This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multimodal LLM Guided Exploration and Active Mapping using Fisher Information

Wen Jiang*[1] Boshu Lei*[1] Katrina Ashton[1]
Kostas Daniilidis[1,2]
[1] University of Pennsylvania [1,2] Archimedes, Athena RC

Abstract

We present an active mapping system which plans for both long-horizon exploration goals and short-term actions using a 3D Gaussian Splatting (3DGS) representation. Existing methods either do not take advantage of recent developments in multimodal Large Language Models (LLM) or do not consider challenges in localization uncertainty, which is critical in embodied agents. We propose employing multimodal LLMs for long-horizon planning in conjunction with detailed motion planning using our information-based objective. By leveraging high-quality view synthesis from our 3DGS representation, our method employs a multimodal LLM as a zero-shot planner for long-horizon exploration goals from the semantic perspective. We also introduce an uncertainty-aware path proposal and selection algorithm that balances the dual objectives of maximizing the information gain for the environment while minimizing the cost of localization errors. Experiments conducted on the Gibson and Habitat-Matterport 3D datasets demonstrate state-of-the-art results of the proposed method.

1. Introduction

Being able to autonomously explore and map an environment while localizing within that map is a core skill for a mobile robot. This ability could empower embodied artificial intelligence systems with effective 3D scene understanding through use in conjunction with vision-language features [60, 73, 94, 98] and can be used as a basis for language-specified robotics tasks [27, 28, 42, 68]. This task is challenging because it requires general knowledge of the typical layout of an environment to identify coarse targets for long-term exploration, as well as the ability to plan each step to ensure localization accuracy while maximizing information gain or 'surprise' when traversing the environment.

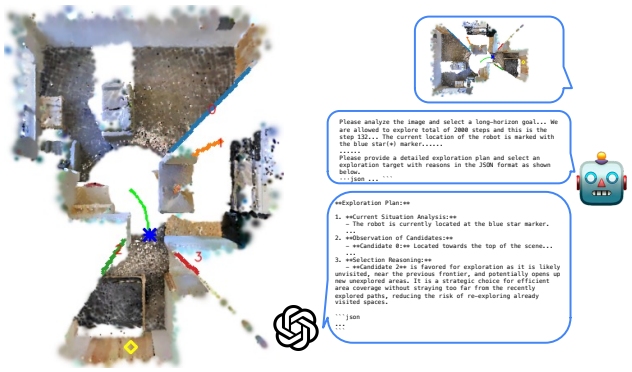


Figure 1. **An overview about the interaction with a multimodal LLM.** The image on the left is a bird's-eye view rendering from our 3DGS representation. A detailed example of the dialog can be found in the supplementary.

We use 3D Gaussian Splatting (3DGS) [36] as scene representation and tackle this problem in two phases: long-horizon planning with a Large Language Model (LLM), and detailed motion planning with our information-based approach by quantifying the Fisher Information for 3DGS parameters and current localization states. As shown in Fig. 1, we prompt a multimodal LLM with our current map state, trajectory, and frontiers [89], used as candidates for long-term exploration to exploit the LLM's knowledge about scene layouts. We then follow the suggestion from the LLM by choosing paths that maximize the Expected Information Gain (EIG) calculated using the Fisher Information about both poses and the scene.

Classical active mapping methods often define objectives to reduce the uncertainty of the state estimate based on its covariance [1, 7, 13, 21, 63, 85]. The covariance is readily available for classical systems which use filters to update the state estimate. However, many recent mapping systems use non-linear optimization to update the state estimate, making the covariance difficult to obtain. In particular, systems with 3DGS for the scene representation [29, 35, 49, 90] have been de-

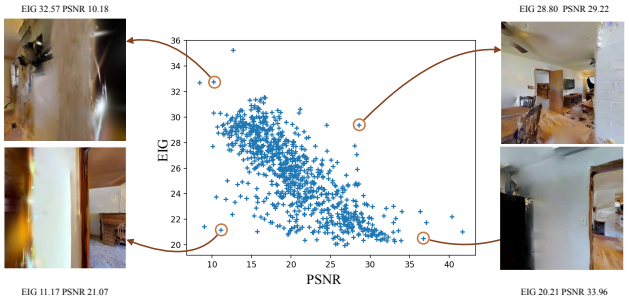


Figure 2. **Scatter Plot of EIG vs. PSNR** We plot the EIG and PSNR at sampled poses in the Cantwell scene of the Gibson dataset. The figure corroborates the intuition that the robot expects to gain little information (low EIG) at well reconstructed region (high PSNR) and gain much information (high EIG) at a poorly reconstructed region (low PSNR).

veloped for high-fidelity rendering of novel views of the scene. Previous methods attempted to quantify the uncertainty of radiance fields for reconstructing scenes from given data [24, 71, 72, 77] for active view selection [33, 41, 56, 77] and active object reconstruction [91] or mapping [33, 34, 39, 92] of scenes with given localization, and for 3D reconstruction of small scenes with an inward-facing camera [95]. However, all the prior methods only model the uncertainty of the scene representation, ignoring another primary source of error for a practical mapping system: the risk of localization failures. Our proposed system could address not only Information Gain during exploration but also consider localization accuracy when the robot explores unknown and texture-less regions.

In terms of long-horizon planning for robot exploration, classical approaches such as frontier-based exploration [89] and A* algorithms are still used in active mapping systems for their efficiency and simplicity [6, 38, 75, 76, 84]. However, algorithms that use only simple heuristics cannot determine the information gain, nor estimate 3D geometry, limiting their objective to simply improving coverage or having minimal travel distances. To address this issue, learnable approaches have been used for exploration tasks [23, 43, 51, 64], but these methods are trained for a limited scene distribution. Recently, LLMs and Vision Language Models have demonstrated extraordinary abilities in visual grounding and logical reasoning and have been studied with various robotic tasks [10, 11, 16, 17, 28, 67, 81]. However, these works do not focus on pure exploration or reconstruction tasks and many do not take advantage of photorealistic 3D representations for multimodal LLMs.

We argue that we leverage the best of both worlds by using a multi-modal LLM to identify long-horizon exploration goals and planning paths to them using our information-based algorithm. Our approach is seamlessly integrated with our 3DGS scene representation, we utilize a high-quality bird's-eye view rendering from the scene representation to prompt the LLM and can directly extract the EIG for path planning from it using Fisher Information. As shown in Fig. 2, our EIG metric reflects the 'surprise' and correlates well with the rendering quality on candidate views, as measured by the peak signal-to-noise ratio (PSNR) between the true and rendered image, without actually taking a picture at the candidate location. Notably, we compute the full Fisher Information matrix for localization parameters without any approximation with our efficient CUDA implementation. We also derived localization uncertainty on candidate paths with the Cramér–Rao bound, and we use this along with the EIG in our final objective to balance the dual problem of exploration and localization. To validate our approach, we evaluate our method on scenes from the Gibson [88] and Habitat-Matterport 3D [61] datasets quantitatively and qualitatively. We show superior reconstruction quality in various metrics compared to several baselines and recent state-of-the-art methods [9, 23, 58, 89, 92].

Our contributions can be summarized as follows:

- We present an active mapping system for ground robots that could autonomously explore the environment and extensively compare our system with previous approaches. To the best of our knowledge, we are the first active mapping system with 3D Gaussian representation that is not dependent on ground truth camera pose readings.
- We provide a way to leverage the zero-shot long-horizon planning ability of LLM into our active mapping system seamlessly.
- We introduce localization uncertainty with active mapping systems and effectively balance the information gain for exploration and the cost of possible localization errors.

2. Related work

Active Mapping and Localization Efficiently exploring an environment in order to map it while being able to localize in that map is a fundamental problem in robotics. There are many methods that address aspects of this problem separately – Simultaneous Localization and Mapping (SLAM) methods [5, 18, 19, 32, 52–54, 79, 82] address the mapping and localization. Many exploration methods [12, 23, 33, 34, 39, 92] address the exploration and mapping aspects while assuming poses are provided. Active SLAM methods [6, 75, 76, 84] consider both of these problems. The exploration in these systems is usually driven by a measure of uncer-

tainty [4, 26, 44, 59]; specific utility functions are often drawn from either Information Theory (IT) [70], or the Theory of Optimal Experimental Design (TOED) [57]. Recently, learning-based approaches have been developed for active mapping and localization; Active Neural SLAM [9] learns policies to drive exploration and estimate the agent pose, more similarly to us NARUTO [22] uses an SLAM backbone to estimate the pose and uses uncertainty to drive exploration. However, they only consider the reconstruction uncertainty, whereas we choose long-horizon planning goals and consider the localization uncertainty as well as the reconstruction uncertainty when planning the best path to the long-horizon goal.

Uncertainty quantification for radiance fields The vast majority of previous work on uncertainty quantification for radiance fields has been for post-processing scenes [24, 71, 72, 77, 80], view selection [33, 41, 56, 77] or active view selection [33, 56, 77], all of which assume the input images are posed. Active neural mapping [39, 92] uses neural variability, that is, the prediction robustness against random weight perturbations, as an estimate of uncertainty to actively map a scene with ground truth poses provided. Fisher-RF [33] also performs active scene mapping with ground truth poses provided, based on an approximation of the Fisher Information of views along candidate paths. Zhan et al. [95] performs active reconstruction without ground truth camera poses. However, they only evaluate small-scale scenes and limit the camera trajectories to be inwards facing and only model scene uncertainty, not localization uncertainty, which is a key consideration for active mapping and exploration.

Robot Planning with Foundation Models LLMs and VLMs have been used widely in robotics as for language-specified tasks, either as high-level planners in conjunction with other methods for low-level control [3, 16, 17, 28, 30, 45, 62, 74], to create intermediate representations which can be planned over [31] or directly outputting actions [46, 97]. Outside of language-driven tasks, VLMs and LLMs have been used in robotics for providing rewards to drive exploration in Reinforcement Learning [83], and for visual localization [50]. More similar to our task, LLMs and VLMs have also been used for goal-driven navigation, such as searching for specific objects [15, 65, 69, 93].

3. Method

We divide the scene exploration task into two phases: (a) long-horizon planning for the coarse direction of exploration that leads to better coverage and understanding (Sec. 3.2), and (b) detailed trajectory proposal and path selection that aims to improve 3D reconstruction (Sec. 3.3) and localization (Sec. 3.4) from a geometrical perspective. The former task relies more on semantic information and prior knowledge about the possible layout of the environment, which is suitable for a generalist model with prior distributions on the scene. The latter task, however, involves detailed motion planning and is better addressed by considering information gain on fine-level geometries. Fig. 3 shows an overview of our method.

3.1. Preliminary

In 3D Gaussian Splatting (3DGS) [36], the scene is represented by a set of 3D Gaussians whose color and opacity are learned via a rendering loss. An image can be rendered by projecting the Gaussians to 2D and using α-blending for the N ordered points on the 2D splat that overlaps each pixel. The Jacobian of the current camera pose $\mathbf{x}$ with respect to the parameters of 3D Gaussians can be computed by defining the (left) partial derivative on the manifold [49]:

$$\frac{\mathcal{D}f(\mathbf{x})}{\mathcal{D}\mathbf{x}} \triangleq \lim_{\tau \to 0} \frac{\mathrm{Log}(f(\mathrm{Exp}(\tau) \circ \mathbf{x}) \circ f(\mathbf{x})^{-1})}{\tau}, \quad (1)$$

where $\tau \in \mathfrak{se}(3)$, $\circ$ is a group composition, and Exp and Log are the exponential and logarithmic mappings between Lie algebra and Lie Group.

Fisher Information is a measurement of the information that a random variable $\mathbf{y}$ carries about an unknown parameter $\mathbf{w}$ of a distribution that models $\mathbf{y}$. In the problem of novel view synthesis, we are interested in measuring the observed information of a radiance field with parameters $\mathbf{w}$ at a camera pose $\mathbf{x}$ using the negative log-likelihood of the image observation $\mathbf{y}$ taken from that pose:

$$-\log p(\mathbf{y}|\mathbf{x};\mathbf{w}) = (\mathbf{y} - f(\mathbf{x},\mathbf{w}))^T(\mathbf{y} - f(\mathbf{x},\mathbf{w})), \quad (2)$$

where $f(\mathbf{x},\mathbf{w})$ is the rendering model. Under regularity conditions [66], the Fisher Information of $-\log p(\mathbf{y}|\mathbf{x};\mathbf{w})$ is the Hessian of Eq. 2 with respect to $\mathbf{w}$, denoted $\mathbf{H}''[\mathbf{y}|\mathbf{x},\mathbf{w}]$. In our formulation $\mathbf{y} \in R^{h\times w\times c}$ is the RGB-D observation, and $\mathbf{w}$ is a tensor for all the 3D Gaussian Parameters $R^{N\times 14}$, where N is the number of Gaussians and the parameters are the 3D means, RGB colors, unormalized rotations represented as quaternions, opacities and per-axis log-scales.

FisherRF [33] uses Fisher Information to address active perception problems such as active view selection and active mapping, although they do not consider localization uncertainty. Given a training set of views D^{train} and a set of candidate trajectories $\{\mathcal{P}_j\}$ Fish-

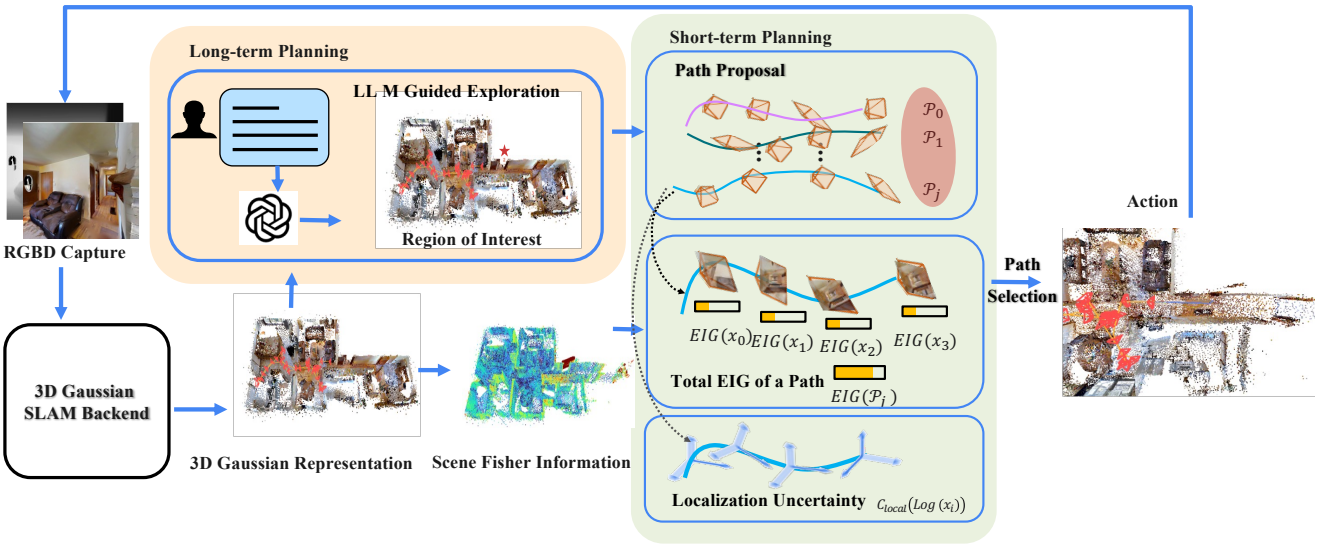


Figure 3. **An Illustration of Our Active Mapping System** Our method first identifies long-horizon targets with a multimodal LLM by using novel-view synthesis from our 3DGS representation. Our information-based path proposal method then proposes and selects short-term action sequences from the region of interest the multimodal LLM identifies. The best path and action along the path is selected concerning both the information gain and localization accuracy.

erRF selects the path which maximizes the following:

$$\mathrm{tr}\left(\left(\sum_{\mathbf{x}_i^{acq}\in\mathcal{P}_j}\mathbf{H}''[\mathbf{y}_i^{acq}|\mathbf{x}_i^{acq},\mathbf{w}]\right)\mathbf{H}''[\mathbf{w}|D^{train}]^{-1}\right) \tag{3}$$

where $\mathbf{w}$ is the initial estimate of model parameters using the current training set. Crucially, the Fisher Information $\mathbf{H}''[\mathbf{y}_i^{acq}|\mathbf{x}_i^{acq},\mathbf{w}]$ does not depend on the label $\mathbf{y}_i^{acq}$ of the acquisition sample $\mathbf{x}_i^{acq}$, thus it is feasible to compute it before visiting the potential view candidate $\mathbf{x}_i^{acq}$. However, the number of optimizable parameters is typically more than 20 million, which means it is impossible to compute without sparsification or approximation. In practice, FisherRF [33] applies a Laplace approximation [14, 47] that approximates the Hessian matrix with its diagonal values plus a log-prior regularizer λI as follows

$$\mathbf{H}''[\mathbf{y}|\mathbf{x},\mathbf{w}] \simeq \mathrm{diag}(\nabla_{\mathbf{w}} f(\mathbf{x},\mathbf{w})^T \nabla_{\mathbf{w}} f(\mathbf{x},\mathbf{w})) + \lambda I. \tag{4}$$

3.2. Long Horizon Exploration with Foundation Model

We propose using a multimodal LLM as a zero-shot long-horizon planner and leaving the detailed path planning to a closed-form uncertainty-aware motion planning algorithm.

As we consider an agent moving in a 2D action plane (e.g., a ground robot), we leverage the view synthesis ability of 3DGS to create expressive bird's-eye view renderings that could provide an overview of the environment. An occupancy grid on the motion plane of our robot is created using our 3D Gaussian representation, which can be used to identify the frontiers of the current environment. The frontiers are defined as points a set of neighboring points on the action plane on the boundary between free space and unobserved space. The agent has many choices of frontiers, especially during the early stages of exploration. We apply Chain-of-Thought Prompting [87] to encourage the multimodal LLM to provide analysis on the candidate frontiers first before selection. We provide contextual information such as the description of the task, the total steps allowed for exploration, and our current step in the textual prompt. The LLM is also allowed to decide whether the robot should not go to a frontier and instead focus on improving existing regions. We annotate the rendered map with the current location, previously visited trajectory and the frontiers which form the candidate long-term navigation goals. We prompt the LLM with both the contextual text information and annotated map as shown in Fig. 1 and extract the long-term goal from its output. By using a set of possible candidates we avoid possible infeasible destinations for motion planning. For example, the unexplored regions we have no reconstruction for at this point might be unreachable from the current space. We can also check that the LLM's output is actually one of the provided options before proceeding, if it is not then we instead select the largest frontier. We use GPT-4o [55] for our long-horizon planning task, but our method is agnostic to the underlying multimodal LLM as the inputs for the

LLM are text prompts and a bird's-eye view rendering. Detailed examples can be found in the Supplementary.

3.3. Expected Information Gain for Path Proposal and Selection

We use the Expected Information Gain (EIG) as both a preliminary method to further refine our candidate poses and as part of our path selection criteria.

After identifying a long-horizon goal with the multimodal LLM, we form an initial set of candidate poses $\mathcal{T}_I$ by sampling points in the coarse region of interest from the LLM. If there are no unvisited boundaries or the LLM suggests we do not need to explore the frontiers, we sample poses across the free space to form $\mathcal{T}_I$. We then evaluate the Expected Information Gain (EIG) for each pose $\mathbf{x}_i^{acq} \in \mathcal{T}_I$, given by

$$\texttt{EIG}(\mathbf{x}_i^{acq}) = \text{tr}\left(\mathbf{H}''[\mathbf{y}_i^{acq}|\mathbf{x}_i^{acq}, \mathbf{w}]\, \mathcal{I}(\mathbf{w})^{-1}\right), \quad (5)$$

as a preliminary selection metric to form our final candidate target poses set $\mathcal{T}_F$. We calculate $\mathcal{I}(\mathbf{w})$ differently to FisherRF [33], which approximates it using $\mathbf{H}''[\mathbf{w}|D^{train}]$ by computing the Hessians on the training set. This is also known as empirical Fisher Information, whose limitations have been widely discussed [40, 48]. In most scenarios, this is a reluctant design choice because the distribution of $\mathbf{x} \sim p(\mathbf{x})$ is unknown (i.e., the distribution of all possible images). However, $\mathbf{x} \sim p(\mathbf{x})$ in our case is tractable because it represents the possible locations where we can take an observation for the environment, i.e. the free space of our map. Therefore, unlike FisherRF [33] as described in Eq. 3, we propose to use Monte-Carlo sampling to compute the Fisher Information of the current model

$$\mathcal{I}(\mathbf{w}) = \mathbb{E}_{\mathbf{x}\sim p(\mathbf{x})}\left[\mathbf{H}''[\mathbf{y}|\mathbf{x}, \mathbf{w}]\right] \simeq \sum_{k=1}^{N} \mathbf{H}''[\mathbf{y}_k|\mathbf{x}_k, \mathbf{w}], \quad (6)$$

where $\mathbf{x}_k$ is drawn from a uniform distribution of camera poses in the free space of the current map which we use to approximate p. We also uniformly initialize 3D Gaussians in the space, which will be subsequently updated with rendering losses for visited regions.

After the set of final candidate poses $\mathcal{T}_F$ has been formed, we construct paths to each candidate pose with the A* algorithm [25] using the occupancy map. The path can be defined as an ordered set of camera poses from the current location x_t at exploration step t to the frontier points x_T^j from the long horizon planning.

$$\mathcal{P}_j = \{x_{t+1}^j, \ldots, x_T^j\} \quad (7)$$

To evaluate how beneficial a path $\mathcal{P}_j$ would be for improving the reconstruction of the scene we consider the EIG for the 3D Gaussian parameters along that path, which can be computed as the sum over the path of the following term [33]:

$$\texttt{EIG}_{\mathcal{P}_j,i}(\mathbf{x}_i) = \text{tr}\left(\mathbf{H}''[\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}]\, \mathcal{I}_{\mathcal{P}_j,i}(\mathbf{w})^{-1}\right) \quad (8)$$

where $\mathcal{I}_{\mathcal{P}_j,i}(\mathbf{w})$ takes the mutual information along the path into account as follows

$$\mathcal{I}_{\mathcal{P}_j,i}(\mathbf{w}) = \mathcal{I}(\mathbf{w}) + \sum_{\mathbf{x}_t \in \mathcal{P}_j, t<i} \mathcal{I}(\mathbf{w}|\mathbf{x}_t). \quad (9)$$

3.4. Localization Uncertainty for Path Selection

If solely maximizing the EIG, the robot will be more likely to explore unvisited regions. However, exploring regions that have not been well reconstructed also means the agent would have the risk of worse localization accuracy due to noise and ambiguities in the unreconstructed regions during pose optimization. The cost of localization must thus be considered during path planning to balance the importance of exploring new environments with maintaining localization accuracy. We propose to use Fisher Information as a measurement for localization uncertainty to address this. During optimization, we essentially optimize on the logarithmic mapping $\tau_i \triangleq \text{Log}(\mathbf{x}_i)$ of our camera pose. By the Cramér–Rao bound, the covariance of $\tau_i \in \mathfrak{se}(3)$ can be lower-bounded with the inverse of Fisher Information matrix $\mathcal{I}(\tau_i)$:

$$\text{Cov}(T(\hat{\tau}_i)) \geq \mathcal{I}(\tau_i)^{-1} \quad (10)$$

where $T(\tau_i)$ is an unbiased estimator for τ_i solved by iteratively optimizing photo-metric loss. Hence, we can define the localization cost C_{local} at a pose $\mathbf{x}_i$ in terms of τ_i as:

$$C_{local}(\tau_i) = -\log\det(\nabla_{\tau_i} f(\tau_i, \mathbf{w})^T \nabla_{\tau_i} f(\tau_i, \mathbf{w})) \quad (11)$$

Matsuki *et al.* [49] computed the Jacobians of the camera pose with respect to the mean and covariances of each gaussian $\frac{\partial \boldsymbol{\mu}_I}{\partial \mathbf{x}}$ and $\frac{\partial \boldsymbol{\Sigma}_I}{\partial \mathbf{x}}$. However, we need to compute the Jacobian of τ_i with respect to the rendering output:

$$\nabla_{\tau_i} f(\tau_i, \mathbf{w}) = \frac{\partial f(\tau_i, \mathbf{w})}{\partial \tau_i} = \begin{bmatrix} \frac{\partial f(\tau_i, \mathbf{w})}{\partial \boldsymbol{\mu}_I} & \frac{\partial f(\tau_i, \mathbf{w})}{\partial \boldsymbol{\Sigma}_I} \end{bmatrix} \begin{bmatrix} \frac{\mathcal{D}\boldsymbol{\mu}_C}{\mathcal{D}\tau_i} \\ \frac{\mathcal{D}\mathbf{W}}{\mathcal{D}\tau_i} \end{bmatrix} \quad (12)$$

Unlike 3DGS parameters, our computation of the localization cost can be directly computed on the full Fisher Information matrix without using the Laplace approximation [14, 47]. That is because the full Jacobian $\nabla_{\tau_i} f(\tau_i, \mathbf{w})$ is linear to the number of Gaussians.

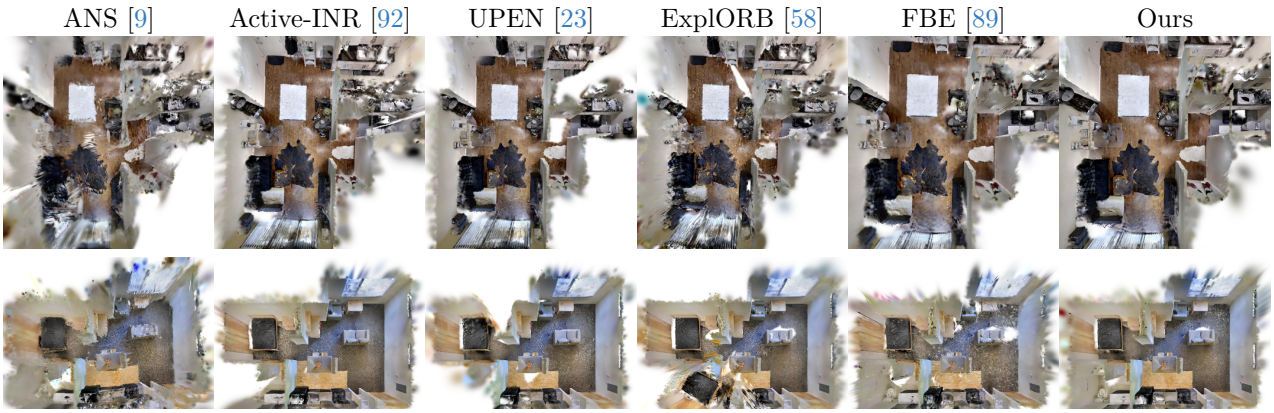


Figure 4. **Qualitative Comparison for Final Scene Reconstruction on Gibson Dataset** Greigsville (top) and Ribera (bottom) scenes. We provide top-down rendering for different methods. Note that UPEN and Active-INR use GT pose in this visualization.



Figure 5. **Qualitative Comparison for Final Scene Reconstruction on Habitat-Matterport 3D Dataset** oPj9qMxrDEa (left) and QKGMrurUVbk (right) scenes. We provide top-down rendering for different methods.

We select the best path by minimizing the total cost for all viewpoints $\mathbf{x}_i$ along a path $\mathcal{P}_j$:

$$\arg\min_{\mathcal{P}_j} \sum_{\mathbf{x}_i \in \mathcal{P}_j} C_{local}(\text{Log}(\mathbf{x}_i)) - \eta \log(\texttt{EIG}_{\mathcal{P}_j,i}(\mathbf{x}_i)) \quad (13)$$

where η is a hyper-parameter controlling the relative importance between EIG and localization accuracy. Our active mapping system constantly updates the map, and we replan using our active path planning algorithm if we detect the agent is getting close to a possible obstacle or upon reaching the end of the previously selected path.

4. Experiments

4.1. Experimental Set-up

Dataset Following previous methods [23, 92], our algorithm is evaluated in the Habitat Simulator [78] on the Gibson [88] and Habitat-Matterport 3D (HM3D) [61] datasets comprised of indoor scenes reconstructed from scans of real houses. We adopt the default start point in the Habitat Simulator as the starting point for exploration in each scene. The total number of steps for each experiment is 2000. The system takes color and depth images at the resolution of 800x800 and outputs a discrete action at each step. The action space consists of MOVE FORWARD by 5cm, TURN LEFT, and TURN RIGHT by 5°. The field of view (FOV) is set to 90° vertically and horizontally. Please refer to the supplement for more details about the evaluation split and other hyper-parameters.

Metrics We evaluated our method using the Peak-signal-to-noise ratio (PSNR), Structural Similarity Index Measure (SSIM) [86], Learned Perceptual Image Patch Similarity (LPIPS) [96] for RGB rendering and mean absolute error (MAE) for depth rendering as metrics for scene reconstruction quality. We calculate these metrics using 2000 points uniformly sampled from the movement plane of the agent in the scene, discarding any points that are not navigable. We argue that the rendering quality reflects both reconstruction quality and pose accuracy because high tracking accuracy would help the training of the 3D Gaussian Splatting model. Meanwhile, misaligned poses will lead to misaligned rendering at test time thus leading to inferior results. Following previous approaches [39, 92], we also use the completion ratio as an evaluation metric. To evaluate the pose estimation accuracy, we use the root mean squared average tracking error (RMSE ATE), but as the trajectories for each method are different, the RMSE ATE should only be considered along with other metrics such as completness.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Depth MAE ↓	RMSE ATE ↓	Completeness (%) ↑
ANS [9]	16.34	0.6818	0.3923	0.3886	0.1105	35.33
UPEN [23]	16.44	0.6678	0.4134	0.4841	0.5158	22.66
ExplORB [58]	18.99	0.7175	0.3994	0.2664	0.2296	30.23
FBE [89]	21.45	0.7618	**0.2126**	0.1028	0.1680	55.87
Ours	**23.28**	**0.8067**	0.2507	**0.0696**	**0.0226**	**84.38**

Table 1. **Quantitative Evaluation on Reconstruction Quality and Tracking Accuracy on Gibson dataset**

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Depth MAE ↓	RMSE ATE ↓	Completeness (%) ↑
UPEN [23]	12.23	0.4795	0.5157	0.7356	0.4393	17.48
ExplORB (gt) [58]	17.81	0.3694	0.6810	0.5071	-	31.92
FBE [89]	15.80	0.5952	0.4392	0.4085	1.2004	22.42
Ours	**19.86**	**0.7127**	**0.4122**	**0.1666**	**0.0336**	**49.76**

Table 2. **Quantitative Evaluation on Reconstruction Quality and Tracking Accuracy on HM3D dataset**

Baselines We compare to two exploration methods which assume ground truth pose: UPEN [23] and Active Neural Mapping (active-INR) [92]. We also compare our method with Active Neural SLAM (ANS) [9], explORB [58] and Frontier Based Exploration (FBE) [89] without ground truth pose provided. Note that we do not run ANS on HM3D as it is not trained on this dataset.

To compare the rendering quality fairly, we run all the baselines using the MonoGS [49] backend for reconstruction. We run UPEN and FBE online, but for ANS, active-INR, and ExplORB, we record and playback trajectories obtained using their source code. Because the forward step size for ANS is much larger than for our method, we interpolate the trajectory so that the forward step size matches our method's to make the steps comparable. For ExplORB, since the official implementation is based on MoveBase, which uses velocity commands, we sample the trajectory at 5 Hz. We also found that ANS, active-INR, and UPEN failed on some scenes due to localization failure of the MonoGS backend. ANS produces a pose estimate (using information from noisy pose sensors not provided to our pipeline), so we set the pose estimate of the MonoGS backend to the one from ANS. As active-INR and UPEN do not produce a pose estimate, we evaluate them using the ground-truth pose.

4.2. Comparison Against Previous Methods

Table 1 shows the results of our method and the baselines for exploration in scenes from the Gibson dataset [88], and Table 2 shows the results on HM3D [8]. Our method outperforms the baselines on all metrics. We further qualitatively compare the reconstruction qualities after active exploration in Fig. 4 and Fig. 5, and the trajectories in Fig. 7. Our method does not have major errors from failed localization, and we have fewer gaps in the scenes than other methods. For example, in the Ribera scene, all methods except for us and FBE miss the bathroom at the bottom left, and FBE misses more of the area around the sofa than us. We show the trajectories taken for the Cantwell scene from the Gibson dataset, which is a relatively large and challenging scene. We show only a few baselines to keep the figure legible. We can see that ANS [9] does not go to the region on the bottom left, Active-INR [92] stays in the center region, and FBE does not go to the region on the bottom right, whereas our method efficiently visits most areas of the scene.

4.3. Ablation Study and Runtime Analysis

To validate the effectiveness of localization uncertainty, hyperparameters, and LLM, we performed an ablative study of each component in our system on the HM3D dataset in Table. 3. As can be seen, the average trajectory error is much lower with the localization uncertainty than without. In addition, after adding LLM, the model performs better in all metrics. We also noticed the tracking error was lower after introducing multimodal LLM to the active mapping system. We hypothesize this improvement is because improved scene reconstruction can also help reduce tracking errors.

We additionally provide a breakdown of the runtime of our method, shown in Fig. 6.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Depth MAE ↓	RMSE ATE ↓	Completeness (%) ↑
w.o. LLM & Localization Uncertainty	16.15	0.6550	0.6193	0.3409	0.2478	35.40
w.o. LLM	16.94	0.6799	0.5847	0.2887	0.1694	37.26
Ours (Llava-7b)	18.46	0.6805	0.4623	0.2033	**0.0159**	17.41
$\lambda = 2 \times 10^{-6}$ †	18.90	0.6976	0.4408	0.1966	0.0479	18.12
$\lambda = 5 \times 10^{-6}$ †	18.90	0.6946	0.4602	0.2145	0.0551	18.05
Ours†	**19.86**	**0.7127**	**0.4122**	**0.1666**	**0.0336**	**49.76**

Table 3. **Ablation Study of Localization Uncertainty Term on Scenes from the HM3D Dataset.** We compare our method with and without the localization uncertainty term to validate that including it improves localization and reconstruction. †: the multimodal LLM is GPT-4o in the experiment.



Figure 6. **Runtime Analysis of the Active Mapping System.** We provide running time statistics by each step and by the percentage of total time consumed on one episode.

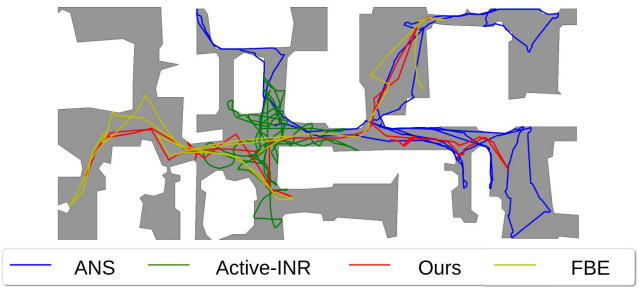


Figure 7. **Qualitative Comparison for Trajectories** on Cantwell from Gibson Dataset.

5. Conclusion

We present an active mapping system that can autonomously explore an environment, the first method to do so using a 3D Gaussian representation without ground truth camera poses. By rendering maps using our scene representation, we can seamlessly query a multimodal Large Language Model for long-horizon planning to leverage its knowledge of scene layouts. This allows the robot to efficiently explore the scene while avoiding setting unreachable navigation goals. We then plan an optimal path to the long-horizon goal using our information-driven path proposal and selection algorithm, which balances the information gain with respect to the acene against the localization uncertainty, ensuring that the robot captures the geometric details of the scene for a high-quality reconstruction. We evaluate our method on scenes from the Gibson [88] and Habitat-Matterport 3D [61] datasets in terms of the rendering quality, completeness, and average tracking error. To enable our method to support more robotics applications, future work could extend it to take advantage of our 3D scene representation and consider movement with higher degrees of freedom (DOF) than the currently supported 3DOF. Our method could also be extended to task-driven exploration such as finding objects [2, 37], leveraging the LLM for efficient search. Incorporating semantic features [60, 73, 98] to allow for grounding language to the scene would also enable many robotics and computer vision applications.

Acknowledgements The authors gratefully appreciate support through the following grants: NSF FRR 2220868, NSF IIS-RI 2212433, NSF TRIPODS 1934960, NSF CPS 2038873. The authors appreciate the help from Prof. Pratik Chaudhari for the insightful discussion and hardware platforms.

References

[1] Arash Asgharivaskasi and Nikolay Atanasov. Semantic octree mapping and shannon mutual information computation for robot exploration. *IEEE Transactions on Robotics*, 39(3):1910–1928, 2023. 1

[2] Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. ObjectNav Revisited: On Evaluation of Embodied Agents Navigating to Objects. *arXiv preprint arXiv:2006.13171*, 2020. 8

[3] Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on robot learning*, pages 287–318. PMLR, 2023. 3

[4] Adam Bry and Nicholas Roy. Rapidly-exploring random belief trees for motion planning under uncertainty. In *2011 IEEE international conference on robotics and automation*, pages 723–730. IEEE, 2011. 3

[5] Carlos Campos, Richard Elvira, Juan J Gómez Rodríguez, José MM Montiel, and Juan D Tardós. Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021. 2

[6] Luca Carlone, Jingjing Du, Miguel Kaouk Ng, Basilio Bona, and Marina Indri. Active slam and exploration with particle filters using kullback-leibler divergence. *Journal of Intelligent & Robotic Systems*, 75:291–311, 2014. 2

[7] Henry Carrillo, Ian Reid, and José A Castellanos. On the comparison of uncertainty criteria for active slam. In *2012 IEEE International Conference on Robotics and Automation*, pages 2080–2087. IEEE, 2012. 1

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision*, 2017. 7

[9] Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *ICLR*, 2020. 2, 3, 6, 7, 16

[10] Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. *arXiv preprint arXiv:2407.05890*, 2024. 2

[11] Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9796–9810, 2024. 2

[12] Tao Chen, Saurabh Gupta, and Abhinav Gupta. Learning exploration policies for navigation. In *International Conference on Learning Representations*, 2019. 2

[13] Yongbo Chen, Shoudong Huang, and Robert Fitch. Active slam for mobile robots with area coverage and obstacle avoidance. *IEEE/ASME Transactions on Mechatronics*, 25(3):1182–1192, 2020. 1

[14] Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux–effortless Bayesian deep learning. In *NeurIPS*, 2021. 4, 5

[15] Vishnu Sashank Dorbala, James F Mullen Jr, and Dinesh Manocha. Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters*, 2023. 3

[16] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*, 2023. 2, 3

[17] Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023. 2, 3

[18] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *European conference on computer vision*, pages 834–849. Springer, 2014. 2

[19] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 2

[20] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996. 14

[21] Hans Jacob S Feder, John J Leonard, and Christopher M Smith. Adaptive mobile robot navigation and mapping. *The International Journal of Robotics Research*, 18(7):650–668, 1999. 1

[22] Ziyue Feng, Huangying Zhan, Zheng Chen, Qingan Yan, Xiangyu Xu, Changjiang Cai, Bing Li, Qilun Zhu,

and Yi Xu. Naruto: Neural active reconstruction from uncertain target observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21572–21583, 2024. 3

[23] Georgios Georgakis, Bernadette Bucher, Anton Arapin, Karl Schmeckpeper, Nikolai Matni, and Kostas Daniilidis. Uncertainty-driven planner for exploration and navigation. In *ICRA*, 2022. 2, 6, 7, 16, 17

[24] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' Rays: Uncertainty quantification in neural radiance fields. *arXiv*, 2023. 2, 3

[25] Peter E Hart, Nils J Nilsson, and Bertram Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. 5

[26] Ruijie He, Emma Brunskill, and Nicholas Roy. Puma: Planning under uncertainty with macro-actions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1089–1095, 2010. 3

[27] Jiawei Hou, Tianyu Wang, Tongying Pan, Shouyan Wang, Xiangyang Xue, and Yanwei Fu. TaMMa: Target-driven multi-subscene mobile manipulation. In *8th Annual Conference on Robot Learning*, 2024. 1

[28] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10608–10615. IEEE, 2023. 1, 2, 3

[29] Huajian Huang, Longwei Li, Cheng Hui, and Sai-Kit Yeung. Photo-slam: Real-time simultaneous localization and photorealistic mapping for monocular, stereo, and rgb-d cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[30] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International conference on machine learning*, pages 9118–9147. PMLR, 2022. 3

[31] 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*, 2023. 3

[32] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, Steve Hodges, Dustin Freeman, Andrew Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 559–568, 2011. 2

[33] Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and mapping with radiance fields using fisher information. In *ECCV*, page 422–440, 2024. 2, 3, 4, 5

[34] Rui Jin, Yuman Gao, Yingjian Wang, Yuze Wu, Haojian Lu, Chao Xu, and Fei Gao. Gs-planner: A gaussian-splatting-based planning framework for active high-fidelity reconstruction. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11202–11209. IEEE, 2024. 2

[35] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat, track & map 3d gaussians for dense rgb-d slam. *arXiv preprint arXiv:2312.02126*, 2023. 1

[36] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuhler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 1, 3, 14

[37] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16373–16383, 2024. 8

[38] Ayoung Kim and Ryan M Eustice. Active visual slam for robotic area coverage: Theory and experiment. *The International Journal of Robotics Research*, 34(4-5): 457–475, 2015. 2

[39] Zijia Kuang, Zike Yan, Hao Zhao, Guyue Zhou, and Hongbin Zha. Active neural mapping at scale. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7152–7159. IEEE, 2024. 2, 3, 6

[40] Frederik Kunstner, Philipp Hennig, and Lukas Balles. Limitations of the empirical fisher approximation for natural gradient descent. *Advances in neural information processing systems*, 32, 2019. 5

[41] Soomin Lee, Le Chen, Jiahao Wang, Alexander Liniger, Suryansh Kumar, and Fisher Yu. Uncertainty guided policy for active robotic 3d reconstruction using neural radiance fields. *IEEE Robotics and Automation Letters*, 7(4):12070–12077, 2022. 2, 3

[42] Peiqi Liu, Yaswanth Orru, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Ok-robot: What really matters in integrating open-knowledge models for robotics. *arXiv preprint arXiv:2401.12202*, 2024. 1

[43] Xu Liu, Ankit Prabhu, Fernando Cladera, Ian D Miller, Lifeng Zhou, Camillo J Taylor, and Vijay Kumar. Active metric-semantic mapping by multiple aerial robots. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3282–3288. IEEE, 2023. 2

[44] Iker Lluvia, Elena Lazkano, and Ander Ansuategi. Active mapping and robot exploration: A survey. *Sensors*, 21(7):2445, 2021. 3

[45] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *CoRL*, 2024. 3

[46] Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. Discuss before moving: Visual language navigation via multi-expert discussions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 17380–17387. IEEE, 2024. 3

[47] David J. C. MacKay. Bayesian Interpolation. *Neural Computation*, 4(3):415–447, 1992. 4, 5

[48] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. 5

[49] Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian Splatting SLAM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1, 3, 5, 7, 14

[50] Reihaneh Mirjalili, Michael Krawez, and Wolfram Burgard. Fm-loc: Using foundation models for improved vision-based localization. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1381–1387. IEEE, 2023. 3

[51] David Morilla-Cabello, Lorenzo Mur-Labadia, Ruben Martinez-Cantin, and Eduardo Montijano. Robust fusion for bayesian semantic mapping. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 76–81. IEEE, 2023. 2

[52] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5): 1255–1262, 2017. 2

[53] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[54] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *2011 international conference on computer vision*, pages 2320–2327. IEEE, 2011. 2

[55] OpenAI. Gpt-4o system card, 2024. https://openai.com/index/gpt-4o-system-card/. 4

[56] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *ECCV*, pages 230–246. Springer, 2022. 2, 3

[57] Andrej Pázman. *Foundations of optimum experimental design*. Springer Dordrecht, 1986. 3

[58] Julio A Placed, Juan J Gómez Rodríguez, Juan D Tardós, and José A Castellanos. Explorb-slam: Active visual slam exploiting the pose-graph topology. In *Iberian Robotics conference*, pages 199–210. Springer, 2022. 2, 6, 7, 16

[59] Julio A Placed, Jared Strader, Henry Carrillo, Nikolay Atanasov, Vadim Indelman, Luca Carlone, and José A Castellanos. A survey on active simultaneous localization and mapping: State of the art and new frontiers. *IEEE Transactions on Robotics*, 2023. 3

[60] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 1, 8

[61] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 2, 6, 8

[62] Krishan Rana, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian D Reid, and Niko Suenderhauf. Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. *CoRR*, 2023. 3

[63] María L Rodríguez-Arévalo, José Neira, and José A Castellanos. On the importance of uncertainty representation in active slam. *IEEE Transactions on Robotics*, 34(3):829–834, 2018. 1

[64] Julius Rückin, Federico Magistri, Cyrill Stachniss, and Marija Popović. An informative path planning framework for active learning in uav-based semantic mapping. *IEEE Transactions on Robotics*, 39(6):4279–4296, 2023. 2

[65] Arjun P S, Andrew Melnik, and Gora Chand Nandi. Cognitive planning for object goal navigation using generative AI models. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024. 3

[66] M.J. Schervish. *Theory of Statistics*. Springer New York, 2012. 3

[67] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652. IEEE, 2024. 2

[68] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. CLIP-fields: Weakly supervised semantic fields for robotic memory. In *ICRA2023 Workshop on Pretraining for Robotics (PT4R)*, 2023. 1

[69] Dhruv Shah, Michael Robert Equi, Błażej Osiński, Fei Xia, Brian Ichter, and Sergey Levine. Navigation with large language models: Semantic guesswork as a heuristic for planning. In *Conference on Robot Learning*, pages 2683–2699. PMLR, 2023. 3

[70] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27 (3):379–423, 1948. 3

[71] Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations. *CoRR*, abs/2109.02123, 2021. 2, 3

[72] Jianxiong Shen, Antonio Agudo, Francesc Moreno-Noguer, and Adria Ruiz. Conditional-flow nerf: Ac-

curate 3d modelling with reliable uncertainty quantification. In *ECCV*, 2022. 2, 3

[73] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 2024. 1, 8

[74] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023. 3

[75] Cyrill Stachniss, Dirk Hahnel, and Wolfram Burgard. Exploration with active loop-closing for fastslam. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, pages 1505–1510. IEEE, 2004. 2

[76] Zezhou Sun, Banghe Wu, Cheng-Zhong Xu, Sanjay E Sarma, Jian Yang, and Hui Kong. Frontier detection and reachability analysis for efficient 2d graph-slam based active exploration. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2051–2058. IEEE, 2020. 2

[77] Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9370–9376. IEEE, 2023. 2, 3

[78] Andrew Szot, Alexander Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Singh Chaplot, Oleksandr Maksymets, et al. Habitat 2.0: Training home assistants to rearrange their habitat. *Advances in neural information processing systems*, 34:251–266, 2021. 6

[79] Yuezhan Tao, Xu Liu, Igor Spasojevic, Saurav Agarwal, and Vijay Kumar. 3d active metric-semantic slam. *IEEE Robotics and Automation Letters*, 9(3): 2989–2996, 2024. 2

[80] Yuezhan Tao, Dexter Ong, Varun Murali, Igor Spasojevic, Pratik Chaudhari, and Vijay Kumar. Rt-guide: Real-time gaussian splatting for information-driven exploration. *arXiv preprint arXiv:2409.18122*, 2024. 3

[81] Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024. 2

[82] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34: 16558–16569, 2021. 2

[83] Eleftherios Triantafyllidis, Filippos Christianos, and Zhibin Li. Intrinsic language-guided exploration for complex long-horizon robotic manipulation tasks. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7493–7500. IEEE, 2024. 3

[84] Darko Trivun, Edin Šalaka, Dinko Osmanković, Jasmin Velagić, and Nedim Osmić. Active slam-based algorithm for autonomous exploration with mobile robot. In *2015 IEEE International Conference on Industrial Technology (ICIT)*, pages 74–79. IEEE, 2015. 2

[85] Teresa Vidal-Calleja, Andrew J Davison, Juan Andrade-Cetto, and David William Murray. Active control for single camera slam. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 1930–1936. IEEE, 2006. 1

[86] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[87] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 4

[88] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 2, 6, 7, 8

[89] Brian Yamauchi. A frontier-based approach for autonomous exploration. In *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*, pages 146–151. IEEE, 1997. 1, 2, 6, 7, 16, 17

[90] Chi Yan, Delin Qu, Dong Wang, Dan Xu, Zhigang Wang, Bin Zhao, and Xuelong Li. Gs-slam: Dense visual slam with 3d gaussian splatting. *arXiv preprint arXiv:2311.11700*, 2023. 1

[91] Dongyu Yan, Jianheng Liu, Fengyu Quan, Haoyao Chen, and Mengmeng Fu. Active implicit object reconstruction using uncertainty-guided next-best-view optimization. *IEEE Robotics and Automation Letters*, 2023. 2

[92] Zike Yan, Haoxiang Yang, and Hongbin Zha. Active neural mapping. In *ICCV*, 2023. 2, 3, 6, 7, 14, 16

[93] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 42–48. IEEE, 2024. 3

[94] Justin Yu, Kush Hari, Kishore Srinivas, Karim El-Refai, Adam Rashid, Chung Min Kim, Justin Kerr, Richard Cheng, Muhammad Zubair Irshad, Ashwin Balakrishna, et al. Language-embedded gaussian splats (legs): Incrementally building room-scale representations with a mobile robot. In *IROS*, 2024. 1

[95] Huangying Zhan, Jiyang Zheng, Yi Xu, Ian Reid, and

Hamid Rezatofighi. Activermap: Radiance field for active mapping and planning, 2022. 2, 3

[96] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[97] Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7641–7649, 2024. 3

[98] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21676–21685, 2024. 1, 8