XGMAP: EVIDENCE-GROUNDED ACTIVE MAPPING WITH FOUNDATION GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Active 3D mapping requires geometry that reliably guides view selection, not just accurate static reconstruction. While geometric foundation models predict dense depth from monocular RGB, deploying them in closed-loop exploration suffers from multi-view inconsistency and destructive free-space over-carving from monocular depth errors. We present XGMAP, an RGB-only active mapping framework that couples sliding-window depth fusion with an evidence-grounded planner. To prevent depth hallucinations from misleading view selection, XGMAP weights predicted surface coverage by cross-view geometric consistency. A tri-state observation model (surface, free space, unknown) retains surface support despite conflicting free-space updates and uses regions lacking predicted-depth support as exploration targets. Across outdoor landmarks and indoor environments, XGMAP outperforms existing RGB-only baselines, matches or exceeds sensor-privileged RGB-D systems, and substantially improves 3D reconstruction fidelity. Ablations confirm that explicit observation memory and consistency-weighted planning are essential for autonomous visual exploration.

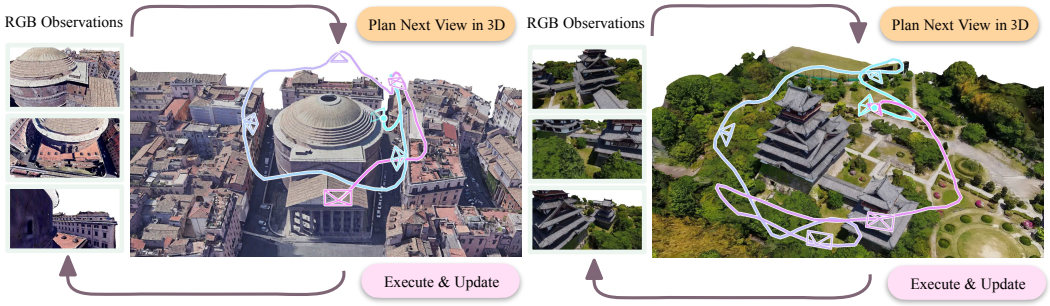


Figure 1: **RGB-only active mapping with XGMAP.** The agent integrates RGB observations into a 3D map, plans the next viewpoint, and updates the map as it explores.

1 INTRODUCTION

Active mapping requires an autonomous agent to decide where to look next while reconstructing an unfamiliar environment. It operates in a closed loop: the current map guides the next action, and the resulting observation updates the map. Under a limited motion budget, the agent must select views that reveal previously unseen parts of the scene. Performing this task using RGB images acquired along a monocular camera trajectory is particularly attractive for large outdoor environments, where reliable depth measurements may be unavailable (Guédon et al., 2023). We study this problem assuming known camera parameters in a common metric coordinate frame, focusing on how to reconstruct reliable scene geometry from images and use it to guide further exploration.

Feedforward 3D foundation models offer an efficient way to recover the scene geometry required for this task. VGGT and MAPANYTHING are scene-agnostic networks that predict dense geometry from images in a single forward pass (Wang et al., 2025a; Keetha et al., 2025). Recent advances

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.

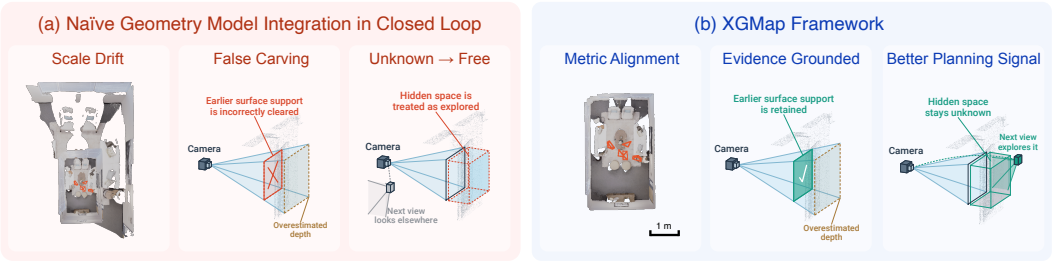


Figure 2: **Foundation geometry for active mapping. (a)** Scale-inconsistent predictions can duplicate surfaces and cause erroneous ray carving, while sparse predictions provide little incentive to explore occluded regions. **(b)** XGMAP integrates geometry using known metric poses and tracks surface, free-space, and unknown evidence to balance coverage and exploration.

further improve the multiview consistency of predictions and support for streaming reconstruction (Wang et al., 2025b; Zhuo et al., 2026; Lin et al., 2025; Chen et al., 2026). These capabilities make foundation models a promising source of scene geometry for active mapping, where the agent, through exploration, repeatedly updates the observation.

To select where to move next, active mapping systems estimate which viewpoint reveals the most previously unobserved surface—a criterion known as surface coverage gain (Guédon et al., 2023; Li et al., 2026). Although 3D foundation models efficiently predict dense geometry, directly deploying them in closed-loop exploration raises two critical vulnerabilities. First, monocular predictions across changing viewpoints can drift and disagree on surface locations, degrading the planner's ability to estimate future visibility. Second, foundation models predict sparse or no geometry in occluded or unobserved regions (Figure 2a). For example, an unvisited courtyard viewed through a fence gap receives sparse returns, but that sparse prediction does not mean the courtyard has been explored. Without an explicit record distinguishing surface support, free-space support, and regions lacking observation support, naive planners can conflate unobserved space with empty air and fail to explore occluded structures (Figure 2b).

We introduce XGMAP, an RGB-only active mapping framework that integrates feed-forward geometric foundation models with evidence-grounded planning. We fuse pose-conditioned depth predictions across sliding camera windows to build a consistent metric map, while simultaneously tracking tri-state observation support (surface, free space, and unknown). Our planner pairs learned surface coverage predictions with an explicit reward for unobserved space lacking ray support. This couples global structural anticipation with an incentive to explore regions lacking predicted-depth support.

We evaluate XGMAP across 15 outdoor landmark scenes on Macarons++ and 18 indoor apartments on Replica. Across both benchmarks, XGMAP achieves higher coverage efficiency than existing RGB-only and RGB-D baselines, while significantly improving 3D reconstruction fidelity over prior camera-only methods. In summary, our contributions are threefold:

- **Geometric foundation-model integration:** A streaming pipeline that fuses depth predictions across sliding camera windows and weights coverage rewards by multi-view consistency to suppress hallucinations.
- **Evidence-grounded active planning:** A dual-objective planner balancing learned coverage with explicit tri-state spatial records (surface, free space, unknown) to prevent false carving and retain exploration incentive.
- **Effective RGB-only exploration:** Rigorous evaluation on Macarons++ and Replica showing state-of-the-art coverage efficiency, with matched ablations validating each planning component.

2 RELATED WORK

Active exploration and view planning. Yamauchi (1997) introduced frontier exploration to target the boundary of known space, while hierarchical systems such as FUEL (Zhou et al., 2021) and TARE (Cao et al., 2021) organize exploration through multi-level planning. Volumetric next-

best-view methods score unresolved space (Bircher et al., 2016) or information gain accounting for uncertainty and visibility (Bai et al., 2016; Isler et al., 2016). PC-NBV learns view gains from point clouds (Zeng et al., 2020), SCONE integrates predicted occupancy and visibility (Guédon et al., 2022), and Guédon et al. (2023) learn RGB reconstruction and coverage prediction online in MACARONS. NextBestPath predicts accumulated coverage toward long-term goals (Li et al., 2025a), while MAGICIAN renders imagined Gaussians for efficient trajectory scoring (Li et al., 2026). We build on MAGICIAN's imagined-Gaussian planner, integrating repeated depth predictions and using their consistency and observation support to guide exploration.

Scene representations and observation evidence. Occupancy maps distinguish occupied, free, and unknown space (Hornung et al., 2013), while NeRF and 3D Gaussian Splatting support rendering-based reconstruction (Mildenhall et al., 2020; Kerbl et al., 2023). ActiveNeRF and NARUTO use learned uncertainty for view acquisition (Pan et al., 2022; Feng et al., 2024); FisherRF uses Fisher information (Jiang et al., 2024), and ActiveGAMER couples Gaussian mapping with coarse-to-fine exploration (Chen et al., 2025). Hybrid representations combine reconstruction and spatial exploration: ActiveGS pairs Gaussian confidence with a voxel map (Jin et al., 2025), while Li et al. (2025b) combine dense Gaussians with sparse workspace topology in ActiveSplat. Our pairing uses predicted-depth evidence and a learned occupancy prior, separating observation support from predicted structure while encouraging exploration of unknown regions.

Feed-forward and streaming geometry. Wang et al. (2024) introduced DUSt3R to predict pairwise point maps from unposed images, followed by MASt3R for matching (Leroy et al., 2024). VGGT generalized feed-forward reconstruction to multiview sequences via alternating attention (Wang et al., 2025a). π^3 removed dependence on anchor views with permutation-equivariant geometry learning (Wang et al., 2025c). MapAnything introduced camera-conditioned metric reconstruction across arbitrary camera models (Keetha et al., 2025). VGGT-Ω scaled feed-forward reconstruction with register attention to cut training memory by 70% (Wang et al., 2026). For sequential observations, streaming architectures extend these models to continuous online mapping: Spann3R (Wang & Agapito, 2025) and CUT3R (Wang et al., 2025b) propagate recurrent spatial state across camera trajectories, while StreamVGGT (Zhuo et al., 2026) and LingBot-Map (Chen et al., 2026) maintain cached geometric context for streaming multi-frame updates. Our focus is integrating geometry into action selection: we reconcile repeated depth predictions across overlapping views and combine their observation evidence with learned coverage.

3 METHODOLOGY

XGMAP is an RGB-only active mapping framework that integrates geometric foundation models into closed-loop view planning (Figure 3). It combines foundation-model depth predictions with streaming multi-view fusion (Section 3.2) to build a consistent metric map, while an evidence-grounded planner (Section 3.5) balances imagined-Gaussian coverage (Section 3.4) against explicit observation memory of unresolved space (Section 3.3). Implementation details and parameter settings are reported in Appendix A.

3.1 PRELIMINARIES

Problem formulation. We consider an agent actively mapping an unknown, static 3D environment under an action budget of B steps. The agent carries a single monocular camera with known intrinsics K and known metric camera-to-world pose $T_t \in \mathrm{SE}(3)$, receiving no physical depth measurements. At step t, the observation history is $\mathcal{H}_t = \{(I_k, T_k)\}_{k=0}^{t}$. From its current pose, the agent selects an action $a_t \in \mathcal{A}_t$ leading to next pose T_{t+1} and observation I_{t+1}, aiming to maximize newly observed surface area while building an accurate 3D reconstruction. Because decisions rely solely on history $\mathcal{H}_t$, exploration requires anticipating which reachable viewpoints will reveal unobserved surfaces before moving to them.

Coverage gain and imagined Gaussians. Ideal surface coverage gain measures the unobserved surface area revealed by a candidate viewpoint v (Guédon et al., 2022; Li et al., 2026). For visited

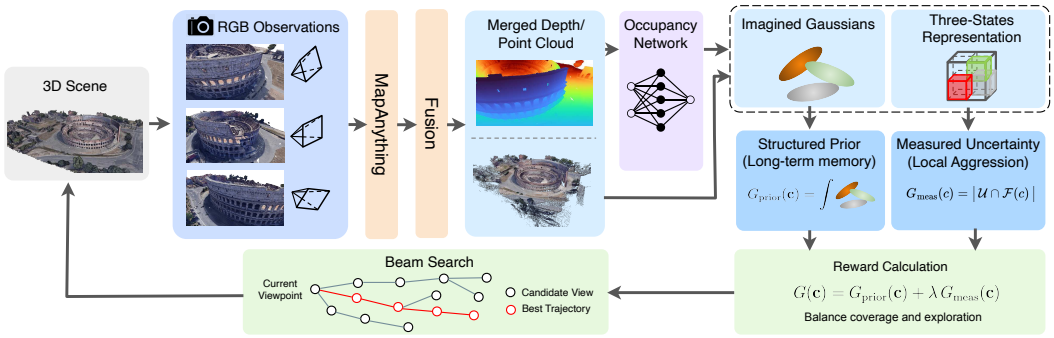


Figure 3: **Overview of XGMAP.** MapAnything predicts depth from RGB images and known metric poses; predictions from overlapping windows are fused into a surface point cloud. Occupancy predictions define imagined Gaussians for estimating unseen surface coverage, while three-state observation evidence identifies unresolved space. These complementary rewards guide receding-horizon beam search. The agent executes the first action, acquires a new observation, and replans.

poses $\mathcal{C}_t = \{T_0, \ldots, T_t\}$ and scene surface $\mathcal{S}$, this ideal gain is:

$$G(v \mid \mathcal{C}_t) = \int_{\mathcal{S} \cap \mathrm{VF}(v)} o(x, v)\, \gamma(x \mid \mathcal{C}_t)\, dA(x), \tag{1}$$

where $\mathrm{VF}(v)$ is the view frustum, $o(x, v)$ indicates visibility, and $\gamma(x \mid \mathcal{C}_t)$ denotes that surface point x is previously unobserved. Because true surface $\mathcal{S}$ is unknown during exploration, MAGICIAN (Li et al., 2026) approximates this integral by instantiating *imagined Gaussians* from a learned occupancy prior and rendering them via fast GPU rasterization. In Section 3.4, we extend this formulation to make it robust to feed-forward depth uncertainty.

3.2 STREAMING MULTI-VIEW DEPTH FUSION

We maintain an incremental surface point cloud $\mathcal{P}_t$ by fusing repeated depth predictions across sliding camera windows.

Sliding-window depth prediction. At step t, we form a causal window $\mathcal{W}_t$ of the N most recent frames. Conditioning a frozen MAPANYTHING model (Keetha et al., 2025) on the window's images, camera intrinsics K, and known metric poses T_i yields projective depth D_i^t and per-pixel confidence C_i^t for each frame $i \in \mathcal{W}_t$.

Confidence-weighted fusion and deferred commitment. Because frame i appears in multiple overlapping windows $\mathcal{W}_k$, its depth is repeatedly re-estimated from varying baselines. We fuse these predictions pixelwise via a confidence-weighted average and compute their cross-window variance:

$$\bar{D}_i(u) = \frac{\sum_{k: i \in \mathcal{W}_k} C_i^k(u) D_i^k(u)}{\sum_{k: i \in \mathcal{W}_k} C_i^k(u)}, \qquad \sigma_i^2(u) = \frac{\sum_{k: i \in \mathcal{W}_k} C_i^k(u) (D_i^k(u))^2}{\sum_{k: i \in \mathcal{W}_k} C_i^k(u)} - \bar{D}_i(u)^2. \tag{2}$$

The variance $\sigma_i^2(u)$ quantifies multi-view depth consistency, which we later use to weight planning rewards (Section 3.4). To allow subsequent views to refine estimates before committing to the map, frame i is finalized after a fixed delay of d steps, at time $t = i + d$. After applying depth-validity, background, and confidence-quantile masks, accepted pixels are unprojected to metric 3D points $p = T_i[\bar{D}_i(u) K^{-1} \tilde{u}; 1]$ and appended to the persistent surface point cloud $\mathcal{P}_t$.

3.3 MEASURED UNCERTAINTY: ACCUMULATED OBSERVATION EVIDENCE

While surface point clouds $\mathcal{P}_t$ record detected geometry, they cannot distinguish cleared free space from occluded or unexplored volume. To guide the agent into uninspected regions, we maintain an explicit volumetric observation record across spatial proxy points $\mathcal{X} = \{x_j\}_{j=1}^M$ sampling the environment. By accumulating ray observations from acquired depth estimates, we identify regions lacking predicted-depth support, defining our *measured-uncertainty* exploration signal.

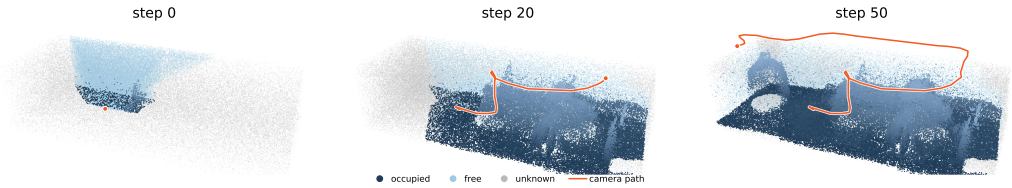


Figure 4: **Accumulated observation evidence in Pisa.** Proxy-point states at exploration steps 0, 20, and 50: occupied (blue, lighter with height), free (light blue), and unknown (gray), with the camera path so far in orange. As depth evidence accumulates, surface and free-space support expand, while unresolved regions remain unknown and provide an exploration signal.

Surface and free-space support. For proxy point $x_i \in \mathcal{X}$ viewed from camera v, let $\delta(x_i, v) = z_v(x_i) - D_v(\pi_v(x_i))$ denote its signed ray-depth residual, where $z_v(x_i)$ is the point's depth and $D_v(\pi_v(x_i))$ is the predicted surface depth along that ray. Points near the predicted surface receive surface support, while points in front receive free-space support:

$$n_{\text{surf}}(x_i) \mathrel{+}= \mathbf{1}[|\delta(x_i, v)| \le \beta], \qquad n_{\text{free}}(x_i) \mathrel{+}= \mathbf{1}[\delta(x_i, v) < -\beta], \tag{3}$$

where β is a narrow support band; occluded points behind the surface ($\delta > \beta$) receive neither. We classify each proxy point into three discrete states based on accumulated evidence:

$$\text{state}(x_i) = \begin{cases} \text{OCCUPIED} & n_{\text{surf}}(x_i) > 0, \\ \text{FREE} & n_{\text{surf}}(x_i) = 0,\ n_{\text{free}}(x_i) > 0, \\ \text{UNKNOWN} & n_{\text{surf}}(x_i) = n_{\text{free}}(x_i) = 0. \end{cases} \tag{4}$$

Crucially, surface support takes precedence over free space: once a proxy point receives surface evidence ($n_{\text{surf}} > 0$), subsequent free-space observations cannot clear it. This asymmetry protects previously supported surfaces from conflicting free-space updates (Figure 2), but does not guarantee correct geometry. Erroneous surface support persists, and points incorrectly assigned free-space support cannot return to the unknown state, although later surface evidence can mark them occupied. Figure 4 illustrates how these evidence states accumulate during exploration.

Unknown space as an exploration signal. Following volumetric next-best-view planning (Bircher et al., 2016), we score candidate viewpoint v by counting unknown proxy points within its viewing frustum $\text{VF}(v, r)$ up to planning range r:

$$G_{\text{meas}}(v) = \sum_{x_i \in \mathcal{X}} \mathbf{1}[\text{state}(x_i) = \text{UNKNOWN}]\mathbf{1}[x_i \in \text{VF}(v, r)]. \tag{5}$$

This score favors views observing unresolved volume without candidate-view occlusion tests, directly quantifying missing observation support.

3.4 Structured prior: imagined-Gaussian coverage score

While observation evidence (Section 3.3) identifies unresolved space, it does not indicate where new surfaces are likely to lie. To anticipate visible surface gain, we instantiate *imagined Gaussians* from a learned occupancy prior (Li et al., 2026), weighting their rendered rewards by multi-view geometric consistency to suppress monocular depth hallucinations.

Gaussian instantiation. A frozen occupancy network (Guédon et al., 2022) takes reconstructed points $\mathcal{P}_t$ as input and predicts occupancy probabilities $p_t(x) \in [0, 1]$ over spatial query points $x \in \mathcal{X}$, extrapolating plausible structure beyond visible surfaces. Proxy points exceeding an occupancy threshold form isotropic imagined Gaussians $\mathcal{G}_t$ with opacities $p_t(x)$ (Li et al., 2026), refreshed as $\mathcal{P}_t$ expands.

Observation novelty from visibility checks. To prevent the agent from repeatedly exploring familiar areas, each imagined Gaussian tracks an observation record $\nu_g \in [0, 1]$ (initialized to 0). Following Li et al. (2026), we render the imagined scene from visited camera poses and project

each Gaussian center x_g into the resulting depth image: if g falls within the field of view and its depth matches the rendered surface within tolerance ϵ_d, it is deemed visible and updated via $\nu_g \leftarrow \max(\nu_g, \mathbf{1}[g \text{ is visible}])$. Thus, $1 - \nu_g$ represents remaining novelty: previously observed Gaussians yield zero reward ($\nu_g = 1$), while unobserved surfaces retain full exploration incentive ($\nu_g = 0$). During hypothetical planning rollouts, we adapt the direction-dependent update from GAVIS (Xue et al., 2026) to reward viewing surfaces from novel angles in subsequent rollout steps (see Appendix A for details).

Geometric quality weighting from depth consistency. Conditioning occupancy on feed-forward depth can produce spurious, floating surfaces. To deter the agent from pursuing these hallucinations, we derive a geometric reliability weight $q_g \in [0, 1]$ from multi-view depth consistency. Each surface point $y_j \in \mathcal{P}_t$ carries cross-window depth variance $\sigma^2_{i_j}(u_j)$ (Equation 2). We rank negative depth variance within each frame and normalize it as $Q_j = (r_j - 1)/(n_{i_j} - 1)$, where r_j is the ascending rank of $-\sigma^2_{i_j}(u_j)$ among the n_{i_j} pixels in frame i_j. Higher scores indicate greater relative depth consistency within that frame. We then transfer this reliability score to each imagined Gaussian g via nearest-neighbor lookup in $\mathcal{P}_t$:

$$q_g = \begin{cases} Q_{j^*} & \text{if } \|x_g - y_{j^*}\| \le r_q, \\ 1 & \text{otherwise,} \end{cases} \tag{6}$$

where $j^* = \arg\min_j \|x_g - y_j\|$ selects the closest reconstructed point. Consequently, Gaussians near inconsistent geometry are down-weighted ($q_g \to 0$) to suppress hallucinations. Conversely, setting $q_g = 1$ when no reconstructed points lie within radius r_q ensures that genuinely unexplored frontiers retain full exploration incentive.

Composite reward and candidate rendering. Having established both observation novelty $(1 - \nu_g)$ and geometric reliability q_g, we assign each imagined Gaussian $g \in \mathcal{G}_t$ a scalar exploration reward $\kappa_g = (1 - \nu_g)q_g$ (encoded as its splat color attribute). The reward combines observation novelty with relative depth consistency, while retaining full quality weight for Gaussians beyond the reconstructed map. For candidate pose v, we render the imagined Gaussians $\mathcal{G}_t$ using Gaussian splatting (Kerbl et al., 2023), yielding a rendered reward map $\widehat{\kappa}_v(u)$ and depth map $\widehat{z}_v(u)$ across image pixels $u \in \Omega$. We sum this reward over valid-depth pixels with quadratic distance attenuation:

$$G_{\text{prior}}(v) = \sum_{u:\widehat{z}_v(u)>0} \widehat{\kappa}_v(u)\, w_d(\widehat{z}_v(u)), \qquad w_d(z) = \min\left(1, \left(\frac{z}{z_0}\right)^2\right). \tag{7}$$

The distance weight $w_d(z)$ counters perspective footprint bias, with reference depth z_0 set relative to the scene extent. The resulting score $G_{\text{prior}}(v)$ provides the structured-prior signal for viewpoint planning.

3.5 Combining measured uncertainty and the structured prior

Measured uncertainty (G_{meas}, Section 3.3) and the structured prior (G_{prior}, Section 3.4) provide complementary signals: the former identifies unobserved space, while the latter anticipates where surfaces lie. We combine both within a receding-horizon beam search, directing the agent toward unexplored regions predicted to contain new scene geometry.

Candidate-relative normalization. Rendered Gaussian coverage (G_{prior}) and spatial proxy counts (G_{meas}) operate on disparate physical units and numerical scales, making fixed-weight combinations brittle across environments. For reachable candidate viewpoints $\mathcal{V}_t$, we compute the empirical mean $\overline{G}_j = \frac{1}{|\mathcal{V}_t|}\sum_{v' \in \mathcal{V}_t} G_j(v')$ for each term $j \in \{\text{prior}, \text{meas}\}$ and normalize with a numerical guard ϵ_G:

$$\mathcal{N}_j(v) = \begin{cases} \dfrac{G_j(v)}{\overline{G}_j} & \overline{G}_j > \epsilon_G, \\ 0 & \text{otherwise,} \end{cases} \qquad G_{\text{blend}}(v) = \begin{cases} \mathcal{N}_{\text{prior}}(v) + \lambda \mathcal{N}_{\text{meas}}(v) & \overline{G}_{\text{prior}} > \epsilon_G, \\ \lambda \mathcal{N}_{\text{meas}}(v) & \text{otherwise.} \end{cases} \tag{8}$$

This normalization yields scale-free scores where $\mathcal{N}_j(v) = 1.0$ represents an average candidate. When the prior provides no reward ($\overline{G}_{\text{prior}} \le \epsilon_G$, e.g., before occupancy beliefs differentiate),

Table 1: Macarons++ results over 15 scenes and five starts per scene, with a 100-action budget. Shaded RGB-D rows use simulator GT depth; RGB-only rows use predicted depth.

Method	Input	AUC ↑	Final cov. ↑
FisherRF (Jiang et al., 2024)	RGB-D	0.352 ± 0.109	0.492 ± 0.132
SCONE (Guédon et al., 2022)	RGB-D	0.541 ± 0.101	0.679 ± 0.098
ActiveGS (Jin et al., 2025)	RGB-D	0.187 ± 0.216	0.206 ± 0.244
MAGICIAN (Li et al., 2026)	RGB-D	0.703 ± 0.104	0.892 ± 0.116
MACARONS (Guédon et al., 2023)	RGB only	0.700 ± 0.071	0.856 ± 0.080
XGMAP	RGB only	0.718 ± 0.093	0.882 ± 0.106

G_{blend} falls back entirely to observation evidence. The blend weight λ is held fixed across all benchmarks and scenes (Appendix A).

Receding-horizon action selection. To avoid short-sighted decisions, we evaluate multi-step trajectories using a beam search of width W over lookahead horizon S (Li et al., 2026). For candidate trajectory $v_{1:S} = (v_1, \ldots, v_S)$, the cumulative objective is:

$$J_t(v_{1:S}) = s_t G_{\text{blend}}(v_1) + \sum_{h=2}^{S} G_{\text{prior}}(v_h \mid v_{1:h-1}), \tag{9}$$

where $s_t = \overline{G}_{\text{prior}}$ (for $\overline{G}_{\text{prior}} > \epsilon_G$, else 1) scales the normalized root blend to match the native magnitude of subsequent lookahead steps. The blended reward is applied only at root v_1 because ray carving requires depth observations, whereas imagined Gaussians can be rendered hypothetically across lookahead poses $v_{2:S}$. The agent executes action a_t transitioning to optimal viewpoint v_1^*, updates its map and evidence upon receiving I_{t+1}, and replans in a receding-horizon loop.

4 EXPERIMENTS

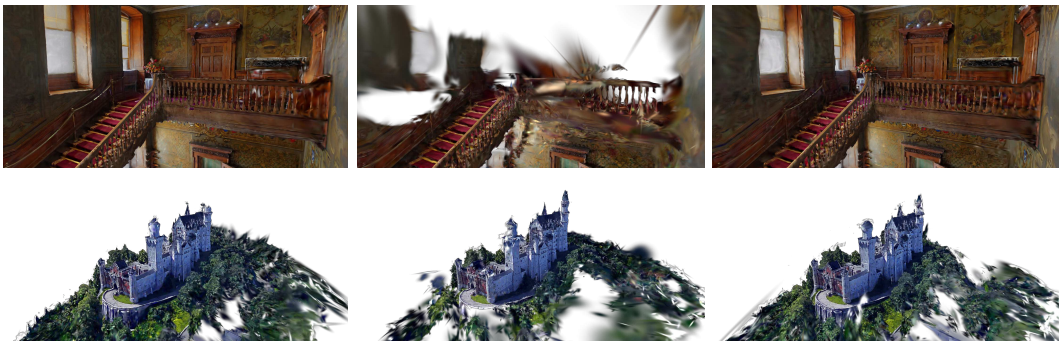
We evaluate XGMAP across outdoor and indoor benchmarks, addressing four questions: **Q1:** How does XGMAP compare with RGB-only and RGB-D methods in coverage and reconstruction? **Q2:** How do predicted coverage (G_{prior}) and measured uncertainty (G_{meas}) interact? **Q3:** Does distinguishing surface, free, and unknown space improve view selection? **Q4:** Can foundation depth simply be plugged into existing planners? Q1 is evaluated on Macarons++ and Replica; matched ablations answer Q2–Q4.

4.1 BENCHMARKS, INPUTS, AND PROTOCOL

We evaluate active mapping under a 100-action budget across two complementary domains. Macarons++ (Li et al., 2026) spans 15 large-scale outdoor landmarks and complex interiors, evaluated from five matched starting poses per scene. Replica (Straub et al., 2019) contains 18 furnished indoor apartments, evaluated from three matched starting poses per scene. All methods share identical initial viewpoints and a common offline evaluator; Appendix B provides protocol details. Experiments run on AMD MI210 GPUs with 64 GB VRAM.

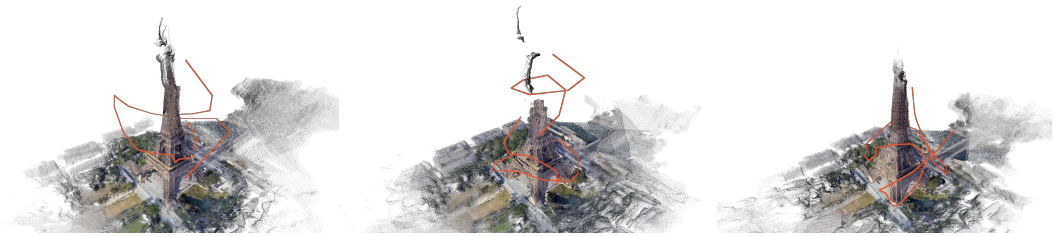
Baselines. Methods are separated by online depth access. The RGB-only group comprises MACARONS (Guédon et al., 2023) and XGMAP, which infer geometry from images alone. The RGB-D group comprises FisherRF (Jiang et al., 2024), SCONE (Guédon et al., 2022), ActiveGS (Jin et al., 2025), and MAGICIAN (Li et al., 2026), which receive simulator ground-truth depth as an upper-bound reference.

Evaluation metrics. Following prior work (Guédon et al., 2023; Li et al., 2026), we measure exploration efficiency via coverage AUC and completeness via final surface coverage, evaluated against ground-truth geometry at reached poses independently of predicted depth. On Replica, we additionally report bidirectional Chamfer distance CD_{L1} and F-score F_1 on accumulated point clouds to assess 3D reconstruction fidelity.



(a) MAGICIAN (b) MACARONS (c) XGMAP (Ours)

Figure 5: **Qualitative reconstruction on Macarons++.** Novel-view renderings from 3D Gaussian Splatting models fitted post-hoc using MILo (Guédon et al., 2025) on images collected along each method's trajectory: (a) MAGICIAN, (b) MACARONS, and (c) XGMAP. Our RGB-only method achieves visual completeness comparable to RGB-D MAGICIAN, while MACARONS leaves substantial regions missing, particularly in the indoor scene.



(a) MAGICIAN (b) MACARONS (c) XGMAP (Ours)

Figure 6: **Point-cloud reconstructions on Macarons++.** Reconstructions of the Eiffel Tower scene from (a) MAGICIAN, (b) MACARONS, and (c) XGMAP, with exploration trajectories overlaid in orange. Our RGB-only method recovers a more complete tower structure than MACARONS, with completeness comparable to RGB-D MAGICIAN.

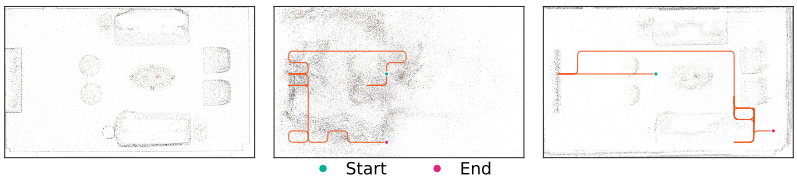
4.2 BENCHMARK COMPARISONS

Macarons++ results (outdoor landmarks). As shown in Table 1, XGMAP achieves the highest coverage efficiency across all evaluated methods, outperforming the RGB-only baseline MACARONS and surpassing even the sensor-privileged RGB-D baselines, while approaching MAGICIAN in final surface coverage. This efficiency stems from our dual-objective planner: explicit observation evidence (G_{meas}) drives the agent into unobserved space where occupancy priors alone stall, accelerating early coverage without physical depth sensors. In contrast, local radiance-field baselines (ActiveGS, FisherRF) get trapped in local extrema on these expansive landmark scenes. Visual reconstructions corroborate these gains: in novel-view 3D Gaussian renderings (Figure 5), XGMAP attains visual completeness on par with RGB-D MAGICIAN, whereas MACARONS leaves substantial structural gaps. Similarly, point-cloud reconstructions (Figure 6) show XGMAP circumnavigating complex landmarks to recover full vertical structure, where MACARONS halts with missing geometry. Notably, XGMAP operates zero-shot using frozen MapAnything, avoiding the benchmark pre-training required by MACARONS (Guédon et al., 2023).

Replica results (indoor apartments). On the indoor Replica benchmark (Table 2), XGMAP leads all evaluated methods in both coverage efficiency and final coverage, outperforming MACARONS and exceeding the RGB-D baselines. Beyond exploration completeness, our streaming multi-view depth fusion and geometric filtering translate these trajectories into high-fidelity 3D reconstruction, dramatically improving Chamfer distance and F_1 score over MACARONS and approaching depth-sensor performance. Top-down reconstructions (Figure 7) reflect these trends: guided by tri-state observation evidence, XGMAP traverses the entire apartment layout to reconstruct sharp, contiguous

Table 2: Replica results across 18 indoor scenes (mean ± SD over three starts; 100-action budget). Shaded rows use simulator GT depth; unshaded rows use predicted depth.

Method	Input	AUC ↑	Final cov. ↑	CD_{L1} ↓	F_1 ↑
FisherRF (Jiang et al., 2024)	RGB-D	0.567 ± 0.096	0.777 ± 0.076	1.282 ± 0.688	0.834 ± 0.057
SCONE (Guédon et al., 2022)		0.742 ± 0.039	0.873 ± 0.038	0.773 ± 0.302	0.889 ± 0.028
ActiveGS (Jin et al., 2025)		0.707 ± 0.050	0.801 ± 0.063	1.772 ± 2.215	0.816 ± 0.067
MAGICIAN (Li et al., 2026)		0.687 ± 0.064	0.884 ± 0.083	0.867 ± 1.119	0.909 ± 0.070
MACARONS (Guédon et al., 2023)	RGB only	0.683 ± 0.049	0.850 ± 0.054	2.486 ± 0.617	0.431 ± 0.043
XGMAP		0.763 ± 0.030	0.905 ± 0.050	1.001 ± 0.608	0.804 ± 0.034



(a) Ground truth (b) MACARONS (c) XGMAP (Ours)

Figure 7: **Qualitative reconstruction on Replica.** Top-down point clouds of `room_0` with trajectories in orange (green/magenta: start/end). XGMAP recovers contiguous room layouts, whereas MACARONS leaves severe structural voids.

room boundaries, whereas MACARONS remains confined to localized clutter near the starting pose with severe omissions.

4.3 Ablation Studies

We evaluate XGMAP's core components across 75 matched trajectories on Macarons++ (Table 3).

Complementary rewards (Q2). Both terms are essential: omitting G_{meas} drops coverage AUC ($\Delta\text{AUC} = -0.071$), confirming observation evidence accelerates early discovery, while removing G_{prior} reduces both efficiency and final coverage.

Three-state evidence (Q3). Replacing tri-state ray carving with binary frustum marking degrades both AUC and final coverage by falsely clearing occluded space.

Direct integration (Q4). Supplying MapAnything depth directly to MAGICIAN yields lower coverage AUC and final coverage, confirming the benefit of our combined framework.

Table 3: Macarons++ planning ablations across 75 trajectories.

Variant	AUC ↑	Final cov. ↑
XGMAP (full control)	**0.718**	**0.882**
w/o G_{meas} reward	0.646	0.857
w/o G_{prior} reward	0.611	0.763
w/o three-state mask	0.667	0.847
Naive MapAnything	0.657	0.836

5 Conclusion

We presented XGMAP, an RGB-only active mapping framework integrating feed-forward foundation geometry into closed-loop view planning. By coupling tri-state ray carving for predicted-depth observation support with imagined Gaussians for long-horizon coverage lookahead, XGMAP outperforms both RGB-only and RGB-D baselines across outdoor and indoor benchmarks. Ablations validate that explicit observation memory and consistency weighting are essential to unlock foundation depth for autonomous exploration.

Limitations and future work. Our approach assumes known poses, static environments, and simulated collision checks; predicted depth noise remains a limitation compared to real sensors. Erroneous surface or free-space assignments may persist in observation memory, hindering full exploration. Future work includes active pose estimation and deployment on physical robots.

AI USE STATEMENT

In this work, we used large language models (specifically Codex, Claude Code, and Pi) for writing assistance—including drafting and line-editing text for conciseness, checking citations against published proceedings, and scripting benchmark automation tools. AI tools were not used for theoretical formulation, synthetic data generation, or result fabrication; all quantitative metrics reflect executed simulation runs. All AI-assisted text and code were verified by the authors, who take full responsibility for the contents of this paper.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our findings, the core methodology, geometry integration pipeline, and planning objectives are formalized in Section 3. Implementation parameters, including inference window size, commitment delay, and beam search parameters, are provided in Appendix A, Table 4. Benchmark settings, coordinate conversions, and evaluation protocols for Macarons++ and Replica are documented in Section 4 and Appendix B; full per-scene Macarons++ results appear in Appendix C.

ETHICS STATEMENT

This research investigates active visual mapping and autonomous view planning for mobile robots in simulated environments (Macarons++ and Replica). The research does not involve human participants, private personal data, or physical robotic platforms that present direct safety risks. The intended applications include autonomous inspection, environmental monitoring, and search-and-rescue operations. We do not foresee direct negative ethical implications or societal harms arising from this work.

REFERENCES

Shi Bai, Jinkun Wang, Fanfei Chen, and Brendan Englot. Information-theoretic exploration with bayesian optimization. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1816–1822, 2016. 3

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. 3, 5

Chao Cao, Hongbiao Zhu, Howie Choset, and Ji Zhang. TARE: A hierarchical framework for efficiently exploring complex 3D environments. In *Robotics: Science and Systems*, 2021. 2

Lin-Zhuo Chen, Jian Gao, Shangzhan Zhang, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, and Yinghao Xu. LingBot-Map: Geometric context transformer for streaming 3D reconstruction. *arXiv preprint arXiv:2604.14141*, 2026. 2, 3

Liyan Chen, Huangying Zhan, Kevin Chen, Xiangyu Xu, Qingan Yan, Changjiang Cai, and Yi Xu. ActiveGAMER: Active GAussian mapping through efficient rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16486–16497, 2025. 3

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*, pp. 21572–21583, 2024. 3

Antoine Guédon, Pascal Monasse, and Vincent Lepetit. SCONE: Surface coverage optimization in unknown environments by volumetric integration. In *Advances in Neural Information Processing Systems*, 2022. 3, 5, 7, 9

Antoine Guédon, Tom Monnier, Pascal Monasse, and Vincent Lepetit. MACARONS: Mapping and coverage anticipation with rgb online self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 7, 8, 9, 14

Antoine Guédon, Diego Gomez, Nissim Maruani, Bingchen Gong, George Drettakis, and Maks Ovsjanikov. MILo: Mesh-in-the-loop gaussian splatting for detailed and efficient surface reconstruction. *ACM Transactions on Graphics*, 44(6), 2025. doi: 10.1145/3763339. 8

Armin Hornung, Kai M. Wurm, Maren Bennewitz, Cyrill Stachniss, and Wolfram Burgard. OctoMap: An efficient probabilistic 3d mapping framework based on octrees. *Autonomous Robots*, 2013. doi: 10.1007/s10514-012-9321-0. 3

Stefan Isler, Reza Sabzevari, Jeffrey Delmerico, and Davide Scaramuzza. An information gain formulation for active volumetric 3D reconstruction. In *IEEE International Conference on Robotics and Automation*, pp. 3477–3484, 2016. 3

Wen Jiang, Boshu Lei, and Kostas Daniilidis. FisherRF: Active view selection and mapping with radiance fields using fisher information. In *European Conference on Computer Vision*, pp. 422–440, 2024. 3, 7, 9

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. doi: 10.1109/LRA.2025.3555149. 3, 7, 9

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 1, 3, 4, 14

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 2023. 3, 6

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision*, 2024. 3

Shiyao Li, Antoine Guédon, Clémentin Boittiaux, Shizhe Chen, and Vincent Lepetit. NextBestPath: Efficient 3D mapping of unseen environments. In *International Conference on Learning Representations*, 2025a. 3

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*, 2026. 2, 3, 4, 5, 7, 9, 12, 13, 14

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, 2025b. doi: 10.1109/LRA.2025.3580331. 3

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 2

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. In *European Conference on Computer Vision*, pp. 405–421, 2020. 3

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, 2022. 3

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Yuheng 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 A. Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint*, 2019. 7, 14

Hengyi Wang and Lourdes Agapito. 3D reconstruction with spatial memory. In *International Conference on 3D Vision*, 2025. 3

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a. 1, 3

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. 3

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10510–10522, 2025b. 2, 3

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024. 3

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025c. 3

Shangjie Xue, Jesse Dill, Dhruv Ahuja, Frank Dellaert, Panagiotis Tsiotras, and Danfei Xu. Uncertainty-driven 3D Gaussian Splatting active mapping via anisotropic visibility field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5014–5026, 2026. 6, 13

Brian Yamauchi. A frontier-based approach for autonomous exploration. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pp. 146–151, 1997. 2

Rui Zeng, Wang Zhao, and Yong-Jin Liu. PC-NBV: A point cloud based deep network for efficient next best view planning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020. 3

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. doi: 10.1109/LRA.2021.3051563. 2

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4D visual geometry transformer. In *International Conference on Learning Representations*, 2026. 2, 3

A Implementation Details

This section details the architectural choices, parameters, and algorithms underlying XGMap, covering causal streaming depth fusion (Section 3.2), structured occupancy priors (Section 3.4), and evidence-grounded view planning (Section 3.5). All hyperparameters remain strictly fixed across all scenes and benchmarks without per-environment tuning. Table 4 summarizes the complete parameter settings grouped by component. Unless otherwise specified, we retain MAGICIAN's implementation settings (Li et al., 2026).

Runtime. On an AMD MI210 GPU, our method takes approximately 10–14 minutes for a 100-action Macarons++ trajectory, averaging 6–8 seconds per exploration step. These estimates are derived from worker-run durations rather than isolated planner timings.

Table 4: Implementation parameters, grouped by component. Planning hyperparameters remain unchanged across benchmarks and scenes.

Parameter	Symbol	Value
Geometry		
Inference window size	N	5 frames
Surface commitment delay	d	4 steps
Surface confidence quantile	—	0.15
Surface-support band	β	0.5 scene units
Structured prior		
Occupancy threshold	—	0.5
Rendering depth reference	z_0	$s_{\text{scene}}/2$
Quality-transfer radius	r_q	1 runtime scene unit
Reference-point sampling cap	—	120,000 points
Planning		
Maximum planning range	r	70 runtime scene units
Measured-uncertainty blend weight	λ	1
Beam width	W	10
Search horizon	S	5 steps
Directional-novelty sharpness	η	3
Normalization threshold	ϵ_G	10^{-9}

Geometry. When unprojecting depth pixels into 3D surface points, we filter out unreliable predictions by discarding simulator background (e.g., sky), out-of-range depths, and the bottom 15% lowest-confidence pixels. Free-space updates use depth predictions passing the validity and background masks, without 15% confidence filtering, to assign free-space support to points in front of predicted surfaces. Macarons++ scenes use nonmetric coordinates; we apply a fixed dataset-level factor of 4.588 to calibrate them to metric scale for MapAnything inference. Replica scenes are already metric and use a factor of 1.0.

Structured prior. To prevent nearby surfaces from dominating rendered coverage due to perspective expansion, rendered rewards are distance-weighted up to reference depth $z_0 = s_{\text{scene}}/2$, where s_{scene} is the average bounding-box dimension of the scene. To assign geometric quality weights q_g, we convert cross-window depth variance into a normalized percentile rank in $[0, 1]$, where the most consistent (lowest-variance) pixels receive weights near 1. When cross-window variance is not yet available for newly observed frames, we rank model confidence as a fallback, defaulting to $q_g = 1$ if neither exists. Each imagined Gaussian inherits this quality weight from the nearest reconstructed surface point within search radius $r_q = 1$. To keep nearest-neighbor lookups fast as the map expands, the reference point cloud is uniformly subsampled to a maximum of $120{,}000$ points.

Planning. We adopt a search horizon of five steps ($S = 5$) rather than MAGICIAN's ten (Li et al., 2026) to limit the compound accumulation of monocular prediction errors during rollouts. At the root expansion, candidate-relative scores are blended and rescaled by $\overline{G}_{\text{prior}}$ when this mean exceeds ϵ_G, as specified in Section 3.5. Otherwise, the root score is $G_{\text{blend}}(v) = \lambda \mathcal{N}_{\text{meas}}(v)$, retaining the observation-evidence reward when the prior provides no reward. During beam expansion, the planner prioritizes collision-free unvisited neighbors, considering feasible revisits only when all unvisited candidates are blocked. Across all 75 Macarons++ and 54 Replica benchmark trajectories, no run encountered empty candidate sets or terminated early.

Directional novelty in planning rollouts. Surfaces observed only from a single perspective (e.g., the front of a landmark) still hold high information value when viewed from novel angles (e.g., the back or side). To reward viewing surfaces from unobserved directions without maintaining expensive spherical harmonics (Xue et al., 2026), we summarize past viewing history using directional statistics. For each Gaussian g, let n_g denote the number of past frames that observed it, and let $m_g = \sum_{k=1}^{n_g} u_k$ be the vector sum of unit rays u_k from the surface toward each observing camera center. The mean viewing direction is $\widehat{m}_g = m_g / \|m_g\|$, and the concentration is $R_g = \|m_g\| / \max(n_g, 1) \in [0, 1]$. We set $\widehat{m}_g = 0$ when $m_g = 0$; without past observations,

$R_g = 0$ as well. When past views are concentrated from one side, the unit vectors align and $R_g \approx 1$; conversely, when views surround the surface from multiple directions, opposing vectors cancel out and $R_g \to 0$.

During planning rollouts, we evaluate candidate viewpoint v along direction $d_g(v)$ via the directional factor:

$$a_g(v) = (1 - R_g) + R_g \exp\left[\eta\left(d_g(v)^\top \widehat{m}_g - 1\right)\right], \tag{10}$$

where $\eta = 3$ controls angular sharpness. After scoring v, child rollout nodes update their observation records as:

$$\nu'_g = \begin{cases} 1 & g \text{ passes the rendered-depth visibility test at } v, \\ a_g(v)\nu_g & \text{otherwise.} \end{cases} \tag{11}$$

Visible Gaussians are marked observed. For Gaussians not visible from v, the directional factor reduces their observedness when the candidate direction differs from concentrated past viewing directions. Familiar directions or dispersed viewing histories give $a_g(v) \approx 1$, leaving their observedness approximately unchanged. This adjustment affects rewards at subsequent rollout steps, not the score of the current candidate. Direction statistics (m_g, R_g) are held fixed during rollouts.

B Benchmark Protocols and Metrics

As summarized in Section 4, we evaluate on Macarons++ (Li et al., 2026) (15 scenes, 5 matched starts, 456×256 resolution) and Replica (Straub et al., 2019) (18 scenes, 3 matched starts, 320×240 resolution), sharing a common camera field of view of $60° \times 91.6°$ (vertical $\times$ horizontal). Across all methods, initial poses and the 100-action budget are matched. Following MAGICIAN (Li et al., 2026), the agent explores within a predefined workspace. Simulator geometry supplies background-validity masks and collision checks and supports offline evaluation; ground-truth depth values are not used for our online reconstruction or reward computation.

Coverage and reconstruction protocols. Following MAGICIAN (Li et al., 2026), reference surfaces for coverage evaluation are constructed from ground-truth depth rendered from accessible viewpoints, avoiding penalties for unobservable surfaces. All methods are evaluated against this common reference. On Replica, geometric reconstruction metrics (CD_{L1} and F_1) are evaluated on point clouds accumulated from each method's online depth predictions (or simulator depth for RGB-D baselines), separating reconstruction fidelity from trajectory coverage. For any early-terminating trajectory, coverage is held constant for the remainder of the budget.

Ablation protocol. Component ablations (Table 3) use the Macarons++ protocol with matched starts, inputs, and search horizon. Removing measured uncertainty sets its weight $\lambda = 0$. Removing imagined Gaussians disables G_{prior} across all rollout depths. The three-state ablation replaces our ray-depth carving with a binary record tracking points that have entered a visited camera frustum. The direct-integration baseline feeds pose-conditioned MapAnything depth (Keetha et al., 2025) directly into MAGICIAN (Li et al., 2026) using its native 10-step horizon, without repeated depth fusion or our evidence-based reward.

Evaluation metrics and aggregation. Following standard active exploration protocols (Guédon et al., 2023; Li et al., 2026), trajectory surface coverage measures the fraction of evaluator-only ground-truth surface within threshold $\epsilon = 1.0$ of visited sensor observations, with coverage AUC computed via trapezoidal integration over the 100-action budget. We report the mean of per-scene means and the root-mean-square of within-scene sample standard deviations across starting poses, weighting scenes equally.

C PER-SCENE MACARONS++ RESULTS

Table 5 provides per-scene coverage AUC and final coverage on Macarons++, complementing the aggregate comparison in Table 1.

Table 5: Per-scene Macarons++ coverage AUC and final coverage (mean ± SD over five starts; higher is better). The best mean in each row is bold. RGB-D methods use simulator GT depth, whereas RGB-only methods predict depth.

	RGB-D				RGB only	
Scene	FisherRF	SCONE	ActiveGS	MAGICIAN	MACARONS	XGMAP
Coverage AUC						
Alhambra	0.200 ± 0.092	0.416 ± 0.196	0.091 ± 0.144	0.680 ± 0.141	0.684 ± 0.048	**0.727** ± 0.015
Bannerman	0.285 ± 0.097	0.381 ± 0.162	0.154 ± 0.201	**0.762** ± 0.046	0.735 ± 0.042	0.725 ± 0.024
Barts	0.677 ± 0.082	0.800 ± 0.087	0.532 ± 0.456	0.820 ± 0.046	0.738 ± 0.058	**0.832** ± 0.057
Bridge	0.327 ± 0.123	0.646 ± 0.126	0.297 ± 0.159	0.737 ± 0.070	0.776 ± 0.042	**0.804** ± 0.011
Colosseum	0.293 ± 0.157	0.478 ± 0.071	0.125 ± 0.224	0.749 ± 0.044	0.658 ± 0.073	**0.750** ± 0.018
Dunnottar	0.167 ± 0.123	0.553 ± 0.118	0.089 ± 0.135	0.712 ± 0.062	0.557 ± 0.112	**0.718** ± 0.037
Eiffel	0.394 ± 0.089	0.707 ± 0.053	0.053 ± 0.094	0.764 ± 0.060	0.740 ± 0.036	**0.798** ± 0.047
Fushimi	0.382 ± 0.177	0.620 ± 0.075	0.291 ± 0.261	0.796 ± 0.041	0.770 ± 0.018	**0.798** ± 0.032
Liberty	0.194 ± 0.095	0.555 ± 0.048	0.197 ± 0.124	0.755 ± 0.050	0.722 ± 0.052	**0.757** ± 0.025
Neuschwanstein	0.221 ± 0.056	0.317 ± 0.122	0.121 ± 0.154	0.536 ± 0.187	**0.619** ± 0.077	0.551 ± 0.246
Pantheon	0.166 ± 0.053	0.235 ± 0.091	0.103 ± 0.081	0.415 ± 0.229	**0.471** ± 0.168	0.401 ± 0.192
Pisa	0.249 ± 0.058	0.439 ± 0.077	0.185 ± 0.115	0.685 ± 0.052	0.681 ± 0.029	**0.710** ± 0.016
Redeemer	0.472 ± 0.085	0.745 ± 0.059	0.182 ± 0.223	0.800 ± 0.021	**0.874** ± 0.016	0.832 ± 0.051
Sestino Museum	0.598 ± 0.161	0.410 ± 0.031	0.044 ± 0.040	0.595 ± 0.170	**0.728** ± 0.065	0.585 ± 0.133
Sofia Church	0.658 ± 0.093	**0.811** ± 0.022	0.342 ± 0.384	0.734 ± 0.038	0.742 ± 0.063	0.776 ± 0.027
Final coverage						
Alhambra	0.302 ± 0.110	0.490 ± 0.196	0.096 ± 0.152	**0.913** ± 0.015	0.861 ± 0.058	0.908 ± 0.018
Bannerman	0.399 ± 0.091	0.509 ± 0.110	0.199 ± 0.292	**0.928** ± 0.027	0.904 ± 0.033	0.921 ± 0.018
Barts	0.957 ± 0.030	0.925 ± 0.084	0.611 ± 0.527	**0.995** ± 0.001	0.882 ± 0.104	0.990 ± 0.009
Bridge	0.578 ± 0.217	0.825 ± 0.136	0.318 ± 0.172	**0.975** ± 0.006	0.937 ± 0.021	0.958 ± 0.023
Colosseum	0.380 ± 0.207	0.593 ± 0.032	0.134 ± 0.245	**0.917** ± 0.009	0.753 ± 0.085	0.917 ± 0.011
Dunnottar	0.243 ± 0.220	0.715 ± 0.099	0.099 ± 0.153	**0.978** ± 0.009	0.722 ± 0.111	0.938 ± 0.031
Eiffel	0.480 ± 0.067	0.883 ± 0.038	0.053 ± 0.094	0.950 ± 0.043	0.909 ± 0.015	**0.961** ± 0.008
Fushimi	0.480 ± 0.197	0.792 ± 0.054	0.305 ± 0.277	**0.960** ± 0.016	0.886 ± 0.023	0.934 ± 0.038
Liberty	0.289 ± 0.147	0.761 ± 0.035	0.220 ± 0.139	**0.917** ± 0.001	0.909 ± 0.003	0.911 ± 0.006
Neuschwanstein	0.276 ± 0.094	0.441 ± 0.105	0.131 ± 0.164	0.775 ± 0.224	**0.782** ± 0.121	0.747 ± 0.278
Pantheon	0.243 ± 0.104	0.318 ± 0.138	0.108 ± 0.086	0.527 ± 0.299	**0.599** ± 0.206	0.539 ± 0.259
Pisa	0.303 ± 0.060	0.601 ± 0.128	0.201 ± 0.128	**0.894** ± 0.008	0.864 ± 0.036	0.882 ± 0.014
Redeemer	0.643 ± 0.097	0.890 ± 0.033	0.191 ± 0.236	0.984 ± 0.004	**0.985** ± 0.007	0.968 ± 0.024
Sestino Museum	**0.933** ± 0.062	0.494 ± 0.061	0.044 ± 0.040	0.786 ± 0.240	0.930 ± 0.039	0.729 ± 0.133
Sofia Church	0.878 ± 0.032	**0.954** ± 0.017	0.374 ± 0.427	0.888 ± 0.045	0.913 ± 0.031	0.924 ± 0.020

D ADDITIONAL QUALITATIVE RESULTS

Figure 8 shows reconstructions from views selected by XGMAP across all 15 Macarons++ scenes. Figure 9 provides corresponding qualitative results across the 18 Replica scenes.

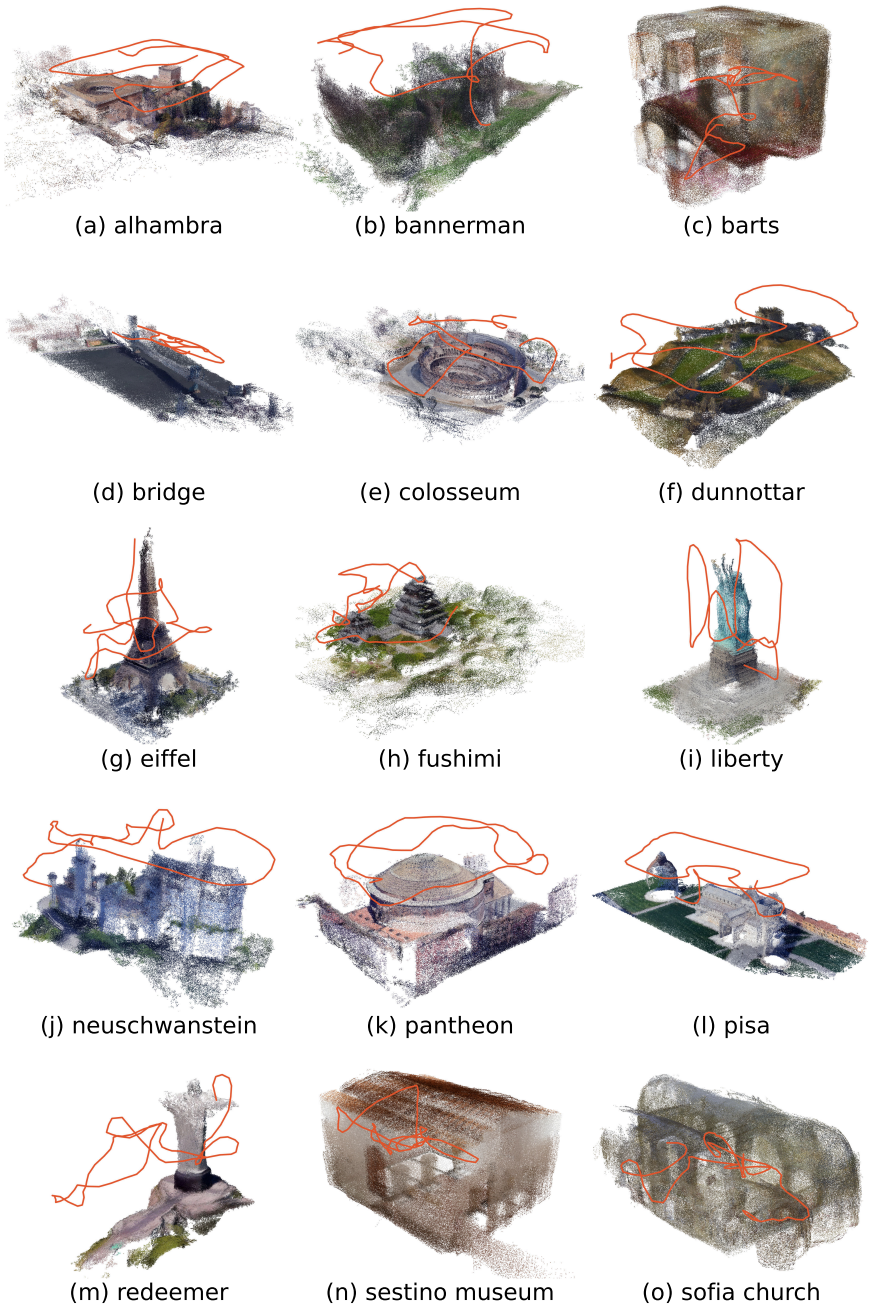


Figure 8: **Additional qualitative results on Macarons++.** Point clouds are reconstructed offline with our depth model from images acquired along XGMAP trajectories, shown in orange. For visualization, object-centric outdoor scenes are cropped to the evaluation region with a margin and low-density outliers are removed; trajectories are smoothed for legibility.

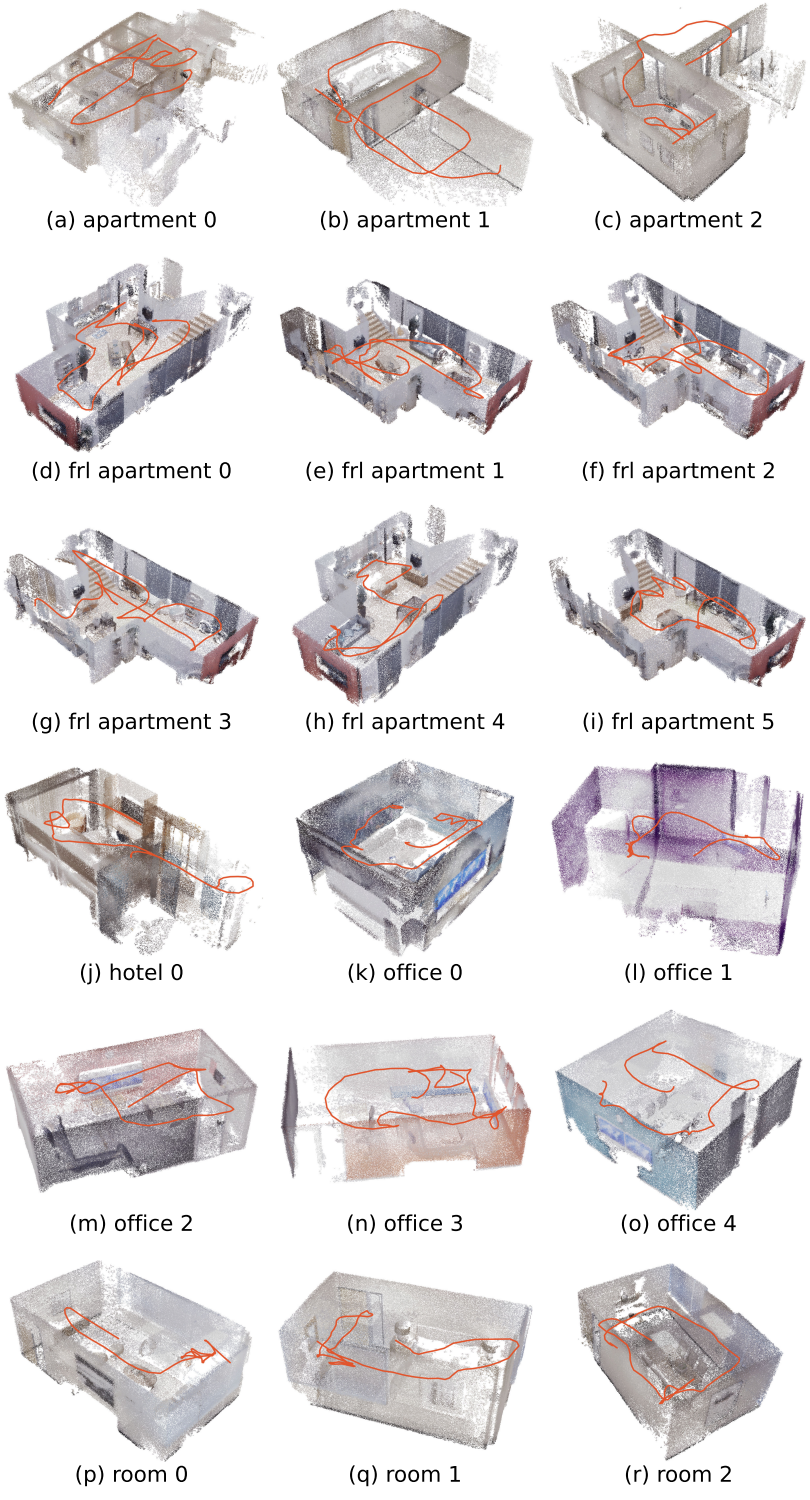


Figure 9: **Additional qualitative results on Replica.** Point clouds are reconstructed offline from the RGB images and camera poses acquired by XGMAP, without ground-truth depth. Orange curves show exploration trajectories, smoothed for visualization; ceilings are removed to reveal interior structure.

E FAILURE CASES

Figure 10 illustrates a representative failure mode on Replica's `apartment_0`, characterized by incomplete exploration in partitioned multi-room environments. While XGMAP thoroughly maps the main living quarters with sharp structural boundaries, it fails to traverse the narrow connecting hallway into the lower wing before the 100-action budget is exhausted, leaving that portion of the apartment unmapped.

This failure highlights two primary challenges in monocular active mapping. First, a finite lookahead horizon ($S = 5$) can bias the planner toward abundant unresolved surfaces within the current room rather than committing actions to low-reward transit corridors that lead to distant spaces. Second, monocular depth uncertainty near doorframes can lead to conservative local view planning around narrow openings. Coupling local evidence-guided beam search with a topological or hierarchical global frontier goal represents a promising direction to prevent agents from stalling in room-level bottlenecks.

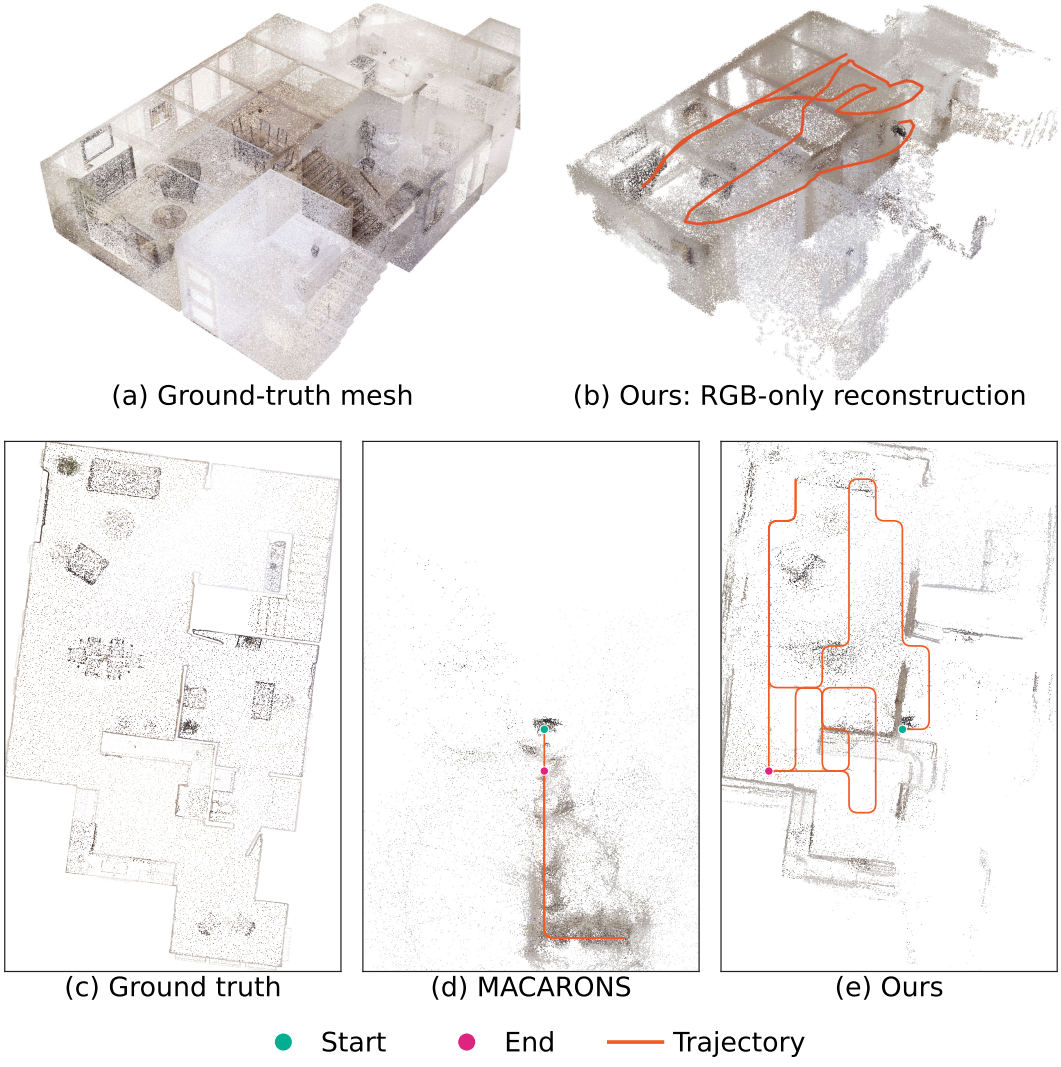


Figure 10: **Incomplete exploration on Replica `apartment_0`.** Top: (a) ground-truth mesh and (b) our RGB-only reconstruction from a shared viewpoint. Bottom: top-down views of (c) ground truth, (d) MACARONS, and (e) ours. Ceilings are removed for visibility, and trajectories are smoothed for display. Panels (b) and (e) illustrate a run where the agent explores the main living area but leaves the lower wing unvisited.