# Appendix

## A. More evaluation metrics

We benchmark models using traditional language metrics, as done in previous work [8, 15, 18], for quick and automated evaluations. However, these metrics often fall short in contextual understanding compared to GPT-4 and human evaluators, which are better at assessing the quality and relevance of responses in longer sequences, such as planning. Advised by the reviewers, we supplement our evaluation with GPT-4 and human assessments. We prompt GPT-4 to compare model outputs with ground truth answers, provide a rating with an explanation, and normalize the score to a percentage, similar to human scoring. We also show the variance of 5 experiments with different random seeds.

In Tab. 6, both GPT-4 scores (all episodes) and human evaluations (100 episodes from each dataset) exhibit a notable correlation with traditional language metrics, further validating the effectiveness of our approach and our method tend to produce consistent responses across multiple tests.

| Methods | XR-QA | | | XR-EmbodiedPlanning | | |
|---|---|---|---|---|---|---|
| | CIDEr | GPT4-score | Human-score | CIDEr | GPT4-score | Human-score |
| Chat-Scene$^{\#}$ | $114.10 \pm 0.52$ | $39.96 \pm 0.15$ | 39.85 | $46.18 \pm 0.95$ | $45.88 \pm 0.52$ | 56.48 |
| Leo$^{*}$ | $112.09 \pm 0.36$ | $40.10 \pm 0.31$ | 36.85 | $39.45 \pm 0.81$ | $46.11 \pm 0.46$ | 54.76 |
| Ll3da | $112.80 \pm 0.73$ | $41.81 \pm 0.24$ | 36.08 | $35.96 \pm 1.23$ | $43.78 \pm 0.23$ | 60.71 |
| **LSceneLLM(Ours)** | **117.21** $\pm 0.31$ | **44.79** $\pm 0.18$ | **42.15** | **63.08** $\pm 0.76$ | **53.88** $\pm 0.24$ | **69.04** |

Table 6. 3D large scene understanding results.

## B. More Details on Generation of XR-Scene

**Generation of Cross-Room Scenes** HM3D [31] contains several cross-room, multi-floor 3D scenes. In SceneVerse [20], annotations are generated for each room in the HM3D scenes, including object properties and spatial relationships with surrounding objects. As shown in Fig. 4(a). For a given scene, we leverage the ground-true central positions of each room in HM3D. We randomly sample one room and calculate the Euclidean distances between other rooms, the nearest $N$ rooms are selected to form a cross-room scene.

**XR-QA Generation** For each cross-room scene containing $N$ rooms, we retrieve object annotations from SceneVerse [20] for these $N$ rooms and **filter out objects that appear exactly once** in the scene to ensure uniqueness corresponding to the question. For each annotated object, we use GPT-4 to generate two types of questions: object properties and spatial relationships with surrounding objects based on the annotations.

**XR-Planning and XR-EmbodiedPlanning Generation** The embodied planning task requires the model to understand the objects in the scene and their specific locations. Given a high-level task, the model needs to use the objects in the scene to generate a series of subtasks. In contrast to single-room scenes, embodied planning in cross-room scenes is more complex for the model, as it needs to understand the relationships between objects and the rooms, not just the relationships between the objects themselves.

Table 7. Ablation study of XR-Scene dataset on OpenEQA.

| Training Data | CIDEr | GPT-score |
|---|---|---|
| Single-Room Scene Data | 29.44 | 35.45 |
| XR-Scene | **41.89** | **40.35** |

The scene captioning task requires the model to provide a general description of the current scene, including the relationships between objects and their attributes. In larger scenes, scene captioning demands a stronger spatial understanding of the model. The model not only needs to perceive the positional relationships between objects but also pay attention to the areas to which the objects belong. Our tasks will include generating captions for the entire large scene as well as requiring the model to caption only a specific room. Furthermore, the model needs to infer room attributes based on the objects present, making scene captioning in cross-scene Scenes more challenging than in single-room Scenes.

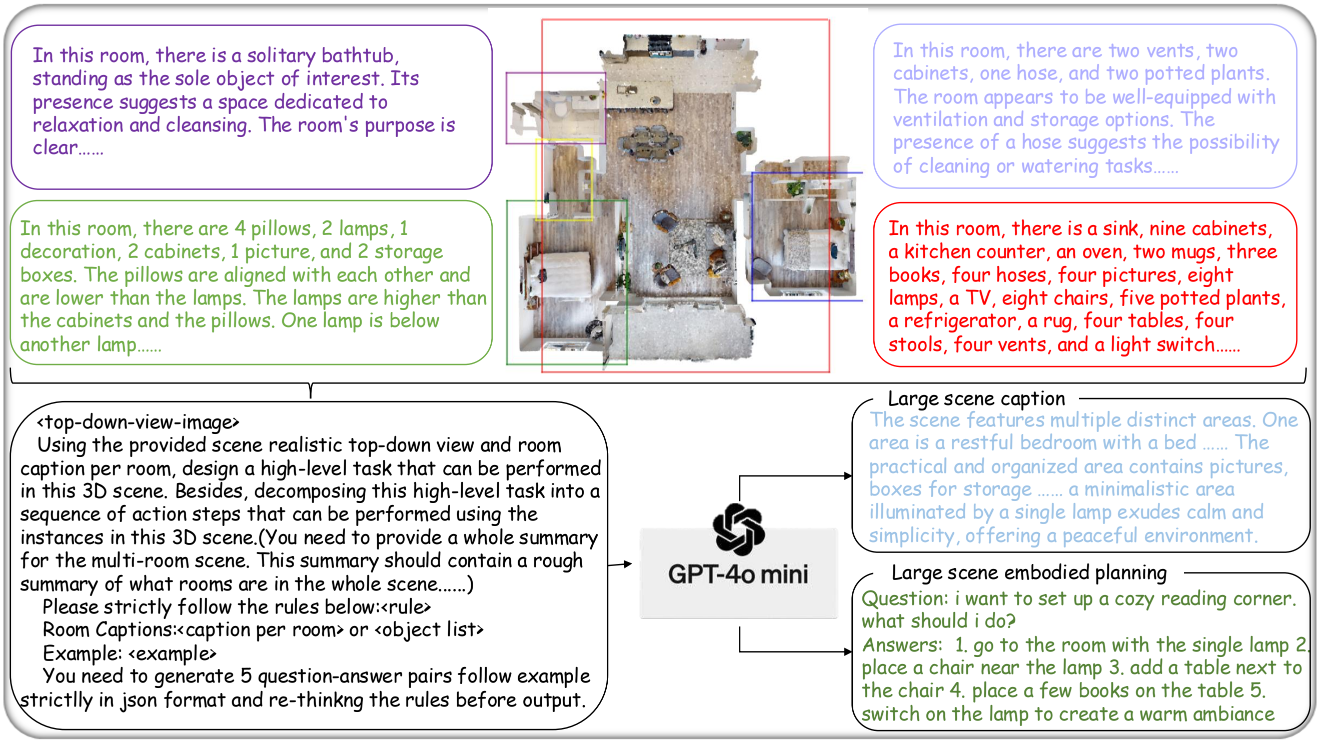
We generate the top-down view of the cross-room scene and use bounding boxes to specify that a certain annotation corresponds to a specific room. Follow Leo [18], We use prompt engineering to guide GPT-4o in understanding the scene and generating scene captions and QA pairs for embodied planning. Additionally, we provide the model with a real RGB-rendered top-down view of the scene to further reduce model hallucinations, as shown in Fig. 6.

## C. Real-world applications and long-term benefits of XR-Scene

Understanding large scenes is a critical ability for many real-world applications, such as robot navigation and AI glasses. For example, humans would wear AI glasses and go across rooms, requiring the AI system to understand larger scenes to assist. Our proposed XR-Scene, which covers a large space, helps to develop such an AI ability. To validate this, we trained two LSceneLLMs, one on single-room data and the other on XR-Scene data, and evaluated them using the OpenEQA [27] benchmark for first-person video understanding, relevant to AI glasses applications. In Tab. 7, the model trained on XR-Scene effectively handles human movement in large open spaces, benefiting real-world VR, AR, and robotics applications.

## D. Ablation Study

**Selection Threshold of Attention Weight.** We also explored the threshold for the confidence of text tokens to



In this room, there is a solitary bathtub, standing as the sole object of interest. Its presence suggests a space dedicated to relaxation and cleansing. The room's purpose is clear......

In this room, there are 4 pillows, 2 lamps, 1 decoration, 2 cabinets, 1 picture, and 2 storage boxes. The pillows are aligned with each other and are lower than the lamps. The lamps are higher than the cabinets and the pillows. One lamp is below another lamp......

In this room, there are two vents, two cabinets, one hose, and two potted plants. The room appears to be well-equipped with ventilation and storage options. The presence of a hose suggests the possibility of cleaning or watering tasks......

In this room, there is a sink, nine cabinets, a kitchen counter, an oven, two mugs, three books, four hoses, four pictures, eight lamps, a TV, eight chairs, five potted plants, a refrigerator, a rug, four tables, four stools, four vents, and a light switch......

<top-down-view-image>
Using the provided scene realistic top-down view and room caption per room, design a high-level task that can be performed in this 3D scene. Besides, decomposing this high-level task into a sequence of action steps that can be performed using the instances in this 3D scene.(You need to provide a whole summary for the multi-room scene. This summary should contain a rough summary of what rooms are in the whole scene......)
Please strictly follow the rules below:<rule>
Room Captions:<caption per room> or <object list>
Example: <example>
You need to generate 5 question-answer pairs follow example strictlly in json format and re-thinkng the rules before output.

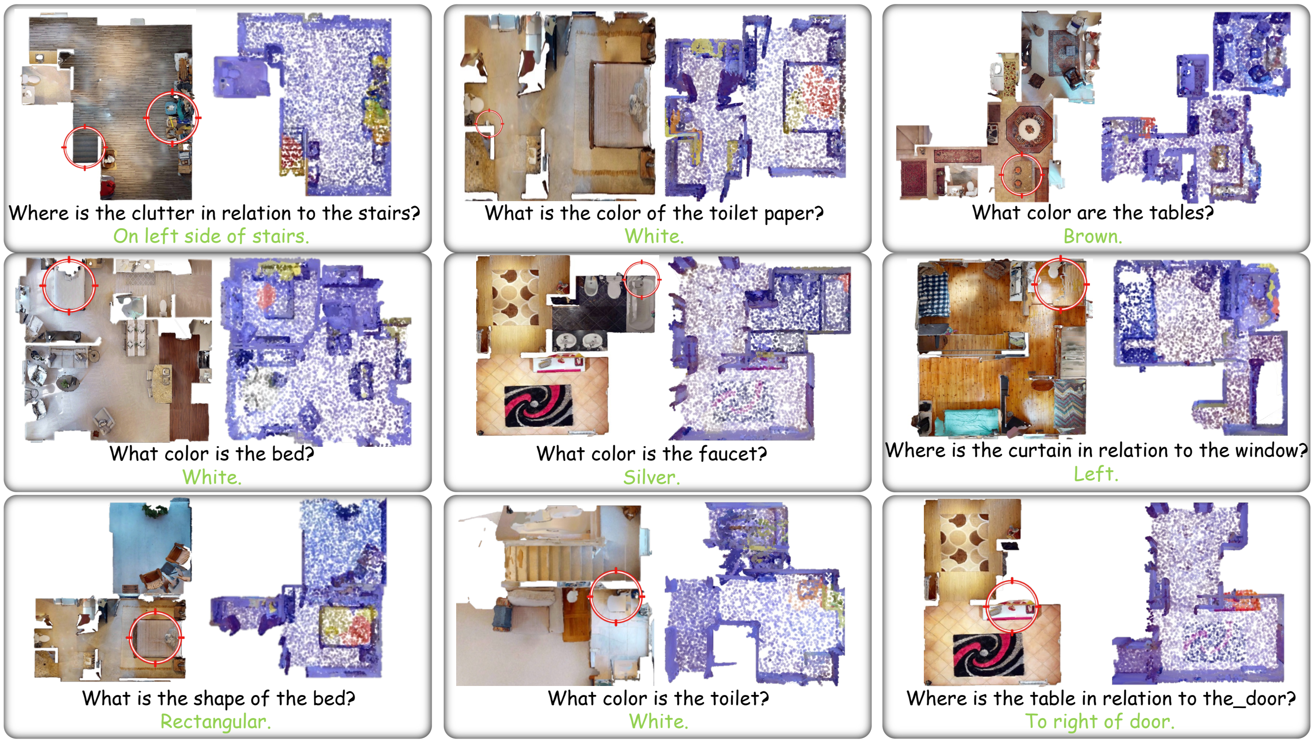
Large scene caption

The scene features multiple distinct areas. One area is a restful bedroom with a bed ...... The practical and organized area contains pictures, boxes for storage ...... a minimalistic area illuminated by a single lamp exudes calm and simplicity, offering a peaceful environment.

Large scene embodied planning

Question: i want to set up a cozy reading corner. what should i do?
Answers: 1. go to the room with the single lamp 2. place a chair near the lamp 3. add a table next to the chair 4. place a few books on the table 5. switch on the lamp to create a warm ambiance

Figure 6. **Generation pipeline of XR-SceneCaption and XR-EmbodiedPlanning**.



Where is the clutter in relation to the stairs?
On left side of stairs.

What is the color of the toilet paper?
White.

What color are the tables?
Brown.

What color is the bed?
White.

What color is the faucet?
Silver.

Where is the curtain in relation to the window?
Left.

What is the shape of the bed?
Rectangular.

What color is the toilet?
White.

Where is the table in relation to the_door?
To right of door.

Figure 7. More Attention Visualization of LSceneLLM.

Table 8. Ablation studies of selection threshold

| Threshold | Activate Token Ratio | ROUGE | METEOR | CIDEr |
|---|---|---|---|---|
| 64 | 40% - 50% | 37.68 | 19.07 | 114.69 |
| 96 | 10% - 20% | **38.18** | **19.30** | **117.21** |
| 127 | 3% - 5% | 37.89 | 19.26 | 115.92 |

Table 9. Ablation studies of the number of vision tokens

| Vision Token Num | Scene Magnifier Module | ROUGE | METEOR | CIDEr |
|---|---|---|---|---|
| 512 | ✗ | 37.27 | 18.80 | 112.89 |
| 128 | ✗ | 36.58 | 18.65 | 109.92 |
| 128 | ✓ | **38.18** | **19.30** | **117.21** |

Table 10. Ablation studies of dense token

| Dense Token Num | ROUGE | METEOR | CIDEr |
|---|---|---|---|
| 2 | 37.91 | 19.14 | 115.32 |
| 4 | **38.18** | **19.30** | **117.21** |
| 6 | 37.54 | 19.03 | 115.14 |

Table 11. Ablation studies of selection strategies

| Select Strategy | ROUGE | METEOR | CIDEr |
|---|---|---|---|
| Attention Map | **38.18** | **19.30** | **117.21** |
| Random | 37.64 | 19.18 | 115.66 |

Table 12. Start layer of scene magnifier module

| Start Layer | ROUGE | METEOR | CIDEr |
|---|---|---|---|
| 1 | 18.51 | 36.51 | 110.25 |
| 4 | 18.78 | 37.09 | 112.82 |
| 8 | **19.30** | **38.18** | **117.21** |

vision tokens in the attention map. We normalized the attention weight of a text token to all vision tokens to a range of 0-255. The experimental results show that the model performs best when the chosen threshold is 96, meaning 10%-20% of the vision tokens are selected to interact with the corresponding fine-grained scene features. If too many tokens are selected, the model cannot accurately focus on the local areas, while if too few tokens are selected, the fine-grained scene information provided is insufficient, offering limited help in understanding the scene, as shown in Tab. 8.

**Numbers of Dense Vision Token Interact With Sparse Vision Token.** This ablation experiment investigates the optimal number of dense vision tokens with which each sparse vision token should interact. We sample a certain number of point cloud features around the center point of the sparse vision token from the dense point cloud features and then aggregate them. As shown in Tab. 10, using 4 dense vision tokens to represent the fine-grained features of a local region provides the greatest benefit to the model.

**The number of Vision Tokens** We first explored whether sampling more visual information from the environment would improve the model's performance. As shown in Tab. 9, although using four times the number of vision tokens does lead to some performance improvement, the enhancement is not as significant as the improvement achieved by incorporating the LSceneLLM module, which validates the efficiency of our approach.

**Dense Vision Token Selection Strategy.** We conducted ablation experiments to verify that the attention map in the self-attention module reflects the visual information the model focuses on when answering questions. As shown in Tab. 11, the selection strategy based on attention weight outperforms the random selection strategy, demonstrating that the information about the regions that the model is currently focusing on aids in understanding the scene, while the random selection strategy provides little benefit to the model.

**Start Layer of Scene Magnifier Module.** We conduct an ablation study for $N_{SA}$ and visualize the attention map of different layers, as shown in Tab. 12. In Fig. 8, the attention map in the early layers is dispersed, with similar values across most regions. Applying a threshold at this stage may lead to randomly selected focus areas, introducing unpredictability in the self-attention input, thereby reducing the signal-to-noise ratio and hindering the model's learning ability.

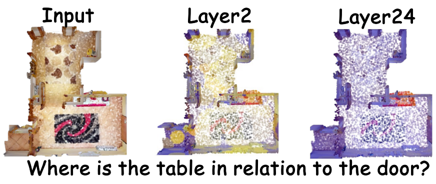


Figure 8. Attention visualization of different layers.

## E. More Scene Understanding Results on ScanNet

We also test our method on scene caption, embodied planning, and embodied qa, these datasets are sourced from the ScanNet part of 3D-LLM [15] and organized by Ll3da [8]. Embodied QA requires the model to answer questions from the perspective of an agent, considering the agent's position and orientation within the environment. All of these tasks demand the model to have a holistic understanding of the entire scene. As shown in Tab. 13, our method outperforms the current state-of-the-art approaches on most metrics, demonstrating that the proposed approach not only captures fine-grained details in the scene but also achieves an accurate overall understanding of the entire scene.

Table 13. More 3D scene understanding results. * means do not identify the question-related objects for the model.

| Method | Scene Caption | | | Embodied Planning | | | Embodied QA | | |
|---|---|---|---|---|---|---|---|---|---|
| | ROUGE | CIDEr | METEOR | ROUGE | CIDEr | METEOR | ROUGE | CIDEr | METEOR |
| Leo* [18] | 1.80 | 20.84 | 13.29 | 46.40 | 204.78 | 19.86 | 30.89 | 86.14 | 18.81 |
| Chat-Scene [16] | **3.67** | 21.05 | 12.60 | 40.03 | 210.86 | 20.71 | 34.23 | 99.01 | 18.48 |
| Ll3da [8] | 1.44 | **24.62** | 12.93 | 45.34 | 186.13 | 19.60 | 33.75 | 95.53 | 19.81 |
| **LSceneLLM(Ours)** | 3.07 | 21.88 | **14.79** | **47.05** | **214.63** | **21.05** | **36.00** | **104.98** | **21.26** |

## F. More Attention Visualization of LSceneLLM on XR-QA

We provide more attention map visualization results when LSceneLLM deals with different instructions on XR-QA. Experiment results show that our proposed method can accurately locate the task-relevant visual features using adaptive visual preferences from LLM.

## G. Computational Complexity Analysis

In Fig. 9, the performance of our method increases significantly with the increase of the flops. Compared with existing methods, our method achieves better performance with fewer flops.

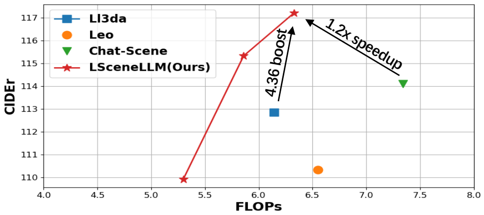


Figure 9. More Attention Visualization of LSceneLLM.