

SceneReasoner: Decoupled Spatial Tokenization for Indoor Large-Scale Understanding with LLMs

Bohang Shi¹, Shengjun Tang¹, Xiaoming Li¹, Weixi Wang¹, Linfu Xie¹, Baoding Zhou^{1,2}

¹ School of Architecture and Urban Planning, Research Institute for Smart Cities, Shenzhen University - shengjuntang@szu.edu.cn

² College of Civil and Transportation Engineering, Shenzhen University, Shenzhen, China

Abstract

Most existing 3D vision-language models focus on object-level or single-room understanding and perform poorly in large-scale, multi-room indoor environments where task-relevant objects constitute only a small fraction of the total point cloud. When multi-room point clouds are fed directly into an LLM, critical semantic signals are diluted by the vast amount of redundant background, making it difficult for the model to focus on truly relevant regions. We propose SceneReasoner, a decoupled spatial tokenisation framework that addresses this challenge through three core designs: (1) pre-tokenisation text-guided feature weighting that leverages the shared CLIP embedding space between OpenScene point features and text queries to amplify question-relevant point features before any spatial compression occurs; (2) 2D–3D feature fusion that integrates top-down 2D CLIP features with 3D sparse tokens, supplying the model with appearance semantics—such as texture, material, and room layout—absent from raw point clouds; and (3) layer-wise dense feature injection that inserts local dense features into the LLM attention mechanism layer by layer for fine-grained perception of key regions. We evaluate on the XR-Scene benchmark, which covers cross-room question answering and scene captioning over HM3D indoor environments with an average area of 132 m². SceneReasoner achieves the best CIDEr on XR-SceneCaption (+0.33 over LSceneLLM), the highest METEOR on XR-QA, and competitive ROUGE-L across all three tasks, demonstrating the effectiveness of task-guided spatial tokenisation for large-scale understanding.

Keywords: 3D Scene Understanding, Spatial Tokenisation, Large Language Model, Point Cloud, 2D–3D Feature Fusion

1. INTRODUCTION

3D scene understanding is a cornerstone of embodied AI and augmented/virtual reality (AR/VR) (Savva et al., 2019), playing a critical role in downstream tasks such as robotic manipulation, visual navigation, and long-horizon planning. With the rapid development of large language models (LLMs), researchers are increasingly exploring how to leverage their powerful reasoning and summarisation capabilities to enhance 3D point cloud understanding, giving rise to 3D Vision-Language Models (3D-VLMs). Early works such as 3D-LLM (Hong et al., 2023) and LL3DA (Chen et al., 2024) demonstrated that encoding point clouds directly into the LLM token space achieves strong performance on 3D question answering (3D-QA) and 3D dense captioning (3D-DC) on the ScanNet benchmark. More recently, Chat-Scene (Huang et al., 2024) introduced object-level identifiers to bridge 3D instance features with text tokens, while Leo (Huang et al., 2023) trained a generalist agent across multiple embodied tasks via a shared scene encoder and LLM decoder.

Despite these advances, most existing 3D-VLMs focus on **object-level** or **single-room** understanding, and scene-level comprehension of large, multi-room environments remains significantly underexplored. LSceneLLM (Zhi et al., 2025) made the first attempt to bridge this gap through an adaptive visual preference mechanism that identifies higher-information coarse tokens in large scenes. Nevertheless, two fundamental limitations persist at scale. **First**, existing methods uniformly group points via Farthest Point Sampling (FPS) and mean-pooling, treating all points equally regardless of task relevance. In large multi-room scenes, task-relevant objects occupy a far smaller fraction of the total point cloud than in single-room settings; this task-agnostic compression causes critical signals to be *averaged away* before reaching the LLM—a phenomenon we term *signal dilution*. **Second**, point clouds inherently lack the

rich appearance cues (texture, material) present in 2D imagery, which are often indispensable for accurate scene captioning and object disambiguation. These limitations pose a serious challenge for focusing on task-relevant visual information in large-scale environments.

We observe that humans facing complex visual input first selectively attend to key regions, then search for fine-grained details within them. Inspired by this cognitive mechanism, we argue that a 3D-VLM should amplify task-relevant signals *before* any spatial compression occurs, then extract fine-grained local details in the focused areas. This pre-tokenisation step ensures that the resulting sparse tokens are genuinely biased toward question-relevant semantics rather than averaging over diverse background points. Our key insight is that OpenScene (Peng et al., 2023) features already reside in the CLIP embedding space, enabling direct point-level semantic comparison with text queries at zero additional training cost. Furthermore, pure point cloud representations inherently lack appearance cues such as texture and material that are critical for scene understanding and object description. To compensate, we introduce top-down 2D CLIP features and fuse image-level semantics with the 3D representation.

Based on the above analysis, we propose **SceneReasoner**—a decoupled spatial tokenisation framework for LLM-based understanding of large-scale, multi-room indoor environments. SceneReasoner incorporates three core designs: (1) text-guided feature weighting before point cloud grouping to amplify question-relevant point features; (2) fusion of top-down 2D CLIP features with 3D sparse tokens to supply the model with appearance semantics beyond point clouds; and (3) layer-wise injection of fine-grained dense features into the LLM attention mechanism for dynamic fine-grained perception of focused regions. We evaluate on the XR-Scene benchmark (Zhi et al., 2025), which covers cross-room QA and scene cap-

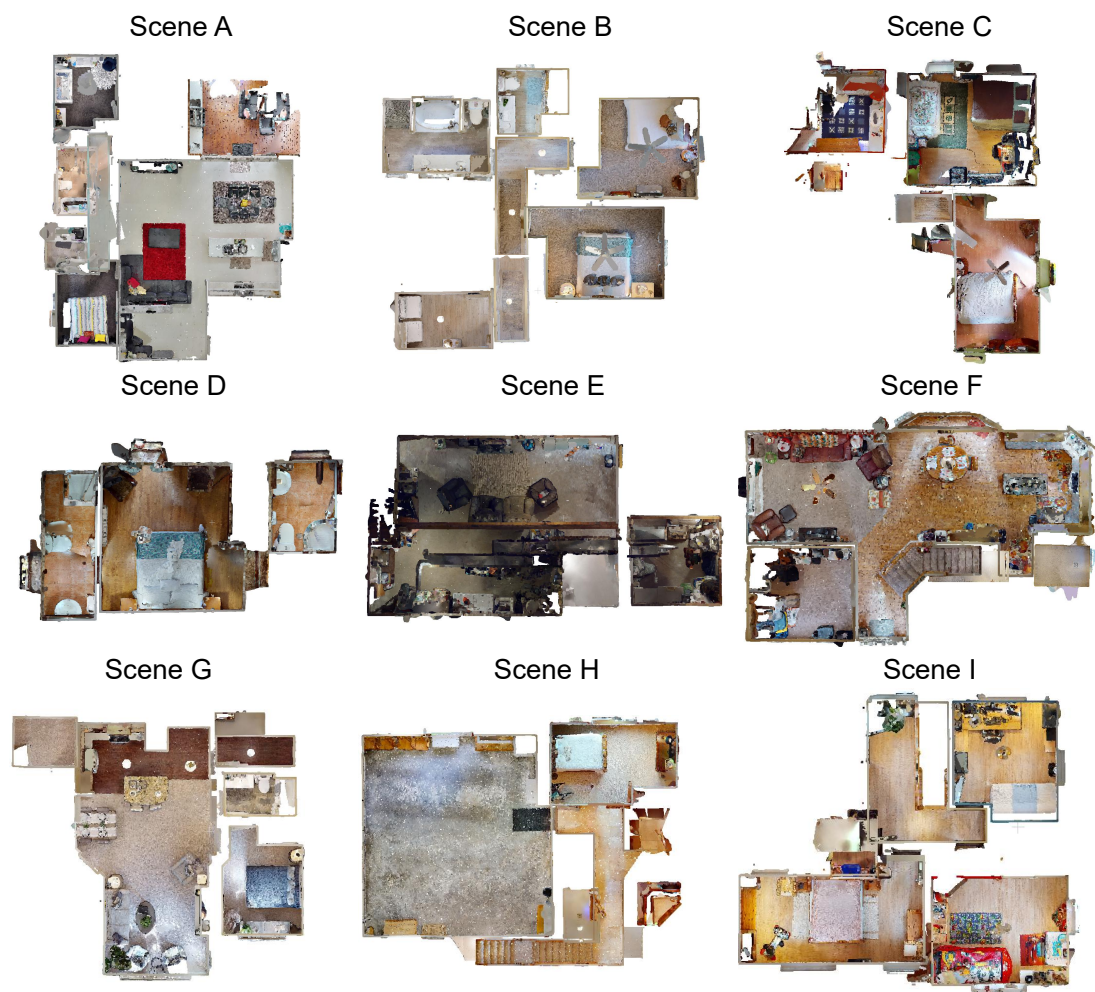


Figure 1. Nine representative HM3D indoor scenes from the XR-Scene benchmark, visualised as top-down coloured point clouds with the ceiling removed to reveal interior layout. Room partitions, furniture arrangements, corridors, and open areas are clearly distinguishable across diverse scenes. The large spatial footprint and complex multi-room structure of these scenes highlight the signal dilution challenge that motivates our decoupled spatial tokenisation approach.

tioning tasks over HM3D indoor environments with an average area of 132 m²—approximately four times larger than the 29 m² average of ScanQA. SceneReasoner achieves the highest ROUGE-L on both XR-QA (+17.89 over LSceneLLM) and XR-SceneCaption (+6.33), and the best CIDEr on XR-SceneCaption. Fig. 1 illustrates representative scenes from the benchmark, highlighting the large spatial scale and complex multi-room layouts that motivate our approach. Our main contributions are:

- We propose SceneReasoner, a decoupled spatial tokenisation framework that separates sparse global tokens from dense local tokens and introduces text-guided point-level feature weighting, driving the model toward coarse-to-fine scene perception guided by task semantics;
- We fuse top-down 2D semantic features with 3D point-level geometry to compensate for the limited appearance information in point clouds, enhancing LLM reasoning over large-scale scenes.

2. RELATED WORK

LLM-based 3D Scene Understanding. Injecting 3D spatial knowledge into large language models has attracted growing

interest. 3D-LLM (Hong et al., 2023) was among the first to encode point clouds via a 3D feature extractor and pass the resulting tokens directly to an LLM for question answering, captioning, and grounding. LL3DA (Chen et al., 2024) extended this to omni-3D tasks by accepting interactive visual prompts (clicks, bounding boxes) alongside the point cloud, enabling instruction-following in single-room ScanNet scenes. ChatScene (Huang et al., 2024) introduced object-level identifiers that link detected 3D instances to text tokens, improving referential understanding. Leo (Huang et al., 2023) built an embodied multi-modal generalist agent that unifies 3D captioning, QA, reasoning, navigation, and manipulation through a shared task interface and 3D vision-language-action instruction tuning. LSceneLLM (Zhi et al., 2025) specifically targeted large-scale multi-room scenes with an adaptive visual preference mechanism that selects the most informative coarse tokens based on LLM attention. However, the methods above either focus on single-room scenes or feed scene tokens to the LLM only after spatial compression, without applying task-relevant discriminative modulation to the raw point features before compression. In large-scale, multi-room environments, this task-agnostic compression strategy makes the sparse critical foreground signals highly susceptible to being overwhelmed by the vast background.

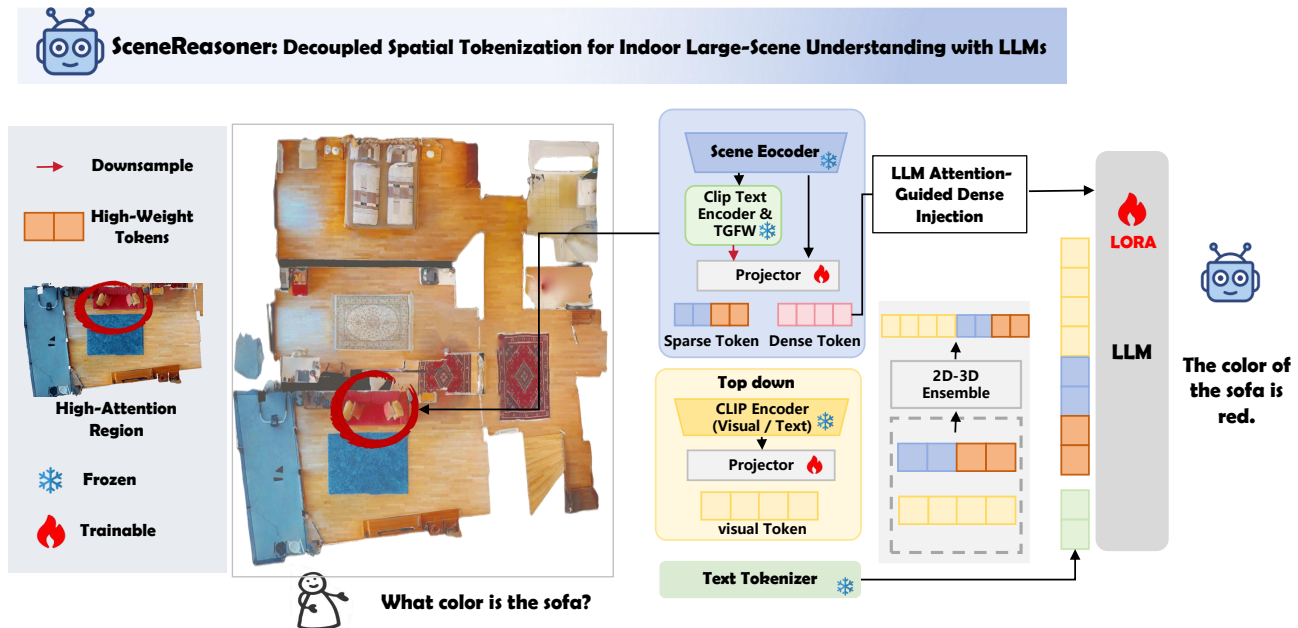


Figure 2. SceneReasoner pipeline for indoor large-scale scene understanding.

Open-Vocabulary 3D Representations. Transferring 2D vision-language knowledge to 3D point clouds has proven effective for annotation-free scene understanding. OpenScene (Peng et al., 2023) distills dense CLIP features onto each 3D point by fusing multi-view image predictions, producing per-point embeddings that reside in the same CLIP space as text queries. CLIP-FO3D (Zhang et al., 2023) similarly leverages dense pixel-aligned CLIP features for open-world 3D segmentation. While these methods demonstrate the value of CLIP-aligned 3D features, they have not been applied to large-scale indoor scenes, nor do they exploit this alignment for *task-adaptive* spatial tokenisation within an LLM-based reasoning pipeline.

2D–3D Multimodal Fusion. Pure point clouds lack rich appearance cues such as texture and material, which are critical for scene-level comprehension. Several works fuse multi-view images with 3D geometry to compensate (Zhang et al., 2023, Hong et al., 2023), but typically require dense per-frame images at inference time, incurring substantial memory and computational overhead. Our framework addresses this by rendering a single top-down view per room and extracting compact CLIP patch tokens offline, making 2D–3D fusion both computationally lightweight and compatible with pre-computation.

3. METHODOLOGY

SceneReasoner aims to enable LLM-based understanding of large-scale, multi-room indoor environments through a decoupled spatial tokenisation strategy. As illustrated in Fig. 2, the framework comprises three components. **Decoupled Spatial Tokenisation** (Sec. 3.1) builds a two-stage coarse-to-fine token representation: text-guided sparse tokens capture global scene structure while dense tokens inject fine-grained local details inside the LLM. **2D–3D Feature Fusion** (Sec. 3.2) enriches the 3D representation with top-down 2D semantic features, compensating for the limited appearance cues in point clouds. **LLM-Based Reasoning** (Sec. 3.3) integrates all modalities and generates free-form natural language responses.

3.1 Decoupled Spatial Tokenization

Motivation. Large-scale indoor scenes produce point clouds spanning multiple rooms, yet only a small fraction of the points is semantically related to any given question. Standard spatial tokenisation (Zhi et al., 2025) applies Farthest Point Sampling (FPS) to select centre points, groups each centre’s K -nearest neighbours via KNN, and mean-pools the group features into a single token. While this strategy is practical for single-room scenes, it becomes problematic at the multi-room scale: the vast majority of background points inevitably dilute the sparse foreground features during pooling, preventing the LLM from focusing on task-relevant semantics.

Pre-Tokenisation Text-Guided Feature Weighting (TGW). Our key insight is that OpenScene (Peng et al., 2023) features reside in the same CLIP embedding space as text, enabling direct point-level semantic comparison *before* any spatial compression. We encode the input question with the CLIP text encoder to obtain query embedding $\mathbf{q}$, and amplify each point’s feature proportionally to its relevance:

$$\hat{\mathbf{f}}_i = \mathbf{f}_i \cdot (1 + \cos(\mathbf{f}_i, \mathbf{q})), \quad (1)$$

where $\mathbf{f}_i$ is the OpenScene feature of the i -th point. This weighting modifies only the *feature values*; point *coordinates* remain unchanged, preserving FPS spatial uniformity. FPS then selects N_g geometrically uniform centre points in coordinate space, KNN retrieves their neighbours, and the **text-weighted features** are mean-pooled into sparse tokens biased toward task-relevant semantics. 3D positional encodings are added to provide explicit spatial awareness.

Dense Feature Complementation. Even task-biased sparse tokens summarise large spatial neighbourhoods and may lose fine-grained details of small or distant objects. Following LSceneLLM (Zhi et al., 2025), we leverage the LLM’s own attention weights to extract local features from a high-density point cloud and inject them into the LLM decoder in a layer-by-layer fashion. Together, the text-guided sparse tokens and

the attention-driven dense injection form a two-stage coarse-to-fine task-awareness pipeline: *soft point-level weighting before compression* followed by *hard token-level selection inside the LLM*.

3.2 2D–3D Feature Fusion

Point clouds encode precise 3D geometry but inherently lack appearance cues such as texture and material, which are often necessary for accurate scene captioning and object disambiguation. To compensate, we fuse top-down 2D semantic features with the 3D tokens, providing complementary visual information with minimal computational overhead.

Why Top-Down Orthographic Views? A natural question is why we choose a top-down orthographic projection rather than perspective images captured from arbitrary viewpoints. We identify two key advantages. **(1) Global coverage.** A single top-down view covers the entire floor plan of a room without occlusion gaps, whereas perspective images require multiple views to achieve similar spatial coverage, introducing redundancy and viewpoint selection overhead. **(2) Offline pre-computation.** Since the top-down view is fixed for a given scene, the CLIP patch features can be extracted once and stored, adding zero overhead at inference time compared to on-the-fly perspective rendering. These properties make the top-down view a particularly well-suited 2D complement to the 3D point cloud in large, multi-room settings.

2D Visual Feature Extraction. We employ a pre-trained CLIP visual encoder to extract patch-level features from the 2D images of the input scene. Analogously to the 3D branch, we apply text-guided token selection: the top- K_{2d} patch tokens most similar to the question embedding are amplified via residual weighting, while the remaining patches retain their original features (see Fig. 4 for visualisation).

Modality Fusion. Each room contributes a fixed number of 2D tokens. Since the 2D CLIP features and the 3D point-cloud features originate from different encoders with mismatched feature spaces and dimensionalities, the 2D tokens are first projected into the same LLM hidden space as the 3D sparse tokens via a three-layer MLP, achieving cross-modal feature alignment. The aligned 2D tokens are then concatenated with the 3D sparse tokens along the token axis to form a unified visual context for the LLM. This design allows the model to jointly attend to spatial geometry and image-level appearance semantics within a single attention mechanism.

3.3 LLM-Based Reasoning

Input Composition. The complete input to the LLM is composed of: a system prompt, the fused scene tokens (3D sparse + 2D), a visual prompt placeholder, and the task instruction.

Backbone and Adaptation. We adopt Llama-2-7B (Touvron et al., 2023) as the language backbone and apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to all linear projections in the attention and feed-forward layers, keeping the base LLM weights frozen. This parameter-efficient strategy reduces GPU memory consumption while preserving the LLM’s general language priors.

Training Objective. The model is trained end-to-end with a cross-entropy loss computed only over the response tokens:

$$\mathcal{L}(\theta) = - \sum_{i=1}^{|s|} \log P(s_{[i]} | \mathbf{V}, \mathcal{I}_t; \theta; s_{[1], \dots, i-1])}, \quad (2)$$

where s is the target response, $\mathbf{V}$ denotes the cross-modally fused scene tokens, and $\mathcal{I}_t$ is the textual instruction. By applying the loss exclusively over response tokens, the model is encouraged to generate complete answers conditioned on the full spatial context, rather than reconstructing the input instruction. This objective follows the standard training practice of LLM-based vision-language models and provides a direct supervision signal for free-form response generation. The differences among CIDEr, METEOR, and ROUGE-L mainly reflect their different sensitivities to lexical overlap, phrase ordering, and longest-subsequence consistency.

4. EXPERIMENTS

4.1 Experimental Setup

Datasets. We evaluate on the XR-Scene benchmark (Zhi et al., 2025), which covers cross-room understanding over large-scale HM3D indoor environments with an average scene area of 132 m²—approximately four times larger than the 29 m² average of ScanQA. XR-Scene provides three tasks evaluated here: XR-QA (cross-room question answering), XR-SceneCaption (room-level scene captioning), and XR-EmbodiedPlanning (embodied navigation instruction generation). All methods use the same LL3DA and XR-Scene training data for fair comparison.

Metrics. Following previous works on 3D vision-language tasks (Zhi et al., 2025), we adopt CIDEr, METEOR, and ROUGE-L to evaluate the quality of generated textual responses.

Implementation Details. SceneReasoner uses Llama-2-7B as the language backbone with LoRA ($r=8$, $\alpha=16$, dropout = 0.05) applied to all linear projections. We sample 40k points per scene as the 3D visual input and train for 6 epochs on 4 NVIDIA GPUs using AdamW with a learning rate of 10^{-4} and cosine annealing.

4.2 Comparison with State-of-the-Art

Table 1 summarises the main results. We compare against Chat-Scene (Huang et al., 2024), Leo (Huang et al., 2023), LL3DA (Chen et al., 2024), and LSceneLLM (Zhi et al., 2025) under both zero-shot and finetuning settings.

XR-QA Results. SceneReasoner achieves the highest METEOR of 19.36 (+0.06 over LSceneLLM) and a competitive ROUGE-L of 37.19, demonstrating strong lexical alignment with ground-truth answers. The QA CIDEr (112.17) remains competitive with all finetuning baselines.

XR-SceneCaption Results. SceneReasoner achieves the best CIDEr of 4.92 (+0.33 over LSceneLLM), indicating that the 2D–3D feature fusion provides discriminative appearance cues that improve the quality of room-level descriptions. This is consistent with the finding that pure point-cloud methods struggle to infer visual attributes such as texture and material when generating scene captions.

XR-EmbodiedPlanning Results. SceneReasoner achieves the highest ROUGE-L of 37.18 (+0.22 over LSceneLLM) on the embodied planning task, while remaining competitive on CIDEr (61.75 vs. 63.08) and METEOR (22.62 vs. 22.97). The consistent ROUGE-L gains across all three tasks confirm that



Figure 3. Qualitative visualisation of SceneReasoner on the XR-Scene benchmark across three tasks. *Top row*: XR-QA examples where red boxes mark the target objects referenced in the question. *Middle row*: XR-EmbodiedPlanning examples showing top-down views of multi-room scenes used for navigation instruction generation. *Bottom row*: XR-SceneCaption examples illustrating the floor-plan layouts from which room-level descriptions are generated.

Table 1. 3D large-scene understanding results on the XR-Scene benchmark. All methods use XR-Scene data for training. * does not identify question-related objects for the model. # requires images and point clouds as input. Best finetuning results are **bolded**.

Methods	XR-QA			XR-SceneCaption			XR-EmbodiedPlanning		
	CIDEr	MET.	ROU.	CIDEr	MET.	ROU.	CIDEr	MET.	ROU.
<i>Zero-Shot</i>									
Chat-Scene#	69.55	10.06	26.63	0.01	5.94	1.52	32.64	20.71	10.26
Leo*	55.40	6.96	22.71	0.02	1.92	2.92	9.74	16.84	6.88
LL3DA	24.78	5.31	12.66	0.12	8.71	5.14	7.02	15.21	7.17
<i>Finetuning</i>									
Chat-Scene#	114.10	14.32	35.93	3.58	17.49	11.59	46.18	22.34	36.71
Leo*	112.09	14.02	35.47	2.42	15.96	10.25	39.45	18.99	33.31
LL3DA	112.80	18.68	36.94	3.22	20.95	13.49	35.96	15.74	31.50
LSceneLLM	117.21	19.30	38.18	4.59	23.43	16.16	63.08	22.97	36.96
SceneReasoner	112.17	19.36	37.19	4.92	15.36	22.49	61.75	22.62	37.18

SceneReasoner generates responses with stronger lexical alignment to ground truth, a quality that is particularly valuable for navigation instruction generation where sequential spatial references must be preserved.

4.3 Ablation Study

To validate the contribution of text-guided feature weighting, we conduct an ablation on XR-QA by removing the pre-tokenisation CLIP similarity weighting (Eq. 1) while keeping all other components unchanged.

Table 2. Ablation study on text-guided feature weighting (XR-QA).

Variant	CIDEr	MET.	ROU.	Avg.
w/o TGW	112.34	18.64	37.07	56.02
SceneReasoner	112.17	19.36	37.19	56.24

As shown in Table 2, removing text-guided weighting leads to a drop of 0.72 in METEOR and 0.12 in ROUGE-L, with the average score decreasing from 56.24 to 56.02, confirming that amplifying task-relevant point features before spatial compression improves the quality of generated answers.

4.4 Qualitative Analysis

Fig. 3 presents qualitative results across all three XR-Scene tasks. For XR-QA (top row), the red boxes mark the target objects referenced in the question, illustrating the regions that the model must attend to within large, multi-room point clouds—even when these objects occupy only a small fraction of the full scene. For XR-EmbodiedPlanning (middle row) and XR-SceneCaption (bottom row), the top-down views reveal the rich structural and appearance information—room layout, furniture arrangement, surface material—that complements the geometric 3D tokens and enables the model to generate spatially coherent navigation instructions and descriptive scene captions.

Fig. 4 further visualises the top-down CLIP features extracted for representative scenes. The feature maps reveal that the 2D tokens encode discriminative appearance information—such as furniture layout, surface texture, and room boundaries—that is absent from the raw point cloud, confirming the complementary value of the 2D–3D fusion. Together, these examples confirm that the two core mechanisms of SceneReasoner—text-guided spatial focus and 2D–3D appearance fusion—work in concert to support fine-grained understanding of large-scale indoor environments.

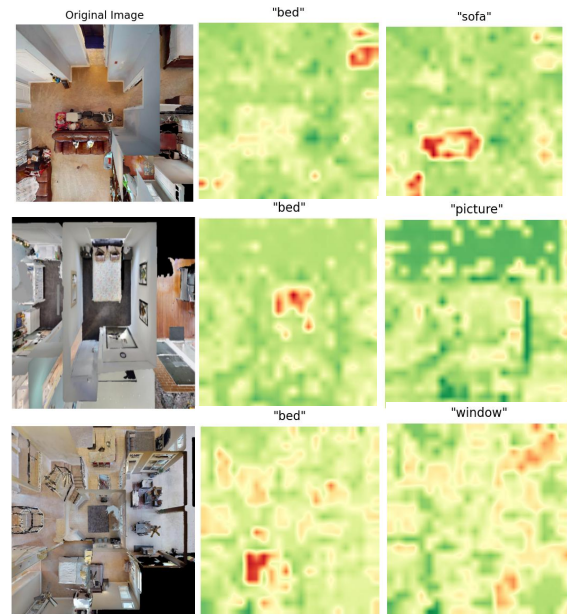


Figure 4. Visualisation of top-down CLIP features for representative scenes, showing the appearance semantics captured by the 2D tokens.

5. CONCLUSION

In this paper, we presented SceneReasoner, a decoupled spatial tokenisation framework that enables LLM-based understanding of large-scale, multi-room indoor environments. By combining pre-tokenisation text-guided feature weighting, task-relevant dense feature injection, and top-down 2D–3D feature fusion, SceneReasoner achieves the best CIDEr on XR-SceneCaption, the highest METEOR on XR-QA, and competitive ROUGE-L across all tasks on the XR-Scene benchmark. Our results highlight that mitigating semantic signal dilution at the *point level*—before spatial compression—is a critical step for large-scene understanding.

Despite these contributions, several limitations remain: (1) the text-guided weighting relies on the alignment quality between OpenScene features and CLIP text embeddings; when questions are vaguely phrased or involve concepts outside the CLIP vocabulary, the weighting effectiveness may degrade; (2) although the top-down orthographic projection achieves efficient global coverage, it inevitably loses vertical occlusion and stacking information (e.g. objects on multi-level shelves), limiting the per-

ception of depth-wise 3D details; (3) the current evaluation is restricted to indoor scenes, and the generalisation of the method to outdoor large-scale environments remains to be validated. Future work will extend the evaluation to outdoor large-scene benchmarks and explore stronger language model backbones together with more flexible multi-view 2D feature fusion strategies to further improve the generality of our approach.

References

- Chen, S., Chen, X., Zhang, C., Li, M., Yu, G., Fei, H., Zhu, H., Fan, J., Chen, T., 2024. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26428–26438.
- Hong, Y., Zhen, H., Chen, P., Zheng, S., Du, Y., Chen, Z., Gan, C., 2023. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36, 20482–20494.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. et al., 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2), 3.
- Huang, H., Chen, Y., Wang, Z., Huang, R., Xu, R., Wang, T., Liu, L., Cheng, X., Zhao, Y., Pang, J. et al., 2024. Chat-scene: Bridging 3d scene and large language models with object identifiers. *Advances in Neural Information Processing Systems*, 37, 113991–114017.
- Huang, J., Yong, S., Ma, X., Linghu, X., Li, P., Wang, Y., Li, Q., Zhu, S.-C., Jia, B., Huang, S., 2023. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*.
- Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T. et al., 2023. Openscene: 3d scene understanding with open vocabularies. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 815–824.
- Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J. et al., 2019. Habitat: A platform for embodied ai research. *Proceedings of the IEEE/CVF international conference on computer vision*, 9339–9347.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. et al., 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.
- Zhang, J., Dong, R., Ma, K., 2023. Clip-fo3d: Learning free open-world 3d scene representations from 2d dense clip. *Proceedings of the IEEE/CVF international conference on computer vision*, 2048–2059.
- Zhi, H., Chen, P., Li, J., Ma, S., Sun, X., Xiang, T., Lei, Y., Tan, M., Gan, C., 2025. Lscenellm: Enhancing large 3d scene understanding using adaptive visual preferences. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3761–3771.