MATISSE: EVIDENCE-SPACE REASONING FOR ACTIVE 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

How can a 3D reconstruction system acquire and retain useful information to understand the geometry of a scene from partial views under a limited computation budget? Existing active view acquisition methods typically estimate uncertainty over observed or instantiated geometry, limiting their ability to reason about unseen structure, while long-horizon reconstruction methods often retain redundant observations. We introduce *Matisse*, a training-free framework that unifies active reconstruction and keyframe selection by leveraging evidence provided by a pretrained generative 3D model. Matisse estimates Evidential Uncertainty from cross-attention evidence associated with 3D latent tokens and derives an Evidential Information Gain to guide both view acquisition and keyframe selection based on the expected reduction in posterior entropy. Matisse supports multi-object scenes through occlusion-aware, object-balanced aggregation and propagates uncertainty through intermediate latents to avoid full reconstruction during planning. Matisse reduces Chamfer distance by 12.7%, 3.8%, and 9.2% on GSO30, YCB-V, and Replica, respectively, relative to the best baseline on each dataset, and achieves a $1.50\times$ end-to-end speedup over the best active reconstruction baseline on GSO30 with the same reconstruction backend. In the GSO30 keyframe selection experiment for long-horizon reconstruction, Matisse achieves comparable Chamfer distance using 14% of the input views compared with Stream3D.

1 INTRODUCTION

How can a 3D reconstruction system (Chen et al., 2026; Xiang et al., 2026a; Zhang et al., 2026; Schmid et al., 2026) acquire and retain useful information to understand the geometry of a scene from partial views under a limited computation budget? This question connects two problems across computer vision and robotics: *active reconstruction*, which selects the next-best-view to acquire, and *keyframe selection for long-horizon reconstruction*, which selects a subset of views to retain as more frames are collected. Although these problems concern different stages of reconstruction, both require assessing how much information an observation contributes beyond the evidence available.

Existing active reconstruction methods represent scenes using occupancy grids (Zhou et al., 2021; Guédon et al., 2022; 2023), neural radiance fields (Lee et al., 2022; Pan et al., 2022; Yan et al., 2023; Feng et al., 2024), or Gaussian splatting (Jiang et al., 2024; Xie et al., 2025; Li et al., 2025; Chen et al., 2025; Jin et al., 2025; Xue et al., 2026; Jeong et al., 2026; Jun-Seong et al., 2026). These methods typically model uncertainty within the current reconstruction and use it to guide subsequent observations. However, uncertainty defined over observed or instantiated geometry, i.e., reconstruction space, can provide an incomplete account of the priors we might have about unobserved space, resulting in suboptimal view selection. Long-horizon reconstruction using generative models refers to incrementally building and updating a 3D scene with a generative model over an extended sequence of observations. This setting poses a complementary challenge: retaining useful observations while keeping memory use and per-update computation bounded as the input sequence grows. Keyframe selection is a longstanding approach in simultaneous localization and mapping (SLAM) (Mur-Artal et al., 2015; Maggio & Carlone, 2026), but remains less explored in generative reconstruction. Approaches that jointly incorporate all frames through weighted multi-diffusion (Bar-Tal et al., 2023; Li et al., 2026a) become computationally expensive as the sequence grows. Stream3D (Zhou et al., 2026) addresses memory growth through *Adaptive Evidential Memory* (AEM). AEM uses

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

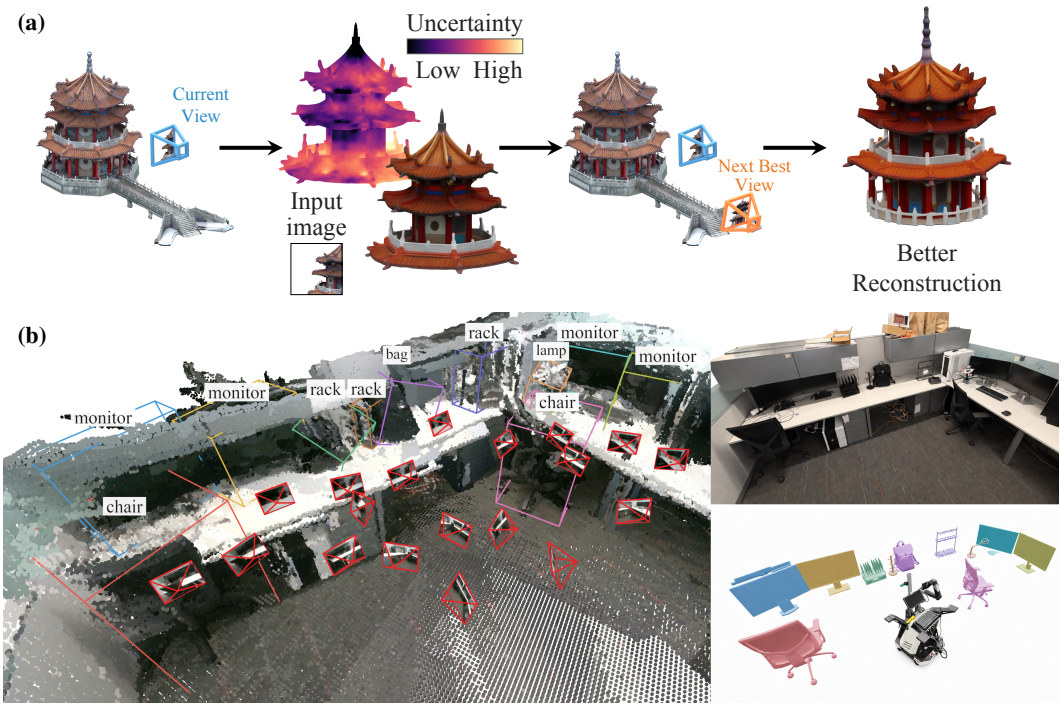


Figure 1: Overview of Matisse. **(a) Active reconstruction workflow.** Given the current observation, Matisse estimates uncertainty over observed and unobserved geometry, selects the next best view, and as more frames are acquired, it decides which keyframes to retain to achieve a more complete reconstruction. **(b) Real robot experiment.** The aggregated depth point cloud with selected camera views and object bounding boxes (left), alongside the real captured scene image (upper right) and reconstructed object meshes (lower right).

cross-attention between image features and 3D query tokens to assign each observation an evidence weight for each token, measuring how strongly that observation supports the token. We refer to these observation-to-token weights, organized over a shared set of spatially aligned 3D tokens, as *evidence space*. This representation makes it possible to aggregate support from different views at corresponding 3D locations. AEM retains a fixed number of high-evidence frame indices per token, bounding the retained evidence while still processing every incoming frame. Selecting keyframes therefore remains an important opportunity for improving efficiency.

These limitations motivate a shared question: Can a unified evidence-space formulation of scene representation, uncertainty, and information gain guide which observations to acquire and retain? We introduce *Matisse*, a training-free framework that represents scenes in the evidence space of a pretrained generative 3D model. From this formulation, Matisse derives uncertainty and information gain to guide both active view selection and long-horizon keyframe selection.

First, to quantify how well the acquired observations support each part of the scene, we introduce *Evidential Uncertainty* (EU). Matisse formulates Adaptive Evidential Memory as a maximum a posteriori (MAP) estimation over 3D latent tokens and derives uncertainty from the posterior covariance. This connects evidence aggregation and uncertainty estimation within a shared, spatially aligned latent representation. Because these tokens represent both observed and generatively completed geometry, EU can characterize uncertainty beyond the currently observed surfaces.

Second, to identify views that provide information beyond the current observations, we introduce *Evidential Information Gain* (EIG). EIG measures the expected reduction in uncertainty from additional evidence. Inspired by FisherRF's information-driven planning (Jiang et al., 2024), Matisse models how additional observations can reduce uncertainty. Whereas FisherRF evaluates information gain through a local approximation over reconstruction-space model parameters, Matisse evaluates it in the evidence space incorporating generative scene hypotheses. This allows acquiring and retaining views, while accounting for evidence relevant to both observed and occluded regions.

Third, we develop efficient *evidence-space view selection*. By defining EU and EIG in evidence space, we score candidate views without expensive mesh generation. Batched GPU-accelerated evaluation further reduces the amortized scoring time to 1.26 ms per candidate.

Fourth, we integrate these components into a scalable reconstruction system. Visibility-aware, object-balanced aggregation extends the information-gain criterion to multi-object scenes, accounting for occlusion while balancing the contributions of individual objects. An occupancy-stability termination test determines when to stop acquiring observations.

We evaluate Matisse on GSO30 single-object reconstruction (Downs et al., 2022), YCB-V tabletop scenes (Calli et al., 2015), and Replica room-scale environments (Straub et al., 2019). Matisse reduces the Chamfer distance (CD) versus the best prior active-reconstruction baselines by 12.7%, 3.8%, and 9.2%, respectively. On GSO30, it reduces P-FID by 9.6% and runs $1.50\times$ faster than the strongest prior active-view planner using the same Stream3D backend. Beyond active reconstruction, the same EIG criterion enables long-horizon keyframe selection, preserving comparable generation quality while using 14% as many input views as Stream3D. These results establish evidence-space reasoning as a shared foundation for deciding where to acquire and which observations to retain.

2 Related Work

Scene Representations for Active Reconstruction. Scene representations have played a central role in active reconstruction. Active reconstruction methods use geometric landmarks and occupancy maps (Sim & Roy, 2005; Bourgault et al., 2002; Stachniss et al., 2005; Carlone et al., 2010; Jadidi et al., 2015; Bircher et al., 2016; Zhou et al., 2021; Guédon et al., 2022; 2023), neural radiance fields (Lee et al., 2022; Pan et al., 2022; Yan et al., 2023; Feng et al., 2024), or Gaussian splatting (Jiang et al., 2024; Xie et al., 2025; Li et al., 2025; Chen et al., 2025; Jin et al., 2025; Xue et al., 2026; Jeong et al., 2026; Jun-Seong et al., 2026) to represent scenes and guide view selection. To reason beyond observed surfaces, learned completion methods predict unobserved geometry (Guédon et al., 2022; 2023; Li et al., 2026b). MAGICIAN (Li et al., 2026b), for example, predicts unseen occupancy and converts it into Gaussian primitives for uncertainty rendering. In contrast, we use a generative 3D foundation model as the reconstruction backbone and maintain the scene as an implicit 3D evidence memory. This formulation provides a learned geometric completion prior for occluded and unobserved regions while permitting efficient uncertainty propagation as new observations arrive.

Uncertainty and Information Gain. Active perception requires an uncertainty measure predicting the utility of candidate observations. Existing methods guide acquisition using uncertainty or confidence measures derived from predictive variance, rendering coverage, transmittance, or geometric support (Pan et al., 2022; Chen et al., 2025; Li et al., 2025; Jin et al., 2025; Xue et al., 2026). However, observing the most uncertain region does not necessarily yield the greatest information gain. Information-driven approaches instead evaluate expected uncertainty reduction (Bourgault et al., 2002; Stachniss et al., 2005; Carlone et al., 2010; Jadidi et al., 2015). FisherRF (Jiang et al., 2024) approximates information gain about reconstruction-model parameters using the curvature of the rendering likelihood, while GauSS-MI (Xie et al., 2025) evaluates mutual information between Gaussian reliability and future observations. However, these formulations remain tied to the instantiated scene representation. FisherRF relies on a local, second-order approximation around the current parameter estimate, while GauSS-MI models the reliability of existing Gaussian primitives. Consequently, they do not explicitly represent the uncertainty or priors over the unobserved portion of the scene. In contrast, we derive uncertainty and information gain from plausible scene completions from a diffusion model. Our formulation is therefore global in scene-hypothesis space, rather than being restricted to a local geometry around the current model parameters.

3 Problem Setup

We study how a 3D reconstruction system can acquire and retain useful information to recover scene geometry from partial views under a limited computation budget. This problem involves two coupled decisions: where to observe next and which observations to retain. At each time step t, a mobile agent acquires an RGB-D observation (I_t, D_t) from camera views v_t with camera pose $c_t \in \mathrm{SE}(3)$. To acquire useful information, the agent selects the next feasible viewpoint c_{t+1} based on its current

estimate of the scene geometry. To retain useful information, the system selects keyframes from the observations collected so far for reconstruction.

4 PRELIMINARY

3D Generative Model. Recent 3D generative models (Xiang et al., 2025; Chen et al., 2026; Xiang et al., 2026a) use a two-stage pipeline: a sparse-structure (SS) stage predicts coarse occupancy, followed by a structured-latent (SLAT) stage that models geometry and appearance. Both stages represent the object using spatially aligned 3D tokens in a canonical coordinate frame. Let $q \in \{1, \dots, Q\}$ index the canonical tokens; for example, $Q = 16^3 = 4096$ in SAM3D.

Adaptive Evidential Memory. Our method builds on Adaptive Evidential Memory (AEM), introduced by Stream3D (Zhou et al., 2026), to combine observations at shared spatial locations in the canonical 3D token grid. For each acquired view $v \in \mathcal{V}_t$, an evidence weight $M_v[q] \geq 0$ measures how strongly the image features support the token at location q. A low weight indicates weak support from that image, as may occur when the corresponding region is occluded or outside the field of view.

For each token, AEM retains up to the D highest evidence weights over all observed views together with the indices of the views that supplied them. These paired entries form the evidence memory $M \in \mathbb{R}^{Q \times D}$ and frame-index memory $F \in \mathbb{N}^{Q \times D}$. To select views for reconstruction, tokens vote for views recorded in their frame-index memory, and these votes are aggregated across tokens to choose K views. The resulting set $\mathcal{V}_t^* \subseteq \mathcal{V}_t$ supplies the features used for reconstruction.

We focus on the sparse-structure stage, which predicts the object's coarse occupancy (Chen et al., 2026). At each token q, the selected views contribute latent features $V_\theta(z_t, v)[q]$, which are combined using their normalized evidence weights:

$$\bar{V}_\theta(z_t)[q] = \sum_{v \in \mathcal{V}_t^*} \bar{M}_v[q] V_\theta(z_t, v)[q], \qquad \bar{M}_v[q] = \frac{M_v[q]}{\sum_{u \in \mathcal{V}_t^*} M_u[q]}. \tag{1}$$

Thus, views with larger evidence weights contribute more strongly to the fused feature at each 3D location. The sparse-structure generator denoises the fused representation and decodes it into an occupancy prediction.

5 METHOD

We propose an evidence-space view selection framework for 3D object reconstruction that builds on Stream3D's canonical evidence memory (Zhou et al., 2026) over sparse 3D tokens. We derive evidential uncertainty from this memory (Section 5.1) and use it to formulate information gain for evaluating candidate views (Section 5.2). We then integrate these quantities into an efficient view-selection system (Section 5.3).

5.1 EVIDENTIAL UNCERTAINTY

Our key idea is to interpret Stream3D's association weights probabilistically: a view that provides stronger support for a 3D token should contribute a more precise estimate of its representation. To formalize this intuition, let $x[q] \in \mathbb{R}^d$ denote the unknown, view-independent representation of canonical token q. We interpret Stream3D's association weight $M_v[q]$ as indicating how strongly we should trust view v's observation of token q. Specifically, consider the Gaussian model

$$p(x[q]) = \mathcal{N}\left(\mu_0, \frac{\sigma^2}{\alpha} I\right), \qquad p(V_\theta(z_t, v)[q] \mid x[q]) = \mathcal{N}\left(x[q], \frac{\sigma^2}{M_v[q]} I\right), \tag{2}$$

where a view with $M_v[q] = 0$ contributes no likelihood term and $\alpha > 0$ is prior pseudo-evidence. This interpretation allows us to derive uncertainty directly from Stream3D's existing evidence memory. With Gaussian conjugacy, we obtain the following result.

Theorem 1: MAP Estimate

Define the accumulated evidence score of token q after observing $\mathcal{V}_t^*$ as

$$E_t[q] := \sum_{v \in \mathcal{V}_t^*} M_v[q], \tag{3}$$

where larger values represent stronger or repeated observational support. Given the prior and measurements with likelihoods as in Eq. (2), the resulting token posterior is

$$\begin{aligned} p(x[q] \mid \mathcal{V}_t^*) &= \mathcal{N}(\mu_t[q], \Sigma_t[q]), \\ \mu_t[q] = \frac{\alpha\mu_0 + \sum_{v \in \mathcal{V}_t^*} M_v[q] V_\theta(z_t, v)[q]}{\alpha + E_t[q]}, &\quad \Sigma_t[q] = \frac{\sigma^2}{\alpha + E_t[q]} I. \end{aligned} \tag{4}$$

Its unique maximum-a-posteriori estimate is $x^{\mathrm{MAP}}[q] = \mu_t[q]$.

The proof is provided in Appendix A.1.

Connection to Stream3D Fusion. Substituting $E_t[q] = \sum_{v \in \mathcal{V}_t^*} M_v[q]$ and Eq. (1) into the posterior mean from Theorem 1 gives the exact relation $\mu_t[q] = \frac{\alpha}{\alpha + E_t[q]}\mu_0 + \frac{E_t[q]}{\alpha + E_t[q]} \bar{V}_\theta(z_t)[q]$. When the prior pseudo-evidence is small relative to the accumulated observation evidence, $\alpha \ll E_t[q]$, this relation reduces to $x^{\mathrm{MAP}}[q] = \mu_t[q] \approx \bar{V}_\theta(z_t)[q]$. Thus, in the evidence-dominated regime, Stream3D's normalized feature fusion is approximately the MAP estimate of the view-independent canonical token representation in Theorem 1.

Beyond the fused representation, the posterior covariance quantifies the remaining uncertainty in each token. Following the A-optimality criterion (Sim & Roy, 2005), we summarize this uncertainty by the covariance trace and normalize it by its prior value.

Definition 2: Evidential Uncertainty

Normalizing the posterior variance by its value before any view is observed defines the bounded token uncertainty

$$U_t[q] := \frac{\operatorname{tr} \Sigma_t[q]}{\operatorname{tr} \Sigma_0[q]} = \frac{\alpha}{\alpha + E_t[q]}. \tag{5}$$

5.2 Evidential Information Gain

For view selection, we evaluate information gain over the predicted occupied canonical tokens, including those representing generatively completed geometry. All subsequent token sums and visibility sets are restricted to these occupied tokens. For each such token, the Gaussian posterior $p(x[q] \mid \mathcal{V}_t^*) = \mathcal{N}(\mu_t[q], \Sigma_t[q])$ allows us to quantify the information gained from a candidate view as the expected reduction in posterior entropy, as formalized below.

Theorem 3: Evidential Information Gain

For a candidate camera c that contributes evidence $M_c[q]$, define the token-wise information gain as its expected reduction in posterior entropy,

$$\mathrm{IG}_t(q, c) := H(x[q] \mid \mathcal{V}_t^*) - H(x[q] \mid \mathcal{V}_t^*, c) = \frac{1}{2} \log \frac{|\Sigma_t[q]|}{|\Sigma_{t+1}[q; c]|} = \frac{d}{2} \log\left(1 + \frac{M_c[q]}{\alpha + E_t[q]}\right). \tag{6}$$

Let $\Omega(c)$ denote the set of canonical tokens visible from candidate camera c. Assume conditionally independent token latents, small incremental candidate evidence, and the binary-visibility approximation $M_c[q] \approx \mathbf{1}[q \in \Omega(c)]$. Then maximizing the candidate information gain is approximated by maximizing

$$\widehat{\mathrm{IG}}_t(c) := \sum_{q \in \Omega(c)} U_t[q]. \tag{7}$$

The proof of Theorem 3 is provided in Appendix A.2. We compute $\Omega(c)$ by projecting the canonical token grid into candidate camera c and applying a depth-buffer visibility test.

5.3 EVIDENCE-SPACE VIEW SELECTION

Building on the evidential uncertainty and information gain derived above, we present a view-selection system that combines multi-object aggregation, efficient candidate scoring, and geometric exploration, and stops acquiring views when the reconstruction stabilizes.

Multi-Object Formulation. For a scene containing N_{obj} reconstructed objects, let $\mathcal{Q}_t^{(i)}$ denote the occupied canonical tokens of object i. We normalize each object's visible information gain by its token count, $\widehat{\text{IG}}_t^{(i)}(c) = \sum_{q\in\Omega^{(i)}(c)} U_t^{(i)}[q]/|\mathcal{Q}_t^{(i)}|$, and average across objects to obtain the scene-level score, $\widehat{\text{IG}}_t^{\text{scene}}(c) = \sum_{i=1}^{N_{\text{obj}}} \widehat{\text{IG}}_t^{(i)}(c)/N_{\text{obj}}$. This gives each object equal weight, preventing denser token grids from dominating view selection. The score requires no additional prediction head: it is computed directly from the evidence memory and decreases as evidence accumulates over a fixed set of visible tokens.

Efficient Evidence-Space View Selection. View selection requires only sparse-structure decoding and online uncertainty updates using Eq. (5), avoiding the expensive SLAT stage. We further batch information-gain calculations across candidate views on the GPU, achieving an amortized scoring time of 1.26 ms per candidate.

Occupancy-Stability Termination. We stop acquiring views once the decoded sparse-structure occupancy stabilizes. Let $\mathcal{P}_t$ and $\mathcal{P}_{t-1}$ be the occupied positions in the canonical 64^3 sparse-structure decoded voxel grid at two consecutive reconstruction steps. For two sets A and B, the symmetric difference $A \triangle B := (A \setminus B) \cup (B \setminus A)$ contains elements belonging to exactly one of the two sets. We measure the relative change as $\Delta_t = |\mathcal{P}_t \triangle \mathcal{P}_{t-1}|/|\mathcal{P}_t \cup \mathcal{P}_{t-1}|$ and terminate if $\Delta_t < \tau_{\text{occ}}$. We also impose a maximum view budget. The test is applied only when a previous reconstruction exists. For a multi-object scene, it is evaluated independently for each object: converged objects are frozen while active objects continue along the shared camera trajectory.

Exploration Factor. Our approximate information-gain score does not explicitly account for proximity to previously acquired camera poses. To encourage exploration, we weight information gain by viewpoint diversity. Let $\mathcal{S}_t$ denote the camera poses acquired by step t and $\rho(c, c')$ the distance between two poses. The distance from candidate c to its nearest acquired pose is $\rho(c, \mathcal{S}_t) = \min_{c'\in\mathcal{S}_t} \rho(c, c')$. We define the exploration factor as $w(c; \mathcal{S}_t) = 1 - \lambda \exp\big(-\rho(c, \mathcal{S}_t)^2/(2\ell^2)\big)$, where λ controls the maximum redundancy penalty and ℓ its distance scale. This factor downweights views near previously acquired poses and approaches one for more distant candidates. Given a user-defined set of feasible camera poses $\mathcal{C}_t$ (see Appendix C.2 for our setup), we select

$$\hat{c}_{t+1} = \arg\max_{c\in\mathcal{C}_t} s_t(c), \qquad s_t(c) := \widehat{\text{IG}}_t^{\text{scene}}(c) w(c; \mathcal{S}_t). \tag{8}$$

Beyond this one-step policy, Matisse also supports receding-horizon planning to account for future rewards, as detailed in Appendix B.

Keyframe Selection. For keyframe selection, we apply the same view-selection rule to a given sequence of posed observations, using the camera poses of unselected frames as the candidate set $\mathcal{C}_t$. Starting from an initial frame, we select the frame with the highest score in Eq. (8) and incorporate its observation into the reconstruction to update the evidence memory and uncertainty. We repeat this process with the remaining frames until the occupancy stabilizes or the view budget is reached, retaining the selected frames as keyframes.

6 EXPERIMENTS

We evaluate *Matisse* on geometric and appearance reconstruction in sparse-view 3D reconstruction. Sections 6.2 and 6.3 present single-object and multi-object reconstruction results, respectively. Section 6.4 presents ablation studies and analyses of evidential uncertainty.

6.1 EXPERIMENTAL SETUP

Implementation. Our experiments are run on two NVIDIA 4090 GPUs. We use SAM3D as the generative backbone, applying its four-step shortcut setting in both SS and SLAT stages. We use

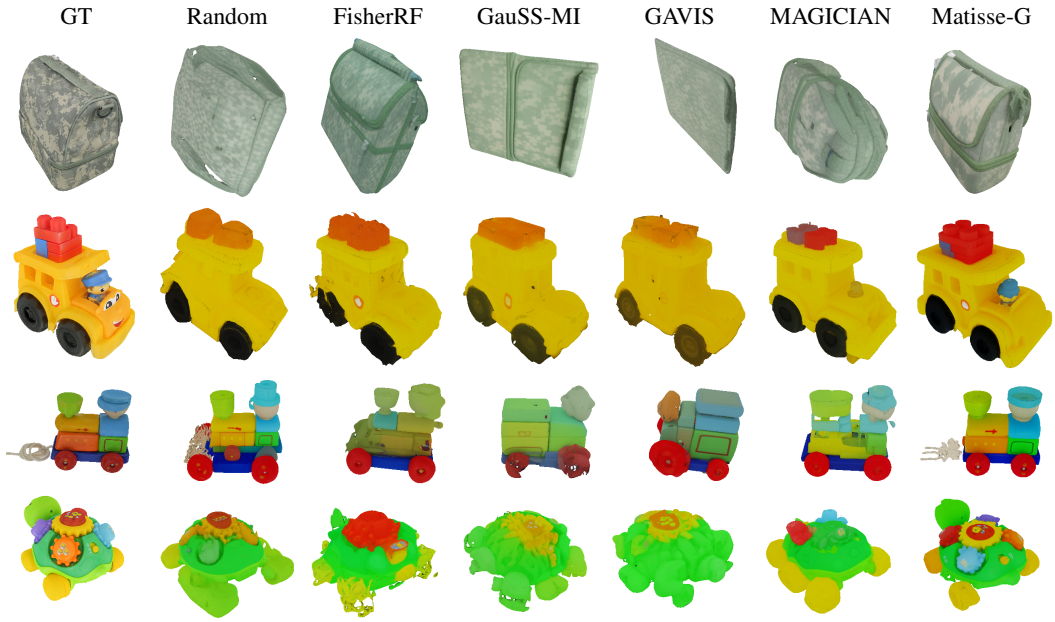


Figure 2: **Qualitative comparisons on GSO30.** Matisse reconstructs geometry and appearance more faithfully than the other active reconstruction baselines.

Stream3D with its standard settings as the multi-view fusion backbone. Each run starts with one view and makes four online updates, each selecting the best next view and fusing the resulting observation into the reconstruction. For YCB-V, we evaluate robustness to initialization with five separate runs. Each uses a five-view budget and starts from one of five initial views evenly spaced along the dataset trajectory. We set $\alpha = 1$; sensitivity to this choice is evaluated in Appendix C.4. For receding-horizon baselines, we use a fixed length $L = 4$, discount $\gamma = 0.9$, and occupancy-stability threshold $\tau_{\text{occ}} = 0.1$. Dataset-specific camera and view-selection settings are provided in Appendix C.2 (Table 5). Matisse-G denotes Matisse with greedy mode and Matisse-RH denotes Matisse with receding-horizon mode. Appendix B has more details on the receding-horizon planner.

Datasets. We use three complementary benchmarks. GSO30 (Downs et al., 2022) contains 30 single-object scenes from Google Scanned Objects. YCB-V (Calli et al., 2015) contains 12 multi-object scenes, with 55 object instances from 21 categories. Following Yu et al. (2022); Ni et al. (2024; 2026), we use room and office scenes from Replica (Straub et al., 2019), manually selecting 54 objects from eight scenes to evaluate active reconstruction in cluttered environments. All methods start from the same initial observation within each trial.

Baselines. We compare Matisse with single-view feedforward methods SAM3D (Chen et al., 2026) and TRELLIS.2 (Xiang et al., 2026a), as well as TRELLIS.2+M.D., which does multi-diffusion (Bar-Tal et al., 2023) without an evidence memory. For the single-view baselines SAM3D and TRELLIS.2, we use only the initial image provided to the active methods. Our active view selection baselines cover different combinations of learned priors and information gain formulations: GAVIS (Xue et al., 2026) is a state-of-the-art method that uses neither, FisherRF (Jiang et al., 2024) and GauSS-MI (Xie et al., 2025) use information gain without a learned prior, and MAGICIAN (Li et al., 2026b) uses a learned prior without an information gain formulation. We use the standard settings for each baseline, with the same 3,500 iterations whenever 3DGS training is required. We additionally include a random walk planner and occupancy-grid planners inspired by Bircher et al. (2016), using occupancy predictions from the sparse-structure stage of SAM3D. For Replica, we disable MAGICIAN's beam search module due to its high computational cost in the dense grid sampling setting.

Metrics. We evaluate geometry with Chamfer distance, mask IoU, and P-FID computed from PointNet++ features of reconstructed point clouds following Zhou et al. (2026). GSO30 and YCB-V use bidirectional CD, whereas Replica uses one-sided CD because only partial ground truth meshes exist. We evaluate appearance with PSNR, SSIM and LPIPS following Ni et al. (2026). We also evaluate the end-to-end system runtime. 3DGS methods use their native open-source backends.

Table 1: GSO30 results. Best/second-best values are green/yellow.

Backend	Method	Geometry			Appearance			Efficiency
		CD (mm) ↓	IoU ↑	P-FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Runtime (s) ↓
Feed-forward	TRELLIS.2 (Xiang et al., 2026a)	108.341	0.554	82.161	13.030	0.821	0.195	26.965±14.716
	TRELLIS.2 M.D. (Xiang et al., 2026a)	115.192	0.571	110.274	13.095	0.829	0.192	62.283±49.878
	SAM3D (Chen et al., 2026)	79.481	0.715	57.440	14.083	0.848	0.173	3.444±0.980
3DGS	FisherRF (Jiang et al., 2024)	76.267	0.583	99.486	12.031	0.765	0.244	49.121±4.834
	GauSS-MI (Xie et al., 2025)	100.909	0.405	119.429	8.243	0.676	0.350	44.200±0.627
	GAVIS (Xue et al., 2026)	92.276	0.528	108.910	10.710	0.754	0.263	100.032±9.184
	MAGICIAN (Li et al., 2026b)	72.139	0.676	71.670	13.022	0.793	0.204	41.852±0.781
Stream3D	Random	62.078	0.736	48.456	14.341	0.846	0.169	16.374±1.566
	Occupancy-F	70.778	0.731	53.528	14.263	0.848	0.168	16.929±1.870
	Occupancy-RH	66.404	0.729	52.263	14.296	0.846	0.171	17.390±1.953
	FisherRF (Jiang et al., 2024)	65.866	0.716	53.443	13.941	0.843	0.176	55.566±4.329
	GauSS-MI (Xie et al., 2025)	88.676	0.686	64.831	13.915	0.842	0.185	53.878±2.785
	GAVIS (Xue et al., 2026)	87.598	0.687	64.394	13.894	0.844	0.187	95.679±7.689
	MAGICIAN (Li et al., 2026b)	63.666	0.733	50.175	14.392	0.848	0.166	23.965±2.626
	Matisse-G (Ours)	54.167	0.765	45.354	14.542	0.849	0.159	16.023±1.397
	Matisse-RH (Ours)	56.228	0.752	48.652	14.357	0.847	0.167	16.922±1.714

6.2 Single-Object Reconstruction

GSO30. Table 1 shows that Matisse-G leads all six reconstruction metrics, reducing CD by 12.7% versus the strongest baselines. Unlike 3DGS-based methods, it completes unobserved geometry without costly optimization before each selection step (Jiang et al., 2024; Xie et al., 2025; Xue et al., 2026), making it 1.50× faster than the fastest prior active-view planner. MAGICIAN (Li et al., 2026b) avoids online Gaussian optimization, but its beam search repeatedly renders and scores future views, taking 3.96 s per object. Figure 2 shows that Matisse more faithfully reconstructs geometry and appearance across four GSO30 objects.

Keyframe Selection. Figure 3 compares Matisse-G with Stream3D on the GSO30 spiral-view experiment (Zhou et al., 2026). Both methods start from the same observation: Matisse-G selects keyframes from the spiral trajectory, while Stream3D processes the incoming views sequentially. The exponential fits summarize their convergence trends. Matisse-G converges after 7 views, compared with 48 for Stream3D, at a comparable Chamfer distance. An additional keyframe-selection experiment on YCB-V is provided in Appendix C.6.

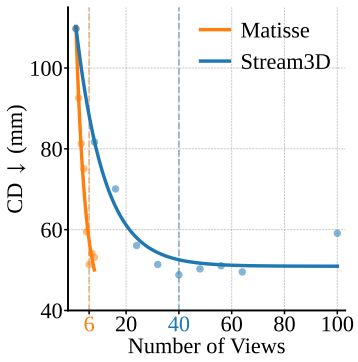


Figure 3: GSO30 keyframe selection with exponential plateau fits.

6.3 Multi-Object Reconstruction

YCB-V. Table 2 shows that Matisse remains effective in multi-object scenes. Matisse-G achieves the best CD, IoU, SSIM, and LPIPS, while Matisse-RH achieves the best P-FID and PSNR. Both variants substantially outperform active view selection baselines with 3DGS backends across all six reconstruction metrics. Qualitative comparisons for two YCB-V multi-object scenes are provided in Appendix C.7 (Figure 5). Figure 4 in Appendix C.5 further compares view budgets on YCB-V. Across budgets ranging from 5 to 50 frames, MAGICIAN's CD remains higher than that of Matisse using only five frames. This highlights a limitation of 3DGS-based reconstruction: adding more views cannot recover fully occluded surfaces, such as those in contact with the ground.

Replica. Table 4 in Appendix C.1 further demonstrates Matisse's effectiveness in cluttered indoor scenes. Matisse-G achieves the best results across all six reconstruction metrics. Compared with the strongest baseline for each metric, it reduces CD and P-FID by approximately 9.2% and 5.2%, respectively, and improves PSNR by 0.39dB. Qualitative comparisons on Replica are provided in Appendix C.8.

Table 2: YCB-V results. Best/second-best values are green/yellow.

Backend	Method	Geometry			Appearance			Efficiency
		CD (mm) ↓	IoU ↑	P-FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	Runtime (s) ↓
Feed-forward	TRELLIS.2 (Xiang et al., 2026a)	16.768	0.665	66.609	18.658	0.915	0.087	99.969±31.493
	TRELLIS.2+M.D. (Xiang et al., 2026a)	18.357	0.634	76.751	18.409	0.914	0.092	145.346±52.187
	SAM3D (Chen et al., 2026)	7.043	0.869	22.561	20.421	0.922	0.065	14.513±3.845
3DGS	FisherRF (Jiang et al., 2024)	18.494	0.332	137.698	16.145	0.915	0.104	109.827±12.062
	GauSS-MI (Xie et al., 2025)	12.482	0.421	116.364	10.145	0.802	0.196	95.851±9.837
	GAVIS (Xue et al., 2026)	12.281	0.681	106.955	14.585	0.884	0.114	143.018±20.217
	MAGICIAN (Li et al., 2026b)	17.515	0.636	90.056	18.006	0.906	0.106	323.777±66.196
Stream3D	Random	6.705	0.877	21.249	20.664	0.923	0.064	88.989±19.588
	Occupancy-F	6.899	0.878	22.419	20.754	0.923	0.064	89.514±20.338
	Occupancy-RH	6.822	0.880	21.731	20.835	0.923	0.063	89.816±20.810
	FisherRF (Jiang et al., 2024)	7.852	0.853	25.680	20.288	0.922	0.069	149.129±21.070
	GauSS-MI (Xie et al., 2025)	6.731	0.877	22.039	20.614	0.922	0.064	139.062±20.414
	GAVIS (Xue et al., 2026)	7.286	0.867	23.875	20.536	0.922	0.066	157.075±18.749
	MAGICIAN (Li et al., 2026b)	7.835	0.853	25.421	20.241	0.921	0.068	92.614±18.497
	Matisse-G (Ours)	6.451	0.885	21.450	20.849	0.923	0.062	88.435±18.932
	Matisse-RH (Ours)	6.474	0.885	21.192	20.912	0.923	0.063	91.185±20.519

Table 3: YCB-V component ablations. Best/second-best values are green/yellow.

Method	Geometry			Appearance		
	CD (mm) ↓	IoU ↑	P-FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
w/o Information Gain	7.403	0.866	23.633	20.489	0.921	0.066
w/o Explorative Factor	6.460	0.884	20.938	20.831	0.923	0.063
Matisse-G (Ours)	6.451	0.885	21.450	20.849	0.923	0.062
Matisse-RH (Ours)	6.474	0.885	21.192	20.912	0.923	0.063

6.4 Ablations and Analysis

Contributions of Information Gain and Exploration. We evaluate the contributions of information gain and exploration on YCB-V (Table 3). Removing information gain leads to the largest overall degradation, highlighting its importance for effective view selection. The variant without the explorative factor remains competitive and achieves the best P-FID. However, the full Matisse-G model achieves the best CD, IoU, SSIM, and LPIPS, while Matisse-RH achieves the best PSNR. These results support combining information gain and exploration to improve reconstruction quality.

Sensitivity to Prior Pseudo-Evidence. We find that $\alpha = 1$ provides the best overall performance, motivating its use in our experiments. Detailed results are provided in Appendix C.4.

Analysis of Evidential Uncertainty. We examine Matisse's predicted uncertainty from three complementary perspectives: its relationship to visibility, object symmetry, and model confidence as visual detail is lost. The corresponding analyses are provided in Appendix D.

7 Conclusion

We presented Matisse, a training-free framework that unifies active 3D reconstruction from sparse views and keyframe selection for long-horizon reconstruction using a generative model. Matisse formulates Adaptive Evidential Memory as MAP estimation, derives Evidential Uncertainty from the posterior covariance, and uses Evidential Information Gain to guide view acquisition and retention. Experiments on GSO30, YCB-V, and Replica demonstrate improved reconstruction quality and computational efficiency across object-level and scene-level settings. For long-horizon 3D generation, Matisse preserves comparable generation quality while retaining only one-seventh of the input views.

Together, these results support our central claim: evidence-space reasoning provides a shared foundation for active acquisition and keyframe selection by quantifying each view's information contribution conditioned on the current 3D memory. Matisse thus offers a scalable path towards active 3D reconstruction. Limitations are provided in Appendix F.

AI USE STATEMENT

We used generative AI tools, including OpenAI ChatGPT and Codex, to brainstorm and refine research ideas, help implement Matisse, design experiments, interpret experimental results, retrieve and identify candidate related work, develop proof sketches and mathematical derivations, draft portions of the manuscript, and revise the writing for clarity and concision. The authors independently evaluated all AI-generated suggestions, verified cited literature against the original sources, checked all mathematical claims and proofs line by line, and validated AI-assisted research artifacts and experimental results against the underlying code and outputs. The authors made all final scientific decisions and take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

We provide the methodological, implementation, dataset, evaluation, and hyperparameter details required to reproduce our experiments in the main paper and appendices. These include the reconstruction and view-selection settings, evaluation protocols, computational resources, and dataset-specific configurations. We use publicly available datasets and identify the external models and baselines used in our evaluation. We plan to publicly release our implementation, experiment configurations, and evaluation scripts following publication.

REFERENCES

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. 2023.

Lucas Berry, Axel Brando, and David Meger. Shedding light on large generative networks: Estimating epistemic uncertainty in diffusion models. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Andreas Bircher, Mina Kamel, Kostas Alexis, Helen Oleynikova, and Roland Siegwart. Receding horizon" next-best-view" planner for 3d exploration. In *2016 IEEE international conference on robotics and automation (ICRA)*, pp. 1462–1468. IEEE, 2016.

Frederic Bourgault, Alexei A Makarenko, Stefan B Williams, Ben Grocholsky, and Hugh F Durrant-Whyte. Information based adaptive robotic exploration. In *IEEE/RSJ international conference on intelligent robots and systems*, volume 1, pp. 540–545. IEEE, 2002.

Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. The YCB object and Model set: Towards common benchmarks for manipulation research. In *Intl. Conf. on Advanced Robotics (ICAR)*, pp. 510–517, Jul. 2015.

Luca Carlone, Jingjing Du, Miguel Kaouk Ng, Basilio Bona, and Marina Indri. An application of kullback-leibler divergence to active slam and exploration with particle filters. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 287–293. IEEE, 2010.

Liyan Chen, Huangying Zhan, Kevin Chen, Xiangyu Xu, Qingan Yan, Changjiang Cai, and Yi Xu. Activegamer: Active gaussian mapping through efficient rendering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16486–16497. IEEE, 2025.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, 2026.

Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google Scanned Objects: A high-quality dataset of 3D scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, 2022. URL `https://arxiv.org/abs/2204.11918`.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21572–21583. IEEE, 2024.

Gianni Franchi, Nacim Belkhir, Dat Nguyen Trong, Guoxuan Xia, and Andrea Pilzer. Towards understanding and quantifying uncertainty for text-to-image generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8062–8072. IEEE, 2025.

Antoine Guédon, Pascal Monasse, and Vincent Lepetit. Scone: Surface coverage optimization in unknown environments by volumetric integration. *Advances in Neural Information Processing Systems*, 35:20731–20743, 2022.

Antoine Guédon, Tom Monnier, Pascal Monasse, and Vincent Lepetit. Macarons: Mapping and coverage anticipation with rgb online self-supervision. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 940–951. IEEE, 2023.

Matthew D Hoffman, Tuan Anh Le, Pavel Sountsov, Christopher Suter, Ben Lee, Vikash K Mansinghka, and Rif A Saurous. Probnerf: Uncertainty-aware inference of 3d shapes from 2d images. In *International Conference on Artificial Intelligence and Statistics*, pp. 10425–10444. PMLR, 2023.

Eliahu Horwitz and Yedid Hoshen. Conffusion: Confidence intervals for diffusion models. *arXiv preprint arXiv:2211.09795*, 2022.

Maani Ghaffari Jadidi, Jaime Valls Miro, and Gamini Dissanayake. Mutual information-based exploration on continuous occupancy maps. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6086–6092. IEEE, 2015.

Metod Jazbec, Eliot Wong-Toi, Guoxuan Xia, Dan Zhang, Eric Nalisnick, and Stephan Mandt. Generative uncertainty in diffusion models. *arXiv preprint arXiv:2502.20946*, 2025.

Seunghoon Jeong, Eunho Lee, Jeongyun Kim, and Ayoung Kim. Informative object-centric next best view for object-aware 3d gaussian splatting in cluttered scenes. *arXiv preprint arXiv:2602.08266*, 2026.

Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information. pp. 422–440, 2024.

Liren Jin, Xingguang Zhong, Yue Pan, Jens Behley, Cyrill Stachniss, and Marija Popović. Activegs: Active scene reconstruction using gaussian splatting. *IEEE Robotics and Automation Letters*, 10 (5):4866–4873, 2025.

Kim Jun-Seong, Tae-Hyun Oh, Eduardo Pérez-Pellitero, and Youngkyoon Jang. Sa-resgs: Self-augmented residual 3d gaussian splatting for next best view selection. *arXiv preprint arXiv:2601.03024*, 2026.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Siqi Kou, Lei Gan, Dequan Wang, Chongxuan Li, and Zhijie Deng. Bayesdiff: Estimating pixel-wise uncertainty in diffusion via bayesian inference. In *International Conference on Learning Representations*, volume 2024, pp. 17046–17063, 2024.

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.

Baicheng Li, Dong Wu, Jun Li, Shunkai Zhou, Zecui Zeng, Lusong Li, and Hongbin Zha. Mv-sam3d: Adaptive multi-view fusion for layout-aware 3d generation. *arXiv preprint arXiv:2603.11633*, 2026a.

Shiyao Li, Antoine Guédon, Shizhe Chen, and Vincent Lepetit. Magician: Efficient long-term planning with imagined gaussians for active mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21606–21615, 2026b.

Yuetao Li, Zijia Kuang, Ting Li, Qun Hao, Zike Yan, Guyue Zhou, and Shaohui Zhang. Activesplat: High-fidelity scene reconstruction through active gaussian splatting. *IEEE Robotics and Automation Letters*, 10(8):8099–8106, 2025.

Dominic Maggio and Luca Carlone. Vggt-slam 2.0: Real-time dense feed-forward scene reconstruction. *arXiv preprint arXiv:2601.19887*, 2026.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

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.

Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Phyrecon: Physically plausible neural scene reconstruction. 2024.

Junfeng Ni, Yixin Chen, Zhifei Yang, Yu Liu, Ruijie Lu, Song-Chun Zhu, and Siyuan Huang. G4splat: Geometry-guided gaussian splatting with generative prior. In *International Conference on Learning Representations*, volume 2026, pp. 138924–138951, 2026.

Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *European Conference on Computer Vision*, pp. 230–246. Springer, 2022.

Katharina Schmid, Nicolas von Lützow, Jozef Hladkỳ, Angela Dai, and Matthias Nießner. Genrecon: Bridging generative priors for multi-view 3d scene reconstruction. *arXiv preprint arXiv:2605.23888*, 2026.

Lukas Schmid, Michael Pantic, Raghav Khanna, Lionel Ott, Roland Siegwart, and Juan Nieto. An efficient sampling-based method for online informative path planning in unknown environments. *IEEE Robotics and Automation Letters*, 5(2):1500–1507, 2020.

Robert Sim and Nicholas Roy. Global a-optimal robot exploration in slam. In *Proceedings of the 2005 IEEE international conference on robotics and automation*, pp. 661–666. IEEE, 2005.

Cyrill Stachniss, Giorgio Grisetti, and Wolfram Burgard. Information gain-based exploration using rao-blackwellized particle filters. In *Robotics: Science and systems*, volume 2, pp. 65–72, 2005.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *Advances in Neural Information Processing Systems*, 36: 12349–12362, 2023.

Jiacheng Wang, Zhedong Zheng, Wei Xu, and Ping Liu. Rigi: Rectifying image-to-3d generation inconsistency via uncertainty-aware learning. *IEEE Transactions on Image Processing*, 2026.

Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *European conference on computer vision*, pp. 456–473. Springer, 2024.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480. IEEE, 2025.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, et al. Native and compact structured latents for 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14419–14429, 2026a.

Zhengrui Xiang, Jiaqi Wu, Fupeng Sun, Heliang Zheng, and Yingzhen Li. Dvd: Discrete voxel diffusion for 3d generation and editing. *arXiv preprint arXiv:2605.07971*, 2026b.

Yuhan Xie, Yixi Cai, Yinqiang Zhang, Lei Yang, and Jia Pan. Gauss-mi: Gaussian splatting shannon mutual information for active 3d reconstruction. 2025.

Shangjie Xue, Jesse Dill, Pranay Mathur, Frank Dellaert, Panagiotis Tsiotra, and Danfei Xu. Neural visibility field for uncertainty-driven active mapping. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18122–18132. IEEE, 2024.

Shangjie Xue, Jesse Dill, Dhruv Ahuja, Frank Dellaert, Panagiotis Tsiotras, and Danfei Xu. Uncertainty-driven 3d gaussian splatting active mapping via anisotropic visibility field. pp. 5014–5026, 2026.

Zike Yan, Haoxiang Yang, and Hongbin Zha. Active neural mapping. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10947–10958. IEEE, 2023.

Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Hao Zhang, Mohamed El Banani, Jen-Hao Cheng, Paul Zhang, Yi Hua, Ben Mildenhall, Christoph Lassner, Narendra Ahuja, and Gengshan Yang. World tracing: Generative pixel-aligned geometry beyond the visible. *arXiv preprint arXiv:2606.13652*, 2026.

Boyu Zhou, Yichen Zhang, Xinyi Chen, and Shaojie Shen. Fuel: Fast uav exploration using incremental frontier structure and hierarchical planning. *IEEE Robotics and Automation Letters*, 6 (2):779–786, 2021.

Kaichen Zhou, Zeyang Bai, Xinhai Chang, Mengyu Wang, Paul Liang, and Fangneng Zhan. Stream3d: Sequential multi-view 3d generation via evidential memory. *arXiv preprint arXiv:2605.21472*, 2026.

A PROOFS

A.1 PROOF OF THEOREM 1 (MAP ESTIMATE)

By Bayes' rule and conditional independence of the observed view features, the negative log-posterior, up to terms independent of $x[q]$, is

$$\begin{aligned}\mathcal{L}(x[q]) &:= -\log p(x[q] \mid \mathcal{V}_t^*) + \text{const.} \\ &= \frac{\alpha}{2\sigma^2}\|x[q] - \mu_0\|_2^2 + \sum_{v \in \mathcal{V}_t^*} \frac{M_v[q]}{2\sigma^2}\|V_\theta(z_t, v)[q] - x[q]\|_2^2, \\ \nabla_{x[q]}\mathcal{L} &= \frac{1}{\sigma^2}\left((\alpha + E_t[q])x[q] - \alpha\mu_0 - \sum_{v \in \mathcal{V}_t^*} M_v[q]V_\theta(z_t, v)[q]\right).\end{aligned} \tag{9}$$

Setting the gradient to zero gives $x^{\text{MAP}}[q] = (\alpha\mu_0 + \sum_{v \in \mathcal{V}_t^*} M_v[q]V_\theta(z_t, v)[q])/(\alpha + E_t[q]) = \mu_t[q]$. Moreover, $\nabla^2\mathcal{L} = (\alpha + E_t[q])I/\sigma^2 \succ 0$ because $\alpha > 0$; hence this stationary point is the unique MAP solution. The same expression shows that posterior precision grows additively with Stream3D evidence. □

A.2 PROOF OF THEOREM 3 (EVIDENTIAL INFORMATION GAIN)

For a d-dimensional Gaussian, $H(x) = \frac{1}{2}\log((2\pi e)^d|\Sigma|)$. Assimilating candidate c adds $M_c[q]/\sigma^2$ to the precision, and therefore $\Sigma_{t+1}[q; c] = \sigma^2 I/(\alpha + E_t[q] + M_c[q])$. Subtracting the two Gaussian entropies proves Eq. (6). Assuming the token latents are conditionally independent, entropy is additive and the view's total information gain is the sum of its token-wise gains. Define $r_q = M_c[q]/(\alpha + E_t[q])$. When each candidate supplies a small increment relative to the current precision, Taylor expansion gives

$$\begin{aligned}\text{IG}_t(c) &= \frac{d}{2}\sum_q \log(1 + r_q) \\ &= \frac{d}{2\alpha}\sum_q M_c[q]U_t[q] + \mathcal{O}\left(\sum_q r_q^2\right) \\ &\approx \frac{d}{2\alpha}\widehat{\text{IG}}_t(c),\end{aligned} \tag{10}$$

where the last line replaces the unknown future association by binary geometric visibility, $M_c[q] \approx \mathbf{1}[q \in \Omega(c)]$. We compute $\Omega(c)$ by projecting the canonical token grid into candidate camera c and applying a depth-buffer visibility test. Since $d/(2\alpha) > 0$ is constant across candidates, removing it does not change the maximizing view. This proves Eq. (7). Equation (7) should therefore be read as a first-order, visibility-based approximation to Bayesian information gain. It avoids predicting the content and confidence of an unobserved image while retaining the desired diminishing return for tokens that are already well supported. □

B RECEDING-HORIZON PLANNING

To account for future rewards, we use receding-horizon planning (Bircher et al., 2016; Schmid et al., 2020; Li et al., 2026b), at each step searching feasible candidate sequences $b = (c_1, \ldots, c_L)$ of horizon L. A branch is scored by its discounted cumulative gain,

$$R_t(b) = \sum_{h=1}^{L} \gamma^{h-1} s_t^{(h)}(c_h), \tag{11}$$

where $\gamma \in (0, 1]$ is the rollout discount. At depth h, the exploration factor includes the acquired cameras and the preceding branch cameras $(c_1, \ldots, c_{h-1})$, while the uncertainty gain remains fixed to the current evidence field. We execute only the first view of the best branch,

$$c_{t+1} = \text{first}\left[\arg\max_{b=(c_1,\ldots,c_L)} R_t(b)\right], \tag{12}$$

then update Stream3D's evidence memory and replan. Hypothetical views never update the evidence field; only executed observations do so.

C Additional Experimental Details and Results

C.1 Quantitative Results on Replica

Table 4: Results on scene-level Replica dataset.

		Geometry			Appearance		
Backend	Method	CD (mm) ↓	IoU ↑	P-FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
3DGS	FisherRF (Jiang et al., 2024)	45.186	0.278	106.041	20.712	0.964	0.054
	GauSS-MI (Xie et al., 2025)	34.642	0.320	103.467	11.582	0.866	0.121
	GAVIS (Xue et al., 2026)	32.732	0.677	64.571	21.699	0.963	0.038
	MAGICIAN (Li et al., 2026b)	63.802	0.556	79.714	19.061	0.948	0.061
Stream3D	Random	23.820	0.708	45.440	23.816	0.966	0.037
	FisherRF (Jiang et al., 2024)	27.301	0.713	49.320	23.843	0.966	0.037
	GauSS-MI (Xie et al., 2025)	21.028	0.738	39.764	23.876	0.966	0.034
	GAVIS (Xue et al., 2026)	23.395	0.715	43.508	23.789	0.966	0.037
	MAGICIAN (Li et al., 2026b)	23.195	0.746	40.243	24.056	0.968	0.032
	Matisse-G (Ours)	19.103	0.761	37.686	24.448	0.968	0.031

C.2 Dataset-Specific Implementation Settings

Table 5: Dataset-specific implementation settings.

Setting	GSO30	YCB-V	Replica
Action space	1 m cube	30° spherical cap	6 m cube
Resolution	256×256	640×480	640×480
Position sampling	Cartesian grid	Equal-area cap	Cartesian grid
Viewing direction	Toward object center	Toward scene center	Toward object center
Budget	5 views per object	5 views per scene	5 views per object
Multi-object aggregation	No	Yes	No
Candidate poses	$3 \times 3 \times 3$	30	$8 \times 8 \times 8$
Distance $\rho(c, c')$	$\|\mathbf{p}_c - \mathbf{p}_{c'}\|_2$	$\arccos(\mathbf{u}_c^\top \mathbf{u}_{c'})$	$\|\mathbf{p}_c - \mathbf{p}_{c'}\|_2$
Penalty scale ℓ	0.5 m	30°	0.5 m

$\mathbf{p}_c$: camera center; $\mathbf{u}_c = (\mathbf{p}_c - \mathbf{o})/\|\mathbf{p}_c - \mathbf{o}\|_2$, with scene center $\mathbf{o}$. Distances use metres or degrees; candidate counts precede validity filtering.

Each candidate pose consists of a sampled camera position and an orientation directed toward the object or scene center in the aggregated depth pointmap. For Cartesian sampling, camera positions lie on a regular 3D grid within the specified cube. For YCB-V, we sample 30 camera positions over the spherical cap using equal-area sampling.

C.3 Real-World Experiment

For the real-world demonstration in Figure 1(b), we deploy Matisse on an AgileX Tracer 2.0 mobile robot equipped with an Intel RealSense Depth Camera D455. The system actively captures 19 frames to reconstruct 11 objects. Because a single frame may provide evidence for multiple objects, we score candidate views using the scene-level aggregation described under *Multi-Object Formulation* in Section 5.3. We partition the objects using a greedy, order-dependent procedure: the first pending object initializes a group, and each subsequent object joins only if its center lies within 0.5 m of every object center already in the group. The resulting groups are processed sequentially.

C.4 Sensitivity to Prior Pseudo-Evidence

We evaluate four logarithmically spaced values of α under the same GSO30 protocol as the main benchmark. As shown in Table 6, performance varies modestly across the tested values, with $\alpha = 1$ achieving the best overall results. We therefore use $\alpha = 1$ as a practical default for all datasets.

Table 6: Sensitivity analysis to prior pseudo-evidence α on GSO30. Green and yellow mark the best and second-best settings; the selected $\alpha = 1$ row is gray.

	Geometry			Appearance		
α	CD (mm) ↓	IoU ↑	P-FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
0.01	58.855	0.7568	48.952	14.410	0.8483	0.1630
0.1	57.483	0.7561	49.078	14.379	0.8480	0.1633
1.0	54.167	0.7650	45.354	14.542	0.8491	0.1593
10.0	54.673	0.7593	47.009	14.541	0.8482	0.1605

C.5 View-Budget Comparison on YCB-V

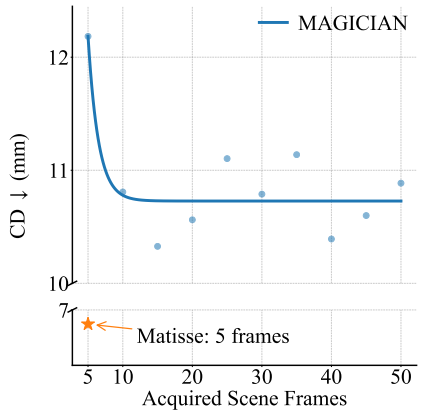


Figure 4: **View-budget comparison on YCB-V scene 48.** Blue points show MAGICIAN's measured Chamfer distance from 5 to 50 acquired frames; the blue curve is a descriptive exponential decay fit with a plateau. The orange star marks Matisse's 5-frame result. The broken vertical axis separates the two CD ranges. Lower CD is better.

C.6 Additional Keyframe-Selection Experiment

We further compare Matisse-G and Stream3D on the fixed BOP19 challenge trajectories of YCB-V. Both methods start from the first frame in the trajectory. Stream3D then processes the next 23 frames sequentially (24 views total), whereas Matisse-G selects three additional views without replacement from the remaining trajectory (4 views total). Table 7 reports the mean CD in each scene. Matisse-G achieves lower CD on nine of the twelve scenes despite using one sixth as many views. Across all 55 object instances, its mean CD is 6.760 mm, compared with 6.994 mm for Stream3D.

Table 7: Additional keyframe-selection results on YCB-V. Entries are scene-wise CD (mm); the final column is the mean over all 55 instances. Lower is better, and bold marks the better method.

	YCB-V scene												
Method	48	49	50	51	52	53	54	55	56	57	58	59	Mean
Stream3D (24 views)	8.775	4.872	**12.294**	6.670	6.833	4.348	7.891	**5.648**	7.436	6.530	6.785	**5.015**	6.994
Matisse-G (4 views)	**6.609**	**4.588**	13.716	**6.562**	**6.722**	**4.023**	**7.699**	5.685	**6.797**	**5.031**	**6.482**	5.983	**6.760**

C.7 Qualitative Results on YCB-V

Figure 5 compares reconstructed geometry and appearance for two multi-object scenes against the ground truth.

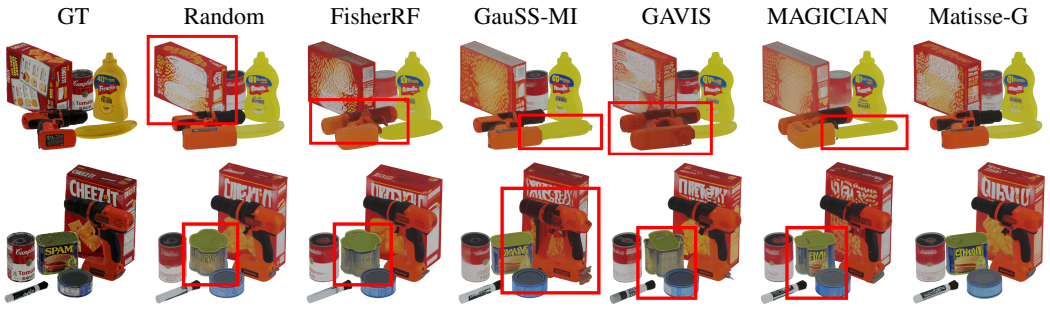


Figure 5: **Qualitative comparisons on YCB-V.** Rows show scene 50 from the fifth camera and scene 59 from the third camera. Columns show ground truth, Random, FisherRF, GauSS-MI, GAVIS, MAGICIAN, and Matisse-G; all reconstruction methods use the Stream3D backend. Predictions are vertex-colored meshes with registration and selected object-orientation corrections for visualization. Ground-truth geometry appears only in the GT column.

C.8 Qualitative Results on Replica

Figure 6 compares reconstructed objects in two Replica office scenes. The common viewpoints show differences in the recovered furniture geometry relative to the ground truth.

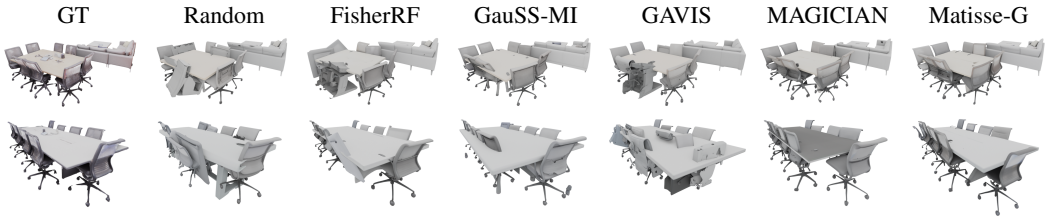


Figure 6: **Qualitative mesh comparisons on Replica.** Rows show office_3 (top) and office_4 (bottom). Columns show ground truth, Random, FisherRF, GauSS-MI, GAVIS, MAGICIAN, and Matisse-G. All reconstruction methods use the Stream3D backend. Registered meshes of the selected objects are shown from a common viewpoint within each row.

D Analysis of Predicted Uncertainty

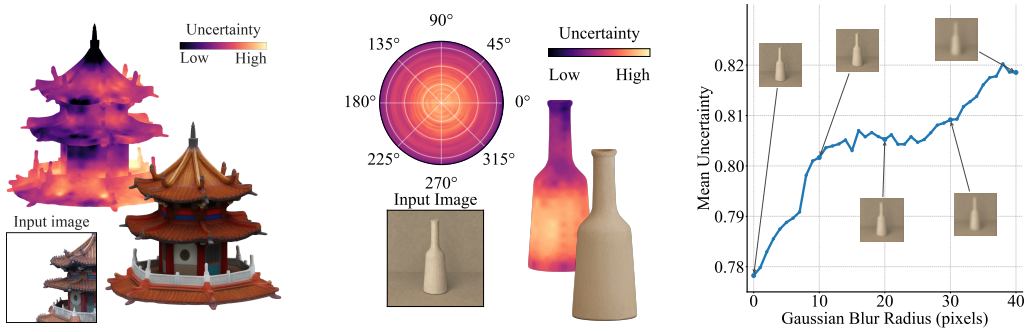
We analyze Matisse's predicted uncertainty from three complementary perspectives: its relationship to visibility, its consistency with object symmetry, and its relationship to model confidence as visual detail is lost.

Visibility. Figure 7(a) illustrates how visibility affects uncertainty. The input image captures only part of the pagoda, with its lower-right region outside the field of view. This unobserved region exhibits higher uncertainty than the visible regions.

Symmetry. We examine whether the predicted uncertainty field reflects the rotational symmetry of a bottle. We compare uncertainty at occupied token locations with that at their counterparts under yaw rotations about the vertical axis. Figure 7(b) shows uncertainty patterns across yaw angles. In the polar plot, the radial axis indexes fixed voxel locations in 3D space, while the angular axis represents the yaw angle. As the object rotates, different object voxels may occupy the same spatial location. Because these voxels are related by the bottle's rotational symmetry, their uncertainty values

remain nearly constant across angles, producing concentric rings in the plot. These results indicate approximate rotational symmetry of the predicted uncertainty field, with small directional differences.

Model Confidence. We examine whether Matisse's uncertainty reflects reduced model confidence as visual detail is lost. We apply progressively stronger Gaussian blur to the same bottle image and rerun the prediction pipeline with fixed inference settings. As shown in Figure 7(c), mean uncertainty over occupied tokens rises overall from 0.778 for the original image to 0.819 at a blur radius of 40 pixels. Although the increase is not strictly monotonic, the overall trend is consistent with reduced model confidence as image quality degrades.



(a) Visibility **(b)** Symmetry **(c)** Model Confidence

Figure 7: **Analysis of Matisse's predicted uncertainty.** (a) The pagoda region outside the input image's field of view exhibits higher uncertainty than visible regions. (b) Uncertainty remains similar at rotationally corresponding locations on the bottle. The concentric rings in the polar plot indicate approximate invariance across yaw angles, consistent with the object's rotational symmetry. (c) Higher mean uncertainty over occupied tokens is consistent with lower model confidence as Gaussian blur removes visual detail from the bottle image.

E Extended Related Work

Scene Representations for Active Reconstruction. Scene representations have played a central role in active perception. Early methods reason about geometric landmarks (Sim & Roy, 2005) or probabilistic occupancy grids (Bourgault et al., 2002; Stachniss et al., 2005; Carlone et al., 2010; Jadidi et al., 2015; Bircher et al., 2016; Zhou et al., 2021; Guédon et al., 2022; 2023). While these representations support efficient visibility and information-gain reasoning, their spatial resolution and memory consumption are constrained by discretization.

More recently, implicit neural fields and Gaussian splatting have enabled photorealistic rendering and high-fidelity novel-view synthesis. Active reconstruction methods based on NeRFs (Lee et al., 2022; Pan et al., 2022; Yan et al., 2023; Feng et al., 2024) and Gaussian splats (Jiang et al., 2024; Xie et al., 2025; Li et al., 2025; Chen et al., 2025; Jin et al., 2025; Xue et al., 2026; Jeong et al., 2026; Jun-Seong et al., 2026) use these representations to evaluate candidate observations. However, their predictions are generally reliable only near observed regions, and updating the scene representation often requires expensive iterative optimization after acquiring new observations.

Several methods introduce learned completion priors to reason beyond current observations (Guédon et al., 2022; 2023). Most closely related to our work, Li et al. (2026b) predict occupancy in unseen space, then convert those predictions into Gaussian primitives for uncertainty rendering, thereby avoiding costly online optimization of a Gaussian field. In contrast, we use a generative 3D foundation model as the reconstruction backbone and maintain the scene as an implicit 3D evidence memory. This formulation provides a learned geometric completion prior for occluded and unobserved regions while permitting efficient uncertainty propagation as new observations arrive.

Uncertainty Modeling for Active Reconstruction. Active perception requires an uncertainty measure that can predict the utility of candidate observations. Early approaches maintain probabilistic occupancy maps and select actions according to expected information gain (Bourgault et al., 2002; Stachniss et al., 2005; Carlone et al., 2010; Jadidi et al., 2015; Bircher et al., 2016). Learned

occupancy-completion methods can additionally estimate uncertainty in unobserved regions (Guédon et al., 2022; 2023).

More recently, radiance-field representations, including implicit NeRFs (Mildenhall et al., 2021) and explicit 3D Gaussian splats (Kerbl et al., 2023), have enabled photorealistic rendering and high-quality novel-view synthesis. Building on these representations, recent neural reconstruction methods derive uncertainty from various properties of the underlying fields and rendering process. NeRF-based methods usually learn an uncertainty field from reconstruction. Pan et al. (2022) predict color variance; Feng et al. (2024) maintain and update an explicit 3D uncertainty field; and Xue et al. (2024) estimate uncertainty from view-dependent transmittance and learn a spatially smooth uncertainty field. Other methods estimate uncertainty from the behavior or information content of the reconstruction parameters. For example, Yan et al. (2023) estimate SDF variability under parameter perturbations. Some methods derive uncertainty directly from quantities involved in the rendering equation. Lee et al. (2022) use the entropy of volumetric rendering weights.

Gaussian-splatting approaches commonly associate uncertainty or confidence with individual primitives. This includes image-space rendering coverage (Chen et al., 2025), combined 2D and 3D coverage (Li et al., 2025), geometrically computed surfel confidence (Jin et al., 2025), measures based on transmittance and directional view similarity (Xue et al., 2026), and uncertainty primarily derived from occupancy predictions (Li et al., 2026b). These approaches use uncertainty to guide view selection. However, observing the most uncertain region does not necessarily yield the greatest information gain: the value of a candidate view depends on how much it is expected to reduce uncertainty, rather than on the current uncertainty alone.

Most closely related to our work, FisherRF (Jiang et al., 2024) estimates parameter-space information from the curvature of the rendering likelihood and approximates the expected information gain of candidate views. Jun-Seong et al. (2026) combine this formulation with a feed-forward reconstruction prior. GauSS-MI (Xie et al., 2025) takes a different information-theoretic approach: it maintains a Bayesian reliability probability for each Gaussian based on rendering residuals and selects views according to the Shannon mutual information between Gaussian reliability and future observations.

However, these formulations remain tied to the currently instantiated scene representation. FisherRF relies on a local, second-order approximation around the current parameter estimate, while GauSS-MI models the reliability of existing Gaussian primitives. Consequently, they do not explicitly represent a distribution over geometries that may appear in the future.

In contrast, we derive uncertainty and information gain from plausible scene completions induced by a diffusion-based generative model. Our formulation is therefore global in scene-hypothesis space, rather than being restricted to a local geometry around the current model parameters. Moreover, the diffusion prior allows the information-gain objective to account for structural predictability, such as symmetry and geometric regularity, and model confidence, such as out-of-distribution uncertainty.

Uncertainty in Generative Models. Uncertainty estimation for generative models generally follows four strategies: training an auxiliary uncertainty predictor (Horwitz & Hoshen, 2022), constructing an ensemble of independently trained models (Berry et al., 2024), estimating a posterior distribution over model parameters (Kou et al., 2024; Jazbec et al., 2025), or querying an external oracle as an uncertainty evaluator (Franchi et al., 2025). These approaches, however, were developed primarily outside 3D shape reconstruction and do not directly address cross-view uncertainty consistency.

In 3D reconstruction, RIGI (Wang et al., 2026) estimates uncertainty in image space and incorporates an uncertainty-aware objective during training, while LatentSplat (Wewer et al., 2024) explicitly predicts the standard deviation of its outputs. Both approaches require uncertainty-specific model training. ProbNeRF (Hoffman et al., 2023) instead estimates pixel-wise uncertainty at inference time by repeatedly sampling latent codes and model weights using Hamiltonian Monte Carlo. Tewari et al. (2023) similarly estimate pixel-wise uncertainty from multiple generations initialized with different noise samples. Because these methods define uncertainty separately in each rendered image, evaluating a new camera view requires multiple additional model evaluations, making uncertainty estimation costly for active perception.

More recently, Xiang et al. (2026b) estimate occupancy-grid entropy at inference time, but do not aggregate uncertainty consistently across multiple observations and viewpoints. In contrast, our method extracts uncertainty from a pretrained 3D generative foundation model without uncertainty-specific

training and fuses it across observations in a shared 3D latent representation. Once constructed, this representation supports consistent and efficient uncertainty evaluation from multiple candidate views. To the best of our knowledge, we are the first to introduce training-free, multi-view uncertainty estimation for a foundational 3D generative model.

F LIMITATIONS

Matisse builds on an underlying pretrained 3D generator and the fusion mechanism of Stream3D (Zhou et al., 2026). Consequently, its reconstruction quality is limited by the learned generator prior and by the ability of Stream3D to associate and fuse observations accurately. Failures in generating plausible geometry, i.e., sparse-structure stage output, or fusing inconsistent observations can propagate to the uncertainty estimates produced by Matisse and, in turn, to its view-selection and keyframe-retention decisions.