6. Appendix

In this supplementary material, we discuss things we left over in our main paper due to page constraints. We provide more background information about 3DGS and the derivative with respect to localization parameters in Sec. 6.1. We discussed more implementation details in Sec. 6.2 and provided the prompt we use for the Multimodal LLM and an example of the interaction in Sec. 6.5 to help reproduce our results. **The source code of this project will be made public soon.** We also include detailed qualitative results in Sec. 6.4.

6.1. Additional Background Information

For the completeness of our method, we also provide the key definition for the 3D Gaussian Splatting backbone [36] and 3D Gaussian SLAM [49]. In 3D Gaussian Splatting [36], the rendered pixel color is calculated by composing all 3D Gaussians projected in a tile.

$$\hat{C}(\mathrm{r}) = \sum_{i=1}^{N_s} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) \mathbf{c}_i \tag{14}$$

$$\alpha_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right) \left(1 - \exp(-\sigma_i \delta_i)\right) \tag{15}$$

$\delta_i = t_{i+1} - t_i$ represents the distance between adjacent samples, and N_s indicates the number of samples. $\mathbf{c}_i$ is the color of each 3D Gaussian given the current view direction $\mathbf{d}$ and σ_i is given by evaluating a 2D Gaussian with covariance Σ.

The Jacobian of the localization parameters are defined as:

$$\frac{\mathcal{D}\boldsymbol{\mu}_C}{\mathcal{D}\mathbf{x}} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{\mu}_C^{\times} \end{bmatrix} \text{ and } \frac{\mathcal{D}\mathbf{W}}{\mathcal{D}\mathbf{x}} = \begin{bmatrix} \mathbf{0} & -\mathbf{W}_{:,1}^{\times} \\ \mathbf{0} & -\mathbf{W}_{:,2}^{\times} \\ \mathbf{0} & -\mathbf{W}_{:,3}^{\times} \end{bmatrix}, \tag{16}$$

where $^{\times}$ denotes the skew symmetric matrix of a 3D vector, and $\mathbf{W}_{:,i}$ refers to the ith column of the matrix.

Unlike PSNR, the EIG can be computed without ground truth images, making it possible to perform view selection during exploration.

6.2. Implementation Details

The 2D occupancy map's resolution is 5cm. For each single frontier pixel on the 2D map, we add 200 3D Gaussians, which are uniformly distributed in the 3D cube above each frontier pixel. Other parameters like color, opacity, and scale are generated uniformly between 0 and 1. When there are frontiers on the 2D map, we choose the next frontier by querying LLM as stated in the main paper. When no frontier exists, we select the top 20% of Gaussians with the highest score. These Gaussians are grouped using DBSCAN [20]. The largest cluster is selected for candidate pose generation. Candidates are uniformly sampled in the range between 0.3m to 1m, facing towards the selected position. Only the poses in free space are kept for path-level selection. The importance factor η in Eq. 13 is set to 5 across all experiments.

We compute the Expected Information Gain (EIG) for each global candidate and use A* to plan a path to each of them. In order to prevent a twisted path, we consider locations 0.15m (3 pixels) away from the current robot position as neighbors and set the robot width to 3 pixels for collision check. However, the path planned by A* might have redundant waypoints, causing unnecessary turns for the robot. Therefore, we smooth the path by finding shortcuts. Specifically, for each waypoint w_i, if the path between waypoint w_{i+2} and w_i is collision-free, then we remove the intermediate waypoint w_{i+1} from the path. Finally, we use a greedy follower for motion planning. If the angle between the heading direction of the robot and the relative next waypoint is larger than 5°, then we turn left or right to decrease the angle. Otherwise, we choose the forward action to approach the next waypoint. In such a way, we get a sequence of actions $\{a_i\}_{i=1}^{T}$ for each path.

Given a sequence of actions $\{a_i\}_{i=1}^{T}$ for each path, we use forward dynamics to compute the future camera poses $\{c_i\}_{i=1}^{T}$. Initially, we use an intermediate variable $\mathbf{H}''_{\text{obs}} \triangleq \mathbf{H}''[\mathbf{w}^*]$ to help compute expected information gain along the path. For each camera pose x_i, we compute its pose Hessian $\mathbf{H}''_{\text{pose}}$ and the current model Hessian matrix $\mathbf{H}''_{\text{cur}} \triangleq \mathbf{H}''[\mathbf{y}|x_i, \mathbf{w}^*]$. $\mathbf{H}''_{\text{cur}}$ is then accumulated, and we update $\mathbf{H}''_{\text{obs}}$ to evaluate the remaining poses on the path. We select the path that minimizes the objective given by Eq. 13 for execution.

6.3. Scenes Used for Evaluation

Following previous literature [92], we use the following scenes for Gibson Dataset: `Greigsville`, `Denmark`, `Cantwell`, `Eudora`, `Pablo`, `Ribera`, `Swormville`, `Eastville`, `Elmira`. For HM3D we use the following scenes: `DBjEcHFg4oq`, `mscxX4KEBcB`, `QKGMrurUVbk`, `oPj9qMxrDEa`, `CETmJJqkhcK`.

6.4. Detailed Version of Qualitative Results

We provide larger versions of the qualitative rendering comparisons from the main paper, Fig. 8 shows the Gibson scenes and Fig. 9 shows the HM3D scenes.

We also present qualitative comparisons on testing views from the Gibson dataset in Fig. 10 and HM3D in Fig. 11.

6.5. Example of Using Multimodal LLM

We provide an example of our interaction with the multimodal LLM in Fig. 12 and additionally provide our full text prompt to the LLM below.

~~~

<system> You are an AI assistant that can analyze images and plan a long-term goal for the exploration task of a ground robot.
You will be given a bird-eye view image of a scene.
The goal is to plan a long-term exploration mission for a robot to traverse the area.
The robot's task is to explore the terrain efficiently, identifying important areas, potential obstacles, and unvisited areas.
Please analyze the image and select a long-term goal from the candidates for the robot to explore the area.
Empty space doesn't always mean they are unvisited regions, sometimes it's just outside the floor plan of this scene.
We are allowed to explore a total of <TOTAL_STEPS> steps and this is step <STEP_ID>.
Therefore, it's better to select a space that is close to the visited regions but still unvisited and not behind the walls.
The current location of the robot is marked with the blue star(*) marker.
The last frontier you selected is marked with a yellow diamond shape.
The visited path is painted as green lines in the image.
Note that you don't have to select the closest point to the robot, but the point that is most likely to be unvisited and important to explore.
As you can see, there are <NUM_FRAME> candidate points to select from.
They are numbered from 0 to <NUM_FRAME - 1> in red color.
If you find all the goals are not necessary to explore and we should instead focus on improving existing reconstruction, please give -1 in the `target` entry of the JSON.
Please provide a detailed exploration plan and select an exploration target with reasons in the JSON format as shown below.


```
{
"target": 2, "reason": "The target is located at an unvisited region of the image and seems to be an unvisited bedroom"
}
```
Do not cut off the JSON and generate the full JSON.
</system>

<user>: I have a bird-eye view image of a scene. The goal is to plan a long-term exploration mission for a robot to traverse the area. Please analyze the attached image and provide the exploration plan first and then an exploration target in the specified JSON format.
</user>
~~~

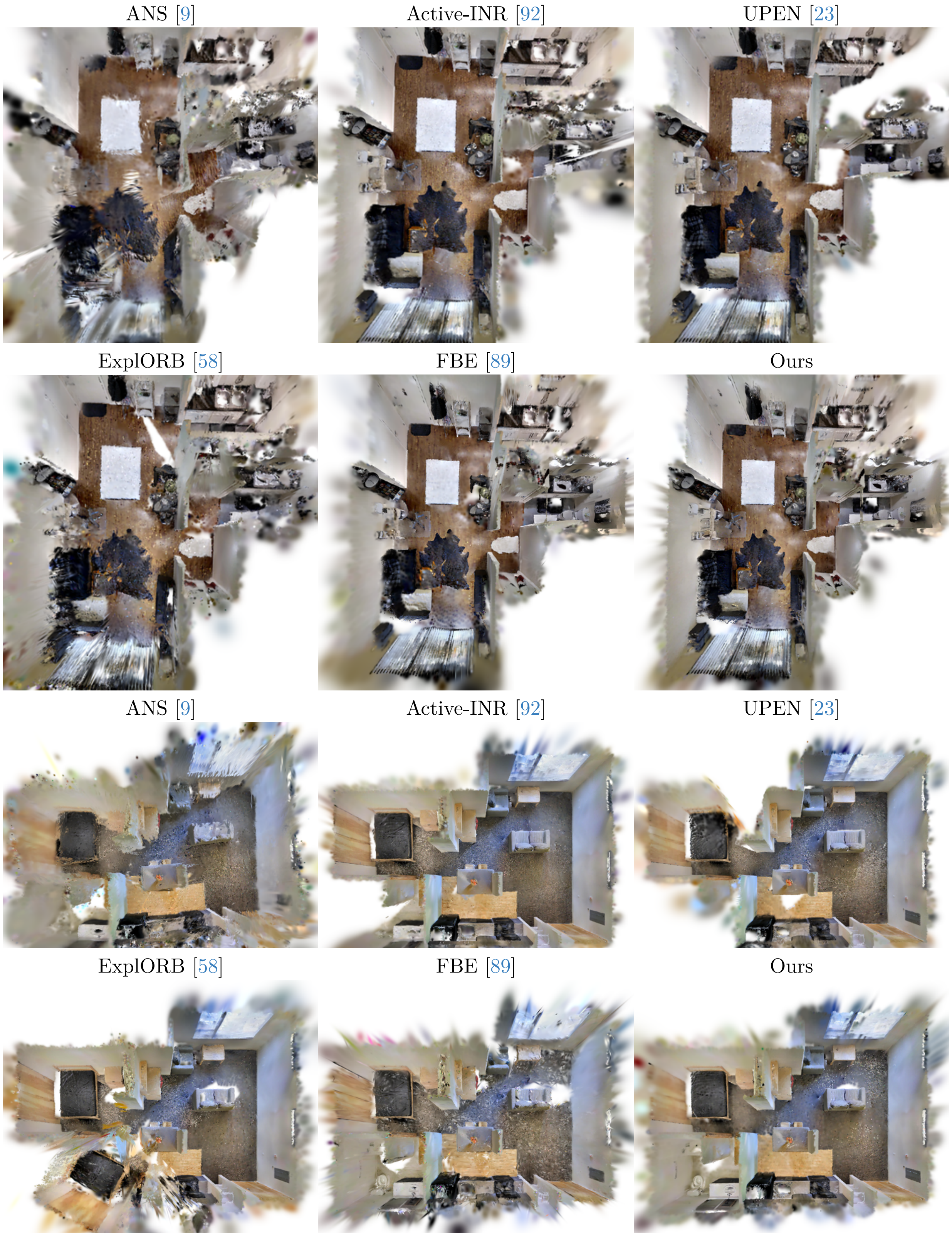

ANS [9]
Active-INR [92]
UPEN [23]
ExplORB [58]
FBE [89]
Ours
ANS [9]
Active-INR [92]
UPEN [23]
ExplORB [58]
FBE [89]
Ours

Figure 8. **Qualitative Comparison for Final Scene Reconstruction on Gibson Dataset** Greigsville (top) and Ribera (bottom) scenes. We provide top-down rendering for different methods. Note that UPEN and Active-INR use GT pose in this visualization.

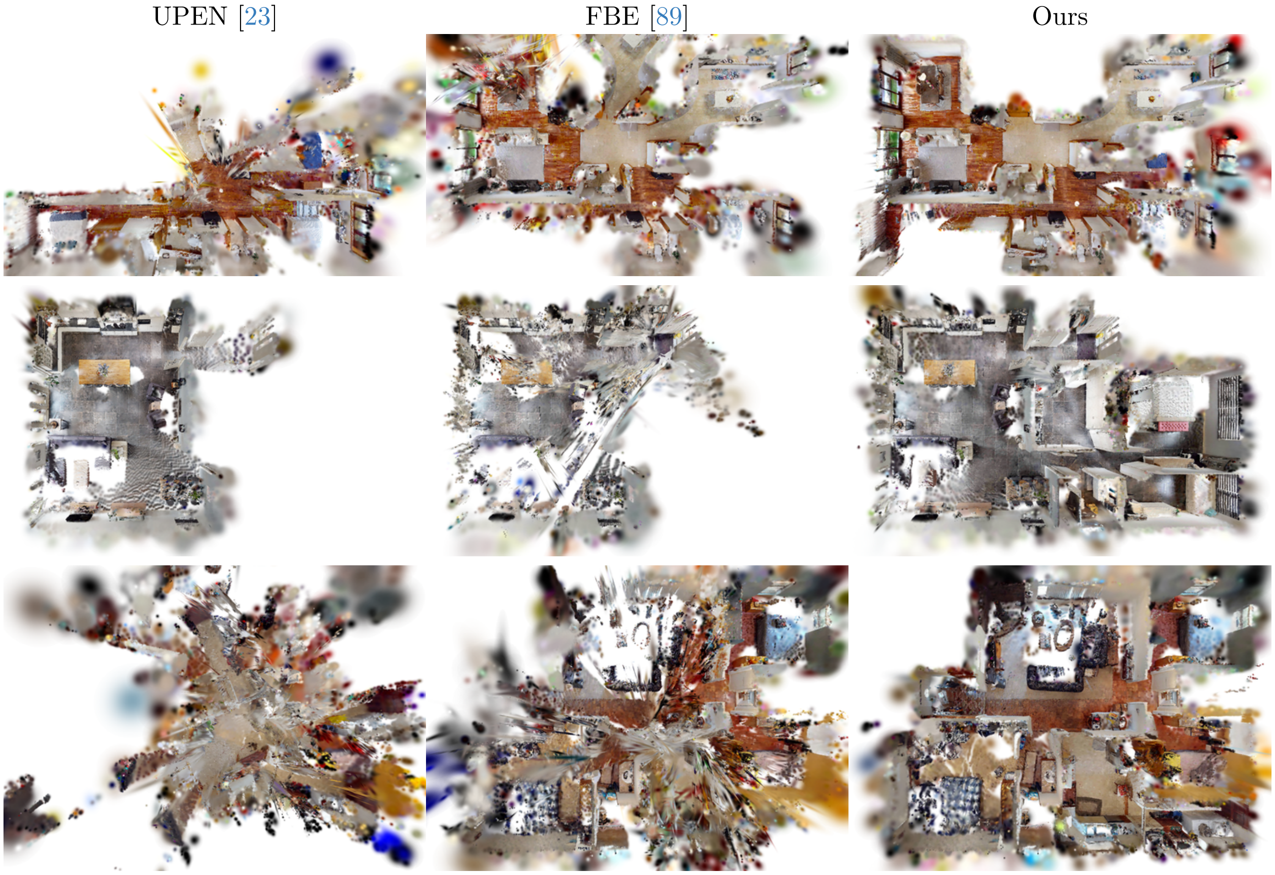


Figure 9. **Qualitative Comparison for Final Scene Reconstruction on Habitat-Matterport 3D Dataset** mscxX4KEBcB (top), oPj9qMxrDEa (middle) and QKGMrurUVbk (bottom) scenes. We provide top-down rendering for different methods.

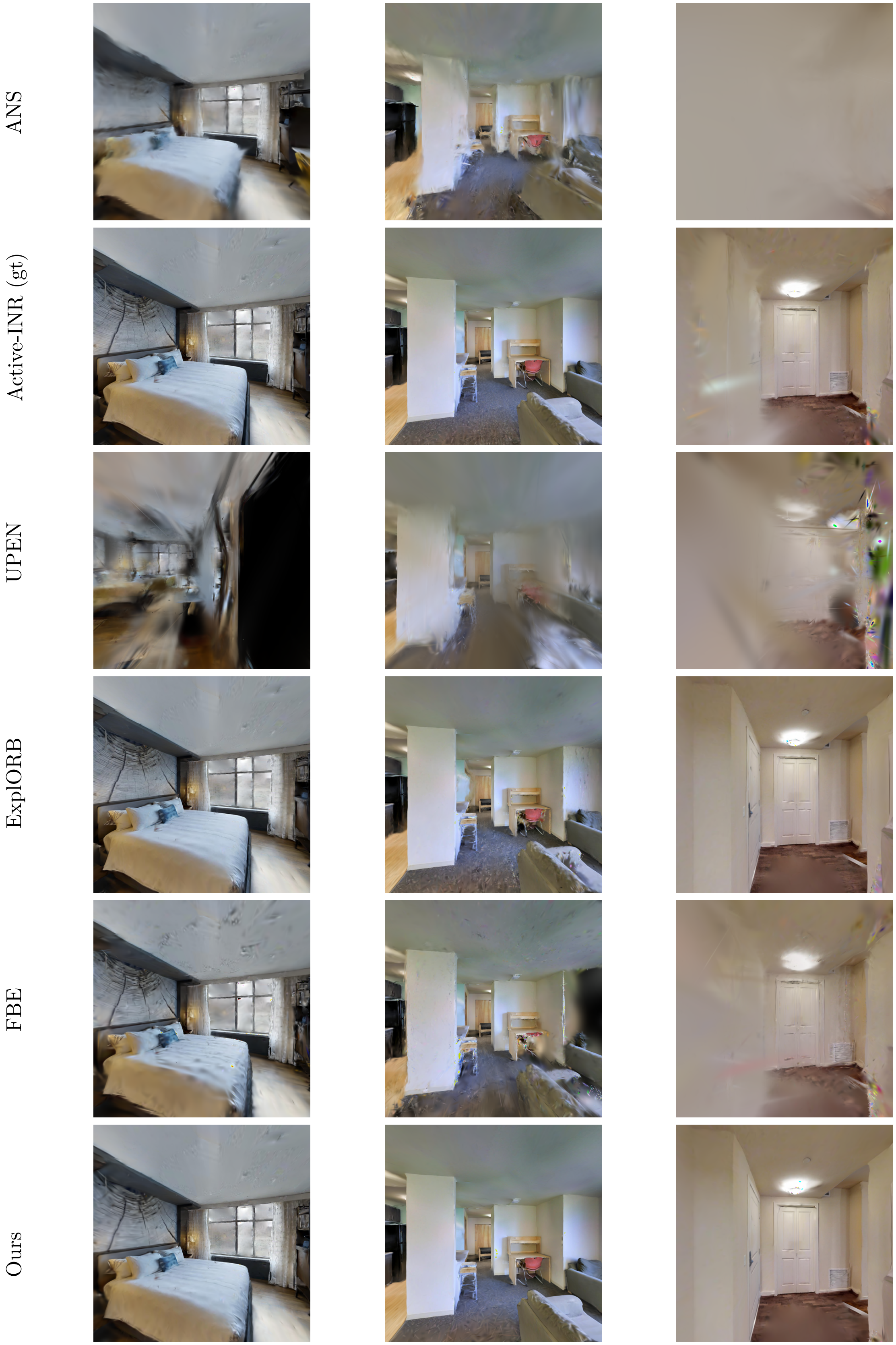


Figure 10. **Test Rendering Qualitative Visualization on Gibson Dataset** All the renderings are from the test view of the Gibson dataset.

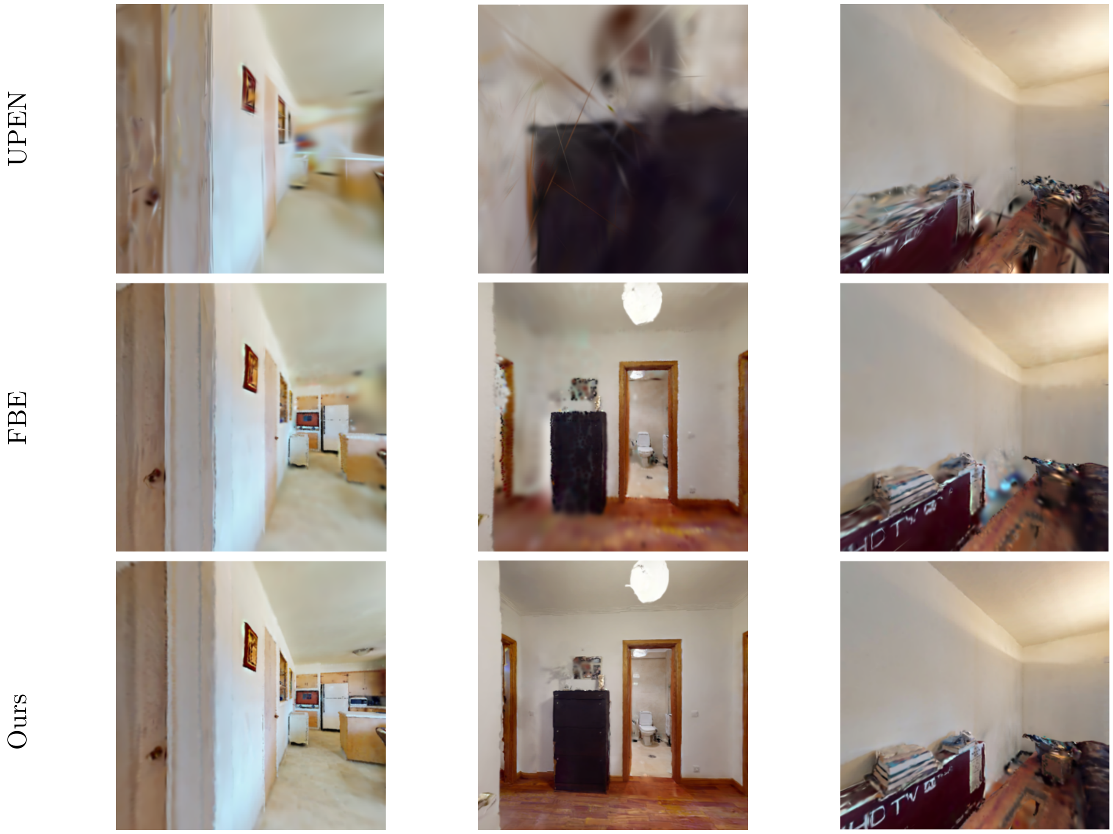
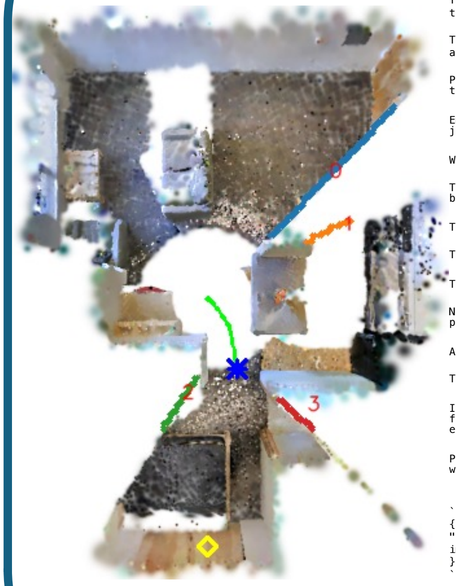


Figure 11. **Test Rendering Qualitative Visualization on Habitat-Matterport 3D Dataset** All the renderings are from the test view of the Habitat-Matterport 3D dataset.



~~~~
System: You are an AI assistant that can analyze images and plan a long-term
goal for the exploration task of a ground robot.

You will be given a bird-eye view image of a scene.

The goal is to plan a long-term exploration mission for a robot to traverse
the area.

The robot's task is to explore the terrain efficiently, identifying important
areas, potential obstacles, and unvisited areas.

Please analyze the image and select a long-term goal from the candidates for
the robot to explore the area.

Empty space doesn't always mean they are unvisited regions, sometimes it's
just outside the floor plan of this scene.

We are allowed to explore a total of 2000 steps and this is step 2.

Therefore, it's better to select a space that is close to the visited regions
but still unvisited and not behind the walls.

The current location of the robot is marked with the blue star(∗) marker.

The last frontier you selected is marked with a yellow diamond(◆) shape.

The visited path is painted as green lines in the image.

Note that you don't have to select the closest point to the robot, but the
point that is most likely to be unvisited and important to explore.

As you can see, there are 4 candidate points to select from.

They are numbered from 0 to 3 in red color.

If you find all the goals are not necessary to explore and we should instead
focus on improving existing reconstruction, please give -1 in the `target`
entry of the JSON.

Please provide a detailed exploration plan and select an exploration target
with reasons in the JSON format as shown below.


```
{
"target": 2, "reason": "The target is located at an unvisited region of the
image and seems to be an unvisited bedroom"
}
```

Do not cut off the JSON and generate the full JSON.


User: I have a bird-eye view image of a scene. The goal is to plan a long-
term exploration mission for a robot to traverse the area. Please analyze the
attached image and provide the exploration plan first and then an exploration
target in the specified JSON format.
~~~~

~~~~
**Exploration Plan:**

1. **Current Situation Analysis:**

   - The robot is currently located at the blue star marker.

   - The previously selected frontier, marked by a yellow diamond, might still have unexplored neighboring areas.

   - So far, the robot has traversed 132 steps and can take a total of 2000 steps.

2. **Observation of Candidates:**

   - **Candidate 0:** Located towards the top of the scene. It appears to be in a separated area from the main path by walls and could be
challenging to access directly.

   - **Candidate 1:** Positioned near the robot's previous path, to the left. The space seems to be partly blocked but may lead to a new room
or area.

   - **Candidate 2:** Situated to the bottom right, close to the previously marked yellow diamond. This area seems unvisited with potential
space for exploration.

   - **Candidate 3:** On the right side, aligned with the robot’s current heading, showing open space but appearing more distant from the
recent paths.

3. **Selection Reasoning:**

   - **Candidate 2** is favored for exploration as it is likely unvisited, near the previous frontier, and potentially opens up new unexplored
areas. It is a strategic choice for efficient area coverage without straying too far from the recently explored paths, reducing the risk of
re-exploring already visited spaces.


```json
{
  "target": 2,
  "reason": "Candidate 2 is favored as it is near the previously marked frontier, indicating potential unexplored area relevant to the mission
objectives. Continuing exploration here aligns with efficient coverage of the space."
}
```
~~~~

Figure 12. **An Example about a Dialog with the Multi-modal LLM.** The chat box on the top is the question from our method, and the bottom is the response from GPT-4o, which we parsed to get a long-term goal for our path proposal. We provided the contextual information in both the textual prompt and the image.