of LLM-generated responses because they fail to capture the flexibility and richness of LLM outputs. This effect is particularly noticeable in the scene captioning task, where CIDEr@0.5 and BLEU-4@0.5 penalize 3DGraphLLM if the model incorporates visual and spatial cues that are missing from the reference descriptions. For example, in the scene shown in Fig. 3, 3DGraphLLM describes a toilet as: "This is a white toilet. It is to the right of the shower curtain." This is a correct description of the object, yet the reference captions use different wording and spatial cues, causing CIDEr@0.5 to assign a score of 0.0 to this description. See Supplementary Materials for a more detailed illustration of this effect.

Quality of instance segmentation. We evaluate how the quality of scene segmentation into objects impacts the performance of 3DGraphLLM. For these experiments, we use the full training pipeline with a pre-training phase on GT instance segmentation on ScanNet data. As shown in

Methods	Instance segmentation	Number of edges	Minimal distance, cm	ScanRefer Acc@0.5↑	Multi3DRef F1@0.5↑
3DGraphLLM-0	GT	0	-	61.5	64.4
3DGraphLLM-2	GT	2	0	66.9	69.9
3DGraphLLM-0	Mask3D	0	-	52.0	55.1
3DGraphLLM-2	Mask3D	2	0	55.6	58.2
3DGraphLLM-2	Mask3D (+ NMS)	2	0	55.7	58.6
3DGraphLLM-2	Mask3D (+ NMS)	2	1	56.2	58.7
3DGraphLLM-0	OneFormer3D	0	-	50.0	52.8
3DGraphLLM-2	OneFormer3D	2	0	52.8	55.8
3DGraphLLM-2	OneFormer3D (+NMS)	2	1	54.6	57.2

Table 5. Ablation study on semantic edges role depending on quality of instance segmentation.

Methods	Instance segmentation	Relations as triplets	Number of edges	ScanRefer Acc@0.5↑	Multi3DRef F1@0.5↑
3DGraphLLM-0	Mask3D	✗	0	52.0	55.1
3DGraphLLM-2	Mask3D	✗	2	54.2	56.3
3DGraphLLM-2	Mask3D	✓	2	54.3	57.3

Table 6. Ablation study on subgraph representation.

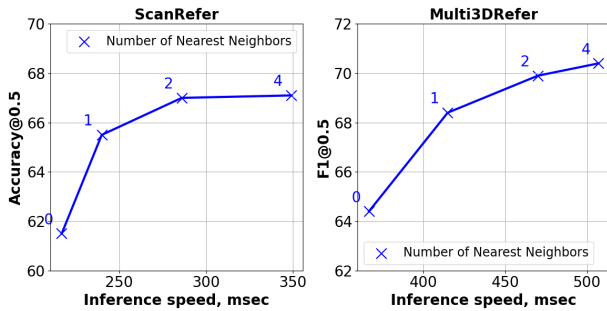


Figure 4. Dependence of inference speed and visual grounding quality on the number of nearest neighbors in the object subgraph. This experiment utilizes the GT instance segmentation.

Tab. 5, even with noisy neural network segmentation, representing the scene as a graph with semantic relationships is still more effective than using a simple list of objects. We conduct experiments with different object proposal methods, including OneFormer3D [33] and Mask3D [46], but we found that Mask3D segmentation shows better results for our tasks. Therefore, in subsequent experiments, we use the Mask3D method to maintain consistency with the baseline Chat-Scene approach.

The analysis of objects selected as nearest neighbors reveals a high number of duplicate objects among the chosen neighbors. To address this issue, we propose two filters. First, we add an NMS filter to remove duplicates between the potential neighbors for an object, using a threshold of $IoU = 0.99$. Second, we introduce a minimum distance filter of 1 cm to the nearest neighbor to prevent selecting duplicates of the original object as its neighbors.

Adding the NMS filter improves the performance of the visual grounding task when using Mask3D instance segmentation (see Tab. 5). The additional minimum distance filter further enhances visual grounding quality. The combination of filters is also effective for OneFormer3D [33] scene

instance segmentation, as shown in Tab. 5.

Number of nearest neighbors. We examine how the number of nearest neighbors affects the quality of visual grounding and the speed of model inference, as adding more connections increases the number of tokens used to describe each object. This experiment was performed using ground-truth scene segmentation, as this setup provides the highest quality embeddings for semantic relations between objects. We vary the number of nearest neighbors in powers of two, capping it at 4 due to GPU memory constraints during training. As shown in Fig. 4, increasing the number of nearest neighbors enhances visual grounding quality with a slight increase in inference time.

Subgraph representation. In our work, we use an object-centric graph representation, where relationships between objects are represented as triplets $\{F_N^v, F_{Nk_1}^e, F_{k_1}^v\}$. We conduct an experiment in which we remove duplicate vertex tokens from the subgraph-based object description. As a result, object N is described by the following sequence: $\{< OBJN > F_N^{2d}, F_N^v, F_{Nk_1}^e, F_{Nk_2}^e\}$. We do not perform the pretraining phase on GT instance segmentation in this experiment. Tab. 6 shows that the object-centric graph representation using triplets improves the performance of the visual grounding task.

We include additional experimental results from ablation studies on scene captioning and visual question answering tasks in the Supplementary Materials.

5. Conclusion

In this paper, we propose a new learnable approach to using a 3D semantic scene graph for a large language model to solve 3D vision-language tasks. Detailed experiments demonstrate the effectiveness of this approach, which explicitly takes into account semantic relations between objects represented as 3D point clouds. Our method, called 3DGraphLLM, surpasses the baseline approach without semantic relationships on popular ScanRefer, Multi3DRefer, Scan2Cap, ScanQA, and SQA3D datasets. Moreover, 3DGraphLLM achieves state-of-the-art performance in the object grounding task, matching the quality of methods that require five times more inference time.

A limitation of the method is a significant increase in resource consumption with an increase in the edge number for each graph node. At the same time, we showed that taking into account only two edges for each object demonstrates an acceptable trade-off between performance and model quality.

For further development of the work, it seems appropriate to search for methods to reduce token usage for encoding object relationships in our graph representation. Another important aspect for further work is the creation of methods for generating semantic relations between objects that are robust to imperfections in the instance segmentation of the scene point cloud.

Acknowledgments

The study was supported by the Ministry of Economic Development of the Russian Federation (agreement with MIPT No. 139-15-2025-013, dated June 20, 2025, IKG 000000C313925P4B0002).

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 *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020. 5
- [2] AI@Meta. Llama 3 model card. 2024. 6
- [3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022. 1, 2, 5, 6
- [4] Haonan Chang, Kowndinya Boyalakuntla, Shiyang Lu, Siwei Cai, Eric Jing, Shreesh Keskar, Shijie Geng, Adeeb Abbas, Lifeng Zhou, Kostas Bekris, et al. Context-aware entity grounding with open-vocabulary 3d scene graphs. *arXiv preprint arXiv:2309.15940*, 2023. 1, 2
- [5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020. 1, 2, 5, 6
- [6] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. *Advances in neural information processing systems*, 35:20522–20535, 2022. 2, 6
- [7] Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning, 2023. 2, 3, 6
- [8] Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Ruiyuan Lyu, Runsen Xu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024. 2, 3, 6
- [9] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3193–3203, 2021. 1, 2, 5, 6
- [10] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-rgpt: Grounded spatial reasoning in vision-language models. *arXiv preprint arXiv:2406.01584*, 2024. 2, 3
- [11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 4, 5
- [12] Alexandros Delitzas, Maria Parelli, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Multi-clip: Contrastive vision-language pre-training for question answering tasks in 3d scenes. *arXiv preprint arXiv:2306.02329*, 2023. 2
- [13] Azade Farshad, Yousef Yeganeh, Yu Chi, Chengzhi Shen, Böjrn Ommer, and Nassir Navab. Scenegenie: Scene graph guided diffusion models for image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 88–98, 2023. 2
- [14] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, Xiangdong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. Free-form description guided 3d visual graph network for object grounding in point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3722–3731, 2021. 1, 2
- [15] Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024. 6
- [16] Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. Graphdreamer: Compositional 3d scene synthesis from scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21295–21304, 2024. 2
- [17] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024. 1, 2, 3
- [18] Zeyu Han, Fangrui Zhu, Qianru Lao, and Huaizu Jiang. Zero-shot referring expression comprehension via structural similarity between images and captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14364–14374, 2024. 2
- [19] Yu He and Kang Zhou. Relation-wise transformer network and reinforcement learning for visual navigation. *Neural Computing and Applications*, pages 1–17, 2024. 1
- [20] Daniel Honerkamp, Martin Büchner, Fabien Despinoy, Tim Welschehold, and Abhinav Valada. Language-grounded dynamic scene graphs for interactive object search with mobile manipulation. *IEEE Robotics and Automation Letters*, 2024. 1, 2, 3
- [21] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *NeurIPS*, 2023. 3, 6
- [22] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 1, 2
- [23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6
- [24] Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3d

- v2: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023. 2, 3, 6
- [25] Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 3, 5, 6
- [26] Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 5, 6
- [27] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15524–15533, 2022. 6
- [28] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pages 417–433. Springer, 2022. 6
- [29] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 2
- [30] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1219–1228, 2018. 2
- [31] Weitai Kang, Haifeng Huang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Robin3d: Improving 3d large language model via robust instruction tuning, 2025. 6
- [32] Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14183–14193, 2024. 1
- [33] Maxim Kolodiaznyy, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20943–20953, 2024. 3, 8
- [34] Sergey Linok, Tatiana Zemskova, Svetlana Ladanova, Roman Titkov, and Dmitry Yudin. Beyond bare queries: Open-vocabulary object retrieval with 3d scene graph. *arXiv preprint arXiv:2406.07113*, 2024. 2, 3
- [35] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 5, 6
- [36] Taiki Miyanishi, Daichi Azuma, Shuhei Kurita, and Motoaki Kawanabe. Cross3dvg: Cross-dataset 3d visual grounding on different rgb-d scans. In *2024 International Conference on 3D Vision (3DV)*, pages 717–727. IEEE, 2024. 2, 4, 5
- [37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 4
- [38] Ege Özsoy, Tobias Czempel, Felix Holm, Chantal Pellegrini, and Nassir Navab. Labrad-or: lightweight memory scene graphs for accurate bimodal reasoning in dynamic operating rooms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 302–311. Springer, 2023. 1
- [39] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6
- [40] Jiaming Pei, Kaiyang Zhong, Zhi Yu, Lukun Wang, and Kuruvu Lakshmana. Scene graph semantic inference for image and text matching. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(5):1–23, 2023. 2
- [41] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023. 2
- [42] Itthisak Phueaksri, Marc A Kastner, Yasutomo Kawanishi, Takahiro Komamizu, and Ichiro Ide. An approach to generate a caption for an image collection using scene graph generation. *IEEE Access*, 2023. 2
- [43] Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. *arXiv preprint arXiv:2501.01428*, 2025. 6, 7
- [44] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4
- [45] Antoni Rosinol, Andrew Violette, Marcus Abate, Nathan Hughes, Yun Chang, Jingnan Shi, Arjun Gupta, and Luca Carlone. Kimera: From slam to spatial perception with 3d dynamic scene graphs. *The International Journal of Robotics Research*, 40(12-14):1510–1546, 2021. 1
- [46] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 3, 5, 6, 7, 8
- [47] Hengcan Shi, Munawar Hayat, and Jianfei Cai. Open-vocabulary object detection via scene graph discovery. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4012–4021, 2023. 2
- [48] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 2
- [49] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

- [50] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 4, 5
- [51] Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong Ma, Yiheng Liu, Xuhui Wang, et al. Large language models for robotics: Opportunities, challenges, and perspectives. *arXiv preprint arXiv:2401.04334*, 2024. 1
- [52] Ziqin Wang, Bowen Cheng, Lichen Zhao, Dong Xu, Yang Tang, and Lu Sheng. VI-sat: Visual-linguistic semantics assisted training for 3d semantic scene graph prediction in point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21560–21569, 2023. 1, 2, 4, 5
- [53] Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024. 1, 2, 3
- [54] Zizhao Wu, Haohan Li, Gongyi Chen, Zhou Yu, Xiaoling Gu, and Yigang Wang. 3d question answering with scene graph reasoning. In *ACM Multimedia 2024*, 2024. 2
- [55] Jianing Yang, Xuweiyi Chen, Shengyi Qian, Nikhil Madaan, Madhavan Iyengar, David F Fouhey, and Joyce Chai. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7694–7701. IEEE, 2024. 2, 3
- [56] Sibe Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4145–4154, 2019. 2
- [57] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10685–10694, 2019. 2
- [58] Zhihao Yuan, Jinke Ren, Chun-Mei Feng, Hengshuang Zhao, Shuguang Cui, and Zhen Li. Visual programming for zero-shot open-vocabulary 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20623–20633, 2024. 2, 3, 6
- [59] Guangyao Zhai, Evin Pinar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonsences: Generating commonsense 3d indoor scenes with scene graphs. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [60] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15225–15236, 2023. 2, 5, 6
- [61] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021. 2, 6
- [62] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023. 6
- [63] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023. 4
- [64] Kang Zhou, Chi Guo, Huyin Zhang, and Bohan Yang. Optimal graph transformer viterbi knowledge inference network for more successful visual navigation. *Advanced Engineering Informatics*, 55:101889, 2023. 1
- [65] Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, and Qing Li. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2911–2921, 2023. 2, 6
- [66] Ziyu Zhu, Zhuofan Zhang, Xiaojian Ma, Xuesong Niu, Yixin Chen, Baoxiong Jia, Zhidong Deng, Siyuan Huang, and Qing Li. Unifying 3d vision-language understanding via promptable queries. In *European Conference on Computer Vision*, pages 188–206. Springer, 2025. 6