3D-Belief: A Generative 3D World Model for Embodied Reasoning and Planning

Yifan Yin [1] **Zehao Wen** [1] **Jieneng Chen** [1] **Zehan Zheng** [1] **Nanru Dai** [1] **Haojun Shi** [1] **Suyu Ye** [1] **Aydan Huang** [1] **Zheyuan Zhang** [1] **Alan Yuille** [1] **Jianwen Xie** [2] **Ayush Tewari** [3] **Tianmin Shu** [1]

Abstract

Recent advances in visual generative models have shown the promise of learning generative world models. However, prior work has largely focused on rendering novel views of observed scenes or predicting future frames. While these models achieve impressive visual quality, they are not optimized to support downstream embodied reasoning and planning tasks. In this work, we theorize what properties generative world models should have to enhance embodied agents. As a first step, we focus on modeling an agent's beliefs over the 3D world—a crucial foundation for a broad range of embodied tasks—and identify several key capabilities. These include spatially consistent scene memory, multi-hypothesis belief sampling, sequential belief updating, and semantically informed future prediction. We then propose 3D-Belief, a generative 3D world model that instantiates these capabilities. Unlike prior work, 3D-Belief predicts unseen regions in an explicit, actionable 3D representation from partial observations and updates this belief online as new observations arrive. It enables embodied agents to reason about the 3D world under partial observability and make sequential decisions based on up-to-date beliefs. We evaluate 3D-Belief on a novel 3D imagination benchmark, 3D-CORE, and challenging object navigation tasks. Experimental results show that robots driven by 3D-Belief outperform those using state-of-the-art models in both simulations and the real world.[1]

1. Introduction

Recent advances in video generation models have shown promising results on learning a generative world model that can predict future frames based on a single, multi-view, or streaming image inputs (Song et al., 2025; Bar et al., 2025; Ren et al., 2025; Yu et al., 2024; 2025; Gu et al., 2025; Ma et al., 2025; Wu et al., 2025b; Huang et al., 2025; Cao et al., 2025). These models have demonstrated the remarkable ability to render novel views of unobserved parts of the environments or the change of the seen parts of the environments caused by agents' actions.

However, there remains a substantial gap between what these models are trained to do and what a world model must provide for embodied decision making under partial observability. In unfamiliar environments, an agent must infer what lies in unobserved regions (i.e., its beliefs of the world) based on partial observation, decide where to explore, and update its beliefs based on the new observation acquired, closing the loop of perception and planning.

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.

As summarized in Table 1, existing methods typically cover only a subset of these capabilities. Pose-controlled or action-conditioned video generation models (Song et al., 2025; Wu et al., 2025a; Bar et al., 2025) provide strong 2D imagination by predicting future frames, and can be conditioned on streaming observations. However, their outputs remain in pixel space and do not yield an explicit, actionable 3D belief. There have been recent works that augment visual generative models with 3D caches (Ren et al., 2025; Yu et al., 2024; 2025), introducing a form of scene memory that can be updated sequentially. However, the 3D component largely serves as a reconstruction cache rather than supporting diverse 3D imagination. Conversely, feedforward novel-view

[1]Johns Hopkins University [2]Lambda [3]University of Cambridge. Correspondence to: Yifan Yin <yyin34@jhu.edu>.

Preprint. February 26, 2026.

[1]Code and videos are available at https://3d-belief.github.io/.

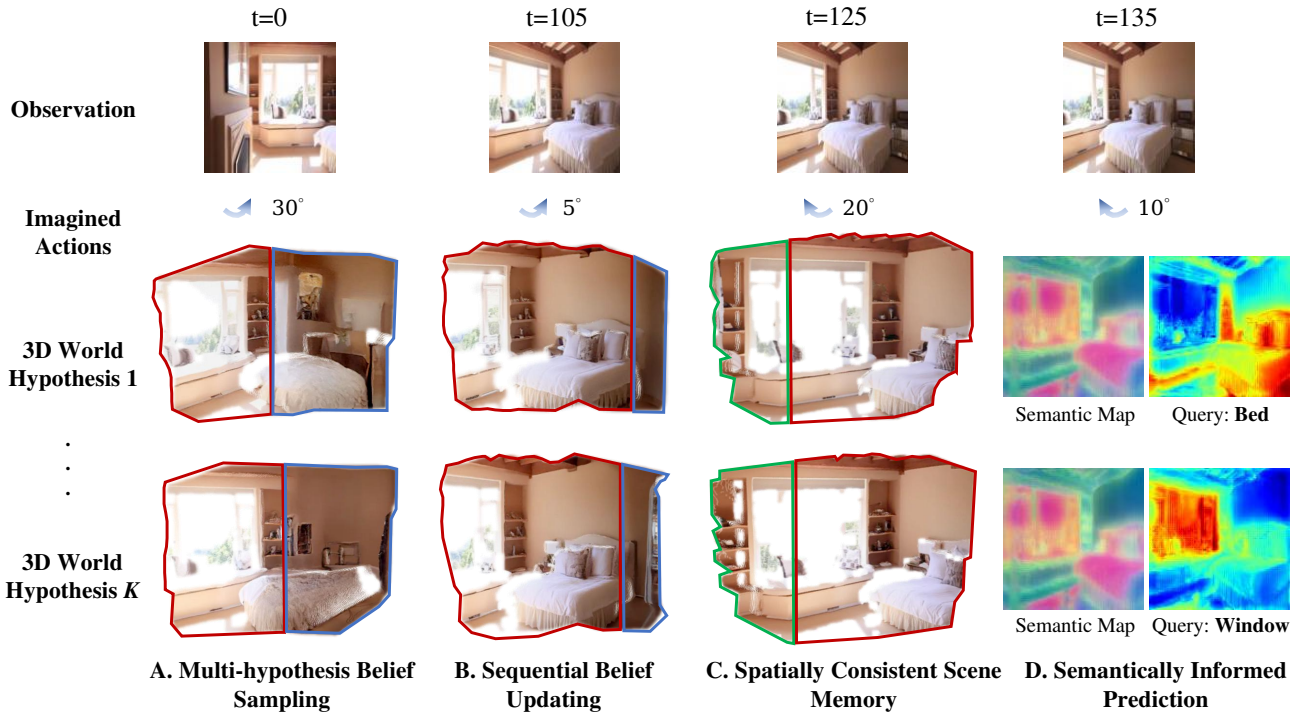


Figure 1. **Key capabilities of a generative 3D belief model.** An agent performs sequential mental simulation in an unknown environment. Top: selected RGB observations; second row: imagined actions; bottom: two sampled 3D world hypotheses. Regions supported by current observations are marked in red, 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*).

synthesis methods (Ye et al., 2024; Charatan et al., 2024; Chen et al., 2024; Jiang et al., 2025) maintain explicit 3D scene representations and can synthesize additional views. Yet, they lack 3D imagination beyond the observed scene. DFM (Tewari et al., 2023) supports diverse, view-consistent imagination by sampling an underlying 3D scene consistent with the observed view. However, its 3D content is represented implicitly in a NeRF-style field, making it difficult to convert into an explicit, actionable 3D structure suitable for embodied planning and reasoning. Recent 3D reconstruction foundation models and their streaming invariants (Wang et al., 2024; 2025a; Zhuo et al., 2025) further strengthen scene memory and online updates, and CUT3R (Wang et al., 2025b) can additionally infer some unobserved structure via state readouts. Nevertheless, these approaches do not provide uncertainty-aware multi-hypothesis imagination in 3D. Finally, across these families, semantic grounding is largely missing: existing methods do not generate semantic predictions of the 3D scenes. Because of this limitation, it is impossible to directly apply existing world models to embodied tasks that require semantic understanding of the scenes without additional models such as VLMs.

To bridge the gap, we envision a generative belief model, ***3D-Belief***, as an instantiation of a generative world model to capture 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 necessary geometric, spatial, and semantic information about the 3D world, which enables an online closed-loop path planning in a coherent 3D space that reflects both recent and past observations.

To evaluate 3D-Belief, we test it on challenging embodied tasks. We first construct a new embodied reasoning benchmark, 3D-CORE (3D COntextual REasoning). Unlike prior reasoning benchmarks, 3D-CORE is designed to probe key abilities to explicitly reason about unseen or partially observed 3D objects and space based on egocentric 2D observations. The proposed reasoning tasks provide a diagnostic evaluation of the four key capabilities of embodied agents' beliefs. We then apply 3D-Belief to model-based planning object navigation tasks, where a robot is instructed to search for a target object in a previously unseen household environ-

Table 1. Comparison of generative world model capabilities. We use ✓ to indicate the capability is supported and ✗ otherwise. "Scene Memory" indicates the method maintains a representation for observed portions of the scene. "2D Diverse Imag." indicates pixel-space multi-hypothesis imagination beyond observed frames. "Explicit 3D Imag." indicates multi-hypothesis imagination of *explicit* 3D representations beyond observed geometry (e.g., meshes, voxels, Gaussian splatting, or point clouds). "Sequential" indicates the representation can be updated online with streaming observations without rebuilding from scratch. "Semantic" indicates the representation carries language-aligned semantic features/labels (e.g., categories, text features).

Models	Scene Memory	2D Diverse Imag.	Explicit 3D Imag.	Sequential	Semantic
DFoT (Song et al., 2025)	✗	✓	✗	✓	✗
NWM (Bar et al., 2025)	✗	✓	✗	✓	✗
GEN3C (Ren et al., 2025)	✓	✓	✗	✓	✗
ViewCrafter (Yu et al., 2024)	✓	✓	✗	✗	✗
DFM (Tewari et al., 2023)	✓	✓	✗	✓	✗
MVSplat (Chen et al., 2024)	✓	✗	✗	✗	✗
VGGT (Wang et al., 2025a)	✓	✗	✗	✗	✗
CUT3R (Wang et al., 2025b)	✓	✗	✗	✓	✗
3D-Belief	✓	✓	✓	✓	✓

ment. We evaluated a mobile manipulator (Stretch robot) in both simulated environments and in the real world. In all experiments, 3D-Belief outperforms all baselines by a large margin while maintaining a low computational cost.

In sum, our main contribution includes (1) a novel conceptual framework of generative 3D belief models; (2) 3D-Belief, a new 3D generative world model that instantiates the key capacities identified in this framework; (3) 3D-CORE, a benchmark for evaluating belief reasoning in 3D world models under partial observability; and (4) an algorithm for model-based planning with 3D-Belief validated on both simulated and real-world robot navigation tasks.

2. Related Work

Visual Generative Models. Recent video diffusion and view-synthesis models can generate realistic, high-resolution videos and support controllable camera motion (Song et al., 2025; Huang et al., 2025; Yu et al., 2024; 2025). Extensions toward 3D/4D generation have also emerged (Zhen et al., 2025). These models, however, primarily generate frames in pixel space and typically require separate reconstruction modules to obtain an explicit 3D scene representation. In contrast, we learn a generative model that directly predicts an explicit, actionable 3D belief.

World Model-Based Planning. Video world models have been used for action-conditioned rollouts and planning via simulation and ranking (Du et al., 2023; 2024; Zhou et al., 2024; Zhang et al., 2024; Bar et al., 2025; Hafner et al., 2025). Prior work demonstrates long-horizon decision making by coupling video generation with search or policy learning (e.g., VLP, NWM, Dreamer) (Du et al., 2023; Bar et al., 2025; Hafner et al., 2025). Unlike these approaches that plan in 2D rollout, our planner leverages a *3D* generative world model to support open-vocabulary object navigation.

Semantically-Embedded 3D Scene Representations. Another line of work augments 3D representations with language-aligned semantics for open-vocabulary querying and task-directed reasoning (Kerr et al., 2023; Gu et al., 2024; Liu et al., 2025). LERF associates text-aligned features with radiance fields for free-form language grounding (Kerr et al., 2023), while ConceptGraph and DynaMem build persistent, online-updated spatial-semantic memories (Gu et al., 2024; Liu et al., 2025). These methods largely focus on representing what has been observed. In contrast, we introduce a semantically informed 3D belief that *predicts* unobserved regions and supports sequential updates.

3. 3D-Belief

In this work, we propose a generative 3D world model, 3D-Belief, that learns to predict and sequentially update explicit 3D representations of a scene online. We first formalize the 3D-Belief model in Section 3.1 and then introduce a model architecture to implement 3D-Belief in Section 3.2 - 3.4.

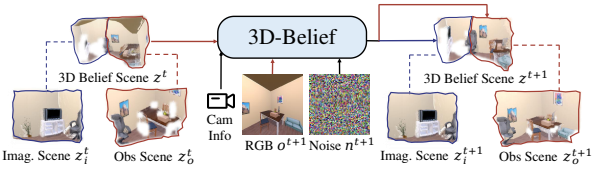
3.1. Formulation

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., 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

(a) Information flow of 3D-Belief for one belief transition.



(b) 3D-Belief model architecture: 3D Scene Diffusion.

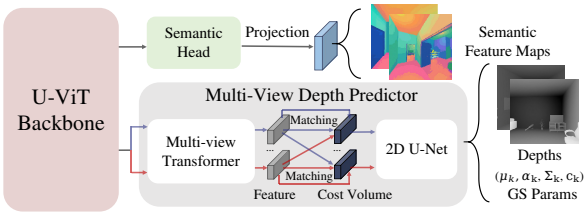


Figure 2. Overview of 3D-Belief. We represent the 3D representation of a sampled scene at step t as z^t, which includes the observed and imagined Gaussians z_o^t and z_i^t. Given z^t, a new observations o^{t+1}, and a sampled noise image n^{t+1}, 3D-Belief samples a new 3D scene representation at step $t+1$, i.e., z^{t+1}, to update the belief of the 3D world.

and semantic information of the scene necessary for downstream embodied tasks, and (2) can be parametrized so 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, SH appearance, semantic embedding). We split $z^t = z_o^t \cup z_i^t$ into observed and imagined Gaussians. For sequential, 3D-consistent belief updates, we rewrite Eq. (1) as an autoregressive update:

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

discarding z_i^t since imagined content may conflict with new evidence. This formulation (i) enables a simple training procedure that requires supervision only over short horizons (See 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) maintains constant per-step computational cost independent of the planning horizon.

Given z^t (a set of 3D Gaussian primitives), we render an imagined view from any pose θ as $\hat{o} = \mathcal{G}(z^t, \theta)$ (Kerbl et al., 2023), producing RGB, depth, and a semantic feature map from per-primitive embeddings. These renderings enable model-based mental simulation, and language queries over the semantic map let a planner score and select plans.

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 semantic head linearly projects backbone features into per-pixel semantic maps, trained by distillation from CLIP-style embeddings (Kerr et al., 2023), enabling text-based querying at test time.

3.3. Diffusion Training

Unlike typical video generation models that perform *frame-by-frame* diffusion in pixel space or latent space, we adopt *scene-level 3D diffusion* as introduced in (Tewari et al., 2023), where diffusion is applied to the entire 3D scene, in this case, the Gaussian primitives $z = \{g_k\}_k^K$. This design encourages the model to capture global geometry and maintain multi-view consistency, yielding more effective 3D scene predictions for embodied tasks.

z is estimated by adding supervision with paired context images o^{ctxt} and target images o^{trgt}. 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 the target camera parameters ϕ^{trgt} by predicting a denoised scene state z_τ and then rendering it into the 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 A for the complete formulation.

3.4. Training Objective

We train Φ_θ with paired context and target observations. Let $\hat{o}^{\text{trgt}} = \mathcal{G}(z_0, \phi^{\text{trgt}})$ and $\hat{o}^{\text{ctxt}} = \mathcal{G}(z_0, \phi^{\text{ctxt}})$ be renderings of the predicted scene in the target and context cameras. For brevity, we use a view index $v \in \{\text{trgt}, \text{ctxt}\}$ and let $o^v \in \{o^{\text{trgt}}, o^{\text{ctxt}}\}$ and $\hat{o}^v \in \{\hat{o}^{\text{trgt}}, \hat{o}^{\text{ctxt}}\}$ denote the corresponding observation and rendering in view v.

RGB loss. We use $I(\cdot)$ to denote RGB channels.

$$\mathcal{L}_{\text{rgb}} = \sum_{v \in \{\text{trgt}, \text{ctxt}\}} \| I(\hat{o}^v) - I(o^v) \|_2^2 . \tag{6}$$

Semantic loss. We align the rendered semantic feature map with features extracted from image patches. Let $S(\hat{o}^v) \in \mathbb{R}^{H\times W\times d}$ be a per-pixel semantic feature map rendered from the predicted scene (in view v). For each v, we sample patch centers $\mathcal{P}^v = \{\mathbf{u}_j\}_{j=1}^M$ on $I(o^v)$, crop patches $\pi(o^v, \mathbf{u}_j)$, and compute features using a frozen CLIP image encoder $f_{\text{clip}}(\cdot)$. We supervise the semantic feature at the corresponding pixel:

$$\mathcal{L}_{\text{sem}} = \sum_{v\in\{\text{trgt,ctxt}\}} \frac{1}{M}\sum_{j=1}^{M} \|S(\hat{o}^v)(\mathbf{u}_j) - f_{\text{clip}}(\pi(o^v, \mathbf{u}_j))\|_2^2 . \tag{7}$$

Depth loss (optional). When ground-truth depth is available, we additionally supervise the rendered depth. Let $D(\cdot) \in \mathbb{R}^{H\times W}$ denote the depth channel of an observation. For each view v, define $\hat{d}^v = D(\hat{o}^v)$ and $d^v = D(o^v)$. To handle missing/invalid depth values, let $m^v \in \{0,1\}^{H\times W}$ be a validity mask (1 for valid depth). We use a masked ℓ_1 loss:

$$\mathcal{L}_{\text{depth}} = \sum_{v\in\{\text{trgt,ctxt}\}} \frac{1}{\sum_{\mathbf{u}} m^v(\mathbf{u})} \sum_{\mathbf{u}} m^v(\mathbf{u}) \left|\hat{d}^v(\mathbf{u}) - d^v(\mathbf{u})\right| . \tag{8}$$

4. Model-based Planning with 3D-Belief

To illustrate how 3D-Belief supports embodied navigation, we use it as the agent's mental 3D world model in a closed-loop planner (Figure 3). At each step, the agent updates its 3D belief z^t from the latest egocentric RGB stream and poses. The planner then proposes waypoint goals, computes candidate paths with a classical 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.

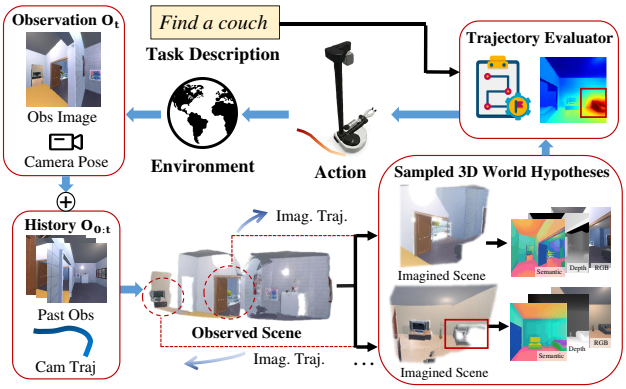


Figure 3. Model-Based Planning with 3D-Belief for Navigation. Dotted circles are imagination regions for visualization, and the red boxes are the potential target objects in imagined renders.

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.

5. Experiments

5.1. Implementation Details

Architecture. The U-ViT (Hoogeboom et al., 2023; 2024) backbone for encoding input images is initialized with DFoT (Song et al., 2025) encoder pretrained weights on the RealEstate10K dataset (Zhou et al., 2018). We use two semantic heads. One for per-primitive semantic feature embeddings, which is distilled using the OpenCLIP (Ilharco et al., 2021) ViT-B/16 model trained on the LAION-2B dataset. Another is for regularization using DINOv2 model (Oquab et al., 2023), following practices in LERF (Kerr et al., 2023). Each of the semantic heads is an MLP with 4 hidden layers. The multi-view depth predictor is initialized with MVSplat (Chen et al., 2024) pretrained weights on the RealEstate10K dataset. Our model is trained on a single L40S GPU with 250k steps for 6 days. We use the AdamW (Loshchilov & Hutter, 2017) optimizer with a learning rate of 2e-5. Appendix C provides more details.

Training Datasets. We use both synthetic videos generated using the AI2-THOR simulator (Kolve et al., 2017) with household environments from ProcTHOR (Deitke et al., 2022) and real-world video datasets, including RealEstate10K, DL3DV (Ling et al., 2024), and Habitat-

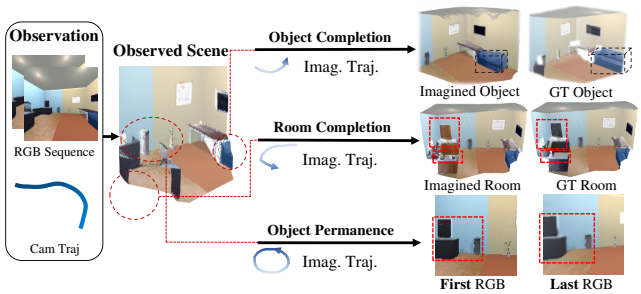


Figure 4. Evaluation of 3D-Belief model on 3D-CORE.

Matterport 3D (Ramakrishnan et al., 2021).

5.2. Experiment 1: Evaluation of 3D Imagination

5.2.1. 3D-CORE Benchmark

In embodied tasks, an agent rarely acts on a single frame. Instead, it must integrate partial observations over time, infer what is currently unobserved, and keep those inferences consistent as it moves. For example, during , an agent must decide whether a partially observed countertop likely contains the target behind occluders, or whether it should exit the room and explore elsewhere. During navigation, a single glance into a room should allow the agent to infer plausible free space and layout to plan safe exploration, rather than repeatedly re-scanning. During manipulation and rearrangement, object permanence is critical. If a mug is briefly visible and then occluded by a cabinet door or a large camera rotation, the agent's belief should remain stable so it can return and act on it.

Despite these needs, many common evaluations of generative world models emphasize visual fidelity or short-horizon prediction (e.g., frame-level quality, distributional realism, or motion consistency (Song et al., 2025; Bar et al., 2025)), which can correlate weakly with whether a model forms a consistent 3D belief under partial observability. To this end, we introduce 3D-CORE (3D COntextual REasoning), a benchmark for evaluating whether 3D world models learn the kinds of belief reasoning that are directly required for embodied decision-making. 3D-CORE is designed to probe the capabilities that matter for downstream embodied planning directly: (1) **spatial expansion** beyond what is currently observed, (2) **semantic reasoning** grounded in 3D structure, and (3) **long-horizon consistency** under large viewpoint changes.

Benchmark Construction. All scenarios are created in AI2-THOR (Kolve et al., 2017) using ProcTHOR (Deitke et al., 2022) houses, providing diverse indoor layouts, and Objaverse (Deitke et al., 2023) assets to increase geometric and appearance diversity of objects beyond the default simulator set. We render egocentric RGB streams with known camera poses and provide ground-truth 3D scene representations (geometry and semantics) for evaluation. The benchmark is model-agnostic: any method that predicts a 3D scene belief from partial observations can be evaluated.

Tasks and Metrics. 3D-CORE contains three complementary tasks that isolate distinct but essential aspects of contextual 3D reasoning. (See Appendix C.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.

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—objects should not drift, or change identity upon revisiting. Metrics: LPIPS and SigLIP similarity.

5.2.2. Evaluation on 3D-CORE

As illustrated in Figure 4, we evaluate whether 3D-Belief learns the belief reasoning capabilities required by embodied decision making on the 3D-CORE benchmark.

Baselines. To the best of our knowledge, no existing method can directly imagine an explicit 3D scene. Thus, we construct a strong baseline, DFoT-VGGT, by combining two SOTA models. DFoT-VGGT is a "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.

Results. Table 2 summarizes the main results. Overall, 3D-Belief consistently outperforms DFoT-VGGT on all three 3D-CORE tasks, indicating stronger belief reasoning for geometry-and-semantic completion, spatial expansion, and persistence under large viewpoint changes.

On **Object Completion** (full results in Appendix B.3.1), 3D-Belief improves both geometric and semantic quality across visibility levels. It achieves higher BEV IoU and 3D IoU and lower Chamfer distance than DFoT-VGGT, with especially notable gains at medium visibility (e.g., 0.55), suggesting more faithful 3D completion rather than simply extrapolating observed surfaces. It also yields higher SigLIP similarity and VLM recognition, indicating better preservation of appearance and category semantics, and thus stronger object-level belief.

Table 2. Results on 3D-CORE: Object Completion, Room Completion, and Object Permanence.

	Object Completion (55% Visibility)					Room Completion			Object Permanence	
Models	**BEV IoU ↑**	**3D IoU↑**	**Chamfer↓**	**SigLIP↑**	**Recognition↑**	**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**

On **Room Completion**, the two models are comparable in scene-level semantic recovery, as reflected by similar object prediction F1. However, 3D-Belief produces a much more accurate occupancy belief: it substantially improves Occ. Acc. on known cells and achieves higher IoU for both free and occupied regions, leading to a markedly higher overall Occ. IoU. These gains translate into more actionable 3D belief maps for embodied planning, where reliable free-space reasoning is critical.

On **Object Permanence**, 3D-Belief shows markedly stronger long-horizon consistency under large camera motions. When returning to the initial viewpoint, it achieves higher perceptual consistency (LPIPS) and higher semantic consistency (SigLIP) than DFoT-VGGT, indicating reduced drift and more stable geometry and object identities over extended rollouts.

Additional results on spatial reasoning tasks are provided in Appendix B.4.

5.3. Experiment 2: Object Navigation in Simulations

Simulation Environment. We conduct all simulated planning experiments following Ehsani et al. (2024) in the AI2-THOR simulator (Kolve et al., 2017), using ProcTHOR house assets (Deitke et al., 2022). Unless otherwise noted, we evaluate on 135 unseen houses for test, ensuring generalization to novel layouts and object configurations.

Task Definition. Following Ehsani et al. (2024), we study open-vocabulary object navigation. Each episode starts the agent at a random house location with a target object category, which may require multi-room exploration. The agent receives egocentric RGB observations and poses sequentially and must plan actions online. Success requires the agent to be within a proximity threshold of the target and to see it in the central region of view.

Metrics. We evaluate task performance using Success Rate (SR) for overall success, Success-weighted by Path Length (SPL) for path efficiency, Success-weighted by Episode Length (SEL) (Eftekhar et al., 2023; Yokoyama et al., 2021) for time efficiency. We also report metrics related to deployment cost, using Inference Time (Inf.) for real-time feasibility, and Token cost for VLM runtime cost and scalability.

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

3D Reconstruction. This group tests whether high-quality **reconstruction of only the observed geometry** is sufficient for efficient search. We use VGGT (Wang et al., 2025a) to build a 3D cache from the agent's RGB history. Variants differ only in the goal-selection module: a classical frontier-based explorer (VGGT w/ frontier) and VLM-driven waypoint selection (VGGT w/ GPT-5 mini or Gemini 3.0).

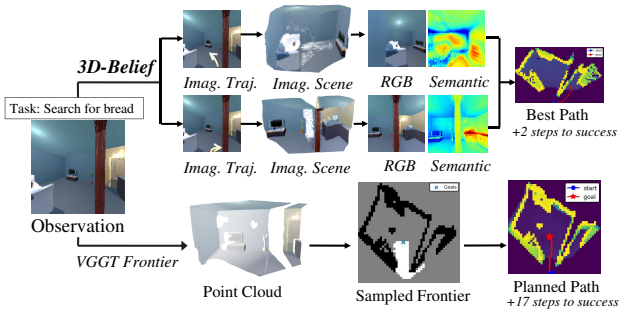


Figure 5. Comparison of planning with 3D-Belief and VGGT (w/ frontier) over the same observation of the same task.

Lifted 2D Imagination. This group evaluates methods that use VGGT (Wang et al., 2025a) to lift 2D imagination to 3D. We include two baselines, Diffusion Forcing Transformer (Song et al., 2025) with VGGT (Wang et al., 2025a) (DFoT-VGGT) and Navigation World Model (Bar et al., 2025) with VGGT (NWM-VGGT).

VLM Agents. This group uses a VLM as the end-to-end policy that outputs navigation decisions directly from the current observation (and interaction history), without any explicit 3D scene representation or belief state. We report GPT-5 mini, Gemini 3.0, and Qwen3-VL-8B-Instruct as representative VLM agents.

Results. Results in Table 3 show that 3D-Belief consistently outperforms all baseline categories, highlighting the benefit of an explicit, updatable 3D belief for planning. It achieves the best SR and strongest SPL/SEL, indicating higher success with more efficient paths and shorter completion times. 3D reconstruction baselines (VGGT variants) are limited by incomplete beliefs about unseen regions. Imagination-then-lift pipelines (DFoT-VGGT, NWM-VGGT) improve some efficiency metrics but suffer from decoupled stages and long-horizon inconsistency. VLM agents can succeed via semantic reasoning but are less stable and far more expensive, with worse SPL and high token costs.

Table 3. Simulated object navigation results. Higher is better for SR, SPL, SEL; lower is better for inference time, and VLM tokens.

Models	SR% ↑	SPL% ↑	SEL% ↑	Inf. (s/step) ↓	Token (/step) ↓
3D Reconstruction					
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
Lifted 2D Imagination					
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
VLM Agents					
GPT-5m	23.33	17.66	20.19	11.10	6644.90
Gemini 3.0	39.79	30.44	**33.57**	7.58	7286.49
Qwen3-VL-8B-Instruct	18.33	14.08	16.94	**2.30**	220.29
3D-Belief	**45.83**	**36.43**	32.99	13.34	**0**

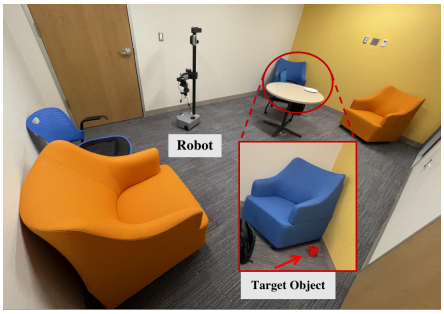


Figure 6. Example setup for the real-world object navigation tasks with 3D-Belief. The small image shows a zoom-in view near the blue couch with a target object hidden.

5.4. Experiment 3: Object Navigation in the Real World

We evaluate 3D-Belief on a real mobile manipulation platform (Hello Robot Stretch) in a mock apartment environment. The environment contains typical household furniture and objects, forming realistic object navigation scenarios (as shown in Figure 6). We select 5 common target objects and randomize each episode by (1) sampling a target object and (2) initializing the robot from one of 3 predefined starting locations to induce different exploration requirements. Note that the environments, objects, and target descriptions are all unseen during training, making this a real-world open-vocabulary object navigation setting.

Task and Success Criteria. At the beginning of each episode, the robot is given the name of the target object to be searched for (e.g., red mug), and must explore the environment and stop when it is found. An episode is marked as successful if (1) the robot reaches within a distance threshold of the target and (2) the target is within the central region of the egocentric view with a facing angle within a tolerance.

Baseline. We compare against a strong VLM agent based on Gemini 3.0, which selects actions directly from the current egocentric observation and interaction history without maintaining an explicit 3D belief.

Table 4. Real-world object navigation results.

Models	SR% ↑	SEL% ↑	Token ↓
3D-Belief	55.56	35.91	0
Gemini 3.0	23.08	13.55	2317.09

Result. Results are shown in Table 4. Across real-robot episodes, 3D-Belief enables more reliable and efficient performance than the Gemini 3.0 agent. It achieves a higher success rate and improved efficiency, demonstrating that explicit, online-updatable 3D beliefs transfer to real-world settings and can robustly support embodied decision making under sensor noise.

6. Discussion

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 7, VGGT-based occupancy quickly accumulates duplicated structure—spurious obstacles appear by t=20, and maps often break by t=60, trapping the agent. In contrast, 3D-Belief remains coherent and steadily expands over much longer horizons (up to 200 steps), 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).

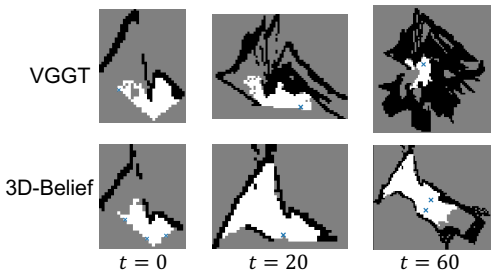


Figure 7. Comparison of inferred occupancy map generated by 3D-Belief and VGGT (w/ frontier) over time in the same task.

Table 7 in Appendix B.2.1 further shows that removing multi-hypothesis sampling causes an $\sim$10% drop, indicating that maintaining multiple plausible beliefs is important for robust decision making (see Figure 5). Together with internal semantic prediction, this lets the agent compare candidate actions conditioned on task or language specifications, rather than relying mainly on waypoint sampling. It also reduces deployment cost: compared to attaching an external VLM for semantics, which adds token and compute overhead (Table 3), 3D-Belief provides semantics within the model.

How does 3D imagination affect 2D rendering quality?

While 3D-Belief is designed for explicit 3D reasoning, it also improves *2D* rendering quality (Appendix Table 5). 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 B.1.2), 2D-only baselines often blur and drift under large camera motion, whereas 3D-Belief better preserves rigid structure and object permanence across revisits.

7. Conclusion

In this work, we studied how generative world models can better support embodied reasoning and planning, 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 contextual reasoning and object navigation have shown that 3D-Belief improved both success rate and efficiency over existing generative world models.

Limitations and Future Work. Our 3D-Belief model assumes a static world. In the future, we intend to incorporate dynamic world modeling into 3D-Belief, which can support broader embodied tasks. It is also possible to further extend the controllability of the 3D imagination by conditioning the hypothesis sampling on high-level guidance, such as scene-level imagination (e.g., language descriptions of scene-graphs of the imagined 3D world).

Acknowledgment

TS acknowledges the computing resources provided by NAIRR Pilot.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bar, A., Zhou, G., Tran, D., Darrell, T., and LeCun, Y. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15791–15801, 2025.

Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., and Fu, Y. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Carreira, J. and Zisserman, A. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Charatan, D., Li, S. L., Tagliasacchi, A., and Sitzmann, V. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19457–19467, 2024.

Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.-J., and Cai, J. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pp. 370–386. Springer, 2024.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., 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.

Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Ehsani, K., Salvador, J., Han, W., Kolve, E., Kembhavi, A., and

Mottaghi, R. Procthor: Large-scale embodied ai using procedural generation. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5982–5994. Curran Associates, Inc., 2022.

Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., and Farhadi, A. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13142–13153, 2023.

Du, Y., Yang, M., Florence, P., Xia, F., Wahid, A., Ichter, B., Sermanet, P., Yu, T., Abbeel, P., Tenenbaum, J. B., et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023.

Du, Y., Yang, S., Florence, P., Xia, F., Wahid, A., brian ichter, Sermanet, P., Yu, T., Abbeel, P., Tenenbaum, J. B., Kaelbling, L. P., Zeng, A., and Tompson, J. Video language planning. In *The Twelfth International Conference on Learning Representations*, 2024.

Eftekhar, A., Zeng, K.-H., Duan, J., Farhadi, A., Kembhavi, A., and Krishna, R. Selective visual representations improve convergence and generalization for embodied ai. *arXiv preprint arXiv:2311.04193*, 2023.

Ehsani, K., Gupta, T., Hendrix, R., Salvador, J., Weihs, L., Zeng, K.-H., Singh, K. P., Kim, Y., Han, W., Herrasti, A., 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.

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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.

Gu, Q., Kuwajerwala, A., Morin, S., Jatavallabhula, K. M., Sen, B., Agarwal, A., Rivera, C., Paul, W., Ellis, K., Chellappa, R., 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.

Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

Hafner, D., Yan, W., and Lillicrap, T. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Hirose, N., Shah, D., Sridhar, A., and Levine, S. Sacson: Scalable autonomous control for social navigation. *IEEE Robotics and Automation Letters*, 9(1):49–56, 2023.

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

Hoogeboom, E., Mensink, T., Heek, J., Lamerigts, K., Gao, R., and Salimans, T. Simpler diffusion (sid2): 1.5 fid on imagenet512 with pixel-space diffusion. *arXiv preprint arXiv:2410.19324*, 2024.

Huang, T., Zheng, W., Wang, T., Liu, Y., Wang, Z., Wu, J., Jiang, J., Li, H., Lau, R., Zuo, W., 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.

Ilharco, G., Wortsman, M., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., et al. Openclip. *Zenodo*, 2021.

Jiang, L., Mao, Y., Xu, L., Lu, T., Ren, K., Jin, Y., Xu, X., Yu, M., Pang, J., Zhao, F., et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *ACM Transactions on Graphics (TOG)*, 44(6):1–16, 2025.

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

Karnan, H., Nair, A., Xiao, X., Warnell, G., Pirk, S., Toshev, A., Hart, J., Biswas, J., and Stone, P. Socially compliant navigation dataset (scand): A large-scale dataset of demonstrations for social navigation. *IEEE Robotics and Automation Letters*, 7(4):11807–11814, 2022.

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

Kerr, J., Kim, C. M., Goldberg, K., Kanazawa, A., and Tancik, M. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19729–19739, 2023.

Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., 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.

Liu, P., Guo, Z., Warke, M., Chintala, S., Paxton, C., Shafiullah, N. M. M., and Pinto, L. 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.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Ma, B., Gao, H., Deng, H., Luo, Z., Huang, T., Tang, L., and Wang, X. You see it, you got it: Learning 3d creation on pose-free videos at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2016–2029, 2025.

Melnik, A., Ljubljanac, M., Lu, C., Yan, Q., Ren, W., and Ritter, H. Video diffusion models: A survey. *arXiv preprint arXiv:2405.03150*, 2024.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Oshima, Y., Iwasawa, Y., Suzuki, M., Matsuo, Y., and Furuta, H. Worldpack: Compressed memory improves spatial consistency in video world modeling. *arXiv preprint arXiv:2512.02473*, 2025.

Qiu, R.-Z., Yang, G., Zeng, W., and Wang, X. Feature splatting: Language-driven physics-based scene synthesis and editing. *arXiv preprint arXiv:2404.01223*, 2024.

Ramakrishnan, S. K., Gokaslan, A., Wijmans, E., Maksymets, O., Clegg, A., Turner, J. M., Undersander, E., Galuba, W., Westbury, A., Chang, A. X., Savva, M., Zhao, Y., and Batra, D. 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.

Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., and Gao, J. 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.

Shah, D., Eysenbach, B., Kahn, G., Rhinehart, N., and Levine, S. Rapid exploration for open-world navigation with latent goal models. *arXiv preprint arXiv:2104.05859*, 2021.

Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., and Sitzmann, V. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

Tewari, A., Yin, T., Cazenavette, G., Rezchikov, S., Tenenbaum, J., Durand, F., Freeman, B., and Sitzmann, V. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *Advances in Neural Information Processing Systems*, 36:12349–12362, 2023.

Triest, S., Sivaprakasam, M., Wang, S. J., Wang, W., Johnson, A. M., and Scherer, S. 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.

Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. Fvd: A new metric for video generation. 2019.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., and Novotny, D. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

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

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., and Revaud, J. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20697–20709, 2024.

Wu, H., Wu, D., He, T., Guo, J., Ye, Y., Duan, Y., and Bian, J. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025a.

Wu, J. Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M. Z., Fidler, S., Gojcic, Z., and Ling, H. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26024–26035, 2025b.

Xiao, Z., Lan, Y., Zhou, Y., Ouyang, W., Yang, S., Zeng, Y., and Pan, X. Worldmem: Long-term consistent world simulation with memory. *arXiv preprint arXiv:2504.12369*, 2025.

Yang, Y., Liu, J., Zhang, Z., Zhou, S., Tan, R., Yang, J., Du, Y., and Gan, C. Mindjourney: Test-time scaling with world models for spatial reasoning. *arXiv preprint arXiv:2507.12508*, 2025.

Ye, B., Liu, S., Xu, H., Li, X., Pollefeys, M., Yang, M.-H., and Peng, S. No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. *arXiv preprint arXiv:2410.24207*, 2024.

Yokoyama, N., Ha, S., and Batra, D. 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.

Yu, M., Hu, W., Xing, J., and Shan, Y. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025.

Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.-T., Shan, Y., and Tian, Y. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

Zhang, H., Wang, Z., Lyu, Q., Zhang, Z., Chen, S., Shu, T., Dariush, B., Lee, K., Du, Y., and Gan, C. Combo: compositional world models for embodied multi-agent cooperation. *arXiv preprint arXiv:2404.10775*, 2024.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. 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.

Zhen, H., Sun, Q., Zhang, H., Li, J., Zhou, S., Du, Y., and Gan, C. Tesseract: Learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

Zhou, J., Gao, H., Voleti, V., Vasishta, A., Yao, C.-H., Boss, M., Torr, P., Rupprecht, C., and Jampani, V. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025.

Zhou, S., Du, Y., Chen, J., Li, Y., Yeung, D.-Y., and Gan, C. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

Zhou, T., Tucker, R., Flynn, J., Fyffe, G., and Snavely, N. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Zhuo, D., Zheng, W., Guo, J., Wu, Y., Zhou, J., and Lu, J. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025.

A. Diffusion Training Formulation

The forward process is defined as:

$$q(o_\tau^{\text{trgt}} \mid o_{\tau-1}^{\text{trgt}}) = \mathcal{N}(o_\tau^{\text{trgt}}; \sqrt{1-\beta_\tau} o_{\tau-1}^{\text{trgt}}, \beta_\tau I). \tag{9}$$

In the reverse process, we reconstruct o^{trgt} conditioned on o^{ctxt} and camera parameters ϕ^{trgt}:

$$p_\theta(o_{0:T}^{\text{trgt}} \mid o^{\text{ctxt}}; \phi^{\text{ctxt}}, \phi^{\text{trgt}}) \tag{10}$$

$$= p(o_T^{\text{trgt}}) \prod_{\tau=0}^{T} p_\theta(o_{\tau-1}^{\text{trgt}} \mid o_\tau^{\text{trgt}}, o^{\text{ctxt}}; \phi^{\text{ctxt}}, \phi^{\text{trgt}}), \tag{11}$$

where $p_\theta(o_{\tau-1}^{\text{trgt}} \mid o_\tau^{\text{trgt}}, o^{\text{ctxt}}; \phi^{\text{ctxt}}, \phi^{\text{trgt}})$ is implemented by first predicting an intermediate clean 3DGS scene z_τ and then mapping it to a denoised observation using the Gaussian Splatting rendering function $\mathcal{G}(\cdot)$:

$$z_{\tau-1} = \Phi_\theta(o^{\text{ctxt}}, o_\tau^{\text{trgt}}; \tau, \phi^{\text{ctxt}}, \phi^{\text{trgt}}), \tag{12}$$

$$\hat{o}_{\tau-1}^{\text{trgt}} = \mathcal{G}(z_{\tau-1}, \phi^{\text{trgt}}), \tag{13}$$

$$o_{\tau-1}^{\text{trgt}} \sim \mathcal{N}(o_{\tau-1}^{\text{trgt}}; C_{\tau-1}\hat{o}_{\tau-1}^{\text{trgt}}, \hat{\beta}_\tau I). \tag{14}$$

Here, $\hat{o}_{\tau-1}^{\text{trgt}}$ serves as an estimate of the clean observation. The constants $C_{\tau-1}$ and $\hat{\beta}_{\tau-1}$ are specifically chosen to ensure the noise at time $\tau-1$ aligns with the total noise introduced during the forward process. Φ_θ is a neural network shown in Figure 2b. During test time, the scene is generated by iteratively applying Eqs. (12)-(14), starting from an initial distribution $p(o_{\tau=T}^{\text{trgt}}) \sim \mathcal{N}(0, I)$.

The model defines a generative process over the Gaussian primitives, formalized as:

$$p_{\theta,\phi^{\text{trgt}}}(z_{0:T} \mid o^{\text{ctxt}}; \phi^{\text{ctxt}}) = \prod_{\tau=1}^{T} p_\theta(z_{\tau-1} \mid o_\tau^{\text{trgt}}, o^{\text{ctxt}}; \phi^{\text{ctxt}}, \phi^{\text{trgt}}). \tag{15}$$

B. Extended Results

B.1. Vision Results

B.1.1. Quantitative Results

We evaluate all models on a fixed set of 200 base trajectories with aligned temporal ranges. Each trajectory specifies a start and end frame, defining a common prediction window. While different models adopt different conditioning strategies (e.g., frame subsampling or initial-frame conditioning), all predictions are generated over the same future time horizon, enabling consistent quantitative evaluation.

2D visual metrics. We employ standard image and video quality metrics to quantitatively assess visual fidelity. **Pixel-level fidelity** is measured via Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), where higher values denote better reconstruction quality. To evaluate **perceptual similarity**, we use LPIPS (AlexNet backbone) (Zhang et al., 2018), which computes the distance in deep feature space; lower values indicate higher perceptual alignment with ground truth. Furthermore, Fréchet Inception Distance (FID) (Heusel et al., 2017) and Fréchet Video Distance (FVD) (Unterthiner et al., 2019) are reported to quantify the distributional discrepancy between generated and real sequences at the image and video levels, respectively.

Computation details. For 2D metrics (PSNR, SSIM, LPIPS), predicted frames are resized to match the ground-truth resolution using bilinear interpolation. PSNR and SSIM are computed frame-wise and averaged over the prediction window. LPIPS is computed using the AlexNet backbone. FID is calculated using the TorchMetrics implementation with Inception-v3 features. For FVD, we employ an I3D backbone pretrained on Kinetics-400 (Carreira & Zisserman, 2017). Video sequences are spatially resized to 224×224 via bilinear interpolation and scaled to $[-1, 1]$ prior to feature extraction. All metrics are computed on RGB frames.

Table 5. 2D quantitative results. Comparison on standard video quality metrics. Best results on the Novel split are highlighted in **bold**. The Seen split is included as an upper-bound reference where future keyframes are visible. Note that FVD is not reported for 3D-Belief on the Seen split due to insufficient sequence length. In addition, NWM evaluation is based on 177 sequences, as the model failed to generate valid outputs for the remaining 23 cases.

Model	Split	#Vid	FVD↓	FID↓	LPIPS↓	PSNR↑	SSIM↑
Reference							
3D-Belief	Seen	200	–	37.95	0.064	29.79	0.925
Comparison on Novel Split							
NWM	Novel	177	299.39	52.24	0.456	16.05	0.628
DFOT	Novel	200	205.68	49.59	0.294	20.07	0.726
3D-Belief	Novel	200	**136.08**	**35.13**	**0.101**	**26.96**	**0.882**

Table 6. 3D evaluation results. 3D metrics are reported only when applicable.

Model	Metric	Setting	Value
VGGT	AUC@30	Rotation only	0.7848
VGGT	AUC@30	Rotation + Translation	0.4103

Conditioning protocols. While all methods are evaluated on the identical set of base trajectories, their input conditions vary reflecting their distinct inference paradigms. For 3D-Belief, conditioning frames are subsampled keyframes within the prediction range. DFOT conditions on the initial frames of the specified range in a sliding-window manner. NWM adopts a strict future-prediction setup, conditioning solely on the first frame. Consequently, quantitative comparisons should be interpreted within the context of these differing visibility protocols: we evaluate each model in its canonical configuration rather than imposing a unified but potentially suboptimal conditioning scheme.

Quantitative comparison. Table 5 presents the evaluation results on the standard Novel split, where no future information is available during inference. 3D-Belief consistently outperforms competitive baselines across all metrics. In terms of perceptual quality, 3D-Belief achieves an LPIPS score of **0.101**, significantly lower than DFOT (0.294) and NWM (0.456), indicating generated frames with sharper details and fewer artifacts. For temporal consistency, **3D-Belief** achieves the lowest FVD (**136.08**), demonstrating that our explicit geometric conditioning effectively reduces the temporal jitter and drift often observed in baseline methods (e.g., NWM with FVD 299.39). We additionally report results on the Seen split as an oracle reference; the narrow gap between our Novel and Seen results (e.g., SSIM 0.882 vs. 0.925) further validates the robustness of our model in unseen scenarios.

3D metric evaluation. To evaluate 3D geometric consistency, we report the Area Under the Curve (AUC) of the cumulative error distribution up to 30° (AUC@30). Given the scale ambiguity inherent in monocular video generation, predicted trajectories are first aligned to the ground truth via Sim(3) alignment. For Rotation, we measure the geodesic distance between rotation matrices. For Translation, we compute the angular error between the aligned translation vectors. The combined Rotation + Translation metric is defined as the element-wise maximum of these errors. As shown in Table 6, VGGT achieves an AUC@30 of 0.7848 for rotation, confirming its capability to synthesize geometrically consistent camera trajectories. Note that other baselines (e.g., DFOT, NWM) function purely as 2D video generators and do not yield explicit camera poses, rendering this metric inapplicable to them.

B.1.2. Qualitative Results

Figure 8 visualizes the generated sequences on the test dataset we generated in AI2THOR (Kolve et al., 2017). Compared to baselines, 3D-Belief demonstrates superior visual fidelity and long-term geometric consistency. Baseline methods (NWM and DFOT) often suffer from severe *texture blurring* and *geometric drift* as the prediction horizon increases.

For instance, in the large-motion scenario of **Scene 114** (Figure 8a), NWM fails to maintain object permanence, resulting in frames that rapidly degrade into incoherent noise. Similarly, DFOT produces distorted room structures where walls and furniture lose their rigidity. In contrast, our model leverages explicit 3D geometric conditioning to preserve sharp textures and structural coherence even under substantial viewpoint changes. This advantage is further evident in the turning scenario of **Scene 87** (Figure 8b), where 3D-Belief correctly synthesizes the geometric perspective of the room, whereas

baselines struggle to infer the correct 3D layout from 2D context alone. Overall, these qualitative comparisons corroborate our quantitative findings, highlighting the necessity of explicit 3D representations for robust long-term video generation.

B.2. Extended Planning Results

B.2.1. Ablation Studies

We report ablations for the 3D-Belief model in the simulated object navigation task in Table 7.

Table 7. Ablation studies for simulated object navigation tasks.

Models	SR% ↑	SPL% ↑	SEL% ↑	Token (/step) ↓
3D-Belief	45.83	36.43	32.99	0
w/o geometry	17.50	13.25	14.75	0
single hypothesis	35.14	28.85	26.81	0

B.2.2. Use VLMs for Semantics

To study the impact of the built-in open-vocabulary semantics in 3D-Belief, we performed extra experiments with VLMs as the model semantic support. Results are reported in Table 8. We find that with VLMs, 3D-Belief model can perform better in navigation tasks, indicating that VLMs can help improve the semantic performance in terms of embodied decision making (e.g. selecting among different candidate paths). The performance of the built-in semantic module is largely affected by the CLIP-style vision-language alignment foundation models.

Table 8. Extended results with VLMs for semantic representations in simulated object navigation tasks.

Models	SR% ↑	SPL% ↑	SEL% ↑	Token (/step) ↓
3D-Belief	45.83	36.43	32.99	0
w/o semantics (w/ GPT-5m)	49.58	40.10	41.01	79.42
w/o semantics (w/ Gemini 3)	46.22	36.34	38.02	340.67

B.3. Full Results on 3D-CORE

B.3.1. Object Completion

We report the full object completion results with different visibilities in Table 9.

Table 9. Results on Object Completion across different visibility.

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**

B.3.2. Room Completion

We report room completion results with more metrics in Table 10

B.4. Spatial Reasoning QA

B.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

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**

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., 2025) 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 VLM as context for answering questions. We compare 3D-Belief against two video world models, **SWM** (Yang et al., 2025) and **SVC** (Zhou et al., 2025), 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 object-centric dynamics remain challenging and may require stronger object-level temporal modeling or higher-fidelity appearance tracking.

	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

Table 11. Results on SAT-Real.

C. Implementation Details

C.1. Model Training

C.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.

C.1.2. Training Settings

3D-Belief We build on pre-trained weights from DFoT (Song et al., 2025) and MVSplat (Chen et al., 2024), and conduct joint training directly on the full composite dataset. 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. We follow DFM (Tewari et al., 2023) to apply loss on the denoised target view image. We calculate L1 loss

and LPIPS loss with a weight of 1.0 between the predicted image and the ground-truth. Although deep supervision is not employed, the model incorporates a depth smoothness loss with a weight of 0.1 as a regularization term. Additionally, the semantic loss based on CLIP (Ilharco et al., 2021) and DINOv2 (Oquab et al., 2023) is calculated separately, with a weight of 0.1. Besides the target view, we also compute all the aforementioned loss for intermediate frames to enhance consistency. To further enhance the model's capabilities, we align the prediction and ground truth at the feature level with a weight of 5.0, while simultaneously aligning the prediction features using the pre-trained VGGT (Wang et al., 2025a) model with a weight of 2.0. Training is completed by performing a weighted sum of all losses.

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, batch size 4 with gradient accumulation of 2, and runs for 48 epochs. 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.

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.

C.2. Planning

C.2.1. Object Navigation Tasks

In all planning experiments, we use 3 hypotheses for 3D-Belief sampling at each decision step.

C.2.2. Real-World Experiments

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.

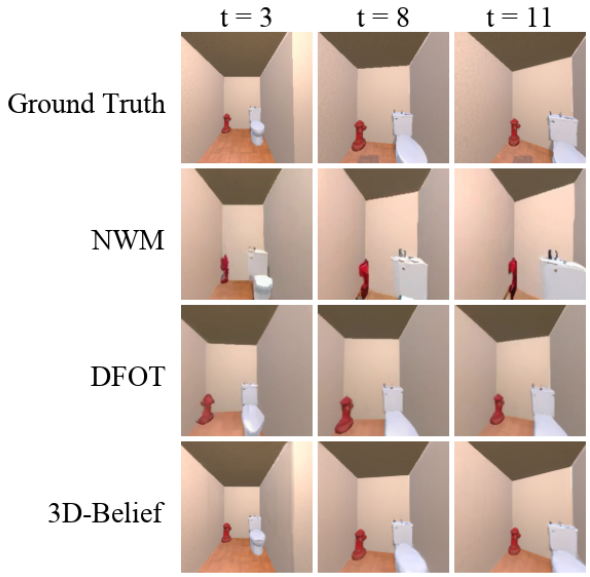
We mount an RGB-D camera (*Intel RealSense* D455) on the robot and use the RGB stream as the visual input. The robot's pose is estimated using onboard wheel-encoder odometry. Together, the RGB observations and associated poses form an egocentric stream that serves as input to our models. Note that 3D-Belief uses only RGB at inference time. 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; 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).

C.3. Reasoning

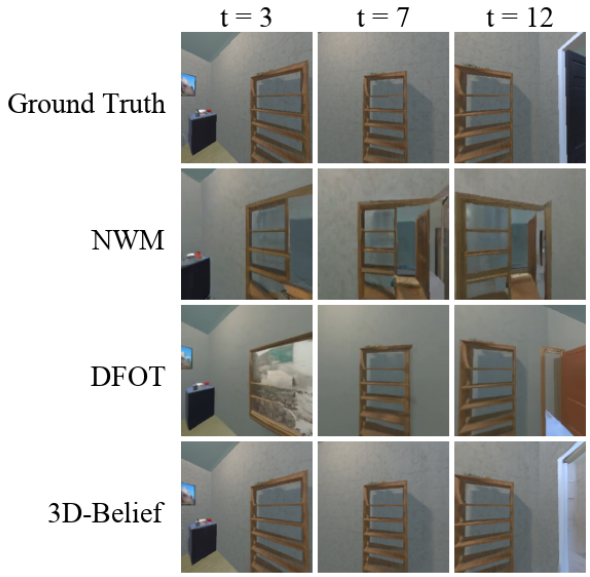
C.3.1. 3D-CORE Metrics

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.

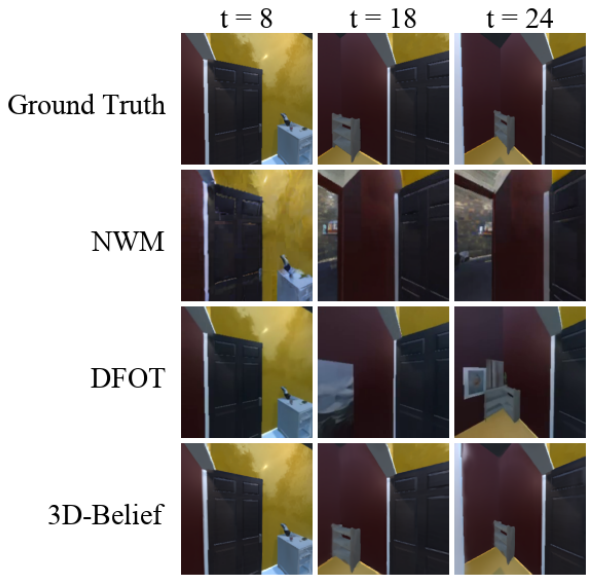
For *Room Completion*, Obj. F1 is calculated by comparing a list of VLM (Gemini 2.5 (Comanici et al., 2025)) recognized objects in imagined views, with the GT object list in that view. Occ. Acc. is defined by the accuracy on cells where both gt and pred are known (0 or 1). IoU Free is the IoU for free 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.



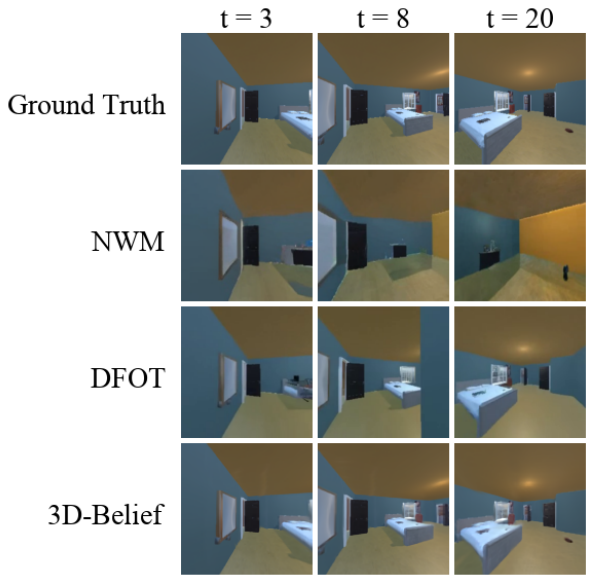
(a) Scene 114 (Large Motion)



(b) Scene 87 (Turning)



(c) Scene 96



(d) Scene 117

Figure 8. **Qualitative results on the Novel split (4 examples).** We compare 3D-Belief against NWM and DFOT across different scenes. Each block shows Ground Truth (top) vs. Predictions (rows) over time (cols). Our model (bottom row in each block) consistently preserves geometry and texture, while baselines degrade significantly.