HISUPER-VLM: MULTI-SCALE SCENE UNDERSTANDING VIA HIERARCHICAL SUPERPOINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

3D Vision-Language Models (3D-VLMs) have shown strong potential for 3D scene understanding. However, scaling to larger, more complex scenes challenges current 3D-VLMs in both efficiency and accuracy. This is largely due to their use of variable-length scene tokens that grow with scene complexity and the decoupling of scene representation and understanding, which limits VLM's adaptive reasoning over scenes with different scales. To address this, we propose HiSuper-VLM, a 3D-VLM framework leveraging hierarchical superpoints to construct compact yet expressive representations, providing a foundation for VLM understanding across multiple scales and hierarchical levels while tightly coupling representation with reasoning. Specifically, for representation, we construct hierarchical superpoints, using fixed-length large-scale superpoints as initial visual tokens for VLMs while retaining fine-scale superpoints for information completeness. For understanding, we perform multi-level queries across VLM layers, from coarse to fine, enhancing reasoning by refining 3D representations during the process. Experimental results demonstrate that HiSuper-VLM significantly outperforms state-of-the-art methods on multiple indoor and outdoor 3D scene understanding benchmarks. Code will be publicly released upon acceptance.

1 INTRODUCTION

Recent advances in extending large language models (LLMs) Iyer et al. (2022); OpenAI (2023) to multimodal domains have spurred growing interest in 3D vision-language models (3D-VLMs) Tang et al. (2024); Xu et al. (2024); Chen et al. (2024a); Zhi et al. (2025); Yuan et al. (2025), which aim to equip LLMs with the ability to perceive and reason about real-world 3D environments. Driven by applications such as embodied assistants Huang et al. (2023b), robotics Huang et al. (2023c); Zhen et al. (2024), and mixed reality Zhao et al. (2024); Song et al. (2025) that require sophisticated interaction with complex environments, the research focus has shifted from early object-level 3D understanding Guo et al. (2023); Tang et al. (2024); Xu et al. (2024); Luo et al. (2023) to more challenging scene-level comprehension Huang et al. (2023a; 2024; 2023b); Hong et al. (2023); Chen et al. (2024a); Wang et al. (2023); Zhi et al. (2025); Yu et al. (2025); Yuan et al. (2025), and further extends toward larger and more complex 3D scenarios.

Similar to image VLMs Chen et al. (2024b); Wang et al. (2024); Li et al. (2024), 3D-VLMs first align 3D scene representations with the embedding space of LLMs, followed by concatenating them with textual instructions and jointly feeding them into LLMs for 3D scene understanding. However, the development of 3D-VLMs is fundamentally challenged by the irregular and sparse nature of 3D point cloud data, which prevents straightforward partitioning into regular grids. To simplify this challenge, existing scene-level 3D-VLMs Huang et al. (2023a; 2024; 2023b); Chen et al. (2024a); Wang et al. (2023); Yu et al. (2025); Yuan et al. (2025) typically adopt instance-level features as vision tokens: they first utilize pre-trained object detection Chen et al. (2023); Misra et al. (2021); Qi et al. (2019) or instance segmentation Schult et al. (2022); Zhang et al. (2023a); Jiang et al. (2020) models to identify all object instances from scene point clouds, and then represent the scene as a sequence of instance features, similar to image patches in 2D vision.

Despite the advancements, such approaches Huang et al. (2023a; 2024; 2023b); Chen et al. (2024a); Wang et al. (2023); Yu et al. (2025); Yuan et al. (2025) encounter two critical limitations when scaling to larger and more complex scenarios (fig. 1(a)). ***First, generalization and scalability are***

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

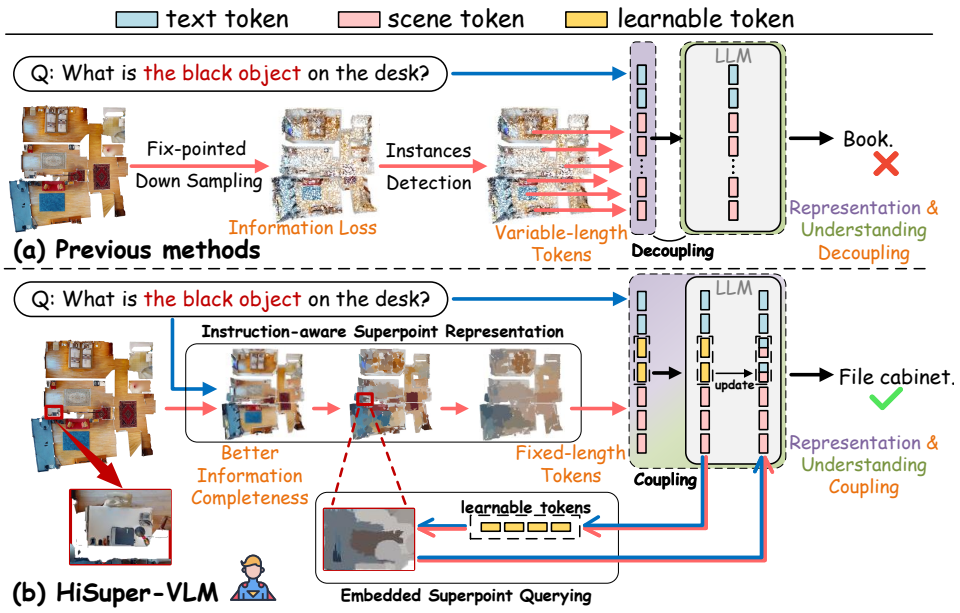


Figure 1: Comparison between previous methods and our HiSuper-VLM. (a) Previous methods suffer from limited scalability and struggle to balance efficiency and accuracy. (b) HiSuper-VLM addresses these issues via hierarchical superpoint representation and embedded information querying.

constrained. Prior works represent scenes using variable-length instance-level tokens, which rely on pre-trained detection models and hinder their application to large-scale complex scenes with numerous objects. In addition, their use of fixed-point downsampling for point cloud preprocessing leads to irreversible information loss. ***Second, balancing efficiency and accuracy is challenging.*** These methods adopt decoupled designs that separate 3D representation (as input to LLMs) from 3D understanding (processed within the LLM) without joint optimization, which limits VLM's adaptive reasoning over scenes with different scales. This decoupling leads to an inevitable trade-off in large-scale complex scenes: simply compressing tokens to improve efficiency will degrade accuracy, while increasing the number of tokens to pursue accuracy often sacrifices efficiency.

Therefore, we aim to develop a unified paradigm that constructs information-preserving scene representations with fixed-length vision tokens, while also coupling representation with understanding to jointly enhance efficiency and accuracy. To achieve this design, we leverage a key property of point clouds: their inherent hierarchical organization from local geometric components to complete objects and the entire scene. By iteratively aggregating geometrically or semantically consistent points into superpoints, a hierarchical structure can be constructed that preserves rich details at lower levels while providing compact semantics at higher levels, effectively addressing the issues of scalability and information loss. Meanwhile, with low-level details preserved, representation and understanding can be seamlessly coupled by allowing the LLM to access critical local details during its reasoning process.

Building on this insight, we propose **HiSuper-VLM**, a novel 3D-VLM framework that leverages hierarchical superpoints for both compact scene representation and fine-grained information querying (fig. 1(b)). HiSuper-VLM introduces two key components: an **Instruction-aware Superpoint Aggregation Module (ISAM)** that constructs a hierarchical superpoint structure to provide both better information completeness and fixed-length vision tokens with enhanced semantic compactness, and an **Embedded Superpoint Query Module (ESQM)** that couples representation and understanding through selective fine-grained information querying. Specifically, ISAM performs instruction-guided feature aggregation and intra-level feature interaction during construction to enhance semantic expressiveness and contextual perception, while ESQM leverages LLMs' attention guidance to accurately extract critical details from low-level superpoints and efficiently supplements fine-grained information with an embedded query design. Through the synergy of ISAM and ESQM, HiSuper-VLM achieves an improved balance between efficiency and accuracy, enabling scalable and precise 3D scene understanding in complex environments (fig. 2). Experimental results demonstrate that

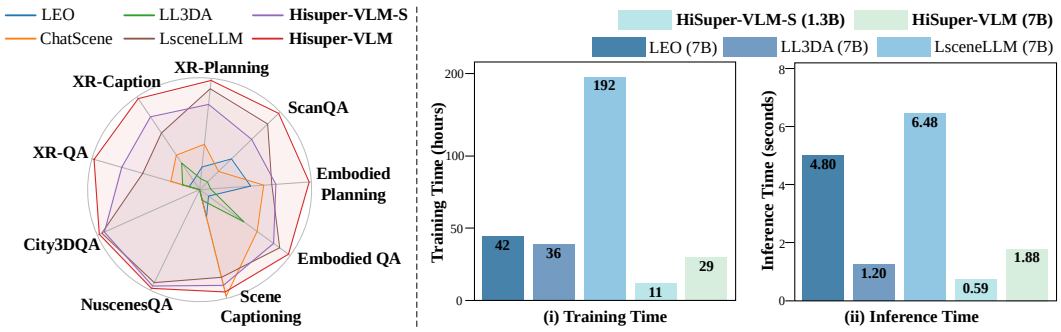


Figure 2: Our method outperforms previous approaches across multiple benchmarks, achieving a better balance between efficiency and accuracy.

HiSuper-VLM achieves state-of-the-art performance across a variety of 3D tasks and scene understanding benchmarks, including indoor single-room, multi-room, outdoor street-level, and city-scale scenarios.

Our main contributions are summarized as follows:

- We propose HiSuper-VLM, a novel 3D-VLM framework, which utilizes hierarchical superpoints for compact scene representation and fine-grained information querying, achieving enhanced semantic expressiveness while reducing computational overhead.
- We design an Instruction-aware Superpoint Aggregation Module for adaptive hierarchical superpoint construction, and an Embedded Superpoint Query Module that couples representation with understanding by integrating LLMs' visual preferences with an embedded query mechanism.
- Our method significantly advances the state-of-the-art performance across multiple indoor and outdoor 3D scene benchmarks, demonstrating the effectiveness and superiority of HiSuper-VLM.

2 Related Work

2.1 3D Vision-Language Models

Recently, substantial progress has been made in 3D-VLMs Xu et al. (2024); Luo et al. (2023); Huang et al. (2023a; 2024; 2023b); Hong et al. (2023); Chen et al. (2024a); Wang et al. (2023); Zhi et al. (2025); Yu et al. (2025); Yuan et al. (2025) by leveraging the powerful reasoning capabilities of LLMs Zhang et al. (2022); Bai et al. (2023); Chiang et al. (2023); Touvron et al. (2023) for 3D scene understanding. To enable LLMs to comprehend irregular scene point clouds, existing methods typically leverage extracted object features to construct scene representations. Chat-Scene Huang et al. (2024) represents scenes as sequences of object embeddings with unique identifier tokens for precise referencing and grounding. LEO Huang et al. (2023b) captures object-level features using PointNet++ Qi et al. (2017) and incorporates a Spatial Transformer Chen et al. (2022) to model inter-object spatial relationships, realizing an embodied multimodal agent. Inst3D-LMM Yu et al. (2025) introduces a multi-view cross-modal fusion module to obtain fine-grained instance-level vision tokens, as well as an instance spatial relation module to capture pairwise spatial relationships among objects. Larger scenes are inherently more complex, posing greater challenges to 3D-VLMs in constructing efficient and expressive scene representations. LL3DA Chen et al. (2024a) adopts a Q-Former Li et al. (2023) mechanism to query task-relevant scene features, improving efficiency but suffering from information loss. Moreover, its Q-Former Li et al. (2023) relies on additional language models (e.g., BERT Turc et al. (2019)) for language modeling, and requires extensive training data for specialized vision-language pre-training Yao et al. (2024). LSceneLLM Zhi et al. (2025) introduces a scene magnifier that improves scene encoding through progressive amplification. However, it implicitly assumes expandable computation, and its layer-wise sampling and processing of dense point cloud features incur substantial computational cost, thereby limiting inference efficiency. Different from previous work that primarily focuses on improving scene encoding quality while struggling to balance efficiency and accuracy, we redefine the problem as maximizing infor-

mation accessibility under a fixed computation budget. We introduce superpoint aggregation and query modules that construct hierarchical superpoints for compact and expressive scene representation and fine-grained information querying, achieving precise and efficient 3D reasoning.

2.2 SUPERPOINT-BASED POINT CLOUD ANALYSIS

The strategy of partitioning large 3D point clouds into groups of adjacent and homogeneous points, termed superpoints, has been successfully applied across various point cloud analysis tasks Hui et al. (2021); Landrieu & Boussaha (2019); Landrieu & Simonovsky (2018); Sun et al. (2023); Robert et al. (2023); Zhang et al. (2023b); Robert et al. (2024). SPG Landrieu & Simonovsky (2018) proposes organizing 3D point clouds into superpoints and constructing superpoint graphs, encoding relational features between adjacent superpoints and propagating context via graph convolutions for large-scale semantic segmentation. SPFormer Sun et al. (2023) groups features from point clouds into superpoints and directly predicts instances through query vectors, achieving end-to-end 3D instance segmentation. SPT Robert et al. (2023) introduces a superpoint transformer and leverages hierarchical superpoint partitions for efficient segmentation of large-scale 3D scenes. GrowSP Zhang et al. (2023b) introduces a purely unsupervised 3D semantic segmentation method that guides feature learning via progressive growing of superpoints, using superpoints for semantic primitive clustering and pseudo-label generation. SuperCluster Robert et al. (2024) reformulates panoptic segmentation as a scalable graph clustering problem by predicting superpoint semantic categories and target consistency between adjacent superpoints. Instead of treating superpoints merely as preprocessing outputs, we draw upon the inherent hierarchical organization of point clouds to construct a multi-level superpoint hierarchy, which not only provides compact high-level scene semantics but also enables the LLM to selectively query low-level details during reasoning.

3 METHOD

3.1 OVERVIEW

In this paper, we propose HiSuper-VLM, which constructs efficient and expressive 3D scene representations through a hierarchical superpoint structure and enables fine-grained information querying within the LLM's reasoning process. As shown in fig. 3, HiSuper-VLM comprises two key modules: i) the Instruction-aware Superpoint Aggregation Module (ISAM) for representation and ii) the Embedded Superpoint Query Module (ESQM) for understanding. Raw point clouds are first preprocessed (see section 3.2) and then fed into ISAM, which adaptively aggregates instruction-relevant superpoint features and enables intra-level interactions among them (see section 3.3). The resulting multi-level hierarchical superpoint structure effectively adapts to diverse 3D scenes, preserving both expressiveness and efficiency: low-level superpoint features retain fine-grained details, while compact high-level features are employed as a fixed number of vision tokens and then input into the LLM. Within the LLM, ESQM selectively extracts critical details from low-level superpoint features through the attention map, and injects them into learnable tokens to efficiently supplement instruction-related details (see section 3.4), thereby coupling 3D representation and understanding. Subsequently, the enriched representation is processed by the remaining LLM layers to produce the final response.

3.2 PREPROCESSING

The raw point cloud $\mathbf{P} \in \mathbb{R}^{N \times C}$ of a 3D scene, where N denotes the number of points and C is typically 6, with three dimensions for spatial coordinates and three for color, is first fed into a scene encoder Peng et al. (2023) to extract per-point features $\mathbf{f} \in \mathbb{R}^{N \times F}$, where F is the feature dimension. Then the point cloud is processed into initial superpoints $\mathbf{s}_1 = [\mathbf{c}_1, \mathbf{f}_1]$ using a fast superpoint generation algorithm Landrieu & Obozinski (2017), which clusters the N points into N_1 level-1 superpoints to form $\mathbf{s}_1 = [\mathbf{c}_1, \mathbf{f}_1]$. $\mathbf{c}_1 \in \mathbb{R}^{N_1 \times 3}$ represents the superpoint centers and $\mathbf{f}_1 \in \mathbb{R}^{N_1 \times F}$ their corresponding features. Unlike prior methods Huang et al. (2024; 2023b); Hong et al. (2023); Chen et al. (2024a); Wang et al. (2023); Zhi et al. (2025); Yu et al. (2025) that rely on fixed-size downsampling (e.g., sampling 40k points), our adaptive superpoint representation compresses the raw point cloud by several orders of magnitude while preserving fine-grained geometric and semantic details.

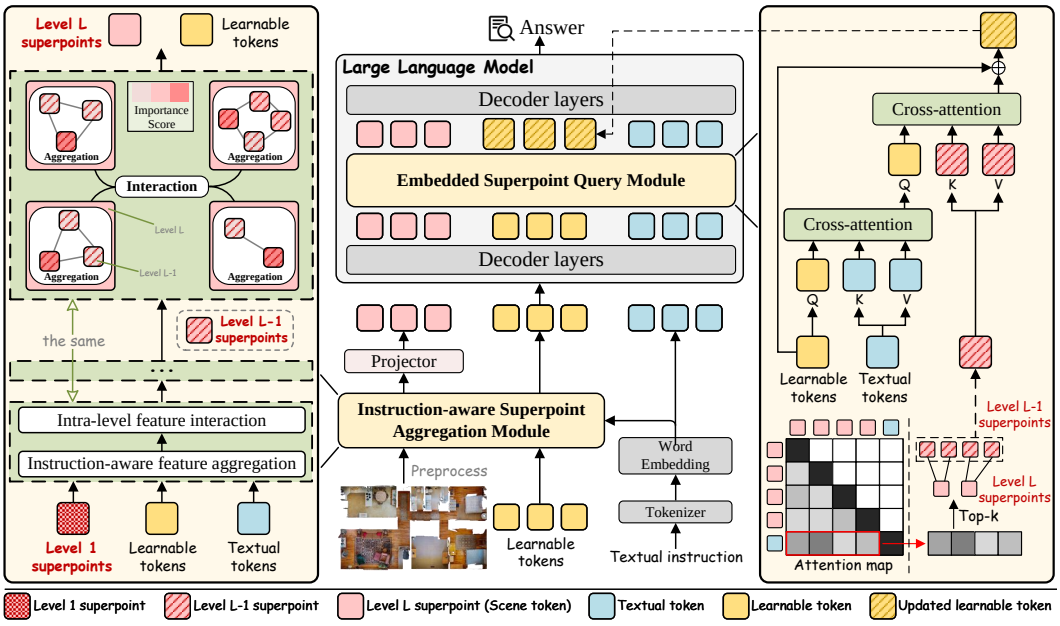


Figure 3: **Overview of HiSuper-VLM.** HiSuper-VLM constructs compact and expressive 3D representations and tightly couples representation with understanding. Left: The Instruction-aware Superpoint Aggregation Module constructs a hierarchical superpoint structure through instruction-guided feature aggregation and intra-level interaction, offering fixed-length vision tokens with better information completeness. Right: The Embedded Superpoint Query Module leverages LLM-guided attention to extract fine-grained superpoints in focused regions and integrates them into the LLM through updated learnable tokens, refining 3D representation during the understanding process.

3.3 Instruction-aware Superpoint Representation

In this section, we propose an Instruction-aware Superpoint Aggregation Module to construct compact and expressive hierarchical superpoints, which performs instruction-aware aggregation (section 3.3.1) to aggregate key semantics, and intra-level interaction (section 3.3.2) to enhance contextual expressiveness among superpoints.

3.3.1 Instruction-aware Feature Aggregation

After obtaining the initial superpoints, the instruction-aware feature aggregation layer further aggregates semantic information to construct a hierarchical superpoint structure. Taking the construction of the i-th level superpoints as an example, we first randomly sample N_i centroid points from the $(i-1)$-th level superpoints, denoted as $\mathbf{c}_i = [\mathbf{c}_{i,1}, \mathbf{c}_{i,2}, \ldots, \mathbf{c}_{i,N_i}]^T \in \mathbb{R}^{N_i \times 3}$. For k-th centroid $\mathbf{c}_{i,k}$, we sample $(i-1)$-th level superpoint features within its neighborhood to form a local group $\mathcal{G}(i,k) = \{\mathbf{f}_{i,k,m} \mid m = 1, 2, \ldots, M\}$, where $\mathcal{G}(i,k)$ denotes the local group corresponding to the k-th superpoint at level i, and M is the group size. We then perform instruction-aware feature aggregation to generate the i-th level superpoint features. To achieve this, we first extract instruction information using learnable tokens via cross-attention:

$$\mathbf{l}' = \text{softmax}\left(\frac{(\mathbf{l}\mathbf{W}_q)(\mathbf{t}\mathbf{W}_k)^T}{\sqrt{d_k}}\right)(\mathbf{t}\mathbf{W}_v) \tag{1}$$

$\mathbf{l} \in \mathbb{R}^{l \times d}$ denotes the learnable tokens and $\mathbf{t} \in \mathbb{R}^{t \times d}$ represents the textual tokens, where l and t are the numbers of learnable and textual tokens respectively, and d is the hidden dimension of the LLM. $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are projection matrices, and d_k is the projected dimension. Next, for each local group $\mathcal{G}(i,k)$, we compute the similarity between each superpoint feature $\mathbf{f}_{i,k,m} \in \mathcal{G}(i,k)$ and the instruction-aware learnable tokens $\mathbf{l}'$. These similarities are then normalized via group-wise softmax to ensure relevance weighting within the group:

$$\alpha_{i,k,m} = \text{softmax}_{\mathbf{f}_{i,k,m} \in \mathcal{G}(i,k)}\left(\frac{(\mathbf{f}_{i,k,m}\mathbf{W}_f)\mathbf{l}'}{\sqrt{d_k}}\right) \tag{2}$$

where $\mathbf{W}_f$ is a feature projection matrix. Using these similarities as importance scores, we perform weighted aggregation to generate the superpoint features of the i-th level, denoted as $\mathbf{f}_i = [\mathbf{f}_{i,1}, \mathbf{f}_{i,2}, \ldots, \mathbf{f}_{i,N_i}]^T \in \mathbb{R}^{N_i \times F}$.

$$\mathbf{f}_{i,k} = \sum_{\mathbf{f}_{i,k,m} \in \mathcal{G}(i,k)} \alpha_{i,k,m} \cdot \mathbf{f}_{i,k,m} \tag{3}$$

Meanwhile, the learnable tokens $\mathbf{l}$ are updated via residual connection to integrate instruction semantics:

$$\mathbf{l} = \mathbf{l} + \text{MLP}(\mathbf{l}') \tag{4}$$

By applying $L-1$ aggregation layers, we construct an L-level hierarchical superpoint structure, providing both better information completeness and fixed-length vision tokens. The superpoint sets corresponding to level $1, 2, \ldots, L$ are denoted as $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_L$, respectively. The features of the top-level superpoints $\mathbf{s}_L$ are used as the vision token, which are then projected and concatenated with the learnable tokens and textual tokens to form the input to the LLM. In our implementation, we set $L = 3$, which achieves the best balance between performance and efficiency as validated in section A.2.3.

3.3.2 Intra-level Feature Interaction

For the superpoints aggregated in each level, we enable intra-level feature interaction through a sparse attention mechanism, allowing each superpoint to incorporate contextual information from its spatial neighbors. This process expands the semantic perception range and refines the overall representation. Specifically, for the feature $\mathbf{f}_{i,k}$ of the k-th superpoint at level i, we take it as the query, while its spatial neighbors $\mathbf{f}_{i,j} \in \mathcal{F}(i,k)$ serve as keys and values. The information propagation among neighboring superpoints is formulated as follows:

$$\mathbf{q}_{i,k} = \theta_q(\mathbf{f}_{i,k}), \mathbf{k}_{i,j} = \theta_k(\mathbf{f}_{i,j}), \mathbf{v}_{i,j} = \theta_v(\mathbf{f}_{i,j}) \tag{5}$$

$$\delta_{k,j} = \theta_p(\mathbf{c}_{i,k} - \mathbf{c}_{i,j}) \tag{6}$$

$$w_{i,j} = \text{softmax}\big(\theta_w(\mathbf{q}_{i,k} - \mathbf{k}_{i,j} + \delta_{k,j})\big) \tag{7}$$

$$\mathbf{f}_{i,k} \overset{+}{=} \sum_{\mathbf{f}_{i,j} \in \mathcal{F}(i,k)} w_{i,j} \cdot \big(\mathbf{v}_{i,j} + \delta_{k,j}\big) \tag{8}$$

In eq. (5)–eq. (6), the query $\mathbf{q}_{i,k}$, key $\mathbf{k}_{i,j}$, and value $\mathbf{v}_{i,j}$ vectors are obtained by linearly projecting the central and neighboring superpoint features through θ_q, θ_k, and θ_v, while $\delta_{k,j}$ encodes their relative spatial relationship via the positional encoder θ_p. Based on query–key similarity and their positional relation, the attention weight $w_{i,j}$ is computed through θ_w as in eq. (7). Finally, neighboring value features $\mathbf{v}_{i,j}$ are aggregated with weights $w_{i,j}$ to update the central feature $\mathbf{f}_{i,k}$ as in eq. (8), where $\overset{+}{=}$ denotes a residual connection and $\mathcal{F}(i,k)$ represents the local (K-nearest) neighborhood of $\mathbf{f}_{i,k}$.

3.4 Embedded Superpoint Querying

In this section, we couple representation and understanding by leveraging the constructed hierarchical superpoint representation to enable fine-grained information querying within the LLM. Specifically, we insert an Embedded Superpoint Query Module (ESQM) into the i-th layer of the LLM, enabling it to leverage visual preferences to extract key details from low-level superpoints and inject fine-grained visual information back into the LLM through learnable tokens. To locate the visual regions of interest, the attention weights from textual tokens to visual tokens are extracted, and the top-k superpoints with the highest scores are selected. For each selected high-level superpoint feature $\mathbf{f}_{L,k}$ in $\mathbf{s}_L$, its corresponding group $\mathcal{G}(L,k)$ is retrieved to extract low-level superpoint features from $\mathbf{s}_{L-1}$. The resulting fine-grained features $\mathbf{f}$ enrich visual details and are integrated into the LLM via learnable tokens $\mathbf{l}$. First, $\mathbf{l}$ interacts with textual tokens through cross-attention to acquire instruction-aware learnable tokens $\mathbf{l}'$:

$$\mathbf{l}' = \text{softmax}\left(\frac{(\mathbf{l}\mathbf{W}_q)(\mathbf{t}\mathbf{W}_k)^T}{\sqrt{d_k}}\right)(\mathbf{t}\mathbf{W}_v) \tag{9}$$

Then, $\mathbf{l}'$ queries $\mathbf{f}$ to acquire instruction-relevant details $\mathbf{l}''$:

$$\mathbf{l}'' = \text{softmax}\left(\frac{(\mathbf{l}'\mathbf{W}_q)(\mathbf{f}\mathbf{W}_k)^T}{\sqrt{d_k}}\right)(\mathbf{f}\mathbf{W}_v) \tag{10}$$

Lastly, $\mathbf{l}''$ is refined by a lightweight MLP and fused with the learnable tokens $\mathbf{l}$:

$$\mathbf{l} = \mathbf{l} + \text{MLP}(\mathbf{l}'') \tag{11}$$

By reusing the fixed-length learnable tokens $\mathbf{l}$ to enhance subsequent reasoning, we efficiently supplement fine-grained visual details with minimal computational overhead, achieving an improved balance between efficiency and accuracy.

4 EXPERIMENTS

4.1 EXPERIMENTS SETUP

4.1.1 DATASETS

We conduct experiments on a diverse set of 3D benchmarks to evaluate both the effectiveness and scalability of our proposed method. For indoor scenes, we train on ScanNet Dai et al. (2017) and HM3D Ramakrishnan et al. (2021) with language annotations from ScanQA Azuma et al. (2022), ScanRefer Chen et al. (2020), Nr3D Achlioptas et al. (2020), the ScanNet subset of 3D-LLM Hong et al. (2023), and XR-Scene Zhi et al. (2025), covering a wide range of 3D tasks such as instance captions, scene descriptions, question answering, and embodied planning. For outdoor street-level scenes, we evaluate on NuScenes-QA Qian et al. (2024) using a model trained specifically for this dataset, focusing on reasoning in sparse point clouds of driving contexts. To further assess performance in even larger scenes, we additionally conduct experiments on City-3DQA Sun et al. (2024) with another model trained independently for this dataset, which targets city-level scene understanding.

4.1.2 METRICS

We report CIDEr Vedantam et al. (2015), METEOR Banerjee & Lavie (2005), and ROUGE Lin (2004) as metrics to measure the quality of generated text responses. For each metric in the table, **bold** and underline numbers indicate the best and the second-best results, respectively.

4.1.3 IMPLEMENTATION DETAILS

We evaluate HiSuper-VLM using two causal LLM backbones: OPT-1.3B Zhang et al. (2022) and LLaMA2-7B Touvron et al. (2023), corresponding to HiSuper-VLM-S and HiSuper-VLM, respectively. We apply a two-stage training paradigm. During the pre-alignment stage, only the instruction-aware superpoint aggregation module and the projector are optimized. In the joint training stage, the entire HiSuper-VLM is optimized, including the instruction-aware superpoint aggregation module, the projector, the learnable tokens, the LLM backbone, and the embedded superpoint query module. We use the AdamW Loshchilov & Hutter (2017) optimizer and a cosine annealing scheduler. More details are provided in the supplementary material.

4.2 MAIN RESULTS

We evaluate HiSuper-VLM across 3D scenes of varying scales, including multi-room, single-room, street-level, and city-level scenarios. Experimental results demonstrate that our method achieves leading performance across all scales, highlighting its strong effectiveness and scalability.

4.2.1 MULTI-ROOM SCENE UNDERSTANDING

As shown in table 1, our HiSuper-VLM demonstrates exceptional capability, achieving significant improvements across all multi-room tasks. Specifically, HiSuper-VLM-S, with the lightweight backbone OPT-1.3B Zhang et al. (2022), achieves performance competitive with the previous best method. Notably, compared under the same backbone, HiSuper-VLM-S significantly outperforms LL3DA Chen et al. (2024a) across all metrics, highlighting the effectiveness of our method even with constrained model capacity. When scaling to a stronger LLM backbone LLaMA2-7B Touvron et al. (2023), our HiSuper-VLM surpasses existing methods by a large margin, improving the state-of-the-art CIDEr scores on XR-QA, XR-SceneCaption, and XR-EmbodiedPlanning tasks by 5.38, 1.57, and 2.07 points, respectively. Chat-Scene Huang et al. (2024) and LEO Huang et al. (2023b) adopt variable-length object-level tokens, suffering from limitations in both expressiveness and efficiency. LL3DA Chen et al. (2024a) employs learnable queries to compress task-relevant object

Table 1: Multi-room scene understanding results. C, R, M denote CIDEr, ROUGE, METEOR, respectively. Note that LEO Huang et al. (2023b) does not identify question-related objects for the model, while ChatScene Huang et al. (2024) requires images and point clouds as input. Blue and red numbers in parentheses denote improvements relative to LL3DA Chen et al. (2024a) and LSceneLLM Zhi et al. (2025), respectively, with the same LLM backbone. Our HiSuper-VLM achieves significant improvements across all tasks.

Method	Backbone	XR-QA			XR-Caption			XR-Planning		
		C	R	M	C	R	M	C	R	M
LEO Huang et al. (2023b)	Vicuna-7B	112.09	35.47	14.02	2.42	15.96	10.25	39.45	33.31	18.99
ChatScene Huang et al. (2024)	Vicuna-7B	114.10	35.93	14.32	3.58	17.49	11.59	46.18	36.71	22.34
LL3DA Chen et al. (2024a)	OPT-1.3B	112.80	36.94	18.68	3.22	20.95	13.49	35.96	31.50	15.74
LSceneLLM Zhi et al. (2025)	LLaMA2-7B	117.21	38.18	19.30	4.59	23.43	16.16	63.08	36.96	22.97
HiSuper-VLM-S	OPT-1.3B	119.53 (+6.73)	38.81 (+1.87)	**19.92** (+1.24)	5.33 (+2.11)	**23.75** (+2.80)	17.22 (+3.73)	58.03 (+22.07)	36.44 (+4.94)	23.39 (+7.65)
HiSuper-VLM	LLaMA2-7B	**122.59** (+5.38)	**41.44** (+3.26)	19.46 (+0.16)	**6.16** (+1.57)	23.53 (+0.10)	**17.75** (+1.59)	**65.15** (+2.07)	**37.59** (+0.63)	**23.49** (+0.52)

Table 2: Single-room task results on ScanQA, Scene Captioning, Embodied Planning, and Embodied QA. Notes follow table 1. C, R, M denote CIDEr, ROUGE, METEOR, respectively.

Method	ScanQA			Scene Captioning			Embodied Planning			Embodied QA		
	C	R	M	C	R	M	C	R	M	C	R	M
LEO Huang et al. (2023b)	80.20	40.24	16.68	1.80	20.84	13.29	204.78	46.40	19.86	86.14	30.89	18.81
ChatScene Huang et al. (2024)	77.75	37.79	15.94	**3.67**	21.05	12.60	210.86	40.03	20.71	99.01	34.23	18.48
LL3DA Chen et al. (2024a)	75.67	37.02	15.37	1.44	**24.62**	12.93	186.13	45.34	19.60	95.53	33.75	19.81
LSceneLLM Zhi et al. (2025)	88.24	40.82	17.95	3.07	21.88	14.79	214.63	47.05	21.05	104.98	36.00	21.26
HiSuper-VLM-S	82.02 (+6.35)	39.42 (+2.40)	16.95 (+1.58)	3.24 (+1.80)	21.28 (-3.34)	13.90 (+0.97)	216.52 (+30.39)	48.25 (+2.91)	22.08 (+2.48)	103.37 (+7.84)	35.81 (+2.06)	22.12 (+2.31)
HiSuper-VLM	**89.56** (+1.32)	**41.79** (+0.97)	**18.29** (+0.34)	3.43 (+0.36)	23.62 (+1.74)	**15.07** (+0.28)	**234.14** (+19.51)	**48.70** (+1.65)	**22.93** (+1.88)	**107.37** (+2.39)	**36.43** (+0.43)	**22.38** (+1.12)

features in the scene, improving efficiency but achieving suboptimal performance due to information loss. LSceneLLM Zhi et al. (2025) supplements details by enlarging regions of interest, yet suffers from inefficiency caused by dense layer-wise sampling. In contrast, our method achieves an improved balance between efficiency and accuracy through hierarchical superpoint representation and embedded superpoint querying.

4.2.2 Single-room Scene Understanding

Single-room scene understanding focuses on compact indoor environments that require the model to capture fine-grained scene semantics and object relationships. We evaluate on ScanQA Azuma et al. (2022) and the ScanNet-based tasks from 3D-LLM Hong et al. (2023) including scene captioning, embodied planning, and embodied question answering. Baseline methods remain the same as those used in the multi-room experiments. Results in table 2 demonstrate that our HiSuper-VLM consistently outperforms all baseline methods across various single-room tasks, further confirming its fine-grained reasoning capabilities in compact single-room scenes.

4.2.3 Street-level Scene Understanding

Beyond indoor settings, we further extend evaluation to street-level outdoor scenes using the NuScenes-QA Qian et al. (2024) benchmark. Outdoor street-level settings present a more expansive spatial scale and sparser visual information, which introduces significant challenges for 3D-VLMs in terms of comprehension and reasoning. As demonstrated in table 3, HiSuper-VLM exhibits exceptional performance, highlighting its proficiency in reasoning within complex outdoor contexts and achieving an accuracy of 57.1%.

4.2.4 City-level Scene Understanding

Finally, we validate scalability to even larger environments on the City-3DQA Sun et al. (2024) dataset in city-wise mode, targeting city-level scene understanding, which remains underexplored among existing 3D-VLMs. We compare our proposed HiSuper-VLM with specialist models that are tailored for specific tasks using task-oriented heads and LLM-based models that are designed for general instructions and responses. As shown in table 4, HiSuper-VLM achieves improvements over the strongest baseline without requiring scene graph input or specialized framework design, demonstrating robust effectiveness and clear scalability to massive point clouds.

Table 3: 3D question answering results on outdoor street-level benchmark NuscenesQA Qian et al. (2024). Evaluated by top-1 accuracy. * denotes specialist model.

Method	Exist			Count			Object			Status			Comparison			Acc
	H0	H1	All	H0	H1	All	H0	H1	All	H0	H1	All	H0	H1	All	
NuscenesQA* Qian et al. (2024)	87.7	81.1	84.1	21.9	20.7	21.3	70.2	45.6	49.2	62.8	52.4	55.9	81.6	68.0	69.2	58.1
LLaVA-Adapter-v2 Gao et al. (2023)	34.2	6.3	19.3	5.0	0.1	2.7	23.7	4.6	7.6	9.8	11.3	10.8	2.6	1.5	1.6	9.6
LLaVA Liu et al. (2023)	38.9	51.9	45.8	7.7	7.6	7.7	10.5	7.4	7.8	7.0	9.9	9.0	64.5	50.8	52.1	26.2
LidarLLM Yang et al. (2025)	79.1	70.6	74.5	15.3	14.7	15.0	59.6	34.1	37.8	53.4	42.0	45.9	67.0	57.0	57.8	48.6
OccLLaMA Wei et al. (2024)	80.6	79.3	79.9	18.6	19.1	18.9	64.9	39.0	42.8	48.0	49.6	49.1	80.6	63.7	65.2	53.4
LSceneLLM Zhi et al. (2025)	86.4	81.3	83.6	19.4	19.8	19.6	64.4	41.3	44.8	58.8	51.2	53.8	81.0	67.5	68.7	56.4
HiSuper-VLM-S	85.4	81.6	83.3	20.2	20.1	20.2	67.3	42.8	46.4	59.0	51.0	53.7	78.5	67.5	68.5	56.6
HiSuper-VLM	**86.6**	**81.9**	**84.0**	**20.5**	**20.2**	**20.4**	**67.6**	**42.9**	**46.5**	**59.5**	**51.7**	**54.3**	**82.0**	**67.9**	**69.2**	**57.1**

Table 4: Results on city-level scene understanding benchmark City3DQA Sun et al. (2024). "PC" means point cloud and " SG" means scene graph. Acc denotes top-1 accuracy.

Method	Input	Acc	CIDEr	ROUGE	METEOR
Specialist Models					
ScanQA	PC	47.33	-	-	-
CLIP-Guided	PC	46.94	-	-	-
3D-VLP	PC	48.74	-	-	-
3D-VisTA	PC	56.74	-	-	-
Sg-CityU	PC+SG	63.76	-	-	-
LLM-based Models					
3D-LLM	PC	56.79	245.55	66.95	34.98
LSceneLLM	PC	64.03	266.43	74.32	40.31
HiSuper-VLM-S	PC	61.93	258.72	72.45	37.06
HiSuper-VLM	PC	**66.63**	**269.98**	**77.36**	**40.46**

4.3 Efficiency Analysis

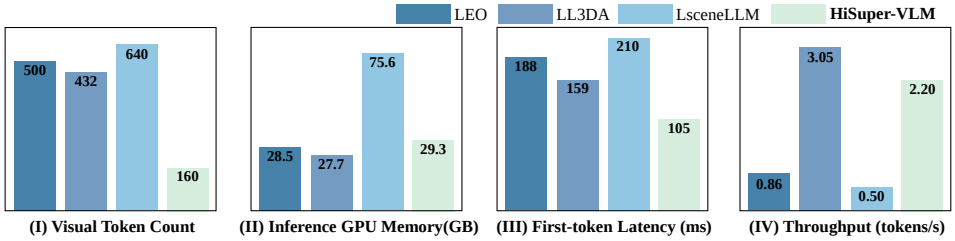


Figure 4: Efficiency analysis.

Fig. 4 compares HiSuper-VLM with competing 3D-VLMs on visual token count, inference GPU memory, first-token latency, and end-to-end throughput under matched settings. Together with Tables 1–4 and Figure 2, the results show that HiSuper-VLM achieves a better efficiency–accuracy trade-off. This comes from two designs: i) hierarchical superpoints produce compact, information-preserving tokens, unlike variable-length instance tokens that grow with scene complexity; ii) ESQM injects fine-grained details efficiently, avoiding the inefficiency of dense processing or the information loss of decoupled designs.

5 Conclusion

In this paper, we propose HiSuper-VLM, a novel 3D-VLM framework that leverages hierarchical superpoints for compact scene representation and fine-grained information querying. Our framework includes an instruction-aware superpoint aggregation module that adaptively perceives and aggregates instruction-relevant semantics to construct a hierarchical superpoint structure, and an embedded superpoint query module that incorporates LLM's visual preferences with an embedded query mechanism to efficiently supplement fine-grained information. Experimental results demonstrate that HiSuper-VLM achieves state-of-the-art performance across multiple indoor and outdoor 3D scene understanding benchmarks. Considering that high-quality and rich 3D visual-language annotations will further improve the model's performance, we regard data collection and construction for large-scale 3D scenes as our next work.

AI USE STATEMENT

In this work, we used generative AI tools for no tasks requiring disclosure. We have not used generative AI tools for any required-disclosure research tasks, and no other required-disclosure tasks are applicable to this work. Additionally, we used generative AI tools for readability and clarity refinement of the manuscript. We have reviewed all AI-assisted work. We manually checked that the edits were limited to wording and presentation clarity and did not alter technical claims, results, citations, figures, or tables. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available 3D scene understanding benchmarks and does not involve new human-subject studies, private data collection, or the release of new datasets. The proposed method is intended to support research on 3D scene representation and reasoning. Since 3D vision-language models may produce inaccurate or incomplete scene interpretations, applications in embodied, robotic, or autonomous driving scenarios should incorporate appropriate safety constraints and human oversight. We comply with the usage terms of all datasets used in our experiments and report results solely for research evaluation.

REPRODUCIBILITY STATEMENT

We provide the full model architecture and training procedure in the main paper, including the hierarchical superpoint construction, the instruction-aware superpoint aggregation module, and the embedded superpoint query module. The experimental setup specifies the datasets, evaluation metrics, backbone models, and training stages used in our experiments, with additional implementation details and ablation studies included in the appendix. All experiments are conducted on publicly available benchmarks, and the reported results can be reproduced using the described preprocessing, training, and evaluation protocols. We will release the code and configuration files upon acceptance to further facilitate reproducibility.

REFERENCES

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 *European conference on computer vision*, pp. 422–440. Springer, 2020.

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*, pp. 19129–19139, 2022.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

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*, pp. 202–221. Springer, 2020.

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.

Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, and Tao Chen. End-to-end 3d dense captioning with vote2cap-detr. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11124–11133, 2023.

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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26428–26438, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6, 2023.

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*, pp. 5828–5839, 2017.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023.

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.

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. *CoRR*, 2023a.

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. *Advances in Neural Information Processing Systems*, 37:113991–114017, 2024.

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*, 2023b.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023c.

Le Hui, Jia Yuan, Mingmei Cheng, Jin Xie, Xiaoya Zhang, and Jian Yang. Superpoint network for point cloud oversegmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5510–5519, 2021.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pp. 4867–4876, 2020.

Loic Landrieu and Mohamed Boussaha. Point cloud oversegmentation with graph-structured deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7440–7449, 2019.

Loic Landrieu and Guillaume Obozinski. Cut pursuit: Fast algorithms to learn piecewise constant functions on general weighted graphs. *SIAM Journal on Imaging Sciences*, 10(4):1724–1766, 2017.

Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4558–4567, 2018.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *Advances in Neural Information Processing Systems*, 36:75307–75337, 2023.

Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2906–2917, 2021.

R OpenAI. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1, 2023.

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*, pp. 815–824, 2023.

Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9277–9286, 2019.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. Nuscenes-qa: A multimodal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 4542–4550, 2024.

Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

Damien Robert, Hugo Raguet, and Loic Landrieu. Efficient 3d semantic segmentation with superpoint transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17195–17204, 2023.

Damien Robert, Hugo Raguet, and Loic Landrieu. Scalable 3d panoptic segmentation as superpoint graph clustering. In *2024 International Conference on 3D Vision (3DV)*, pp. 179–189. IEEE, 2024.

Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. *arXiv preprint arXiv:2210.03105*, 2022.

Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15768–15780, 2025.

Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2393–2401, 2023.

Penglei Sun, Yaoxian Song, Xiang Liu, Xiaofei Yang, Qiang Wang, Tiefeng Li, Yang Yang, and Xiaowen Chu. 3d question answering for city scene understanding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2156–2165, 2024.

Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6617–6626, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962*, 2019.

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*, pp. 4566–4575, 2015.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023.

Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. Occllama: An occupancy-language-action generative world model for autonomous driving. *arXiv preprint arXiv:2409.03272*, 2024.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pp. 131–147. Springer, 2024.

Senqiao Yang, Jiaming Liu, Renrui Zhang, Mingjie Pan, Ziyu Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Hongsheng Li, Yandong Guo, et al. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9247–9255, 2025.

Linli Yao, Lei Li, Shuhuai Ren, Lean Wang, Yuanxin Liu, Xu Sun, and Lu Hou. Deco: Decoupling token compression from semantic abstraction in multimodal large language models. *arXiv preprint arXiv:2405.20985*, 2024.

Hanxun Yu, Wentong Li, Song Wang, Junbo Chen, and Jianke Zhu. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14147–14157, 2025.

Zhihao Yuan, Yibo Peng, Jinke Ren, Yinghong Liao, Yatong Han, Chun-Mei Feng, Hengshuang Zhao, Guanbin Li, Shuguang Cui, and Zhen Li. Empowering large language models with 3d situation awareness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19435–19445, 2025.

Bo Zhang, Jiakang Yuan, Botian Shi, Tao Chen, Yikang Li, and Yu Qiao. Uni3d: A unified baseline for multi-dataset 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9253–9262, 2023a.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zihui Zhang, Bo Yang, Bing Wang, and Bo Li. Growsp: Unsupervised semantic segmentation of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 17619–17629, 2023b.

Youjun Zhao, Jiaying Lin, Shuquan Ye, Qianshi Pang, and Rynson WH Lau. Openscan: A benchmark for generalized open-vocabulary 3d scene understanding. *arXiv preprint arXiv:2408.11030*, 2024.

Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

Hongyan Zhi, Peihao Chen, Junyan Li, Shuailei Ma, Xinyu Sun, Tianhang Xiang, Yinjie Lei, Mingkui Tan, and Chuang Gan. Lscenellm: Enhancing large 3d scene understanding using adaptive visual preferences. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3761–3771, 2025.

A Appendix

A.1 Implementation Details

We adopt a two-stage training paradigm using the AdamW Loshchilov & Hutter (2017) optimizer with a cosine annealing learning rate scheduler. For HiSuper-VLM-S with OPT-1.3B Zhang et al. (2022), the learning rate decays from 1e-4 to 1e-6 with a batch size of 24 during the pre-alignment stage (1 epoch), and from 2e-5 to 1e-6 with a batch size of 16 during the joint training stage (5 epochs). For HiSuper-VLM with LLaMA2-7B Touvron et al. (2023), the learning rate decays from 1e-4 to 1e-6 with a batch size of 8 and 3 gradient accumulation steps during the pre-alignment stage (1 epoch), and from 1e-5 to 1e-6 with a batch size of 6 and 4 accumulation steps during the joint training stage (3 epochs). We employ full-parameter fine-tuning for LLM-backbone. All experiments are conducted on 8 NVIDIA A100 GPUs. All hyperparameters are kept fixed across different LLM backbones and datasets.

A.2 Ablation Study

A.2.1 Ablation on Components

We conduct comprehensive ablation studies using our HiSuper-VLM-S to evaluate the effectiveness of the proposed ISAM and ESQM, and report results on XR-QA. As summarized in table 5, the performance shows consistent gains as more components are incorporated. The baseline (Row 1) is built by directly applying the SA module Qi et al. (2017), which randomly samples points and groups nearby features to construct vision tokens, without constructing hierarchical superpoints or introducing any instruction-aware feature aggregation and intra-level feature interaction. Introducing instruction-aware feature aggregation (Row 2) to construct hierarchical superpoints significantly improves performance, indicating that constructing hierarchical superpoints conditioned on the instruction helps the model focus on task-relevant regions and produce more semantically informative representations. Further incorporating intra-level feature interaction into ISAM (Row 3) brings additional improvement, suggesting that contextual feature interaction within each level facilitates the

capture of cross-object and cross-region relationships and contributes to better global semantic modeling. Finally, the complete model that integrates both ISAM and ESQM (Row 4) achieves the best performance by further injecting fine-grained information within the internal processing of the LLM. Collectively, these results demonstrate that both modules are indispensable for achieving effective and robust 3D vision-language understanding.

Table 5: Ablation study on components. Agg means instruction-aware feature aggregation and Int means intra-level feature interaction.

Method	CIDEr	ROUGE	METEOR
Baseline	109.59	37.25	18.99
+ ISAM-Agg	116.25	38.63	19.79
+ ISAM-Int	117.98	38.77	19.64
+ ESQM	**119.53**	**38.81**	**19.92**

A.2.2 Efficiency Analysis of Hierarchical Superpoints

To investigate the efficiency impact of hierarchical superpoints, we experimentally measure their GPU memory overhead. On average, hierarchical superpoints use 57.5 MB GPU memory (including 53.3 MB for initial superpoints), accounting for $<1\%$ of total memory during training (25967.0 MB) and inference (8611.8 MB), indicating a negligible impact on efficiency. Furthermore, the superpoint generation algorithm for constructing the initial superpoints, which is executed on the GPU, accounts for only 4.33% of the overall inference time, exhibiting a limited efficiency impact.

A.2.3 Ablation on Hierarchical Superpoint Levels

We investigate the impact of the number of hierarchical superpoint levels. As shown in table 6, the 3-level hierarchical structure achieves the best performance, striking an optimal balance between expressiveness and efficiency. When fewer levels are used (e.g., 2 levels), the fixed number of top-level superpoints requires a larger grouping size to cover the entire scene, which merges semantically inconsistent regions and weakens representation expressiveness, leading to inferior performance. Conversely, more levels (e.g., 4 levels) yield no further performance improvement, while introducing additional computational overhead.

Table 6: Ablation on the number of hierarchical superpoint levels.

Hierarchical Levels	CIDEr	ROUGE	METEOR
2	116.10	38.25	19.60
3	**119.53**	**38.81**	**19.92**
4	118.44	38.52	19.63

A.2.4 Effect of Cross-Attention with Learnable Tokens

In ISAM, since the number of text tokens is typically much larger than that of learnable tokens, cross-attention between text tokens and learnable tokens is adopted to reduce visual–text interaction complexity. In ESQM, as learnable tokens encode richer contextual information after multiple LLM layers, cross-attention is reintroduced for task-specific refinement, rather than repeated alignment. As shown in table 7, disabling this mechanism degrades performance, confirming its necessity.

Table 7: Impact of text cross-attention (c-a) in ESQM.

Setting	CIDEr	ROUGE	METEOR
ESQM w/ c-a	**119.53**	**38.81**	**19.92**
ESQM w/o c-a	118.22	38.69	19.75

A.2.5 THE NUMBER OF LEARNABLE TOKENS

We set the number of learnable tokens to 32 and conduct an ablation study to investigate the impact of the number of learnable tokens, reporting the corresponding results on XR-QA. As shown in Table 8, using 32 learnable tokens achieves the best performance. When the number of learnable tokens is insufficient (e.g., 16 tokens), the model struggles to capture sufficient task-relevant information, leading to limited representational capacity. In contrast, an excessive number (e.g., 48 tokens) of learnable tokens increases optimization difficulty and introduces unnecessary complexity, also leading to suboptimal results.

Table 8: Ablation study on the number of learnable tokens.

Learnable Token Num	CIDEr	ROUGE	METEOR
16	115.51	38.39	19.71
32	**119.53**	**38.81**	**19.92**
48	115.43	38.11	19.69

A.2.6 THE K FOR SUPERPOINT SELECTION

We set the k value for superpoint selection to 8 and analyze its effect on the performance of different tasks, with results summarized in Table 9. For the XR-QA task, whose core requirement is to accurately locate specific objects relevant to the question, a smaller k value (e.g., $k = 4$ or 8) can more effectively filter out superpoints directly related to the target object and eliminate interference from irrelevant superpoints, thus improving the accuracy of the QA task. In contrast, the XR-Caption task requires a comprehensive and detailed description of the scene, and a larger k value (e.g., $k = 12$ or 16) can incorporate more superpoints, covering richer object details and providing sufficient context for generating comprehensive scene descriptions. Overall, when k is set to 8, it achieves the optimal balance between the requirements of the two tasks, thereby enabling the model to achieve the best overall performance.

Table 9: Ablation study on the k value for superpoint selection.

Top-k Num	XR-QA			XR-Caption		
	C	R	M	C	R	M
4	**119.82**	**39.14**	19.83	4.05	22.67	16.90
8	119.53	38.81	**19.92**	5.33	23.75	17.22
12	118.74	38.88	19.86	5.56	**24.00**	**17.84**
16	117.97	39.03	19.78	**6.29**	23.91	17.68

A.2.7 THE INSERTION LAYER OF ESQM

We set the insertion layer of ESQM to 8, as the first few layers of the LLM tend to focus on the global information of the vision input Zhi et al. (2025). We further conduct an ablation study to analyze its effect. The results (Table 10) indicate that inserting ESQM at an intermediate LLM layer (e.g., layer 8) performs best, while placing it too early (e.g., layer 2) or too late (e.g., layer 20) leads to slightly degraded performance due to insufficient global context or limited downstream integration, respectively.

Table 10: Ablation study on the ESQM insertion layer.

Layer	CIDEr	ROUGE	METEOR
2	118.46	38.73	19.78
8	**119.53**	**38.81**	**19.92**
20	118.74	38.76	19.85

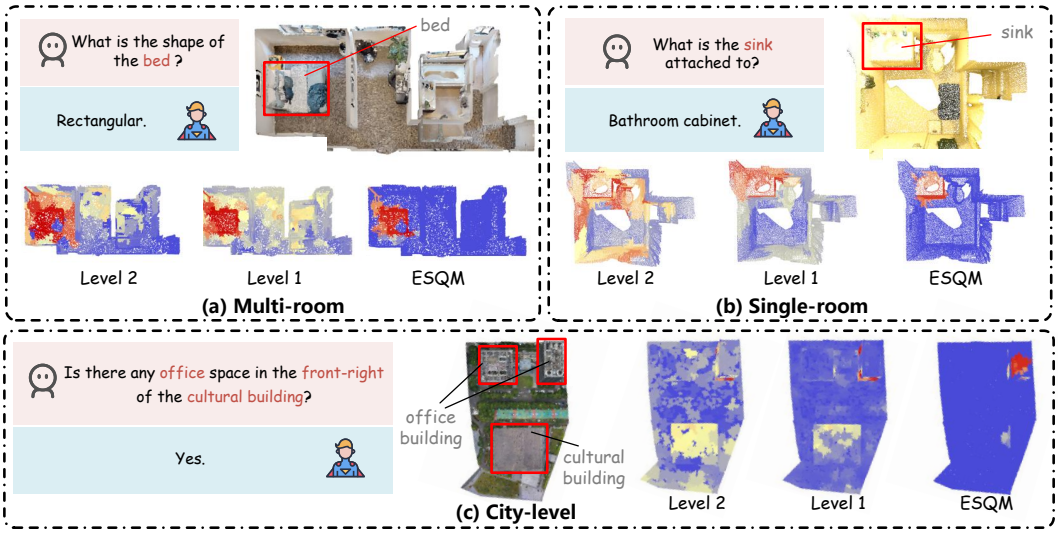


Figure 5: Visualization of attention maps for different level superpoints and ESQM in HiSuper-VLM.

A.3 Qualitative Analysis

For deeper insight into our method's inner workings, we provide a qualitative analysis by visualizing the areas that the model focuses on, including both the attention maps of the generated text of LLM to superpoints at different levels and the attention maps of ESQM. We showcase examples of 3D scenes with varying scales and complexity to demonstrate the model's robustness and scalability. As shown in fig. 5, the instruction-aware semantic modeling of the ISAM enables the model to focus on task-relevant regions, while the ESQM precisely supplements critical fine-grained details.

A.4 Case Study

Fig. 6 shows a failure case of ESQM. Most failures occur in scenes with similar nearby objects or complex spatial relations, where the model attends to relevant but spatially ambiguous regions.



Figure 6: Case study of ESQM.

A.5 Qualitative Results

We present the results and attention map visualizations of HiSuper-VLM on question answering and scene captioning in fig. 7, and on embodied planning and embodied QA in fig. 8. HiSuper-VLM effectively attends to task-relevant visual features and performs precise reasoning, demonstrating superior performance across different 3D vision-language tasks.

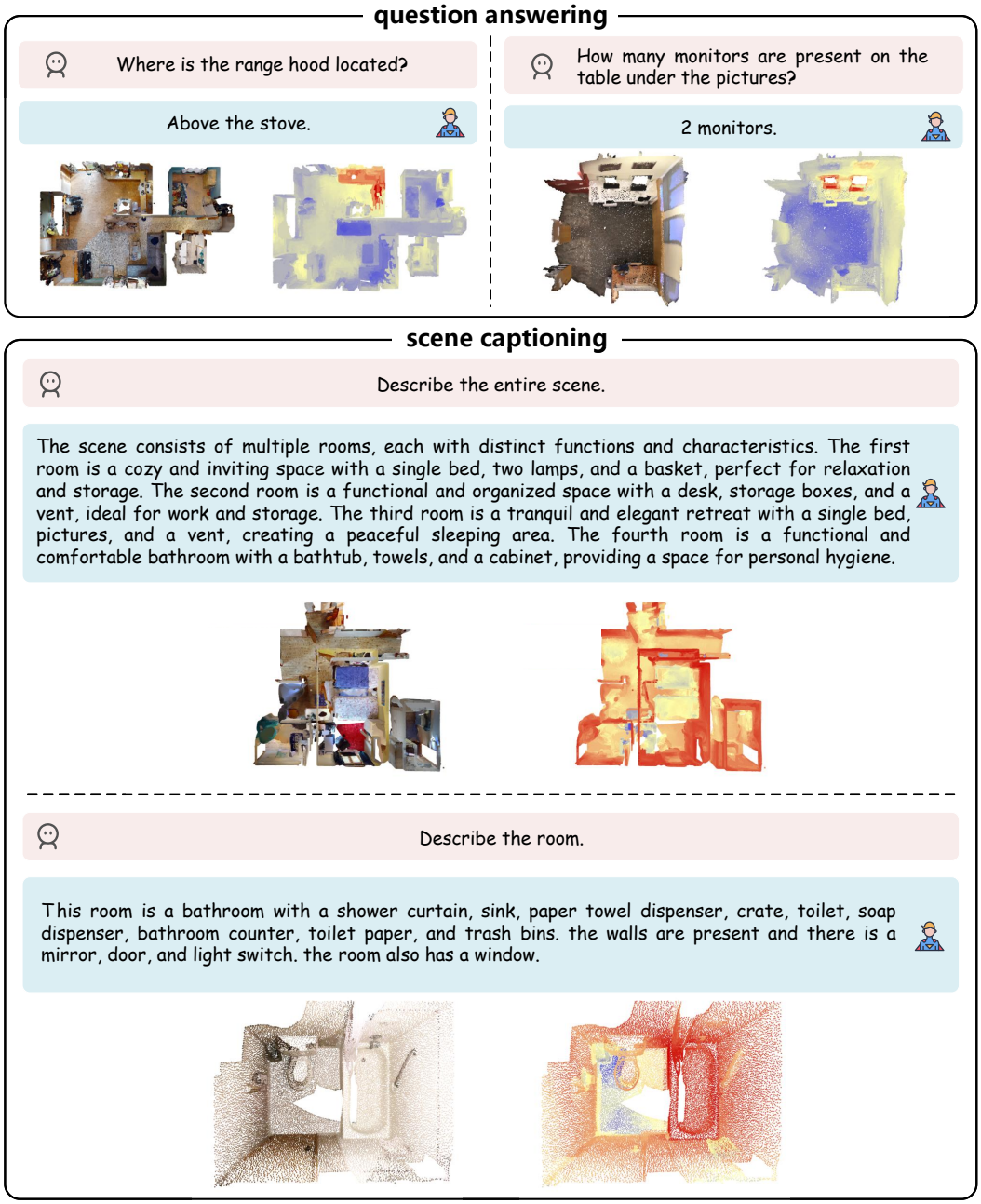


Figure 7: Results and visualization of attention maps for question answering and scene captioning.

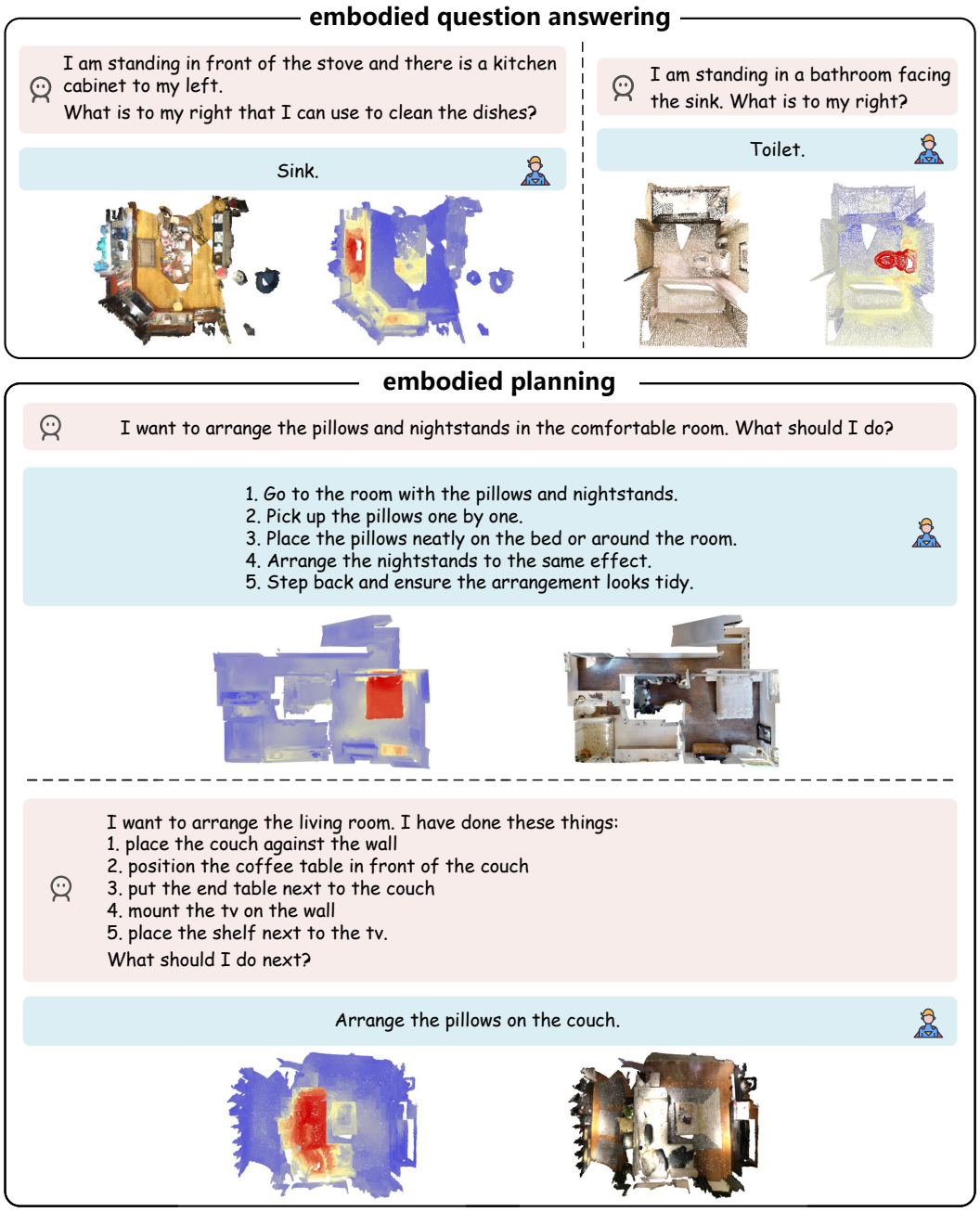


Figure 8: Results and visualization of attention maps for embodied question answering and embodied planning.