3D-BELIEF: EMBODIED BELIEF INFERENCE VIA GENERATIVE 3D WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in visual generative models have highlighted the promise of learning generative world models. However, most existing approaches frame world modeling as novel-view synthesis or future-frame prediction, emphasizing visual realism rather than the structured uncertainty required by embodied agents acting under partial observability. In this work, we propose a different perspective: world modeling as embodied belief inference in 3D space. From this view, a world model should not merely render what may be seen, but maintain and update an agent's belief about the unobserved 3D world as new observations are acquired. We identify several key capabilities for such models, including spatially consistent scene memory, multi-hypothesis belief sampling, sequential belief updating, and semantically informed prediction of unseen regions. We instantiate these ideas in 3D-Belief, a generative world model that infers and updates explicit, actionable 3D beliefs from partial observations. Unlike prior approaches, 3D-Belief updates 3D scene hypotheses by retaining observed geometry, incorporating new evidence, and resampling unseen regions, enabling embodied agents to reason and act under partial observability. We evaluate 3D-Belief on 2D visual quality for scene memory and unobserved-scene imagination, object- and scene-level 3D imagination using our proposed 3D-CORE benchmark, and challenging object navigation tasks in both simulation and the real world. Experiments show that 3D-Belief improves 2D and 3D imagination quality and downstream embodied task performance compared to state-of-the-art methods.[1]

1 INTRODUCTION

Recent advances in video generation models have shown promise for generative world modeling, predicting novel views and future frames from visual inputs (Song et al., 2025; Bar et al., 2025; Ren et al., 2025; Yu et al., 2024; 2025). Recent work further explores camera control and spatial memory for these pixel-space visual predictions (Nam et al., 2026; Wang et al., 2026b; Xu et al., 2026).

However, there remains a substantial gap between what these models are trained to do and what a world model must represent for embodied agents operating from partial observations of a 3D scene. Predicting pixels is not the same as inferring the underlying 3D world: an agent in an unfamiliar environment must form a belief over both observed and unobserved regions, where this belief should capture task-relevant structure rather than only visual appearance. As new observations arrive, the agent should update this belief accordingly, forming an ever-evolving representation of the 3D world.

To achieve this, we theorize that robust embodied agents require a generative world model that has the following key capacities, as illustrated in Figure 1. First, the model needs to have a **spatially consistent scene memory**, which preserves the geometry and semantics of observed parts of the 3D world. Second, the model output should allow **multi-hypothesis belief sampling** to model uncertainty over the unseen parts of the 3D world. Third, the sequent nature of embodied tasks requires **sequential belief updating** conditioned on streaming partial observations. Lastly, it is crucial to produce **semantically informed belief prediction** which provides direct semantic prediction of the unseen parts of the 3D world to drive reasoning and planning.

[1]Videos and demos are available at https://3dbeliefsubmission.github.io/3d-belief-website/.

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

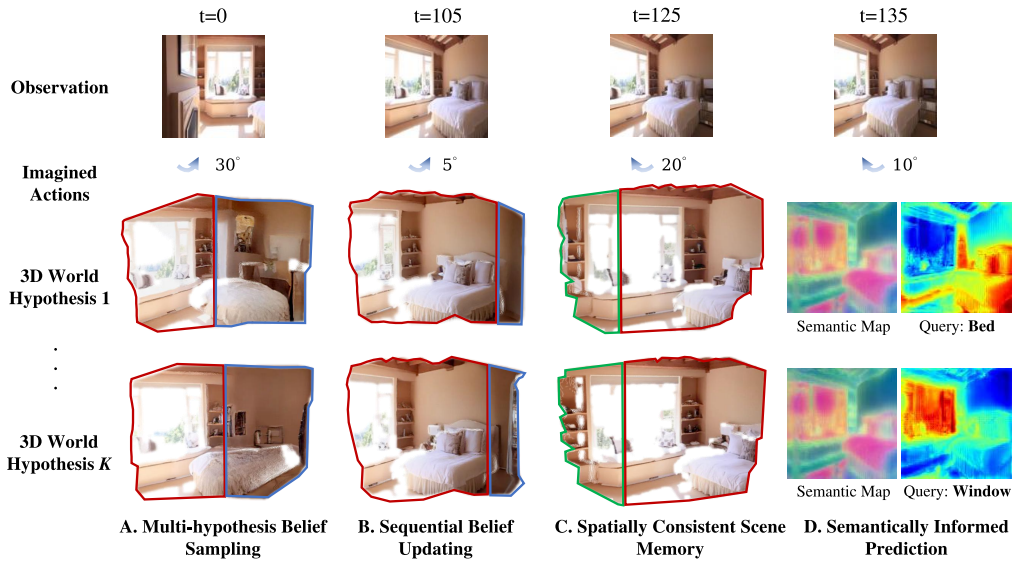


Figure 1: **Key capabilities of a generative 3D belief model.** An agent performs sequential mental simulation in an unknown environment. Top: selected RGB observations; second row: imagined actions; bottom: two sampled 3D world hypotheses. Regions supported by current observations are marked in red, imagined unseen regions in blue, and recalled content from scene memory in green. **A. Multi-hypothesis belief sampling:** at t=0, multiple plausible completions exist for unseen structure (e.g., unseen part of the bed and room). **B. Sequential belief updating:** incorporating observations from t=0 to 105 resolves ambiguity and updates the hypotheses (e.g., the appearance of the bed). **C. Spatially consistent scene memory:** revisiting viewpoints in mental simulation retrieves previously seen structure (e.g., bookshelf) with object permanence. **D. Semantically informed prediction:** semantic maps support task queries (e.g., *bed*, *window*).

Existing methods cover only a subset of these capabilities (Table 5, Appendix A). Pose- or action-conditioned video models provide visual imagination (Song et al., 2025; Bar et al., 2025), while 3D caches improve scene consistency (Ren et al., 2025; Yu et al., 2024). 3D reconstruction models recover explicit geometry and may support online updates (Wang et al., 2025a;b), but they cannot imagine unseen regions. Recent work has advanced image-conditioned 3D scene generation (Tewari et al., 2023; von Lützow et al., 2026; Schwarz et al., 2025; Gao et al., 2026). However, these scene-extension approaches do not support an online belief update that preserves accumulated observations while revising imagined scene content. Finally, semantic grounding remains limited: while semantic 3D scene representations have been explored (Ye et al., 2026), existing generative world models still lack open-vocabulary semantic predictions for unobserved 3D regions.

To bridge the gap, we propose ***3D-Belief***, a generative 3D world model that captures all the key abilities we hypothesized above. Built upon a diffusion model, 3D-Belief learns to predict explicit 3D representations of the full world state (including unseen regions) based on past egocentric visual observations. Critically, 3D-Belief can sequentially update its 3D prediction of the full world state based on new observations while maintaining consistency with all historical observations. The resulting 3D representations contain geometric, spatial, and semantic information about the 3D world, forming a coherent and queryable belief state that can support downstream embodied tasks.

To evaluate 3D-Belief, we assess its ability to predict and update 3D scene beliefs under partial observability. We first evaluate the 2D visual quality of the belief prediction, measuring both scene memorization in observed regions and imagination of unobserved regions from egocentric observations. We then introduce a novel benchmark, 3D-CORE (3D COntextual REasoning), to probe object- and scene-level 3D imagination. The proposed reasoning tasks provide a diagnostic evaluation of the four key capabilities of embodied agents' beliefs in 3D space. Finally, we evaluate whether the learned 3D scene beliefs can support downstream embodied tasks. In particular, we focus on open-vocabulary object navigation, where a robot is instructed to search for a target object in unseen environments. We test the model in both simulation and the real world with a mobile

manipulator. Across these evaluations, 3D-Belief consistently improves 2D visual quality, explicit 3D imagination, and downstream embodied task performance compared to strong baselines.

Our contributions are: (1) a framework for generative 3D belief inference; (2) 3D-Belief, which predicts actionable, semantic, uncertainty-aware 3D beliefs from partial observations; (3) 3D-CORE, a benchmark for object- and scene-level 3D imagination; and (4) evaluations of 2D/3D scene memory and imagination and simulated and real-world object navigation.

2 Related Work

Visual Generative Models. Video diffusion and view-synthesis models enable camera-controlled visual prediction (Song et al., 2025; Huang et al., 2025; Yu et al., 2024; 2025), with extensions to 3D/4D generation (Zhen et al., 2025). DFoT conditions on variable-length observation histories, while GEN3C uses a point-cloud cache to guide generation (Ren et al., 2025). These approaches motivate persistent scene context for imagination. We focus on explicit, semantically queryable scene hypotheses that can be revised as observations arrive.

3D Scene Generation. There have been recent methods that generate 3D scenes based on multi-view images or text prompts. L3DG and GaussianGPT generate Gaussian scenes via latent diffusion and autoregressive token prediction, respectively (Roessle et al., 2024; von Lützow et al., 2026). GGS combines Gaussian feature fields with video diffusion (Schwarz et al., 2025), Gen3R jointly generates appearance and geometry (Huang et al., 2026), and concurrent work PixWorld unifies reconstruction and generation through pixel-space supervision (Gao et al., 2026). Unlike these methods, 3D-Belief maintains and updates a persistent belief: it retains observed Gaussians, incorporates new observations, and resamples imagined regions to support semantic queries and embodied planning.

World Model-Based Planning. Video world models support planning through imagined action outcomes (Du et al., 2023; 2024; Zhou et al., 2024; Zhang et al., 2024). NWM applies action-conditioned prediction to navigation (Bar et al., 2025), and Dreamer 4 trains control policies inside a learned world model (Hafner et al., 2025). 3D-Belief maintains observed geometry, samples plausible unseen structure, and evaluates candidate paths through rendered views and semantic queries, connecting scene completion to open-vocabulary object search.

Semantically-Embedded 3D Scene Representations. LERF grounds language in radiance fields (Kerr et al., 2023), and LangSplat encodes language features in 3D Gaussians for open-vocabulary queries (Qin et al., 2024). ConceptGraphs organizes observations into an open-vocabulary scene graph (Gu et al., 2024), while DynaMem maintains spatial-semantic memory for localization and mobile manipulation (Liu et al., 2025). 3D-Belief also predicts semantic features for imagined regions, allowing the planner to query potential targets before observing them and revise those predictions as evidence arrives.

3 3D-Belief

3D-Belief predicts and sequentially updates explicit 3D scene representations online (Figure 2a). We formalize the model in Section 3.1 and describe its architecture and training in Sections 3.2–3.4.

3.1 Formulation

We follow the definition of the belief in partially observable Markov decision making (POMDP) (Kaelbling et al., 1998). Specifically, belief $b(s^t)$ is a distribution of the state s^t at any given step t conditioned on past observations o^t, i.e., $b(s^t) = P(s^t|o^{1:t})$. Given new observations o^{t+1} at the recent step $t+1$, we can update the belief as

$$b(s^{t+1}) = \sum_{s^t} P(s^{t+1}|o^{t+1}, s^t) b(s^t). \tag{1}$$

Directly predicting the full 3D scene for the first (3D memory) and fourth (semantics) key capabilities can be extremely challenging. Therefore, we can instead predict a 3D representation of the scene, $z^t = \phi(s^t)$. Such representations (1) should preserve the important 3D structures and semantic information of the scene necessary for downstream embodied tasks, and (2) can be parametrized so

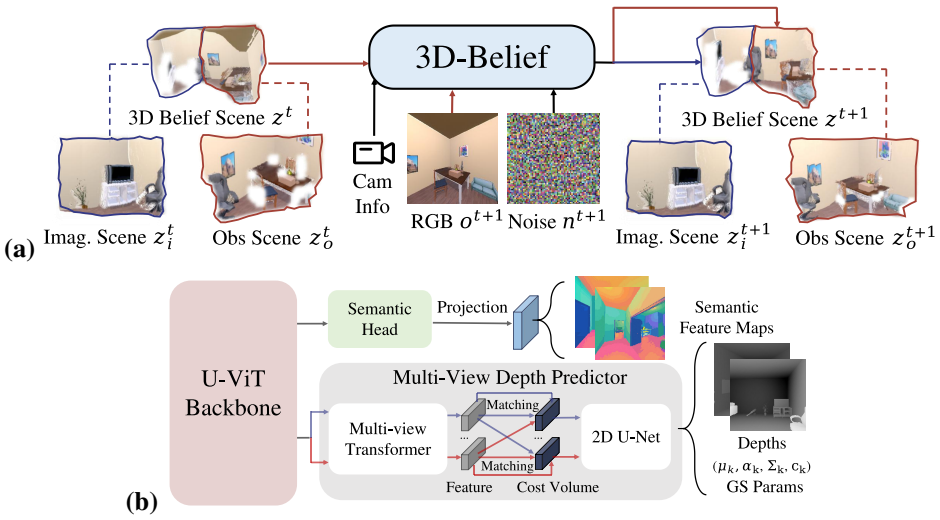


Figure 2: **3D-Belief overview and architecture.** (a) At step t, the sampled scene z^t contains observed Gaussians z_o^t and imagined Gaussians z_i^t. Given z^t, a new observation o^{t+1}, and sampled noise n^{t+1}, 3D-Belief samples z^{t+1} to update the 3D world belief. (b) 3D Scene Diffusion architecture.

that they can be conveniently constructed via model training. In this work, we consider 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023), as it not only provides an explicit 3D representation of a scene and can carry semantic feature embeddings in each primitive (Qiu et al., 2024), but also allows fast rendering to support downstream embodied tasks.

Formally, $z^t = \{g_k\}_{k=1}^K$ with Gaussian primitives $g_k = (\mu_k, \Sigma_k, \alpha_k, S_k, e_k)$ (mean, covariance, opacity, spherical-harmonic appearance, semantic embedding). We split $z^t = z_o^t \cup z_i^t$ by source: real context pixels produce observed Gaussians, and the denoised target branch produces imagined Gaussians. For sequential belief updates, we rewrite (1) as an autoregressive update:

$$z^{t+1} \sim p(z^{t+1} \mid o^{t+1}, z_o^t). \tag{2}$$

In particular, z_o^t is expanded to form z_o^{t+1} by incorporating the new observation, whereas z_i^t is replaced with a new imagination, z_i^{t+1}. The previous imagined content is dropped since it may conflict with new evidence. This formulation (i) enables a simple training procedure that requires supervision only over short horizons (as discussed in Sec. 3.2, we train our model using image pairs), (ii) supports test-time scaling for long-horizon planning via continual belief updates, and (iii) uses fixed-size context for each diffusion step.

Rendering z^t at pose θ as $\hat{o} = \mathcal{G}(z^t, \theta)$ (Kerbl et al., 2023) yields RGB, depth, and semantic maps from per-primitive embeddings for mental simulation and language-based plan scoring.

3.2 Model Architecture

Our model (Figure 2b) uses a shared U-ViT backbone (Song et al., 2025; Hoogeboom et al., 2024) with two heads that jointly predict a 3DGS scene with semantics. For geometric consistency, we employ an MVS-style 3DGS predictor with a multi-view Transformer and cost-volume module, in which the cost volume stores cross-view feature matching scores over discretized depth candidates to guide depth prediction, producing depths and Gaussian parameters that are lifted into primitives for fast differentiable rendering (Chen et al., 2024). A lightweight head regresses continuous 512-dimensional OpenCLIP features (Kerr et al., 2023; Ilharco et al., 2021). At query time, cosine similarity with a text embedding yields an open-vocabulary relevance map.

3.3 Diffusion Training

We adopt a diffusion model that predicts an explicit 3D Gaussian scene at each denoising step (Tewari et al., 2023). We train on image pairs, predicting Gaussian primitives from a context image and

a noisy target image. Rendering these primitives at the target camera estimates the clean target observation. The forward process at diffusion time step τ is defined as:

$$q(o_\tau^{\text{trgt}} \mid o_{\tau-1}^{\text{trgt}}) = \mathcal{N}(o_\tau^{\text{trgt}}; \sqrt{1-\beta_\tau} o_{\tau-1}^{\text{trgt}}, \beta_\tau I). \tag{3}$$

In the reverse process, we reconstruct o^{trgt} conditioned on o^{ctxt} and target camera parameters ϕ^{trgt} by predicting a denoised scene z_τ and rendering it into observation space via the GS renderer $\mathcal{G}(\cdot)$:

$$z_{\tau-1} = \Phi_\theta(o^{\text{ctxt}}, o_\tau^{\text{trgt}}; \tau, \phi^{\text{ctxt}}, \phi^{\text{trgt}}), \tag{4}$$

$$\hat{o}_{\tau-1}^{\text{trgt}} = \mathcal{G}(z_{\tau-1}, \phi^{\text{trgt}}). \tag{5}$$

Here, $\hat{o}_{t-1}^{\text{trgt}}$ serves as an estimate of the clean observation. Φ_θ is the neural network predicting 3D Gaussians, as defined in Sec 3.2. See Appendix B for the complete formulation.

3.4 Training Objective

We train Φ_θ on paired context and target observations. For $v \in \{\text{trgt}, \text{ctxt}\}$, let o^v be the observation and $\hat{o}^v = \mathcal{G}(z_0, \phi^v)$ the rendering of the predicted scene at camera ϕ^v. Let $I(\cdot)$, $S(\cdot) \in \mathbb{R}^{H\times W\times d}$, and $D(\cdot) \in \mathbb{R}^{H\times W}$ denote RGB, semantic features, and depth.

RGB loss. The RGB loss is $\mathcal{L}_{\text{rgb}} = \sum_{v\in\{\text{trgt},\text{ctxt}\}} \|I(\hat{o}^v) - I(o^v)\|_2^2$.

Semantic loss. For each view, we sample M patch centers $\mathbf{u}_j$ and align rendered features with frozen CLIP (Ilharco et al., 2021) embeddings f_{clip} of the corresponding crops $\pi(o^v, \mathbf{u}_j)$: $\mathcal{L}_{\text{sem}} = \sum_{v\in\{\text{trgt},\text{ctxt}\}} \frac{1}{M} \sum_{j=1}^{M} \|S(\hat{o}^v)(\mathbf{u}_j) - f_{\text{clip}}(\pi(o^v, \mathbf{u}_j))\|_2^2$.

Depth loss (optional). When ground-truth depth is available, we compare $\hat{d}^v = D(\hat{o}^v)$ with $d^v = D(o^v)$ using a masked ℓ_1 loss. The binary mask $m^v \in \{0,1\}^{H\times W}$ marks valid depth values: $\mathcal{L}_{\text{depth}} = \sum_{v\in\{\text{trgt},\text{ctxt}\}} \frac{1}{\sum_{\mathbf{u}} m^v(\mathbf{u})} \sum_{\mathbf{u}} m^v(\mathbf{u}) \left|\hat{d}^v(\mathbf{u}) - d^v(\mathbf{u})\right|$.

We additionally use LPIPS and edge-aware depth smoothness losses. The complete objective, view weights, and training-stage settings are specified in Appendix D.

4 Experiments

We evaluate 3D-Belief on 2D visual quality, 3D object and scene imagination, and belief-guided planning with inference-time scaling. Implementation details are in Appendix D.

4.1 Experiment 1: 2D Visual Quality of Belief Prediction

Embodied belief requires scene memory and plausible imagination of unseen space. Following video generation evaluations (Ren et al., 2025), we assess visual, geometric, and semantic quality.

Evaluation protocol. For observed regions, we condition on both ends of a trajectory and predict intermediate views, evaluating paired ground truth with LPIPS (Zhang et al., 2018), PSNR, and SSIM. For unobserved regions, we condition only on the starting frame and generate views beyond its field of view. To account for multimodal predictions, we report distributional metrics FVD (Unterthiner et al., 2019) and FID (Heusel et al., 2017). We test two domains. AI2-THOR (Kolve et al., 2017) navigation data, generated following Ehsani et al. (2024), tests egocentric exploration under partial observability; baselines are DFoT (Song et al., 2025) and NWM (Bar et al., 2025), fine-tuned on the training set. RealEstate10K (Zhou et al., 2018) tests photorealistic novel-view synthesis and video fidelity on real-world trajectories; baselines are pretrained DFoT (Song et al., 2025) and GEN3C (Ren et al., 2025). Figure 3 shows qualitative examples.

Results. Table 1 shows strong performance across both domains. On AI2-THOR, 3D-Belief achieves the best reconstruction and distributional metrics, supporting faithful scene memory and plausible imagination beyond observed views. On RealEstate10K, it achieves the best PSNR, SSIM, LPIPS, and FVD, with competitive FID, demonstrating strong visual prediction on real-world trajectories.

Figure 3 highlights two baseline failure modes. *Semantic drift*: DFoT turns grass into solid green ground in RealEstate10K, while AI2-THOR baselines open doors without plausible opening structure.

Table 1: **2D visual quality of belief prediction across synthetic embodied and real-world domains.** Scene memory is evaluated with paired metrics, while scene imagination is evaluated with distributional metrics. Best results within each domain are highlighted in **bold**.

(a) Observed scene

Domain	Method	PSNR↑	SSIM↑	LPIPS↓
AI2-THOR	NWM	18.75	0.702	0.1876
	DFoT	23.35	0.841	0.1206
	Ours	**28.81**	**0.928**	**0.0502**
RealEstate 10K	DFoT	22.395	0.7167	0.1213
	GEN3C	23.104	0.7626	0.0978
	Ours	**24.878**	**0.8432**	**0.0663**

(b) Imagined scene

Domain	Method	FID↓	FVD↓
AI2-THOR	NWM	89.28	487.4
	DFoT	72.82	429.7
	Ours	**47.24**	**271.8**
RealEstate 10K	DFoT	37.447	127.017
	GEN3C	**24.291**	73.744
	Ours	26.465	**50.993**

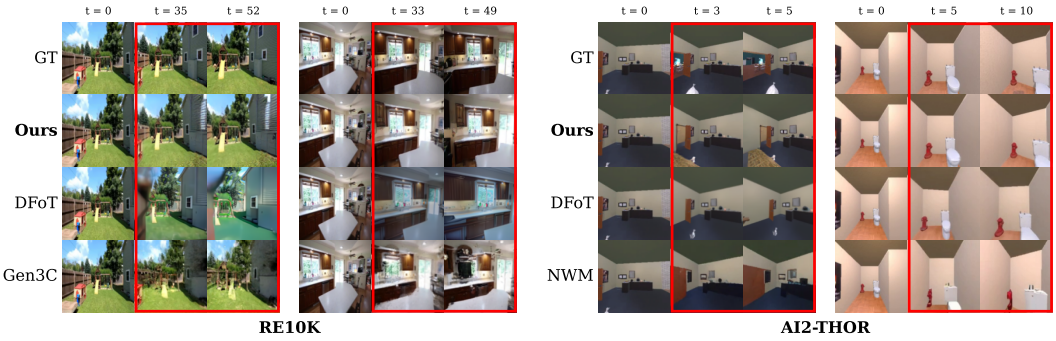


Figure 3: **Qualitative results of 2D visual quality of belief prediction.** Model predictions are highlighted with red boxes. We compare 3D-Belief against baselines on the 2D visual quality of belief prediction across AI2-THOR and RealEstate10K. 3D-Belief better preserves pose alignment, geometry, and texture details in both embodied and real-world trajectories.

Pose and geometric drift: viewpoint changes shift or distort generated content, especially object shapes and room layout in the right-hand AI2-THOR example. 3D-Belief better preserves semantics, pose alignment, and rigid scene structure through its explicit 3D representation.

4.2 Experiment 2: 3D Object and Scene Imagination

4.2.1 3D-CORE Benchmark

Embodied agents must integrate observations, infer unseen structure, and maintain stable beliefs. Visual fidelity and short-horizon prediction metrics (Song et al., 2025; Bar et al., 2025) (Sec. 4.1) do not directly test these abilities. We introduce 3D-CORE (3D COntextual REasoning) to test (1) **spatial expansion** beyond observed regions, (2) **semantic reasoning** grounded in 3D structure, and (3) **long-horizon consistency** across viewpoints.

Benchmark Construction. We build scenarios in AI2-THOR (Kolve et al., 2017) using diverse ProcTHOR (Deitke et al., 2022) houses and Objaverse (Deitke et al., 2023) assets. Egocentric RGB streams include known camera poses and ground-truth 3D geometry and semantics.

Tasks and Metrics. 3D-CORE contains three complementary tasks that isolate distinct but essential aspects of contextual 3D reasoning. (See Appendix D.3.1 for definitions of all metrics.)

Task 1: Object Completion (233 tasks): The model observes a partially visible target object and completes its 3D geometry and appearance, producing plausible shape/texture consistent with the object category. Metrics: BEV IoU, 3D IoU, Chamfer distance, SigLIP similarity, and Recognition.

Task 2: Room Completion (263 tasks): Given a single egocentric view of a room (with pose), the model predicts the unseen layout and semantics to support planning (e.g., plausible navigable regions). Metrics: Object Prediction F1, occupancy accuracy (known cells), and occupancy IoU.

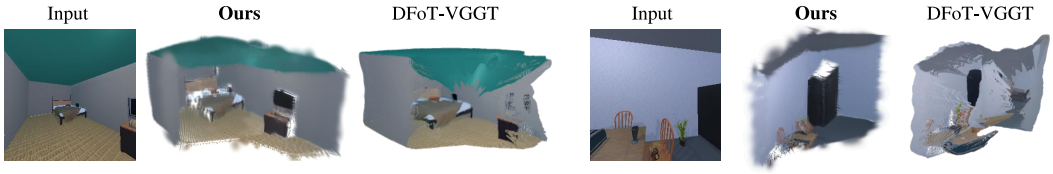


Figure 4: **3D object imagination.** 3D-Belief better infers object semantics and geometry from the observed image than the baseline, producing more complete, consistent 3D predictions.

Table 2: 3D-CORE results (object completion: 55% visibility). Chamfer is in meters; other scores are unitless. Metric definitions: Appendix D.3.1.

Method	Object completion					Room completion			Permanence	
	BEV IoU↑	3D IoU↑	Chamfer (m)↓	SigLIP ↑	Recog. ↑	Obj. F1↑	Occ. Acc.↑	Occ. IoU↑	LPIPS ↓	SigLIP ↑
DFoT-VGGT	0.362	0.243	0.830	0.798	0.767	0.531	0.252	0.110	0.555	0.907
3D-Belief	**0.484**	**0.318**	**0.216**	**0.855**	**0.930**	**0.536**	**0.900**	**0.442**	**0.123**	**0.978**

Task 3: Object Permanence (474 tasks): The model is rolled out along a trajectory with large viewpoint changes that returns to the start. We test whether the 3D belief stays stable, meaning that objects should not drift, or change identity upon revisiting. Metrics: LPIPS and SigLIP similarity.

4.2.2 Evaluation on 3D-CORE

We evaluate 3D-Belief's ability to imagine object- and scene-level 3D structure on the 3D-CORE benchmark. Given an initial RGB observation and an imagination camera trajectory, the agent predicts a completed 3D belief conditioned on the observation. Metrics are computed using semantics, geometry, or rendered views extracted from the predicted 3D representation.

Baseline. We construct a strong baseline, DFoT-VGGT, by combining two models for video generation and 3D reconstruction. DFoT-VGGT is an "imagination-then-lift" baseline that first uses DFoT (Song et al., 2025) to generate imagined observations and then applies VGGT (Wang et al., 2025a) to lift these predictions into a 3D representation. This pipeline provides a two-stage comparison to 3D-Belief's direct generation of explicit 3D scenes.

Results. Table 2 and Figure 4 show stronger object completion than DFoT-VGGT. Across visibility levels, the Chamfer gap over DFoT-VGGT narrows from 0.830 vs. 0.216 m at 55% to 0.189 vs. 0.187 m at 95%, highlighting the role of imagination (Appendix C.3.1). For **Room Completion**, object F1 is similar, while occupancy IoU improves from 0.110 to 0.442 (Appendix C.3.2). For **Object Permanence**, lower LPIPS and higher SigLIP indicate more stable appearance and identity on revisiting a viewpoint.

4.2.3 Scene Uncertainty under Belief Updates

We further evaluated the uncertainty in the belief of the 3D scene produced by 3D-Belief. On AI2-THOR validation trajectories, we draw $K = 5$ belief samples and measure uncertainty as the disagreement in rendered RGB (ℓ_1) and rendered depth (standard deviation) across these samples.

First, we condition on a single observation and autoregressively expand the 3D imagination along the same camera trajectory. We observed that uncertainty grows with angular separation (Table 3a). Second, we fix an evaluation pose and add four observations that progressively approach it through the sequential update in Eq. (2). After each update, we sample beliefs, imagine toward the evaluation pose, and render every sample there. We find that uncertainty decreases as evidence accumulates (Table 3b). These results show that the belief remains uncertain where evidence is missing and contracts as evidence arrives. Detailed protocols and analysis are in Appendix C.5.

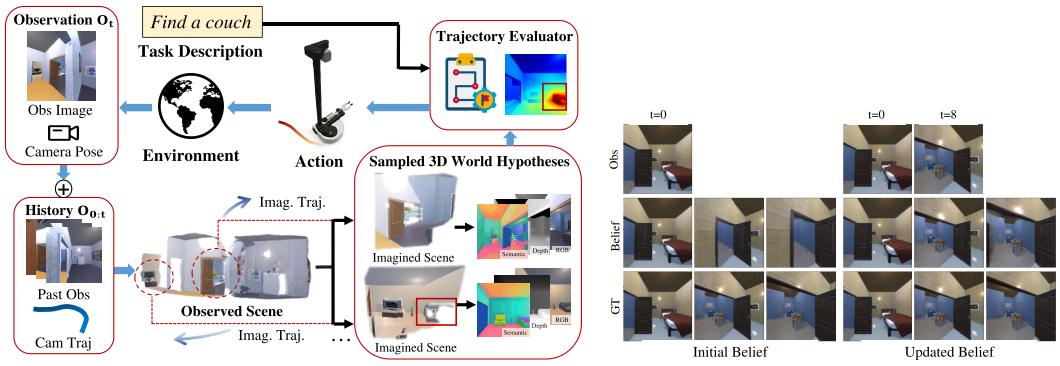
Table 3: Belief uncertainty with $K = 5$ samples on AI2-THOR validation trajectories, measured as the samples' disagreement in rendered RGB (ℓ_1) and rendered depth (standard deviation, in meters).

(a) **Imagination in place.** With the conditioning observation fixed, angles are measured from the observation viewpoint to each belief's rendering viewpoint.

Angle	RGB ℓ_1	Depth std.
0°	0.003	0.027
45°	0.043	0.241
90°	0.062	0.531
135°	0.087	0.798

(b) **Incorporating new observations.** Samples are rendered at the 4th observation's camera pose. Angles are measured from the newest observation to this pose.

# obs.	Angle	RGB ℓ_1	Depth std.
1	135°	0.083	0.605
2	90°	0.065	0.436
3	45°	0.040	0.273
4	0°	0.005	0.053



(a) Belief-guided planning.

(b) Belief updating.

Figure 5: **Belief-guided planning and updating with 3D-Belief.** (a) Dotted circles visualize imagination regions; red boxes mark potential targets in imagined renders. (b) Later views update the initial belief, revising imagined regions to guide navigation.

4.3 Experiment 3: Belief-Guided Planning with Inference-Time Scaling

We evaluate belief-guided navigation under partial observability (Figure 5a). At each step, the agent updates its belief from RGB history and camera poses (Figure 5b), samples plausible 3D scenes, and renders imagined observations along candidate paths. Open-vocabulary semantic queries score these rollouts to guide the agent toward the target. Implementation details are in Appendix D.2.

4.3.1 Object Navigation in Simulations

Simulation Environment. We follow Ehsani et al. (2024) in AI2-THOR (Kolve et al., 2017) and evaluate on 120 object-navigation episodes in unseen houses.

Task Definition. Following Ehsani et al. (2024), episodes start at random locations with a target category and may require multi-room exploration. The agent plans online from RGB observations and poses. Success requires reaching a proximity threshold with the target centered in view.

Metrics. We report Success Rate (SR), Success-weighted by Path Length (SPL), Success-weighted by Episode Length (SEL) (Eftekhar et al., 2023; Yokoyama et al., 2021), and VLM token costs, measuring success, path efficiency, time efficiency, and VLM usage.

Baselines. We compare 3D-Belief against four categories of methods, as shown below:

- 3D Reconstruction. VGGT (Wang et al., 2025a) reconstructs RGB history for frontier-based or VLM-driven waypoint selection, testing search from observed geometry.
- Lifted 2D Imagination. DFoT and NWM predict RGB views, which VGGT lifts into 3D for navigation; GPT-5 mini or Gemini 3.0 selects goals.
- VLM Agents. GPT-5 mini, Gemini 3.0, and Qwen3-VL-8B-Instruct map observations to actions.

Table 4: **Simulated and real-world object navigation.** Full results: Appendix C.2.

Domain	Model	SR%↑	SPL%↑	SEL%↑	Token (/step)↓
Simulation	VGGT (w/ frontier)	27.50	25.82	22.66	**0**
	VGGT (w/ Gemini 3.0)	25.00	24.10	22.66	1448.02
	DFoT-VGGT (w/ Gemini 3.0)	26.05	24.59	23.43	1210.08
	NWM-VGGT (w/ Gemini 3.0)	25.00	23.46	20.76	552.12
	Gemini 3.0	45.00	37.81	**41.47**	7512.70
	Qwen3-VL-8B-Instruct	18.33	14.08	16.94	220.29
	SPOC	31.67	30.97	30.56	**0**
	Ours	**59.17**	**39.07**	40.24	**0**
Real world	Gemini 3.0	23.08	–	13.55	2317.09
	Ours	**55.56**	–	**35.91**	**0**

- Policy Model. SPOC (Ehsani et al., 2024) is a learned object-search navigation policy.

Results. In Table 4, 3D-Belief achieves the best SR and SPL without VLM calls. Its SEL is 40.24% versus Gemini's 41.47%. On the same evaluation episodes, removing geometry lowers SR to 24.17%, and single hypothesis sampling gives 51.67% (Appendix C.2.2). Replacing built-in semantics with GPT-5 mini raises SR to 63.33% at 79.27 tokens/step, exposing an accuracy–cost tradeoff (Appendix C.2.3). We also report how 3D-Belief's scene buffer size, GPU use, and latency over time in Appendix C.7, and evaluate the quality of its learned semantic map in Appendix C.6.

4.3.2 Object Navigation in the Real World

We evaluate 3D-Belief on a real mobile manipulation platform (Hello Robot Stretch) in a mock apartment environment. Appendix D.2.2 provides the environment setup and trial protocol. The environments, objects, and target descriptions are all unseen during training, making this a real-world open-vocabulary object navigation setting.

Success Criteria. At the beginning of each episode, the robot is given a target description (e.g., red mug) and must explore the environment and stop when it finds the target. An episode is successful if the robot reaches within 4.0 m of the target and the target lies in the central region of its egocentric view, with facing alignment within $45°$.

Baseline. We compare against the Gemini 3.0 VLM agent, the best-performing baseline in the simulation experiments.

Result. Table 4 shows that 3D-Belief enables more reliable and efficient navigation than the Gemini 3.0 agent across real-robot episodes. It achieves a success rate of 55.56%, compared with Gemini's 23.08%, and improves SEL from 13.55% to 35.91%. These results demonstrate that explicit, online-updatable 3D beliefs can support object navigation in unseen real-world environments.

5 Conclusion

In this work, we proposed a new framework for embodied belief inference via generative 3D world modeling and identified key capabilities for practical 3D belief modeling. We then proposed 3D-Belief, which predicts unseen regions in an explicit 3D representation from partial observations and updates this belief online. Experimental results on 2D visual quality, contextual reasoning, and object navigation have demonstrated that 3D-Belief improves upon existing generative world models in 2D and 3D scene memory and imagination, as well as downstream embodied task performance.

Limitations and Future Work. 3D-Belief assumes a static world. The observed buffer grows with keyframes, and frozen Gaussians can accumulate pose and geometry drift (Appendix C.7). Future work includes dynamic scenes, buffer merging and online refinement, and language- or scene-graph-guided hypothesis sampling.

AI USE STATEMENT

Generative AI tools assisted with table formatting, grammar checks, proofreading, and literature review during manuscript preparation. Gemini 2.5 supported segmentation and object recognition in the 3D-CORE evaluation, as detailed in Appendix D.3.1. The authors are responsible for the final content of this work.

ETHICS STATEMENT

Potential safety risks from incorrect scene predictions and privacy risks from indoor 3D representations are discussed in Appendix F. Licenses and terms of use for existing assets are documented in Appendix G.

REPRODUCIBILITY STATEMENT

The model formulation and training objectives are described in Section 3, with the diffusion formulation in Appendix B. Appendix D provides training, planning, real-world setup, and evaluation details. Additional results and ablations are reported in Appendix C.2.

REFERENCES

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pp. 370–386. Springer, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5982–5994. Curran Associates, Inc., 2022.

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

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

Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, brian ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Pack Kaelbling, Andy Zeng, and Jonathan Tompson. Video language planning. In *The Twelfth International Conference on Learning Representations*, 2024.

Ainaz Eftekhar, Kuo-Hao Zeng, Jiafei Duan, Ali Farhadi, Ani Kembhavi, and Ranjay Krishna. Selective visual representations improve convergence and generalization for embodied ai. *arXiv preprint arXiv:2311.04193*, 2023.

Kiana Ehsani, Tanmay Gupta, Rose Hendrix, Jordi Salvador, Luca Weihs, Kuo-Hao Zeng, Kunal Pratap Singh, Yejin Kim, Winson Han, Alvaro Herrasti, et al. Spoc: Imitating shortest paths in simulation enables effective navigation and manipulation in the real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16238–16250, 2024.

Sensen Gao, Zhaoqing Wang, Qihang Cao, Dongdong Yu, Changhu Wang, and Jia-Wang Bian. Pixworld: Unifying 3d scene generation and reconstruction in pixel space. *arXiv preprint arXiv:2607.05373*, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028. IEEE, 2024.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Noriaki Hirose, Dhruv Shah, Ajay Sridhar, and Sergey Levine. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56, 2023.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pp. 13213–13232. PMLR, 2023.

Emiel Hoogeboom, Thomas Mensink, Jonathan Heek, Kay Lamerigts, Ruiqi Gao, and Tim Salimans. Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion. *arXiv preprint arXiv:2410.19324*, 2024.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25358–25369, 2026.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44(6): 1–15, 2025.

Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, et al. Openclip. *Zenodo*, 2021.

Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Haresh Karnan, Anirudh Nair, Xuesu Xiao, Garrett Warnell, Sören Pirk, Alexander Toshev, Justin Hart, Joydeep Biswas, and Peter Stone. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7(4): 11807–11814, 2022.

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

Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19729–19739, 2023.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Xinyang Li, Tengfei Wang, Zixiao Gu, Shengchuan Zhang, Chunchao Guo, and Liujuan Cao. Flashworld: High-quality 3d scene generation within seconds. *arXiv preprint arXiv:2510.13678*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Peiqi Liu, Zhanqiu Guo, Mohit Warke, Soumith Chintala, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13346–13355. IEEE, 2025.

Andrew Melnik, Michal Ljubljanac, Cong Lu, Qi Yan, Weiming Ren, and Helge Ritter. Video diffusion models: A survey. *arXiv preprint arXiv:2405.03150*, 2024.

Jisu Nam, Yicong Hong, Chun-Hao Paul Huang, Feng Liu, JoungBin Lee, Jiyoung Kim, Siyoon Jin, Yunsung Lee, Jaeyoon Jung, Suhwan Choi, et al. Worldcam: Interactive autoregressive 3d gaming worlds with camera pose as a unifying geometric representation. *arXiv preprint arXiv:2603.16871*, 2026.

Yuta Oshima, Yusuke Iwasawa, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Worldpack: Compressed memory improves spatial consistency in video world modeling. *arXiv preprint arXiv:2512.02473*, 2025.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20051–20060. IEEE, 2024.

Ri-Zhao Qiu, Ge Yang, Weijia Zeng, and Xiaolong Wang. Feature splatting: Language-driven physics-based scene synthesis and editing. *arXiv preprint arXiv:2404.01223*, 2024.

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

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6121–6132, 2025.

Barbara Roessle, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, Angela Dai, and Matthias Nießner. L3dg: Latent 3d gaussian diffusion. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Katja Schwarz, Norman Müller, and Peter Kontschieder. Generative gaussian splatting: Generating 3d scenes with video diffusion priors. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27510–27520. IEEE, 2025.

Dhruv Shah, Benjamin Eysenbach, Gregory Kahn, Nicholas Rhinehart, and Sergey Levine. Rapid exploration for open-world navigation with latent goal models. *arXiv preprint arXiv:2104.05859*, 2021.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

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

Samuel Triest, Matthew Sivaprakasam, Sean J Wang, Wenshan Wang, Aaron M Johnson, and Sebastian Scherer. Tartandrive: A large-scale dataset for learning off-road dynamics models. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 2546–2552. IEEE, 2022.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Nicolas von Lützow, Barbara Rössle, Katharina Schmid, and Matthias Nießner. Gaussiangpt: Towards autoregressive 3d gaussian scene generation. *arXiv preprint arXiv:2603.26661*, 2026.

Dehui Wang, Congsheng Xu, Rong Wei, Yue Shi, Shoufa Chen, Dingxiang Luo, Tianshuo Yang, Xiaokang Yang, Yusen Qin, Rui Tang, et al. Rein3d: Reinforced 3d indoor scene generation with panoramic video diffusion models. *arXiv preprint arXiv:2604.10578*, 2026a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

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

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026b.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

Jiacong Xu, Hanwen Jiang, Zhixin Shu, Kalyan Sunkavalli, Vishal M Patel, and Yiqun Mei. Wonder: Video world model done better. *arXiv preprint arXiv:2607.26037*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. *arXiv preprint arXiv:2507.12508*, 2025b.

Sheng Ye, Zhen-Hui Dong, Ruoyu Fan, Tian Lv, and Yong-Jin Liu. Semgs: Feed-forward semantic 3d gaussian splatting from sparse views for generalizable scene understanding. *arXiv preprint arXiv:2603.02548*, 2026.

Naoki Yokoyama, Sehoon Ha, and Dhruv Batra. Success weighted by completion time: A dynamics-aware evaluation criteria for embodied navigation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1562–1569. IEEE, 2021.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

Hongxin Zhang, Zeyuan Wang, Qiushi Lyu, Zheyuan Zhang, Sunli Chen, Tianmin Shu, Behzad Dariush, Kwonjoon Lee, Yilun Du, and Chuang Gan. Combo: compositional world models for embodied multi-agent cooperation. *arXiv preprint arXiv:2404.10775*, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: Learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025a.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

Siyuan Zhou, Yilun Du, Yuncong Yang, Lei Han, Peihao Chen, Dit-Yan Yeung, and Chuang Gan. Learning 3d persistent embodied world models. *arXiv preprint arXiv:2505.05495*, 2025b.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

A Extended Related Work

Table 5: Comparison of generative world model capabilities. ✓ indicates a reported capability and ✗ indicates it is not reported under these definitions. "Scene Mem." denotes 3D scene memory of observed content; "2D Imag." denotes pixel-space multi-hypothesis imagination, including rendered 3D samples; "3D Imag." denotes multi-hypothesis generation of explicit 3D representations beyond observed geometry; and "Semantic" denotes predicted semantic features or labels associated with scene elements, rather than appearance features or text conditioning. "Sequential" denotes online processing of newly arriving observations; fixed multi-view conditioning and autoregressive scene extension alone do not qualify. Sequential processing does not by itself establish a belief update that retains observed geometry while revising imagined content.

Models	**Scene Mem.**	**2D Imag.**	**3D Imag.**	**Sequential**	**Semantic**
DFoT (Song et al., 2025)	✗	✓	✗	✓	✗
NWM (Bar et al., 2025)	✗	✓	✗	✓	✗
Persistent EWM (Zhou et al., 2025b)	✓	✓	✗	✓	✗
GEN3C (Ren et al., 2025)	✓	✓	✗	✓	✗
ViewCrafter (Yu et al., 2024)	✓	✓	✗	✗	✗
TrajectoryCrafter (Yu et al., 2025)	✓	✓	✗	✗	✗
MVSplat (Chen et al., 2024)	✓	✗	✗	✗	✗
VGGT (Wang et al., 2025a)	✓	✗	✗	✗	✗
CUT3R (Wang et al., 2025b)	✓	✗	✗	✓	✗
DFM (Tewari et al., 2023)	✓	✓	✗	✓	✗
Lyra 2.0 (Shen et al., 2026)	✓	✓	✓	✗	✗
Rein3D (Wang et al., 2026a)	✓	✓	✓	✗	✗
FlashWorld (Li et al., 2025)	✗	✓	✓	✗	✗
GaussianGPT (von Lützow et al., 2026)	✓	✓	✓	✗	✗
GGS (Schwarz et al., 2025)	✓	✓	✓	✗	✗
PixWorld (Gao et al., 2026)	✓	✓	✓	✗	✗
Gen3R (Huang et al., 2026)	✓	✓	✓	✗	✗
SemGS (Ye et al., 2026)	✓	✗	✗	✗	✓
3D-Belief	✓	✓	✓	✓	✓

B Diffusion Training and Sequential Inference

B.1 Diffusion Training Formulation

Let $\alpha_\tau = 1 - \beta_\tau$ and $\bar{\alpha}_\tau = \prod_{s=1}^{\tau} \alpha_s$. The forward process is

$$q(o_\tau^{\text{trgt}} \mid o_{\tau-1}^{\text{trgt}}) = \mathcal{N}\big(o_\tau^{\text{trgt}}; \sqrt{\alpha_\tau}\, o_{\tau-1}^{\text{trgt}}, \beta_\tau I\big)\,. \tag{6}$$

The reverse process reconstructs o^{trgt} conditioned on o^{ctxt} and the camera parameters:

$$\begin{aligned} &p_\theta(o_{0:T}^{\text{trgt}} \mid o^{\text{ctxt}}; \phi^{\text{ctxt}}, \phi^{\text{trgt}}) \\ &\quad = p(o_T^{\text{trgt}}) \prod_{\tau=1}^{T} p_\theta\big(o_{\tau-1}^{\text{trgt}} \mid o_\tau^{\text{trgt}}, o^{\text{ctxt}}; \phi^{\text{ctxt}}, \phi^{\text{trgt}}\big)\,. \end{aligned} \tag{7}$$

At each step, Φ_θ predicts a 3DGS scene, which is rendered to obtain an estimate of the clean target observation:

$$z_{\tau-1} = \Phi_\theta(o^{\text{ctxt}}, o_\tau^{\text{trgt}}; \tau, \phi^{\text{ctxt}}, \phi^{\text{trgt}}), \tag{8}$$

$$\hat{o}_0^{\text{trgt}} = \mathcal{G}(z_{\tau-1}, \phi^{\text{trgt}}). \tag{9}$$

The rendered estimate and the current noisy observation jointly determine the reverse-step mean:

$$\mu_\tau = \frac{\sqrt{\bar{\alpha}_{\tau-1}}\beta_\tau}{1-\bar{\alpha}_\tau}\,\hat{o}_0^{\text{trgt}} + \frac{\sqrt{\alpha_\tau}(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau}\,o_\tau^{\text{trgt}}, \tag{10}$$

$$o_{\tau-1}^{\text{trgt}} \sim \mathcal{N}\left(\mu_\tau, \tilde{\beta}_\tau I\right), \tag{11}$$

$$\tilde{\beta}_\tau = \frac{1-\bar{\alpha}_{\tau-1}}{1-\bar{\alpha}_\tau}\,\beta_\tau. \tag{12}$$

Here, Φ_θ is the neural network shown in Figure 2b. At test time, sampling starts from $o_T^{\text{trgt}} \sim \mathcal{N}(0, I)$ and iterates from $\tau = T$ to $\tau = 1$ using Eqs. (8)–(11). The sequence of predicted Gaussian primitives is conditioned on the sampled target observations at each reverse step.

B.2 Sequential Inference and 3D Memory Updates

Training uses context–target image pairs without accumulated observed Gaussians. At inference, we use the trained model to update $z^t = z_o^t \cup z_i^t$ as new observations arrive. Here, t indexes observation steps and τ indexes diffusion steps.

Given a new observation o^{t+1}, we retain z_o^t, use o^{t+1} as context, and initialize the target image with Gaussian noise. At each denoising step, Eq. (8) predicts context Gaussians $z_{\tau-1}^{\text{ctxt}}$ and target Gaussians $z_{\tau-1}^{\text{trgt}}$. We express these Gaussians in the shared scene coordinate frame using the camera poses and render the combined scene:

$$\hat{o}_{\tau-1}^{\text{trgt}} = \mathcal{G}\left(z_o^t \cup z_{\tau-1}^{\text{ctxt}} \cup z_{\tau-1}^{\text{trgt}}, \phi^{\text{trgt}}\right). \tag{13}$$

This rendering estimates the clean target observation and guides the next denoising step. Previous observations therefore influence generation through rendering, while the neural predictor takes only the latest context image and the noisy target image. During denoising, z_o^t remains fixed, while the current context and target Gaussians are recomputed at each step.

Denoting the final predictions by z^{ctxt} and z^{trgt}, we update

$$z_o^{t+1} = z_o^t \cup z^{\text{ctxt}}, \tag{14}$$

$$z_i^{t+1} = z^{\text{trgt}}, \tag{15}$$

$$z^{t+1} = z_o^{t+1} \cup z_i^{t+1}. \tag{16}$$

We initialize $z_o^0 = \varnothing$. The previous imagined Gaussians are replaced as new evidence arrives. Here, $\cup$ denotes Gaussian concatenation.

C Extended Results

C.1 Extended Vision Results

C.1.1 More Qualitative Results

We include more qualitative results in Figure 6 and 7.

C.2 Extended Planning Results

C.2.1 Full Navigation Results

We report the full object navigation results in Table 6, including additional baselines and inference speed comparisons.

C.2.2 Ablation Studies

Table 7 evaluates all variants on the same 120 test episodes with the same evaluation code. The full model uses $K = 3$ beliefs and its built-in semantic field, with no external VLM. Removing geometry lowers SR by 35.00 percentage points; using a single hypothesis lowers SR by 7.50 points. These comparisons isolate the contributions of explicit geometry and multi-hypothesis sampling under a matched evaluation protocol.

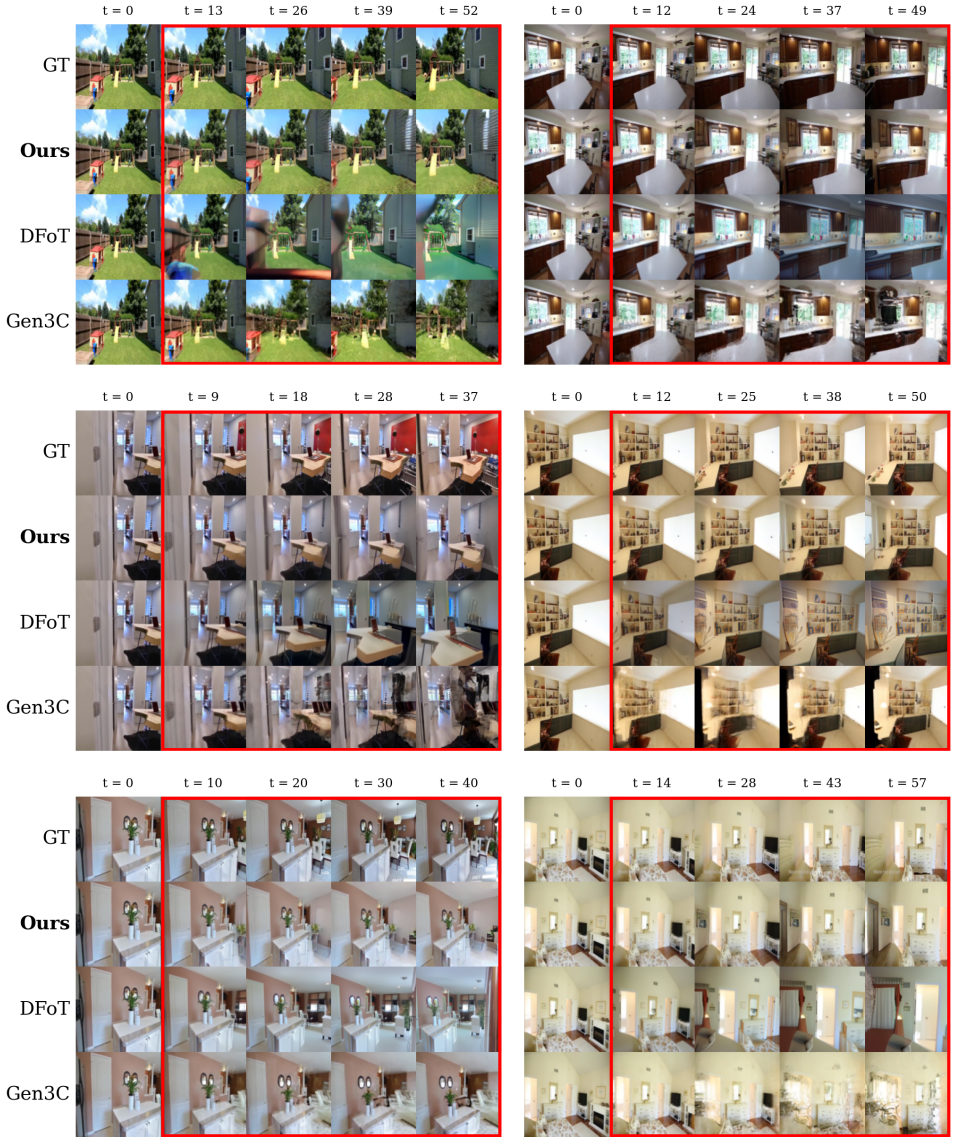


Figure 6: Qualitative results on RE10K.

C.2.3 Built-in Semantics versus External VLMs

Table 8 replaces the built-in semantic head with external VLM scoring on the same 120 test episodes with the same evaluation code as Table 7. The full model achieves 59.17% SR with zero VLM tokens. GPT-5 mini raises SR to 63.33% at 79.27 tokens/step, while Gemini 3.0 gives 60.83% at 335.02 tokens/step. Thus, external scoring can improve accuracy, while built-in semantics permit direct rendering and querying of the 3D field without external VLM calls. The token measurements quantify this deployment tradeoff; they do not measure the local model's total computation.

C.3 Full Results on 3D-CORE

C.3.1 Object Completion

Table 9 reports object completion at different visibility levels. At 95% visibility, DFoT-VGGT and 3D-Belief have similar Chamfer distances (0.189 vs. 0.187 m), whereas the gap is larger at 55% (0.830 vs. 0.216 m), where more geometry must be inferred.

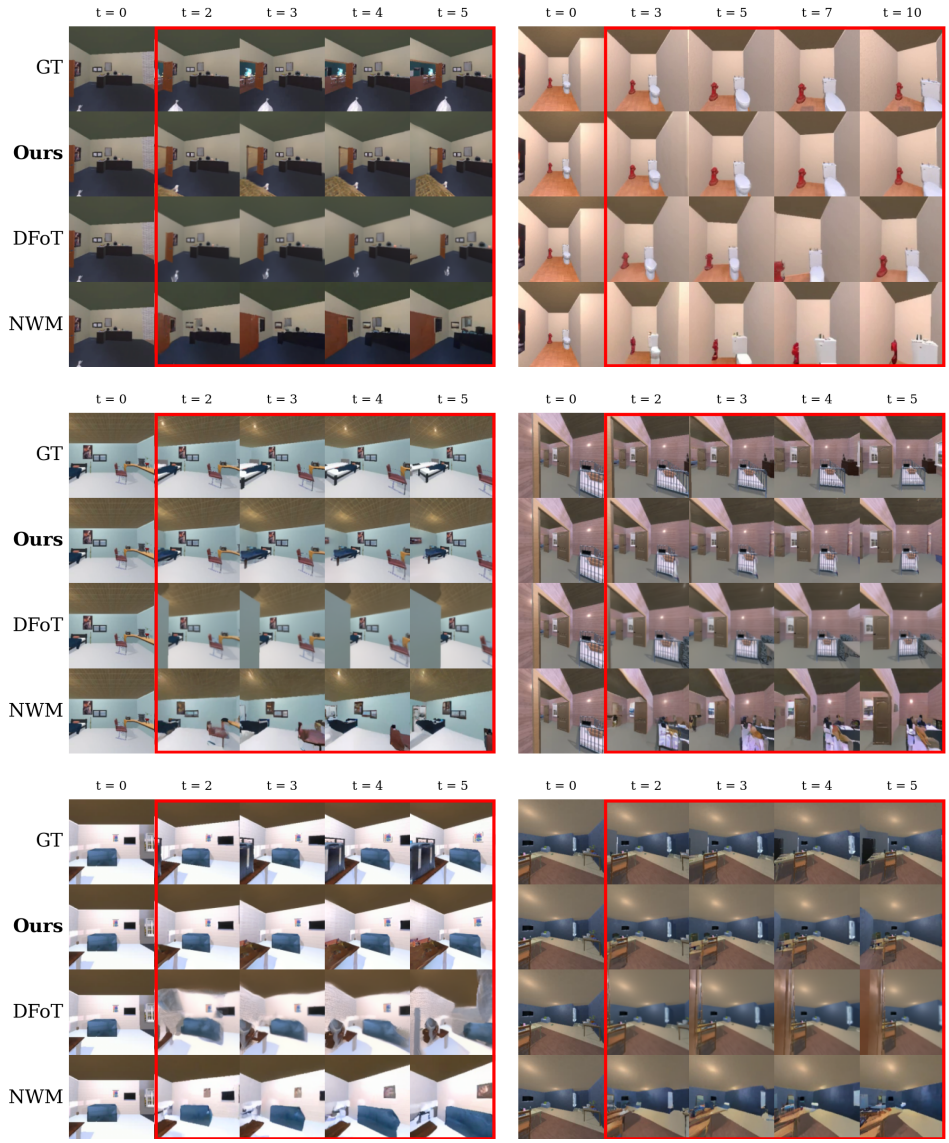


Figure 7: Qualitative results on AI2THOR.

C.3.2 Room Completion

We report room completion results with more metrics in Table 10.

C.4 Spatial Reasoning QA

C.4.1 Evaluation on SAT-Real

To compare 3D-Belief with prior world models on the benefits they provide for reasoning about motion and space, we evaluate on the SAT-Real benchmark, which probes VLM reasoning by presenting two RGB images and asking binary questions over multiple subsets, namely Egocentric Movement (EM), Object Movement (OM), Egocentric-Allocentric Perspective Taking (Pers), Goal Aiming (GA), and Egocentric Action Consequence (EA).

Baselines. We follow MindJourney (Yang et al., 2025b) to perform test-time scaling with a world model: given the current observation, we use beam search to simulate plausible camera motions and generate additional visual evidence using the world model, which is then provided to the

Table 6: **Object navigation results in simulations and the real world.** Higher is better for SR, SPL, and SEL; lower is better for inference time and VLM tokens. Best results are highlighted in **bold**, and second-best simulation results are underlined.

Model	SR% ↑	SPL% ↑	SEL% ↑	Inf. (s/step) ↓	Token (/step) ↓
Simulations					
VGGT (w/ frontier)	27.50	25.82	22.66	25.13	**0**
VGGT (w/ GPT-5m)	25.00	24.17	22.40	17.92	468.66
VGGT (w/ Gemini 3.0)	25.00	24.10	22.66	18.87	1448.02
DFoT-VGGT (w/ GPT-5m)	21.01	20.73	20.48	32.80	423.86
DFoT-VGGT (w/ Gemini 3.0)	26.05	24.59	23.43	45.57	1210.08
NWM-VGGT (w/ GPT-5m)	25.00	23.35	23.00	62.75	315.28
NWM-VGGT (w/ Gemini 3.0)	25.00	23.46	20.76	54.41	552.12
GPT-5m	23.33	17.66	20.19	11.10	6644.90
Gemini 3.0	<u>45.00</u>	<u>37.81</u>	**41.47**	7.58	7512.70
Qwen3-VL-8B-Instruct	18.33	14.08	16.94	2.30	<u>220.29</u>
SPOC	31.67	30.97	30.56	**0.22**	**0**
3D-Belief	**59.17**	**39.07**	<u>40.24</u>	<u>2.14</u>	**0**
Real World					
Gemini 3.0	23.08	–	13.55	4.52	2317.09
3D-Belief	**55.56**	–	**35.91**	**2.02**	**0**

Table 7: Ablations on the same 120 test episodes. The full model uses $K = 3$ beliefs; all rows use built-in semantics and zero VLM tokens.

Models	SR% ↑	SPL% ↑	SEL% ↑	Token (/step) ↓
3D-Belief	59.17	39.07	40.24	0
w/o geometry	24.17	14.61	17.83	0
single hypothesis ($K = 1$)	51.67	33.14	33.98	0

VLM as context for answering questions. We compare 3D-Belief against two video world models, **SWM** (Yang et al., 2025b) and **SVC** (Zhou et al., 2025a), using the same inference protocol and budget (same beam width and rollout length). All methods are paired with the same VLM backbone (Gemini-3.0), and differ only in the auxiliary world model used to generate the additional context.

Results. Table 11 shows that augmenting Gemini-3.0 with 3D-Belief yields the best overall performance (88.7%), outperforming both the base VLM (85.3%) and the best video-world-model baseline (86.7%). The gains are concentrated on subsets that require reasoning about camera motion and 3D space: 3D-Belief improves EM to 100.0% (vs. 95.7%), and substantially boosts EA (97.3% vs. 83.8%) and GA (94.1% vs. 85.3%). These improvements are consistent with 3D-Belief providing an explicit and more geometrically grounded belief update, which better supports motion- and space-related inference under viewpoint changes. Meanwhile, performance on OM and Pers is not the best among methods (e.g., OM 73.9% vs. 78.3%, Pers 75.8% vs. 84.8%), suggesting that fine-grained

Table 8: Semantic-scoring variants on the same 120 test episodes. External VLMs replace the built-in semantic head; the full model uses no external VLM.

Models	SR% ↑	SPL% ↑	SEL% ↑	Token (/step) ↓
3D-Belief	59.17	39.07	40.24	0
w/o semantics (w/ GPT-5m)	63.33	41.99	45.29	79.27
w/o semantics (w/ Gemini 3.0)	60.83	39.55	41.42	335.02

Table 9: Object completion across visibility levels. Chamfer is in meters.

Models	Visibility	BEV IoU↑	3D IoU↑	Chamfer↓	SigLIP↑	Recognition↑
DFoT-VGGT	0.05	0.110	0.064	2.681	0.265	0.126
	0.55	0.362	0.243	0.830	0.798	0.767
	0.95	0.372	0.242	0.189	0.857	0.838
3D Belief	0.05	**0.147**	**0.083**	**2.435**	**0.329**	**0.165**
	0.55	**0.484**	**0.318**	**0.216**	**0.855**	**0.930**
	0.95	**0.535**	**0.369**	**0.187**	**0.884**	**0.909**

Table 10: Results on Room Completion.

Models	Obj. Precision↑	Obj. Recall↑	Obj. F1↑	Occ. Acc.↑	IoU Free↑	IoU Occupied↑	Occ. IoU↑
DFoT-VGGT	0.639	**0.516**	0.531	0.252	0.104	0.115	0.110
3D Belief	**0.678**	0.490	**0.536**	**0.900**	**0.648**	**0.235**	**0.442**

object-centric dynamics remain challenging and may require stronger object-level temporal modeling or higher-fidelity appearance tracking.

Table 11: Results on SAT-Real.

	SAT Real					
	Avg	EM	OM	EA	GA	Pers
Gemini-3.0	85.3	95.7	**78.3**	83.8	85.3	**84.8**
+ SWM	86.7	95.0	70.0	94.1	89.7	81.3
+ SVC	86.0	95.7	**78.3**	86.5	**94.1**	75.8
+ **3D Belief**	**88.7**	**100.0**	73.9	**97.3**	**94.1**	75.8

C.5 Belief Sampling Diversity and Sequential Updating

C.5.1 Diversity Away from Observations

We evaluate AI2-THOR validation trajectories. From a single conditioning frame, we sample $K = 5$ autoregressive beliefs along identical camera trajectories and compare appearance and geometry through RGB ℓ_1 disagreement and depth standard deviation. Table 12 reports the observed view and three imagined keyframes. Disagreement grows monotonically with angular separation: RGB disagreement rises from 0.003 to 0.087, and depth standard deviation from 0.027 to 0.798 m. Samples therefore agree most closely where visual evidence is available and differ more in unseen regions. This measures sample diversity, rather than calibrated uncertainty or reconstruction accuracy.

C.5.2 Contraction as Evidence Arrives

To test updating separately from viewpoint changes, we fix one evaluation pose (a later trajectory frame) per AI2-THOR validation episode. Four observations progressively approach this pose and enter through the sequential update in Eq. (2). After each update, we sample $K = 5$ beliefs, imagine autoregressively toward the evaluation pose, and render all hypotheses at that same pose. Table 13 shows monotonically decreasing disagreement as observations accumulate. RGB disagreement falls from 0.083 to 0.005 and depth standard deviation from 0.605 to 0.053 m; the fourth observation is at the evaluation pose. Together with Table 12, this supports diverse predictions without evidence and contraction as evidence arrives.

Table 12: Diversity of $K = 5$ belief samples from one conditioning frame on AI2-THOR validation trajectories. Angles are relative to the conditioning view.

View	Angle	RGB ℓ_1	Depth std. (m)
Observed	0°	0.003	0.027
Keyframe 0	45°	0.043	0.241
Keyframe 1	90°	0.062	0.531
Keyframe 2	135°	0.087	0.798

Table 13: Belief contraction at a fixed evaluation pose, using $K = 5$ samples after each sequential update. Angle is from the newest observation to the evaluation view.

Observations	Angle	RGB ℓ_1	Depth std. (m)
1	135°	0.083	0.605
2	90°	0.065	0.436
3	45°	0.040	0.273
4	0° (observed)	0.005	0.053

C.6 Direct Evaluation of Open-Vocabulary Querying

Following LERF's relevance-based evaluation (Kerr et al., 2023), we compare bare category names (e.g., "sofa") with unseen free-form descriptions (e.g., "a couch where I can rest") on AI2-THOR evaluation episodes. For the frame where the target is visible, we render its semantic map and evaluate pixel relevance against the ground-truth target mask. Targets covering less than 1% of the image are excluded. Top-1 localization measures whether the maximum-relevancy pixel lies inside the mask; AUROC measures threshold-free ranking of target versus non-target pixels, with chance at 50%.

Table 14 shows that free-form descriptions improve localization by 4.4 percentage points and AUROC by 2.9 points. OpenCLIP supplies the semantic embedding, while DINOv3 regularization encourages spatial coherence.

Table 14: Direct semantic-query evaluation on AI2-THOR evaluation episodes. Query descriptions are unseen; target masks smaller than 1% of the image are excluded.

Query phrasing	Top-1 loc. (%)↑	AUROC (%)↑
Fixed-category	48.9	77.9
Open-vocabulary	**53.3**	**80.8**

C.7 Horizon, Memory, and Latency

We profile up to 100 generation passes on one NVIDIA RTX 5090 (32 GB), with one generation per keyframe and approximately five environment steps between keyframes. Table 15 reports the observed Gaussian count, peak VRAM, and mean latency. The observed buffer grows by approximately 13,222 Gaussians per generation, reaching 1.3×10^6 primitives and 16.16 GiB peak VRAM at 100 keyframes. Latency varies across horizons and is not monotonic. Successful navigation episodes typically require 25–50 generation passes, corresponding to roughly 150–300 environment steps.

Training uses context–target pairs, while inference rolls out autoregressively. Fixed-size diffusion conditioning does not imply constant total system cost: accumulating observed Gaussians increases storage and rendering work. Current opacity/scale pruning limits unnecessary primitives; voxel-level merging and SLAM-style online refinement of frozen observed Gaussians are future directions for controlling memory growth and accumulated pose/geometry drift. The 100-keyframe profile is the tested horizon, not an architectural maximum.

Table 15: Rollout profiling on one RTX 5090 (32 GB). Counts refer to observed Gaussians; latency is reported in seconds per step.

Horizon (keyframes)	Observed Gaussians	Peak VRAM (GiB)	Mean latency (s/step)
5	4.9×10^4	12.06	1.55
25	3.0×10^5	12.91	3.05
50	6.3×10^5	14.00	2.36
100	1.3×10^6	16.16	2.04

D Implementation Details

D.1 Model Training

D.1.1 Datasets

We train our model on a composite dataset, which consists of SPOC (Ehsani et al., 2024), RealEstate10K (Zhou et al., 2018), DL3DV (Ling et al., 2024), and Habitat-Matterport 3D (Ramakrishnan et al., 2021). At each step, we randomly select a subset with equal weights and sample frames. Each sample includes the context view and the target view with randomly spaced frames. To align the datasets, we set the minimum interval to 2. The maximum intervals for SPOC, RealEstate10K, DL3DV, and Habitat-Matterport 3D are 10, 100, 10, and 15, respectively.

D.1.2 Training Settings

3D-Belief The U-ViT backbone (Hoogeboom et al., 2023; 2024) used to encode input images and the depth predictor used to estimate Gaussian primitives are initialized from pretrained weights: the DFoT encoder (Song et al., 2025) and the MVSplat multi-view depth predictor (Chen et al., 2024), respectively, both pretrained on RealEstate10K (Zhou et al., 2018). We use AdamW optimizer with a learning rate of 2×10^{-5} and weight decay of 0.001 for training, applying linear warm-up in the beginning 10K steps followed by cosine decay. The model can still be trained effectively and converge even with a small batch size. We set the batch size to 1 and trained for a total of 250K steps on a single H100 80GB. We use squared ℓ_2 RGB reconstruction and LPIPS losses with weights of 1.0, together with edge-aware inverse-depth smoothness weighted by 0.1. RGB reconstruction, depth smoothness, and available depth supervision are applied to context, target, and intermediate views. LPIPS is applied to context and target views. Semantic supervision includes context, target, and intermediate views. Relative to the target-view losses, context-view losses receive a factor of 0.9 during base training and 0.7 during semantic-head fine-tuning. Intermediate-view losses are averaged across views and weighted by 5.0.

DFoT We use the DFoT model (Song et al., 2025) trained on RealEstate10K (Zhou et al., 2018) as the base model, and finetune it on the SPOC (Ehsani et al., 2024) dataset with light horizontal flip augmentation. Training uses AdamW with learning rate 1×10^{-5}, weight decay 0.01, per-GPU batch size 4 with gradient accumulation of 2, and runs for 48 epochs on 4 $\times$ H100 80GB chips. The diffusion process follows a cosine schedule with sigmoid loss weighting.

NWM We use the official pretrained NWM (Bar et al., 2025) CDiT/XL model, which covers four robotics datasets (SCAND (Karnan et al., 2022), TartanDrive (Triest et al., 2022), RECON (Shah et al., 2021), and HuRoN (Hirose et al., 2023)) and Ego4D (Grauman et al., 2022) videos. Then, we finetune it on SPOC (Ehsani et al., 2024) dataset using AdamW optimizer with a learning rate of 8×10^{-5} and a batch size of 16, and a total of 200K training steps. Training is done on a single H100 80GB chip.

VGGT We utilize the official pretrained VGGT-1B model (Wang et al., 2025a) with both camera and depth heads enabled. We finetune it on the SPOC (Ehsani et al., 2024) dataset on 238×238 inputs (patch size 14) using AdamW (learning rate $1 \times 10^{-}5$, weight decay 0.05) for 60 epochs with an effective batch size of 96. Training is done on 4 $\times$ GH200 96GB chips.

SPOC We finetune the pretrained SigLIP-ViTb-3 CHORES-Nav model (Ehsani et al., 2024) (an EarlyFusionCnnTransformer policy) on the SPOC (Ehsani et al., 2024) RERENDERED mixture, which combines all four ObjectNav task types (ObjectNavType, ObjectNavRoom, ObjectNavRelAttribute, and ObjectNavLocalRef). The policy consumes raw navigation and manipulation camera streams together with the agent's last-action and in-hand-object signals. Training is performed on 4 $\times$ H100 80GB chips with a per-GPU batch size of 6 and a sliding window of 50 steps, using a learning rate of 2×10^{-5} and mixed-precision (fp16) for 50 epochs over a 100,000-sample subset of the dataset.

D.2 Planning

D.2.1 Object Navigation Tasks

We provide additional implementation details for the object navigation experiments. At each step, the agent updates its 3D belief z^t from the latest egocentric RGB stream and poses. A frontier goal sampler then proposes the next possible waypoint goals, computes candidate paths with an A* planner, and performs mental simulation by rendering imagined observations along each path from z^t. These rollouts are scored by goal progress and information gain via semantic-map queries. The agent executes a short prefix of the best plan and repeats.

We describe each key component below.

Waypoint Sampling. Planning directly to the final goal is difficult under partial observability, since large portions of the environment are unseen and therefore uncertain. We instead plan an intermediate waypoint at each step. Waypoints are selected in two cases. If the final goal (a target object or a target location) has already appeared in the observed history $o^{1:t}$, we set the waypoint to the estimated goal position by querying the semantic map. Otherwise, we sample a small set of candidate waypoints in currently unobserved regions within an exploration radius centered at the agent's current position.

Simulating and Evaluating the Paths. For each of the waypoints, we can sample a path and simulate the agent's imagined observations along that path using the predicted scene representation z^t. These rollouts estimate how informative or goal-directed a path is. For example, whether the target object becomes visible in the imagined views or whether the rendered semantics indicate proximity to the goal. Specifically, we score each path by querying the rendered semantic feature maps to assess progress and promise from the imagined observations. If the semantics map indicates that it is likely that the object is on the imagined path, then we will score the path highly.

Execution. We execute the first T steps of the highest-scoring plan, then re-estimate the scene representation from the newly acquired observations and repeat the procedure iteratively until the goal is reached.

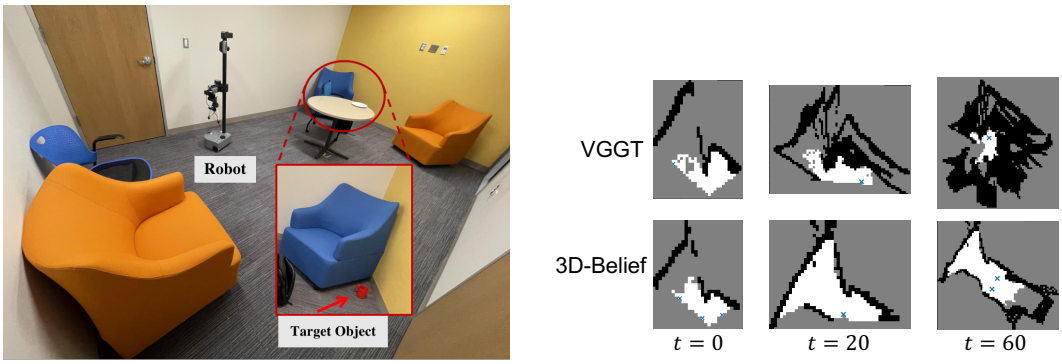
Unless explicitly ablated, planning uses $K = 3$ hypotheses per decision; the diversity and contraction evaluations instead use $K = 5$. The results for the 3D-Belief model and SPOC are obtained using a single H100 80GB chip for inference, while other evaluations rely on a single A100 80GB chip.

D.2.2 Real-World Experiments

We use five target objects and three predefined starting locations, giving 15 trial configurations, each repeated at least three times. 3D-Belief succeeds in 30 of 54 evaluated episodes, and Gemini 3.0 in 12 of 52. Success requires a final target distance below 4.0 m and facing alignment within $45°$, matching the simulation success thresholds. Figure 8a shows the setup. Simulation and real-world results use separately trained models; single-model cross-domain transfer was not evaluated.

We use the Stretch 3 Mobile Manipulator (Hello Robot) as our real-world embodied platform. The 3D-Belief model and all baselines run on an external GPU client (an NVIDIA RTX 4090 desktop), and the robot communicates with the client via ROS 2.

We mount an RGB-D camera (*Intel RealSense* D455) on the robot and use the RGB stream as the visual input. The robot's pose is estimated using onboard wheel-encoder odometry. Together, the RGB observations and associated poses form an egocentric stream that serves as input to our models. The scene predictor takes RGB images and camera poses as input, while sensed depth is used for metric-scale calibration. However, the real-world vision datasets used for training provide camera poses from *Structure-from-Motion (SfM)*, which are defined only up to an unknown global scale;



(a) Real-world setup.

(b) Long-horizon occupancy.

Figure 8: **Real-world setup and long-horizon scene memory.** (a) Example setup for the real-world object navigation tasks with 3D-Belief. The small image shows a zoom-in view near the blue couch with a target object hidden. (b) Comparison of inferred occupancy maps generated by 3D-Belief and VGGT (w/ frontier) over time in the same task.

as a result, the model's depth predictions are not inherently metric. To recover metric-consistent geometry for planning, we align the predicted depth map $\hat{d}$ to the sensed depth d by estimating a per-sequence scalar scale s via robust regression, and apply this scale to the depth maps produced by the Multi-view Depth Predictor (Sec. 3.2).

D.3 Reasoning

D.3.1 3D-CORE Metrics

For *Object Completion*, BEV IoU, 3D IoU, and Chamfer distance are calculated by the extracted point cloud of the imagined object (segmented by Gemini 2.5 (Comanici et al., 2025)) and the GT point cloud (segmented by GT masks). SigLIP is calculated by comparing the predicted and GT views at the same camera poses. Recognition is defined by the success rate of target object recognition by VLMs in the imagined views.

For *Room Completion*, Obj. F1 is calculated by comparing a list of VLM (Gemini 2.5 (Comanici et al., 2025)) recognized objects in imagined views, with the GT object list in that view. Occ. Acc. is defined by the accuracy on cells where both gt and pred are known (0 or 1). IoU Free is the IoU for free cells (class 0), IoU Occupied is the IoU for occupied cells (class 1), Occ. IoU is the average of the previous two. For *Object Permanence*, LPIPS and SigLIP are calculated using the first and the last rendered image, whose camera poses are the same. Unless otherwise specified, reasoning experiments use a single A100 80GB for inference; the horizon profile in Appendix C.7 uses an RTX 5090.

Chamfer is reported in meters; IoU, F1, occupancy accuracy, and recognition are unitless fractions in $[0, 1]$. SigLIP is a unitless image-similarity score, and LPIPS is a unitless perceptual distance. Larger IoU/F1/accuracy/recognition/SigLIP and smaller Chamfer/LPIPS indicate better performance. Comparisons are meaningful within each metric column, not across metrics with different scales.

E Discussion

How does 3D imagination affect 2D rendering quality?

While 3D-Belief is designed for explicit 3D reasoning, it also improves *2D* rendering quality as shown in Section 4.1. Because predictions are structured in an explicit 3D scene, generated frames are constrained by coherent geometry instead of being synthesized purely in pixel space. This acts as a strong prior for cross-view consistency: once surfaces and free space are established in 3D, new viewpoints are rendered by projection, reducing viewpoint ambiguity and limiting texture drift or identity switches over long horizons. This yields better perceptual/distributional metrics

(lower LPIPS/FID/FVD) and higher PSNR/SSIM. Qualitatively (Appendix C.1.1), 2D-only baselines often drift under large camera motion, whereas 3D-Belief better preserves rigid structure and object permanence across revisits.

How do the key 3D belief model capacities affect the embodied task performance?

A key advantage of 3D-Belief is maintaining a spatially consistent scene memory over long rollouts. In Figure 8b, VGGT-based occupancy quickly accumulates duplicated structure, and maps often break afterwards, trapping the agent. In contrast, 3D-Belief remains coherent and steadily expands over much longer horizons, supporting continued exploration and replanning. This failure mode is worse for DFoT-VGGT and NWM-VGGT, where both lifting and video prediction drift under long horizons, corrupting the 3D cache and degrading planning (Xiao et al., 2025; Melnik et al., 2024; Oshima et al., 2025).

On the matched evaluation across 120 test episodes, reducing sampling from $K = 3$ to $K = 1$ lowers SR from 59.17% to 51.67%, while removing geometry lowers it to 24.17% (Appendix C.2.2). The diversity and fixed-pose updating tests in Appendix C.5 separately show sample disagreement growing away from observations and contracting as evidence arrives. Built-in semantics enable zero-token querying; replacing them with external VLM scoring can improve success at additional token cost (Appendix C.2.3).

F Broader Impacts

3D-Belief aims to improve the ability of embodied agents to reason about partially observed environments by maintaining and updating explicit 3D beliefs over both observed and unobserved space. It has potential benefits for assistive robotics, object search, warehouse inspection, and navigation in areas like homes or hospitals. By representing uncertainty explicitly and supporting multi-hypothesis reasoning, it may help agents explore more efficiently and make better-informed decisions.

However, predicting unobserved regions also introduces risks. Plausible but incorrect beliefs could lead to unsafe robot behavior if treated as facts. In indoor settings, explicit 3D scene representations may also capture sensitive information about private spaces and personal belonging, raising privacy and surveillance concerns. This work should therefore be viewed as a research step rather than a deployment-ready system. Practical deployments should include conservative planning, collision checking, privacy-preserving data handling, and clear separation between observed evidence and imagined scene completions.

G Asset Licenses

This section documents the licenses and terms of use for existing assets used in the paper. We do not claim ownership of these third-party assets, and we comply with their corresponding license and usage requirements.

Table 16: Existing assets used in this work and their licenses or terms of use.

Asset	License / Terms
AI2-THOR (Kolve et al., 2017)	Apache-2.0
ProcTHOR / ProcTHOR-10K (Deitke et al., 2022)	Apache-2.0
Objaverse (Deitke et al., 2023)	ODC-By v1.0 dataset license; per-object Creative Commons licenses
RealEstate10K (Zhou et al., 2018)	CC-BY 4.0
OpenCLIP (Ilharco et al., 2021)	MIT
DINOv3 (Siméoni et al., 2025)	DINOv3 License
DFoT (Song et al., 2025)	MIT
Navigation World Models (NWM) (Bar et al., 2025)	CC-BY-NC 4.0
VGGT (Wang et al., 2025a)	Meta Research Materials License
Qwen3-VL-8B-Instruct (Yang et al., 2025a)	Apache-2.0