REMEMBER TO BE CURIOUS: EPISODIC CONTEXT AND PERSISTENT WORLDS FOR 3D EXPLORATION

Anonymous authors
Paper under double-blind review

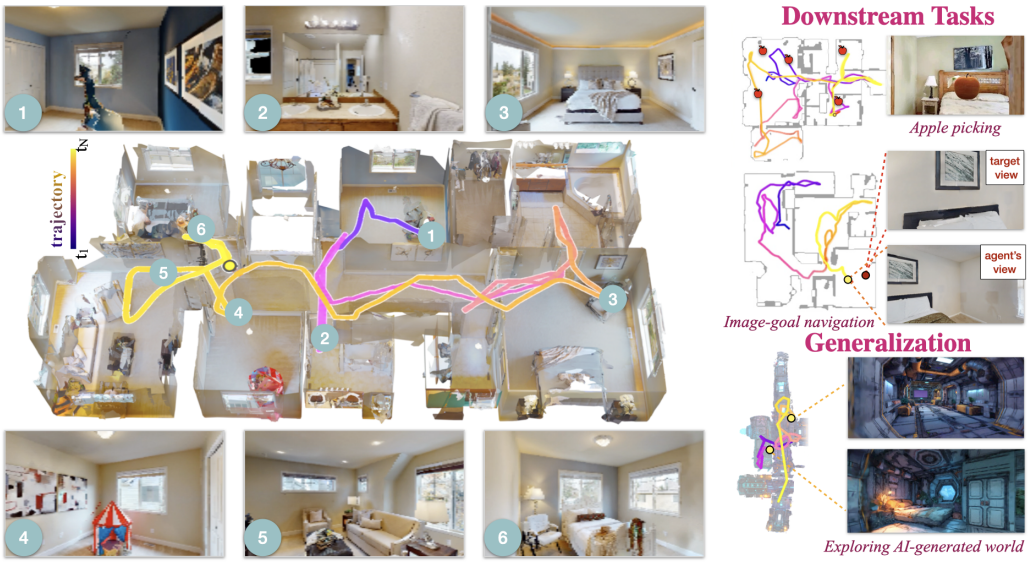


Figure 1: **Remember to be Curious.** Test-time trajectories of our curiosity-driven policy trained with episodic context and a persistent world model. **Left:** Our agent explores an indoor scene purely from an image stream; waypoints show the agent's views along the path. **Right:** Our end-to-end design enables fine-tuning to downstream sparse-reward tasks and zero-shot generalization to AI-generated worlds. Note: the model requires no explicit mapping at test time; trajectories are overlaid on bird's-eye-view for visualization only.

ABSTRACT

Exploration is a prerequisite for learning useful behaviors in sparse-reward, long-horizon tasks, particularly in 3D environments. Curiosity-driven reinforcement learning addresses this via intrinsic rewards derived from the mismatch between the agent's predictive model of the world and reality. However, translating this intrinsic motivation to complex, photorealistic environments remains difficult, as agents can become trapped in local loops and receive fresh rewards for revisiting forgotten states. We show that this failure of prediction-error curiosity stems from a lack of spatial persistence and episodic context. Effective curiosity requires a model of the world that is persistent and continuously updated, paired with an agent that maintains an episodic trajectory history to navigate toward novel regions. We achieve this using an online 3D reconstruction as a persistent model of the world, while the agent policy is parameterized as a sequence model over RGB observations to maintain episodic context. This design enables effective exploration during training while allowing the agent to navigate using solely RGB frames at deployment. Trained purely via curiosity on HM3D, our deployed agent effectively explores 3D scenes using only RGB input. We validate the strength of this curiosity-driven behavior by showing that it outperforms specialized active-mapping policies with matching input and generalizes zero-shot to Gibson and AI-generated worlds. Further, our end-to-end policy adapts efficiently to downstream tasks such as apple picking and image-goal navigation, demonstrating the flexibility and transferability of behavior that emerges from curiosity in 3D worlds.

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.

1 INTRODUCTION

In his seminal work on latent learning, Edward Tolman (1948) demonstrated that agents—such as rats navigating a maze—can acquire complex knowledge of their environment even in the absence of explicit rewards. This suggests that intelligent agents possess an inherent drive to explore, with curiosity serving as a mechanism for engaging with novel and uncertain environments. This phenomenon is intuitively observable in any toddler at a playground: they are never paralyzed by the absence of an explicit goal. Rather than waiting for a sparse external reward, it is far more natural for them to engage in unstructured play, freely wandering and exploring their surroundings.

In the context of artificial agents, this exploratory drive has the practical benefit of densifying sparse rewards in long-horizon tasks, such as visuomotor navigation. Despite the progress in goal-conditioned navigation (Zhu et al., 2017; Wijmans et al., 2020), in real-world scenarios, these tasks fundamentally carry *sparse* success signals. Curiosity-driven exploration offers a compelling alternative, in which the agent derives intrinsic reward from surprise, a signal produced by the prediction error of a world model — trained alongside the agent — to anticipate the consequences of its actions (Schmidhuber, 1991; Pathak et al., 2017). However, scaling curiosity with end-to-end policies to complex photorealistic environments remains challenging, as agents frequently collapse into repetitive behavior driven by cyclic curiosity rewards.

In this work, we show that enabling curiosity requires both a persistent model of the world and an agent equipped with episodic context. This dual requirement addresses a two-fold problem of "amnesiac" exploration: without historical context, agents repeatedly revisit the same locations; simultaneously, without a persistent and continuously updating world model, predictive errors spuriously arise in these revisited areas, yielding false novelty rewards for forgotten states.



To study this, we focus on agents exploring static 3D photorealistic scenes. Ideally, the world model driving curiosity should be dynamic and continuously refined through lifelong agentic experience—a paradigm where action-conditioned video models show promise. However, spatial persistence remains a critical bottleneck for these models; for instance, LingBot-World (Robbyant Team, 2026) fails to maintain scene consistency, generating an entirely different image after a simple in-place 360° turn (see inset figure). We therefore utilize a state-of-the-art online 3D reconstruction method (3DGS) as a proxy, as it enables spatial persistence while continuously updating itself online. Traditional methods like ICM (Pathak et al., 2017) lack this property, as their learned world model acts as a statistical prior over lifelong experience rather than an episodic record of the environment. Indeed, our experiments confirm that when the persistent memory is artificially capped to a short-term window, exploration capability deteriorates significantly.

On the agent side, we build a transformer-based policy that operates purely on a sequence of RGB observations. By conditioning the policy on the agent's recent RGB history, we enable it to form its own internal representation of the episodic context directly from pixels. This is in direct contrast to existing methods that bypass the exploration challenge by embedding *explicit geometric maps* to guide the agent (Chaplot et al., 2020; Ramakrishnan et al., 2020; Chen et al., 2025b), sometimes relying on ground-truth depth observations. While explicit mapping guarantees spatial awareness, such modularity often abstracts away rich semantic information, sacrifices the flexibility of end-to-end learning, and limits generalization to geometrically definable tasks. Our image-sequence agent, conversely, requires no depth sensor or localization at deployment and adapts flexibly to semantically diverse downstream tasks. Furthermore, we show that scheduled, temporary injections of random behavior into RL training suffice to overcome the reward-less stretches of long-horizon exploration (Ecoffet et al., 2021). This enables an emergent behavior where the agent learns to traverse already-seen regions to find novel "branches" of unseen places. Crucially, this is achieved without the imitation learning bootstrapping (Chen et al., 2019), or hierarchical goal selection of prior work (Chaplot et al., 2020; Ramakrishnan et al., 2020; Chen et al., 2025b), highlighting curiosity's potential in sparse-reward settings.

Trained purely from curiosity on Habitat Matterport 3D (HM3D; Ramakrishnan et al., 2021), our agent learns effective exploration behavior using only RGB at deployment. We demonstrate the strength and generality of this behavior by calibrating its exploration performance against specialized RL-based active-mapping policies and evaluating zero-shot transfer to Gibson (Xia et al., 2018) and AI-generated 3D worlds (World Labs, 2024). Crucially, because our policy does not require

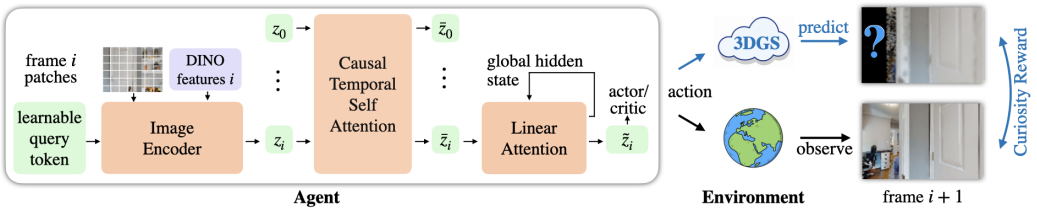


Figure 2: **Method overview.** The agent (left) fuses image-patch and DINOv2 features into frame tokens processed by causal and linear attention before predicting actions and values. The **environment** (right) executes the action and returns the next observation; in parallel, a persistent 3DGS *forward model* renders the same view from accumulated past experience, and the discrepancy with the observed view defines the *curiosity reward*. The forward model is used only at training time (blue path); at deployment, the agent acts from an RGB stream alone.

explicit mapping at test time, it remains fully end-to-end and highly adaptable. Therefore, by fine-tuning briefly on external task rewards, it can outperform policies trained from scratch. Our work demonstrates that coupling a persistent model of the world with an episodic agent provides insights into how curiosity-driven exploration can be scaled to complex, photorealistic 3D environments.

2 Method

Assuming a static 3D scene, we aim to train a camera-agent to efficiently explore its surroundings purely from visual inputs. We wish for novelty-seeking behavior to arise without task-specific rewards or external exploration labels, and so we formulate exploration as a self-supervised RL problem with a curiosity-driven learning signal derived from the agent's own experience. We defer discussion of **related works** and **background** to Section 4 and Appendix C, and describe our method first.

Our method has two key parts: a long-context transformer-based agent architecture which enables long-horizon exploration behavior, and a curiosity module based on 3D reconstruction which provides a stable reward for the agent to optimize with on-policy RL. The agent (Fig. 2, left) takes in *only* a stream of RGB observations from the 3D environment at hand, and outputs local camera motion per-timestep. It maintains internal memory long enough to enable it to discover exploration strategies like backtracking. The curiosity module (Fig. 2, right) builds a photorealistic 3D scene reconstruction from the agent's stream of observations, then uses the disagreement between novel view renders and ground truth observation as reward. To optimize this curiosity-based reward the agent necessarily must learn to visually explore 3D scenes and navigate complex environments to seek novel viewpoints.

Problem setup We frame exploration as interaction with an environment. At each time $1<t<N$, the agent takes action a_t to move through environment $\mathcal{E}$ and receives new observation o_{t+1}, transitioning from its current state to the next. During simulation training, the persistent forward model receives privileged depth and pose as part of $\tilde{o}_{t+1}$. At deployment, the forward model is discarded, and the agent acts solely from its episodic memory over RGB observations and actions.

2.1 Persistent 3D forward model

For the agent to learn exploratory behavior, it should be rewarded for seeking novel regions of the environment. We achieve this via intrinsic curiosity by introducing an auxiliary model of the world during training. This *forward model* is tasked with predicting the next observation conditioned on an action, given the stream of observations so far:

$$\hat{o}_{t+1} = \mathcal{F}(a_t, \tilde{o}_{1:t}). \tag{1}$$

The forward model $\mathcal{F}$ defines a curiosity reward that activates upon encountering novel regions not yet explained by this model, by comparing its prediction to the observed reality:

$$r_t^{\text{cur}} = d(\hat{o}_{t+1}, o_{t+1}), \tag{2}$$

where d is a visual prediction-error metric. A larger discrepancy indicates the new observation is poorly explained by the current forward model, corresponding to higher novelty.

We instantiate the forward model as an online 3D Gaussian Splatting (3DGS) model of the world, capable of incorporating streaming observations while maintaining a persistent model of what has been seen. We later show in Section 3 that this persistence is crucial for a reliable reward signal. We refer the reader to Appendix C.2 for background on 3DGS. Given the training-time privileged stream of observations including RGB image, depth image and camera poses (i.e., $\tilde{o}_{1:t} = \{(I_i, D_i, p_i)\}_{i=0}^{t}$), we maintain a 3DGS representation $\mathcal{G}_t$. For each observed RGB-D frame, a new gaussian per pixel is added to the scene, initialized from its color, depth, and camera pose. At fixed intervals, the representation is optimized on randomly selected past frames to improve reconstruction quality, and pruned and densified following Kheradmand et al. (2024).

At each step, after the agent acts and the camera moves following environment dynamics, the forward model is queried at the new pose p_{t+1} to render a prediction of the incoming observation:

$$\hat{I}_{t+1} = \mathcal{F}(p_{t+1} | I_{1:t}, D_{1:t}, p_{1:t}) = \mathcal{R}(\mathcal{G}_t, p_{t+1}), \tag{3}$$

where $\mathcal{R}$ denotes the differentiable 3DGS renderer. The model incorporates only observations up to the current timestep and has access to exact depth and pose during simulation training, leading to near-perfect reconstruction of previously observed regions. Its prediction error therefore primarily reflects whether the incoming view contains information not yet captured by the forward model, with negligible contribution from reconstruction error. To avoid rewarding residual reconstruction error near high-frequency details in the scene that are not meaningful for exploration, we compute prediction error after low-pass filtering and down-sampling both the prediction and observation:

$$e_t = \frac{1}{|\Omega_s|} \sum_{(j,k)\in\Omega_s} \left\| \mathcal{D}_s(\mathcal{B}(I_{t+1}))^{j,k} - \mathcal{D}_s(\mathcal{B}(\hat{I}_{t+1}))^{j,k} \right\|_2^2, \tag{4}$$

where $\mathcal{B}$ is a low-pass filter, $\mathcal{D}_s$ is down-sampling operator by a factor of s, and Ω_s is the down-sampled pixel set. In our implementation, we threshold the filtered error to obtain a bounded novelty decision: substantially unexplained views receive a positive reward, while explained views receive a small penalty:

$$r_t^{\text{cur}} = \begin{cases} r_{\text{new}}, & e_t > \tau, \\ r_{\text{old}}, & e_t \le \tau, \end{cases} \tag{5}$$

where $r_{\text{new}} > 0$ rewards sufficiently unexplained views, and $r_{\text{old}} < 0$ gives a small penalty for views already explained by the forward model. In Appendix E, we show robustness to whether an error threshold is used and to the choice of its value.

2.2 Agent Architecture

Our agent, decoupled from the forward model, is tasked with seeking novelty and therefore must maintain its own record of past observations to take actions that lead to unseen areas. While a geometric map is an effective abstraction for this, it discards all information beyond scene geometry. Instead we let the agent learn its own internal representation of the world, where both geometry and semantics are preserved. To this end, we use a transformer backbone whose actions are conditioned on the full sequence of past visual observations and actions throughout the exploration episode, hence giving the agent the *episodic memory* needed to learn a novelty-seeking policy π:

$$a_t \sim \pi(\cdot \mid o_{1:t}, a_{1:t-1}). \tag{6}$$

This memory has the further benefit of freeing the agent from being conditioned on any explicit geometric representation that abstracts the observations, making it straightforwardly adaptable to downstream tasks of varied nature that can benefit from exploratory behavior, through fine-tuning.

Concretely, the agent's episodic memory is represented as the sequence of past RGB observations and actions. This sequence is fed into a transformer backbone connected to the actor and critic heads that output an action distribution and a value estimate for the state at each timestep. The architecture, shown in Fig. 2, consists of an image encoder, causal attention layers over a temporal sliding window, and a global linear-attention memory module. At time t, the input is $\{(I_i, a_{i-1})\}_{i=1}^{t}$, where each observation I_i is paired with the preceding action a_{i-1}. Each action is encoded geometrically as a Plücker-ray image (Sitzmann et al., 2021) representing the intended camera transformation,

independent of actual environment dynamics, and concatenated channel-wise with its corresponding RGB observation to form a unified RGB-action input.

This RGB-action image in each timestep is first compressed into a frame token through an image encoder. We also take the RGB image processed by DINOv2 (Oquab et al., 2023) to provide richer visual features. A learnable query token cross-attends to the patch tokens and DINOv2 features:

$$z_i = \mathrm{CrossAttn}(q, [u_i, f_i]), \tag{7}$$

where u_i denotes RGB-action patch tokens, f_i denotes DINOv2 features, and q is the learnable frame query. The resulting frame tokens are processed by sliding-window causal temporal attention:

$$\bar{z}_{1:t} = \mathrm{SelfAttn}_W(z_{1:t}). \tag{8}$$

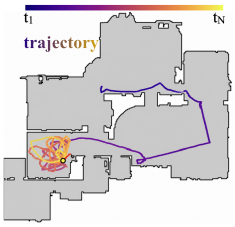
Causality ensures that each token uses only past and current observations, while the window size W keeps computation scalable over long episodes. Sliding-window attention provides efficient direct local context, but long-range information can only propagate indirectly through deeper layers. We therefore interleave selected temporal layers with a global linear-attention memory module inspired by TTT and LoGeR-style long-context architectures (Sun et al., 2025; Zhang et al., 2026). The module reads from a running memory state before updating it with the current token:

$$\tilde{z}_i = \mathrm{LinAttn}(\bar{z}_i, h_{i-1}), \qquad h_i = \mathrm{Update}(h_{i-1}, \bar{z}_i). \tag{9}$$

The memory readout is added back into the temporal stream, giving the policy access to information beyond the sliding attention window without full attention over the episode. Finally, actor and critic heads map the current token to action probabilities $\pi_\theta(a_t \mid o_{1:t}, a_{1:t-1})$ and value estimate $V_\theta(o_{1:t}, a_{1:t-1})$. Actions are sampled using π_θ at train and test time.

2.3 Training and Regularization

We optimize our actor-critic policy using PPO (Schulman et al., 2017). During training, rollouts are collected from the current policy, and the intrinsic reward r_t^{cur} defined in Sec. 2.1 is used as the only reward signal. A practical difficulty in curiosity-driven exploration is maintaining action diversity after rewards become sparse. Once nearby regions have been explored, reaching new parts of the scene may require temporarily moving through already-seen areas without receiving any intermediate reward, such as backtracking to an unseen branch. Regularizing action entropy with an additional loss term alongside the PPO losses helps, but it is sensitive to its coefficient and the optimization landscape: too small a coefficient can be overwhelmed by other loss terms and allow exploration to collapse (see inset), while too large a coefficient can destabilize optimization. Inspired by SAPG (Singla et al., 2024), which collects episode rollouts from a mixture of agents, we occasionally sample actions from a uniform random policy during the rollout collection in addition to a small entropy regularizer.



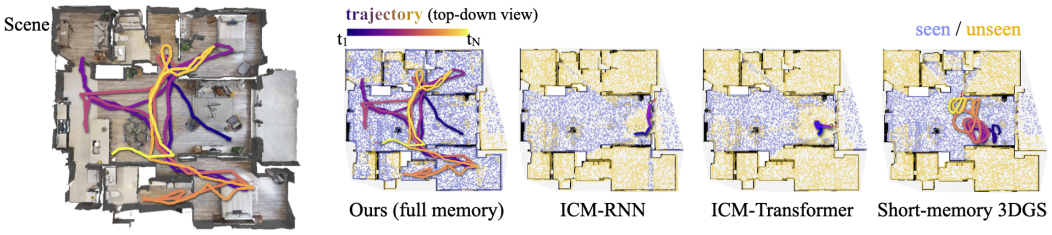
Specifically, at each training step, the executed action is sampled from a mixture of the learned policy and a uniform distribution over the discrete action set $\mathcal{A}$:

$$\mu_\theta(a_t \mid o_{1:t}, a_{1:t-1}) = (1-\beta)\pi_\theta(a_t \mid o_{1:t}, a_{1:t-1}) + \beta\mathcal{U}(a_t), \tag{10}$$

where β is the probability of sampling from the uniform policy and $\mathcal{U}(a_t) = 1/|\mathcal{A}|$. This guarantees persistent exploratory behavior during rollout collection without requiring the learned policy distribution itself to remain highly stochastic. For the PPO update, we account for the mixed behavior distribution in the likelihood ratio by using the probability of the sampled action under the rollout behavior distribution in the denominator i.e. $\rho_t(\theta) = \pi_\theta / \mu_{\theta_{\mathrm{old}}}$. The resulting ratio is then used in the standard PPO clipped surrogate objective as described in Appendix C.1, with the mixing coefficient annealed to zero over training (see Appendix D.2).

3 Experiments

Our experiments ask whether persistence enables curiosity to produce effective 3D exploration behavior. To establish that our agent, trained solely through curiosity, can explore effectively in 3D,



		FM memory			Agent memory				Our performance gain	
Metrics	**Ours**	ICM RNN	ICM transformer	Short Memory (64) 3DGS	Ours RNN	Agent w/ ctx=16	Agent w/ ctx=4	Agent w/ ctx=1	**w/ FM memory**	**w/ agent memory**
Compl. %@256 ↑	**56.50**	32.70	35.25	33.35	50.70	51.80	45.87	40.66	↑ 21–24%	↑ 5–16%
Compl. %@512 ↑	**66.69**	35.52	38.89	36.05	59.52	60.66	53.83	46.09	↑ 28–31%	↑ 6–21%
Compl. %@1024 ↑	**74.94**	37.36	43.44	38.09	67.33	67.77	60.56	50.39	↑ 32–38%	↑ 7–25%
Avg. dist. ↓	**0.14**	1.10	0.94	1.07	0.29	0.28	0.43	0.68	↓ 0.80–0.96	↓ 0.14–0.54

Figure 3: Memory ablations on HM3D (Ramakrishnan et al., 2021) show substantial gains from forward-model (FM) persistence and agent episodic memory: FM persistence is critical for curiosity-driven exploration, while greater agent memory improves coverage and reduces local trajectory loops. Figure shows agents trajectories and top-down projection of 3D points as **seen** or **unseen**.

we evaluate its test-time behavior in *unseen* indoor scenes. We first isolate forward-model persistence and agent memory (Section 3.1), then calibrate the resulting behavior against specialized active-mapping policies (Section 3.2). Finally, we test whether curiosity supplies a reusable search prior for downstream navigation (Section 3.3) and generalizes out of distribution (Section 3.4), assessing learned exploration rather than introducing task-specific benchmarks. Videos of exploration and downstream-task behavior are provided in the supplementary webpage.

Experimental setup All exploration policies train for 110M steps, and all train- and test-time episodes last 1024 steps. We simulate a fixed-height spherical drone in Habitat (Puig et al., 2024) at 1.25m with a 90° forward camera and four actions: `move forward` (0.25m), `look left/right` (15°), and `pause`. Actions are deterministic, and the agent moves horizontally without navmesh constraints. We detect collisions by ray tracing against the scene mesh, and the agent halts in place on collision, avoiding the navmesh-dependent sliding shortcut identified by Monaci et al. (2026). We trained our policy on 8× H100 GPUs for 5.5 days. Appendix D gives optimization, reward, and 3DGS reconstruction details.

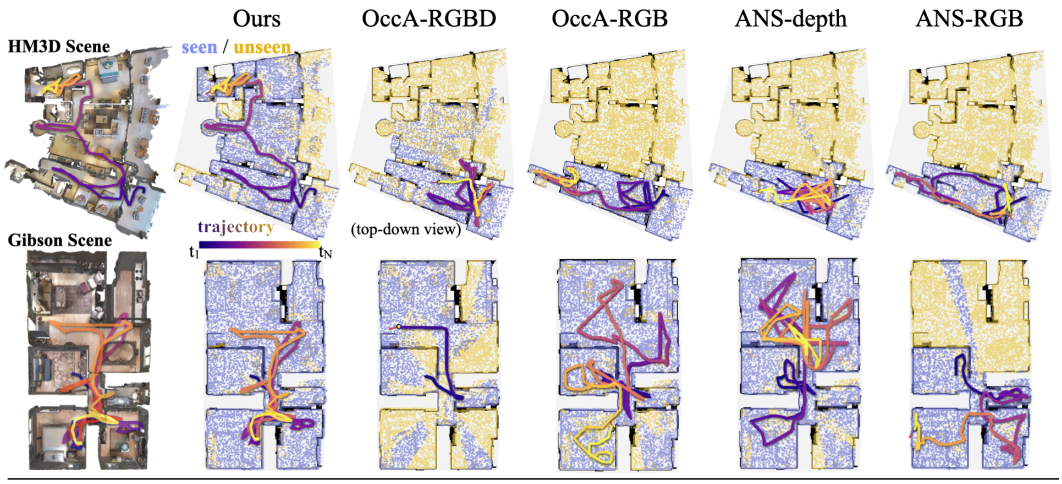
Datasets We train on the HM3D (Ramakrishnan et al., 2021) training set (consisting of 800 scenes), with each episode initialized at a random location. We evaluate on the HM3D validation set (100 scenes, 2 random starting points each, 200 episodes total) and, to assess generalization beyond HM3D, on Gibson (Xia et al., 2018) (86 mostly office-space scenes, 1 starting point each).

Metrics Following Yan et al. (2023), we measure exploration via 3D scene completeness at 3 time horizons (256, 512, and 1024 steps): the percentage of points on the reachable ground truth mesh surface (sampled uniformly, 200k points) whose closest observed point, back-projected using ground-truth depth, lies within 5cm. We also report the average distance from each ground truth point to its nearest observed point.

3.1 Role of Persistence and Episodic Memory — Figure 3

We directly test our central claim that effective curiosity in 3D exploration requires persistence in the forward model and episodic memory in the agent by independently ablating each component's memory capacity. We start from the basic building blocks introduced by ICM (Pathak et al., 2017) and gradually add to the memory capacities of each module.

Baselines Starting from ICM (Pathak et al., 2017), we ablate the world model and agent memory independently. For the world model, we test ICM with its original inverse and forward dynamics modules, paired with either a capacity-matched RNN (**ICM-RNN**) as in Pathak et al. (2017), or our transformer (**ICM-Transformer**), isolating the effect of the world model alone. We then replace the forward model with a 3DGS limited to the last 64 frames (**Short Memory 3DGS**) at each time



Method	Depth-free deployment	HM3D (Ramakrishnan et al., 2021) Completeness % ↑ @256	@512	@1024	Avg. dist. (m) ↓ @1024	Gibson (Xia et al., 2018) Completeness % ↑ @256	@512	@1024	Avg. dist. (m) ↓ @1024
ANS-RGB (Chaplot et al., 2020)	✓	45.28	54.68	65.39	0.41	55.41	64.20	73.14	0.30
ANS-depth (Chaplot et al., 2020)	×	51.02	61.45	69.68	0.34	63.04	72.79	79.89	0.18
OccA-RGB (Ramakrishnan et al., 2020)	✓	47.67	58.32	68.86	0.33	57.33	67.35	77.93	0.16
OccA-RGBD (Ramakrishnan et al., 2020)	×	52.71	64.91	74.62	0.18	63.06	72.96	81.23	0.14
Ours	✓	**56.50**	**66.69**	**74.94**	**0.14**	**66.95**	**75.79**	**82.42**	**0.10**

Figure 4: Our RGB-only agent outperforms all RGB-only baselines and matches or exceeds active-mapping policies that require depth at deployment on HM3D and Gibson.

step, ablating short-horizon world memory. For agent memory, we compare a capacity-matched RNN and transformers whose actor and critic share limited context windows (ctx$= 1, 4, 16$). Appendix F provides the complete results, including asymmetric actor- and critic-context ablations.

Analysis Ablation results show that effective exploration requires both a persistent forward model and an agent with episodic memory. Across evaluation horizons, our model raises scene completeness by **up to** 38 percentage points over the forward model memory ablations: the ICM variants collapse to spinning in place or becoming stuck on geometry, while Short Memory 3DGS produces locally confined trajectories. It similarly raises scene completeness by **up to** 25 percentage points over the agent-memory ablations, with performance degrading progressively as the context shortens. Appendix F further shows that full context in either the actor or critic outperforms restricting both.

3.2 Indoor Scene Exploration — Figure 4

We next calibrate the effectiveness of the resulting exploration behavior against specialized active-mapping policies trained explicitly for floor-plan mapping, demonstrating that curiosity alone can produce practically effective 3D exploration.

Baselines We compare against **Active Neural SLAM (ANS)** (Chaplot et al., 2020) and **Occupancy Anticipation (OccAnt)** (Ramakrishnan et al., 2020), two representative and reproducible RL baselines most closely aligned with our setting. They learn reusable map-conditioned exploration and can be retrained under our primitive-action protocol. We use them to calibrate practical exploration rather than as an exhaustive mapping leaderboard. We match the baselines' training steps, dataset, and simulator setup to ours. Appendix G details adaptation and training.

Analysis At deployment, our RGB-only agent outperforms all RGB-based mapping baselines and matches or exceeds policies requiring depth. Even at earlier horizons, it achieves higher scene completeness than both groups. Figure 4 reports one evaluation seed; Appendix H reports means and standard deviations across six seeds. Baselines also rely on raw camera pose input corrected by a localization module, although this distinction is less critical under deterministic motion. Figure 4 shows characteristic failures: map prediction errors leave OccAnt and ANS-RGB trapped within

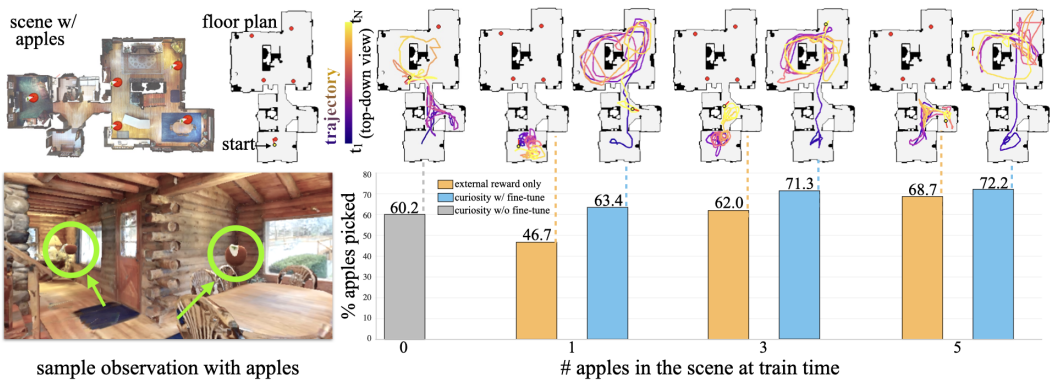


Figure 5: After brief apple-picking fine-tuning, our curiosity-pretrained agent outperforms matched from-scratch training under the same total interaction budget, with larger gains when fewer apples make rewards sparser and require more unrewarded exploration.

falsely closed maps or stuck on unpredicted geometry, while ANS-depth over-invests in local coverage at the expense of long-horizon efficiency. Qualitatively, our agent seeks doorways, traverses corridors, and returns to junctions, while almost never using the `pause` action.

3.3 Task-based fine-tuning — Figure 5 and Figure 6 (left)

We test whether curiosity-learned exploration provides a useful prior for downstream tasks by comparing our fine-tuned policy with an identical policy trained from scratch under the same embodiment, architecture, task reward, success criterion, and total training budget.

We evaluate **apple-picking** (Mirowski et al., 2016), where the agent searches for scattered apples and picks them up by getting close to them, and **image-goal navigation**, where it must get close to the viewpoint depicted by a target image; Appendix D.1 provides task details.

Experimental setting Both tasks use HM3D (Ramakrishnan et al., 2021). For apple-picking in three different experiments, 1, 3, or 5 apples are randomly placed in navigable parts of the scene. In image-goal navigation target pose is sampled such that it is reachable by the agent and sufficiently far from scene geometry. We compare against a matched agent trained from scratch on the task reward only for 118M steps; equal to our 110M pretraining plus 8M fine-tuning with external reward. We also report our non-fine-tuned exploration agent as a zero-shot baseline. We report success rate: average percentage of apples picked and percentage of image goals reached across the test set.

Analysis Operating purely on RGB input without explicit geometric representations, our agent can flexibly adapt to both downstream semantic tasks through fine-tuning alone. In **apple-picking**, as shown in Fig. 5, the baseline degrades under sparser reward (fewer apples at train-time), its search confined to one room, while our fine-tuned agent scales consistently with apple count. Without fine-tuning, pure exploration hits some apples zero-shot. **Image-goal navigation** is a more extreme sparse-reward regime: the baseline agent collapses to a roaming-around behavior, while our fine-tuned model leverages its exploratory prior to find the target view more reliably, as shown in Fig. 6 (left). This is due to the fact that broader exploration increases reward exposure during training, enabling the re-targeting of general exploration to goal-based navigation.

3.4 Out-of-distribution scenes — Figure 6 (right)

To assess generalization beyond indoor Habitat scans, we apply our agent zero-shot to two AI-generated worlds from World Labs (World Labs, 2024) (Hobbit World and Spaceship). These scenes differ not only in aesthetics but also in their rendering pipeline, as they are rendered from a 3DGS representation of these worlds rather than the realistic mesh scans of HM3D. Nevertheless, the agent exhibits coherent exploratory behavior: it navigates corridors, discovers doors to new spaces, and avoids collisions. Specifically, the agent encounters only 2 and 3 collisions over 256 steps in the two worlds, suggesting robust internal representations of navigable space.

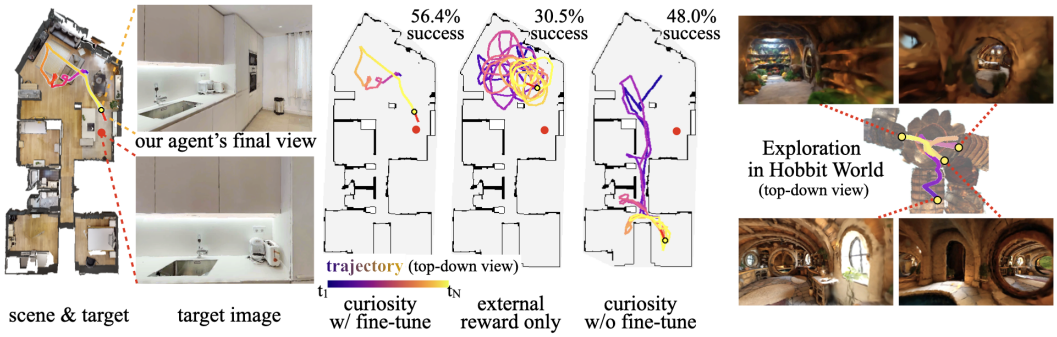


Figure 6: **Left**: Our exploration agent, when fine-tuned briefly on image-goal navigation reward, outperforms an agent trained from scratch on this reward alone, redirecting its general-purpose exploration toward a targeted objective. **Right**: Our agent generalizes zero-shot to AI-generated worlds of different aesthetic and representation.

4 Related Work

We survey the most related works here; a more comprehensive discussion is provided in Appendix B.

Intrinsic Motivation and Curiosity Curiosity-driven exploration often rewards prediction errors from learned forward-dynamics models (Pathak et al., 2017; 2019), fixed-target prediction errors such as RND (Burda et al., 2019b), or episodic visitation proxies (Savinov et al., 2019; Badia et al., 2020; Henaff et al., 2022; Fu et al., 2023). Standard exploration policies are typically reactive or limited to short recurrent contexts, preventing history-based planning (Ecoffet et al., 2021). We address both limitations with a persistent 3D reconstruction and a full-history sequence model for long-horizon novelty seeking.

Active Mapping and Next-Best-View Next-best-view methods greedily select viewpoints to maximize geometric information gain (Yan et al., 2023; Jiang et al., 2024; Li et al., 2026; Jiang et al., 2025), while RL mapping policies typically condition on explicit geometric maps (Chaplot et al., 2020; Ramakrishnan et al., 2020; Chen et al., 2025b) or rely on human demonstrations (Chen et al., 2019; He et al., 2023). By treating mapping as the objective or policy input, these approaches limit adaptation to semantic tasks. Our RGB-only policy instead requires neither maps nor depth at deployment and supports downstream fine-tuning.

3D Representations as World Models Recent video-generative models offer a promising data-driven approach to modeling world dynamics (Ball et al., 2025; Robbyant Team, 2026; Ye et al., 2026), but currently suffer from severe spatial forgetting and open-loop generation (Ma et al., 2026). Alternatively, explicit 3D reconstructions like 3D Gaussian Splatting (Kerbl et al., 2023; Keetha et al., 2024) serve as rudimentary, dynamics-free world models that guarantee exact geometric consistency.

5 Conclusions

We introduced *Remember to be Curious*, a framework for end-to-end curiosity-driven exploration of static realistic 3D scenes, where the agent operates solely from a visual input stream. By using a persistent online 3D reconstruction model as a proxy for the forward model, we demonstrate the necessity of two components for long-horizon exploration: a persistent forward model of the world for a reliable curiosity signal, and an episodic memory architecture for planning toward novelty.

While our use of 3DGS confines the agent to static scenes, we see promise in action-conditioned video models for addressing this limitation. We view our study as a guide for what these future world models require to enable intrinsic motivation. In particular, spatial persistence is key, alongside the ability to continuously update internal representations based on closed-loop online observations. We hope this work opens avenues for curiosity-driven exploration across diverse real-world environments and embodiments, extending to dynamic settings as world modeling matures.

AI USE STATEMENT

We used generative AI tools to edit and polish the manuscript for clarity and readability. AI coding agents also assisted with portions of the software implementation. All AI-assisted text and code were reviewed by the authors, and the resulting implementation was tested for correctness. The authors take full responsibility for the final content of this work.

REFERENCES

Adrià Puigdomènech Badia, Pablo Sprechmann, et al. Never give up: Learning directed exploration strategies. In *ICLR*, 2020.

Seungjae Baek, Brady Moon, Seungchan Kim, Muqing Cao, Cherie Ho, Sebastian Scherer, and Jeong Hwan Jeon. Pipe planner: Pathwise information gain with map predictions for indoor robot exploration. In *IROS*, 2025.

Philip J. Ball, Jakob Bauer, et al. Genie 3: A new frontier for world models, 2025. URL `https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/`.

Roberto Bigazzi, Federico Landi, Silvia Cascianelli, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. Focus on impact: Indoor exploration with intrinsic motivation. In *RA-L*, 2022.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *ICLR*, 2019a.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *ICLR*, 2019b.

Andres Campero, Roberta Raileanu, Heinrich Küttler, Joshua B. Tenenbaum, Tim Rocktäschel, and Edward Grefenstette. Learning with amigo: Adversarially motivated intrinsic goals. In *ICLR*, 2021.

Roger Creus Castanyer, Joshua Romoff, and Glen Berseth. Improving intrinsic exploration by creating stationary objectives. In *ICLR*, 2024.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *ICLR*, 2020.

Liyan Chen, Huangying Zhan, Kevin Chen, Xiangyu Xu, Qingan Yan, Changjiang Cai, and Yi Xu. Activegamer: Active gaussian mapping through efficient rendering. In *CVPR*, 2025a.

Tao Chen, Saurabh Gupta, and Abhinav Gupta. Learning exploration policies for navigation. In *ICLR*, 2019.

Xiao Chen, Quanyi Li, Tai Wang, Tianfan Xue, and Jiangmiao Pang. Gennbv: Generalizable next-best-view policy for active 3d reconstruction. In *CVPR*, 2024.

Xiao Chen, Tai Wang, Quanyi Li, Tao Huang, Jiangmiao Pang, and Tianfan Xue. Gleam: Learning generalizable exploration policy for active mapping in complex 3d indoor scenes. In *ICCV*, 2025b.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O. Stanley, and Jeff Clune. First return, then explore. In *Nature*, 2021.

Yao Fu, Run Peng, and Honglak Lee. Go beyond imagination: maximizing episodic reachability with world models. In *ICML*, 2023.

Georgios Georgakis, Bernadette Bucher, Anton Arapin, Karl Schmeckpeper, Nikolai Matni, and Kostas Daniilidis. Uncertainty-driven planner for exploration and navigation. In *ICRA*, 2022.

Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' Rays: Uncertainty quantification in neural radiance fields. In *CVPR*, 2024.

Antoine Guédon, Tom Monnier, Pascal Monasse, and Vincent Lepetit. Macarons: Mapping and coverage anticipation with rgb online self-supervision. In *CVPR*, 2023.

Yuhang He, Irving Fang, Yiming Li, Rushi Bhavesh Shah, and Chen Feng. Metric-Free Exploration for Topological Mapping by Task and Motion Imitation in Feature Space. In *RSS*, 2023.

Mikael Henaff, Roberta Raileanu, Minqi Jiang, and Tim Rocktäschel. Exploration via elliptical episodic bonuses. In *NeurIPS*, 2022.

Cherie Ho, Seungchan Kim, Brady Moon, Aditya Parandekar, Narek Harutyunyan, Chen Wang, Katia Sycara, Graeme Best, and Sebastian Scherer. Mapex: Indoor structure exploration with probabilistic information gain from global map predictions. In *ICRA*, 2025.

Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and uncertainty quantification for radiance fields using fisher information. In *ECCV*, 2024.

Wen Jiang, Boshu Lei, Katrina Ashton, and Kostas Daniilidis. Multimodal llm guided exploration and active mapping using fisher information. In *ICCV*, 2025.

Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat, track & map 3d gaussians for dense rgb-d slam. In *CVPR*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *SIGGRAPH*, 2023.

Shakiba Kheradmand, Daniel Rebain, Gopal Sharma, Weiwei Sun, Yang-Che Tseng, Hossam Isack, Abhishek Kar, Andrea Tagliasacchi, and Kwang Moo Yi. 3d gaussian splatting as markov chain monte carlo. In *NeurIPS*, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Shiyao Li, Antoine Guedon, Clémentin Boittiaux, Shizhe Chen, and Vincent Lepetit. Nextbestpath: Efficient 3d mapping of unseen environments. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *ICLR*, 2025a.

Shiyao Li, Antoine Guédon, Shizhe Chen, and Vincent Lepetit. MAGICIAN: Efficient Long-Term Planning with Imagined Gaussians for Active Mapping. In *CVPR*, 2026.

Yuetao Li, Zijia Kuang, Ting Li, Qun Hao, Zike Yan, Guyue Zhou, and Shaohui Zhang. Activesplat: High-fidelity scene reconstruction through active gaussian splatting. In *RA-L*, 2025b.

Max Lodel, Bruno Brito, Alvaro Serra-Gómez, Laura Ferranti, Robert Babuška, and Javier Alonso-Mora. Where to look next: Learning viewpoint recommendations for informative trajectory planning. In *ICRA*, 2022.

Ziqi Ma, Mengzhan Liufu, and Georgia Gkioxari. Out of sight, out of mind? evaluating state evolution in video world models. *arXiv preprint arXiv:2603.13215*, 2026.

Andreas Meuleman, Ishaan Shah, Alexandre Lanvin, Bernhard Kerbl, and George Drettakis. On-the-fly reconstruction for large-scale novel view synthesis from unposed images. In *TOG*, 2025.

Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.

Gianluca Monaci, Philippe Weinzaepfel, and Christian Wolf. What does really matter in image goal navigation? In *3DV*, 2026.

Maxime Oquab, Timothée Darcet, et al. Dinov2: Learning robust visual features without supervision. In *TMLR*, 2023.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *ICML*, 2017.

Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *ICML*, 2019.

Xavi Puig, Eric Undersander, et al. Habitat 3.0: A co-habitat for humans, avatars and robots, 2024.

Roberta Raileanu and Tim Rocktäschel. Ride: Rewarding impact-driven exploration for procedurally-generated environments. In *ICLR*, 2020.

Santhosh K. Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation. In *ECCV*, 2020.

Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *NeurIPS*, 2021.

Robbyant Team. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Nikolay Savinov, Anton Raichuk, Raphaël Marinier, Damien Vincent, Marc Pollefeys, Timothy Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *ICLR*, 2019.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*. MIT Press/Bradford Books, 1991.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *ICLR*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. State entropy maximization with random encoders for efficient exploration. In *ICML*, 2021.

Tianchang Shen, Sherwin Bahmani, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026.

Jayesh Singla, Ananye Agarwal, and Deepak Pathak. Sapg: Split and aggregate policy gradients. In *ICML*, 2024.

Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *NeurIPS*, 2021.

Bhavya Sukhija, Stelian Coros, Andreas Krause, Pieter Abbeel, and Carmelo Sferrazza. Maxinforl: Boosting exploration in reinforcement learning through information gain maximization. In *ICLR*, 2025.

Yu Sun, Xinhao Li, et al. Learning to (learn at test time): Rnns with expressive hidden states. In *ICML*, 2025.

Yuezhan Tao, Dexter Ong, Varun Murali, Igor Spasojevic, Pratik Chaudhari, and Vijay Kumar. Rt-guide: Real-time gaussian splatting for information-driven exploration. In *RA-L*, 2025.

Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025b.

Sen Wang, Bangwei Liu, Zhenkun Gao, Lizhuang Ma, Xuhong Wang, Yuan Xie, and Xin Tan. Explore with long-term memory: A benchmark and multimodal llm-based reinforcement learning framework for embodied exploration. In *CVPR*, 2026.

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*, 2024.

Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *ICLR*, 2020.

World Labs. Generating worlds. `https://www.worldlabs.ai/blog/generating-worlds` and `https://www.worldlabs.ai/blog/spark-2.0`, 2024. Accessed: 2025.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025.

Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *CVPR*, 2018.

Zike Yan, Haoxiang Yang, and Hongbin Zha. Active neural mapping. In *ICCV*, 2023.

Seonghyeon Ye, Yunhao Ge, et al. Dreamzero: World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Chen Sun, Ming-Hsuan Yang, Forrester Cole, Trevor Darrell, and Deqing Sun. Loger: Long-context geometric reconstruction with hybrid memory. *arXiv preprint arXiv:2603.03269*, 2026.

Siyuan Zhou, Yilun Du, Yuncong Yang, Lei Han, Peihao Chen, Dit-Yan Yeung, and Chuang Gan. Learning 3d persistent embodied world models. In *NeurIPS*, 2025.

Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J. Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*, 2017.

A Video Results

Video results of our agent exploring diverse 3D environments and performing downstream tasks are provided in the supplementary website archive `supplementary_website.zip`. Open `index.html` locally to view it.

B Related Work Discussion

Reinforcement learning with intrinsic motivation In reinforcement learning (RL), agents learn by interacting with the environment and improving based on the rewards they receive. However, many real-world environments provide sparse or no extrinsic reward, motivating intrinsic-reward methods that encourage exploration through auxiliary objectives such as maximizing state coverage, information gain, or self-generated goals (Seo et al., 2021; Sukhija et al., 2025; Campero et al., 2021).

Curiosity is a form of intrinsic reward, first framed by Schmidhuber (1991) as reward for prediction error in a learned world model. Pathak et al. (2017) proposed a deep RL curiosity framework in which a latent forward dynamics model is trained to predict consequences of actions and the agent is rewarded to seek states with large prediction error. Follow-up works stabilized this curiosity reward further with forward dynamics ensemble disagreement (Pathak et al., 2019) or fixed random network representations (Burda et al., 2019b). However, these rewards lack episodic persistence: the forward dynamics model only updates through gradient steps across training, so the agent can repeatedly revisit surprising states within an episode to accumulate rewards without having them updated (Badia et al., 2020), fixating on a narrow set of states while the forward model receives little diverse data to improve on. Further, the reward target becomes non-stationary as the learned latent representation of states evolve, making the signal noisy and optimization unstable (Burda et al., 2019a; Castanyer et al., 2024).

To address this, follow-up works augment curiosity with episodic visitation counts (Raileanu & Rocktäschel, 2020; Badia et al., 2020; Fu et al., 2023), but count-based methods do not scale to continuous, high-dimensional observations (Bigazzi et al., 2022; Savinov et al., 2019), and efforts such as E3B (Henaff et al., 2022) that use compact learned embeddings inevitably abstract away geometric and semantic structure. Whether grounded in forward dynamics or visitation counts, no method provides a persistent, improving model of the world for a stable curiosity signal in an exploration episode.

Moreover, the exploration policy which is architecturally decoupled from the forward dynamics model (Pathak et al., 2017) is either solely reactive to the current observation or limited to a short recurrent context, preventing the agent from planning exploration based on its accumulated history. While our proposed 3D modeling of the world sacrifices a lifelong learned world model, we use it as a controlled proxy to study the role of persistence in curiosity-driven exploration.

Active mapping and next-best-view selection While curiosity-driven methods learn a policy that seeks novelty as a behavioral objective, active mapping and next-best-view (NBV) methods treat exploration as a means to an end: building a map of the scene. Classical and learned NBV approaches (Yan et al., 2023; Jiang et al., 2024; Goli et al., 2024; Georgakis et al., 2022; Guédon et al., 2023; Li et al., 2025a;b; Chen et al., 2025a; Tao et al., 2025; Ho et al., 2025; Baek et al., 2025) greedily select viewpoints by estimated information gain or coverage; MAGICIAN (Li et al., 2026) extends this to a more comprehensive tree search over information gain maximizing trajectories, and recent works combine LLMs with geometric planning for even longer-horizon goal selection and planning (Jiang et al., 2025; Wang et al., 2026). These methods do not learn from experience, are scene-specific and replan from scratch per scene, making them difficult to extend to downstream tasks beyond mapping.

On the other hand, a line of RL-based methods learns a mapping policy conditioned on an explicit geometric map, giving the agent spatial but not semantic memory. Chen et al. (2019) and Lodel et al. (2022) learn exploration policies conditioned on egocentric depth-derived maps via imitation learning and coverage fine-tuning; ANS (Chaplot et al., 2020) and OccAnt (Ramakrishnan et al., 2020) build hierarchical policies on SLAM-generated maps where a global RL module selects coverage-rewarded goals and a local planner executes them; and GLEAM (Chen et al., 2025b) follows the same paradigm at larger scale with panoramic observations. By conditioning on a geometric map and optimizing for

its coverage, these policies acquire no semantic understanding of the world and therefore can only be extended to tasks that can be defined geometrically, such as point-goal navigation. DeepExplorer (He et al., 2023) forgoes the map via a short recurrent context but still relies on expert demonstrations, and GenNBV (Chen et al., 2024) learns an object-centric NBV policy from occupancy and image features, targeting active object-level reconstruction rather than general exploration.

Our curiosity policy is conditioned on RGB observations alone, requiring no geometric map, depth sensor, or localization at test time, and is flexible to finetuning for downstream tasks such as image-goal navigation where exploratory behavior is a prerequisite.

3D scene representations and world models 3D Gaussian Splatting (Kerbl et al., 2023) reconstructs static scenes as explicit Gaussian primitives with fast differentiable rasterization, with follow-up work improving optimization speed and quality (Kheradmand et al., 2024; Meuleman et al., 2025). Online SLAM methods extend this to real-time incremental reconstruction from an image stream (Keetha et al., 2024). Feed-forward methods (Wang et al., 2024; 2025b; Zhang et al., 2026; Wang et al., 2025a) enable fast dense reconstruction from sequential frames without per-scene optimization, though most output point clouds rather than renderable representations. These reconstruction methods can be viewed as a rudimentary form of a world model: given a camera action, they synthesize the resulting observation, approximating next-state prediction in a geometrically grounded but dynamics-free manner.

Video generative world models (Ball et al., 2025; Robbyant Team, 2026; Ye et al., 2026) offer broader capacity by learning action-conditioned dynamics from data, showing promise for modeling non-stationary worlds for interactive tasks. However, current approaches suffer from spatial forgetting and temporal drift over long horizons (Ma et al., 2026), and approaches that address this with persistent 3D memory (Shen et al., 2026; Zhou et al., 2025; Wu et al., 2025) mostly revert to a static world assumption to maintain consistency. Further, most generative methods condition on previously generated video chunks rather than real observations, leaving planning open-loop.

We show that curiosity-based exploration requires a forward model grounded in real, closed-loop observations with persistent memory. Rather than claiming this is uniquely solved by 3DGS, we use it as a controlled instantiation for static indoor scenes to study what properties a forward dynamics model must have to support reliable curiosity, opening up avenues for curiosity-based exploration in dynamic scenes as generative world models mature.

C Background

C.1 Proximal Policy Optimization

Proximal Policy Optimization (PPO) (Schulman et al., 2017) is an on-policy actor-critic algorithm that optimizes a policy using trajectories sampled from the current policy. Given a policy $\pi_\theta(a_t \mid s_t)$ over actions a_t conditioned on state s_t, a value function $V_\theta(s_t)$, and advantage estimates A_t, PPO updates the policy by maximizing the clipped surrogate objective

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t \left[\min \left(\rho_t(\theta) A_t, \text{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon) A_t \right) \right], \qquad \rho_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}. \tag{11}$$

Here, $\rho_t(\theta)$ is the policy probability ratio between the updated policy and the rollout policy, and ϵ is the clipping threshold. The clipping term limits the size of policy updates, improving training stability. The critic is trained to predict a discounted return target R_t, formed from accumulated future rewards and value bootstrapping (Schulman et al., 2016). Entropy regularization further encourages action diversity, where $\mathcal{H}(\pi_\theta(\cdot \mid s_t))$ denotes the entropy of the policy distribution at state s_t.

$$\mathcal{L}_{\text{critic}}(\theta) = \mathbb{E}_t \left[\left(V_\theta(s_t) - R_t \right)^2 \right], \qquad \mathcal{L}_{\text{entropy}}(\theta) = -\mathbb{E}_t \left[\mathcal{H} \left(\pi_\theta(\cdot \mid s_t) \right) \right]. \tag{12}$$

C.2 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) represents a scene as a set of anisotropic 3D Gaussian primitives. Each Gaussian has a 3D mean $\mathbf{x}_l$, covariance Σ_l, opacity α_l, and view-dependent

color $\mathbf{c}_l$. During rendering, the Gaussians are projected into the image plane and alpha-composited in depth order to produce the pixel color:

$$I(\mathbf{u}) = \sum_{l=1}^{L} T_l(\mathbf{u})\, \alpha_l\, g_l^{2D}(\mathbf{u})\, \mathbf{c}_l, \qquad T_l(\mathbf{u}) = \prod_{j<l} \left(1 - \alpha_j g_j^{2D}(\mathbf{u})\right), \tag{13}$$

where $\mathbf{u}$ is an image-plane pixel location, g_l^{2D} is the projected 2D Gaussian footprint, and T_l is the accumulated transmittance from closer Gaussians. The Gaussian parameters are differentiable with respect to an image reconstruction loss, enabling efficient optimization of an explicit 3D scene representation. In RGB-D settings, depth can be used to initialize Gaussian means directly by unprojecting observed pixels into 3D, reducing geometric ambiguity and making online optimization faster and more stable, as used in RGB-D Gaussian SLAM systems such as SplaTAM (Keetha et al., 2024).

D Training and implementation details

We train our transformer for 110M steps with Adam (Kingma & Ba, 2015) at a learning rate of 10^{-5}. The random-policy regularizer decreases from 20% to zero over 5M steps beginning at 25M, and the action-entropy coefficient decays from 0.1 at a rate of 0.99. We use $r_{new} = 0.5$, $r_{old} = -2 \times 10^{-4}$, and $\tau = 0.01$.

The 3DGS primitives are initialized from depth and color at every step, optimized on 10 random past views every 16 steps with Adam (Kingma & Ba, 2015), and densified via 3DGS-MCMC (Kheradmand et al., 2024). The downsampling factor is 4.

To facilitate reproducibility, we will **publicly release** the complete implementation and training code.

D.1 Downstream task details

In **apple-picking** (Mirowski et al., 2016), the agent must locate and approach apples scattered across the scene. Picked apples are removed, and the agent receives a positive reward per pick and a small step penalty otherwise. Apples are placed at reachable positions sampled with probability proportional to their minimum distance from the already-selected set, avoiding clusters.

In **image-goal navigation**, the agent is given a target RGB image and must reach the corresponding viewpoint. Success requires at least 50% of the target's 3D points to be visible and the agent to be within 1.5m of the target pose. The goal image is passed through the image encoder and provided as an additional token to the temporal module with dense attention to all other tokens.

D.2 Scheduled regularization annealing

Since mixing the learned policy with a uniform random policy shifts the action distribution seen by the critic, it can bias its value estimates. Therefore, we gradually anneal β from 0.2 to zero over a scheduled window during training. This provides additional exploration when the policy is weak and curiosity rewards become sparse locally, while reducing the bias introduced by random actions as the learned policy improves. At test time, we set $\beta = 0$ and act directly from the learned policy.

E Reward robustness

We vary the novelty threshold while otherwise retaining the training setup of the main method. We also test a continuous variant that directly uses the filtered, downsampled prediction error e_t as reward before thresholding. Performance is stable around the default $\tau = 0.01$, and the continuous reward performs slightly better at later horizons. Large thresholds make novelty rewards too sparse and degrade exploration. Overall, performance is robust to the reward formulation and to moderate threshold variation; we use $\tau = 0.01$ for the results reported in the main paper.

Reward	Completeness@256	@512	@1024
Continuous (e_t)	56.22	**66.90**	**75.64**
$\tau = 0.005$	56.17	66.64	74.51
$\tau = 0.01$ (default)	**56.50**	66.69	74.94
$\tau = 0.02$	56.32	64.86	73.00
$\tau = 0.05$	32.01	36.18	40.95

Table 1: HM3D completeness under continuous and thresholded curiosity rewards.

F Complete memory ablations

The complete results below include the asymmetric actor- and critic-context ablations omitted from the compact summary in Figure 3.

Metrics	**Ours**	ICM RNN	ICM transformer	Short Memory (64) 3DGS	Ours RNN	Actor w/ ctx=1	Critic w/ ctx=1	Both w/ ctx=16	Both w/ ctx=4	Both w/ ctx=1
Compl. %@256 ↑	**56.50**	32.70	35.25	33.35	50.70	51.95	47.09	51.80	45.87	40.66
Compl. %@512 ↑	**66.69**	35.52	38.89	36.05	59.52	61.79	56.09	60.66	53.83	46.09
Compl. %@1024 ↑	**74.94**	37.36	43.44	38.09	67.33	70.11	63.72	67.77	60.56	50.39
Avg. dist. ↓	**0.14**	1.10	0.94	1.07	0.29	0.24	0.42	0.28	0.43	0.68

Table 2: Complete HM3D memory-ablation results.

G Baseline implementation and comparison scope

We use the official OccAnt release for both OccAnt and its ANS implementation, keeping the mapping modules, planner, global rewards, PPO setup, and learning hyperparameters unchanged. Both methods are trained from scratch for 110M HM3D steps in our simulator. We do not use released checkpoints because they were trained under different simulator settings, including collision sliding and navmesh-constrained movement, which later work shows can induce shortcut behaviors and limit robustness under more realistic dynamics (Monaci et al., 2026). For ANS, we follow the implementation in OccAnt's official codebase: the local policy is trained with PPO using progress toward the planner's goal, rather than initialized through imitation learning. Together, retraining from scratch under the same simulator dynamics and using PPO rather than an imitation-initialized controller align the embodiment, action interface, and low-level learning setup with ours, enabling a controlled comparison while staying close to the published baseline implementations.

All methods use the same fixed-height spherical embodiment, noise-free sensing and actuation, 0.25m forward and $\pm 15^\circ$ rotation actions, and ray-traced halt-on-collision dynamics. Ground-truth occupancy supervision for the mapping baselines is generated from the scene mesh by ray-casting to surrounding walls, rather than from Habitat's navmesh. This matches our navigation regime, where movement is constrained by physical collisions with scene geometry rather than by the navmesh. RGB/depth variants respectively predict or project into this map, and OccAnt additionally anticipates unobserved occupancy. The unchanged A* local planner operates on accumulated predicted maps. Native global rewards are unchanged: map-area coverage for ANS and anticipation-based map-accuracy improvement for OccAnt.

Despite being several years old, ANS and OccAnt remain the closest reproducible RL systems that learn reusable cross-scene mapping behavior under a low-level Habitat action interface. Recent active-reconstruction systems instead predict uncertainty or coverage and optimize a new goal or path online (Yan et al., 2023; Li et al., 2025a); DeepExplorer uses panoramic observations and expert imitation (He et al., 2023). GLEAM is the closest newer learned policy, but its pipeline integrates depth and pose for global map construction, panoramic observations for long-term goal prediction, and A* reachability verification. A separate major contribution is GLEAM-Bench, which scales training to 1,024 heterogeneous scenes; this emphasis on data scale and cross-dataset generalization is orthogonal to our study of memory in curiosity-driven exploration (Chen et al., 2025b).

H EVALUATION-SEED ROBUSTNESS

Across six evaluation seeds, our fixed RGB-only policy obtains 74.69 ± 0.27, compared with 74.42 ± 0.12 for OccAnt-RGBD. Both show low evaluation variability, with overlapping one-standard-deviation intervals. Estimating training-seed variance is infeasible within our compute budget because each 110M-step run requires 5.5 days on eight 80GB H100 GPUs; we therefore report evaluation-seed variance for the fixed policies. We include OccAnt-RGBD as a sensor-advantaged reference for completeness rather than an input-matched comparison. Matching or exceeding it without requiring depth is a significant practical advantage, removing reliance on depth-sensing hardware and its associated cost, calibration burden, and noisy or missing measurements.

Method	Seed 1	Seed 2	Seed 3	Seed 4	Seed 5	Seed 6	Mean $\pm$ std.
OccAnt-RGBD	74.4737	74.3092	74.3171	74.4842	74.3158	74.6213	74.42 ± 0.116
Ours	74.8800	74.5986	74.8379	74.1374	74.7461	74.9403	74.69 ± 0.270

Table 3: HM3D Completeness@1024 across six evaluation seeds.