Persistent 4D World Models: Enforcing Object Permanence via Latent Set Representations

Ruozhou Lin
linruozhou@yahoo.com

Jinyu Li
jinli66@yahoo.com

Xinghan Chen
xinghanc00@yahoo.com

Chenxi Wang
chenw1999@yahoo.com

Chengchu Xu
xuchengchu@yahoo.com

Yanqing Liu
emelialiuyq@yahoo.com

Yanze Zhang
yanzez@yahoo.com

2025-12-29

***Abstract*—Current generative video models and voxel-based 4D world models frequently suffer from a critical cognitive deficit: a lack of "object permanence." In these systems, representation is often tightly coupled with visibility. Consequently, when an object moves behind an occluder, it is erased from the internal state representation, leading to identity changes, hallucinated replacements, or complete disappearance upon reappearance. This limitation is catastrophic for downstream applications like AR/VR interaction or robotic planning, where consistent physical existence is paramount. In this paper, we propose a Latent Set-Based World Model (LS-WM) that enforces object permanence as a strong inductive bias. Unlike grid-based approaches that suffer from amnesia outside the view frustum, we represent the 4D scene as an unordered set of object-centric latent vectors. A permutation-equivariant dynamics transformer updates these vectors continuously over time—regardless of visibility—while a separate, conditional 3D Gaussian Splatting renderer handles visual occlusion only at the final imaging step. We utilize self-supervised training on complex synthetic datasets (Kubric) to explicitly penalize state amnesia. Experiments demonstrate that our method maintains object identity through heavy, long-duration occlusion significantly better than frame-based or voxel-based baselines, achieving a 26% improvement in re-identification accuracy.**

***Index Terms*—World Models, Object Permanence, Set Representation, 4D Generation, Neural Rendering, Gaussian Splatting**

I. Introduction

Humans possess an innate understanding of *object permanence*—the knowledge that objects continue to exist in time and space even when they cannot be directly perceived [5]. This cognitive prior allows us to predict that a car driving behind a building will emerge on the other side, retaining its color, shape, and velocity.

However, modern generative world models, despite impressive fidelity in image synthesis, often fail this fundamental test of physical reasoning. Video diffusion models, trained primarily on 2D pixel statistics, tend to operate under an "out of sight, out of mind" paradigm. If pixels belonging to an object are overwritten by an occluder, the information is lost. When the object should reappear, the model must hallucinate it anew, frequently violating temporal consistency.

This challenge of maintaining identity over time is well-documented. As explored by **Song et al.** in their work on *Temporal-ID* [1], preserving fine-grained features across long-form generation requires robust memory mechanisms (such as adaptive memory banks) to prevent "identity decay." We extend this philosophy by proposing that identity preservation in 4D environments requires not just memory, but a fundamental decoupling of the simulation state from the rendering process.

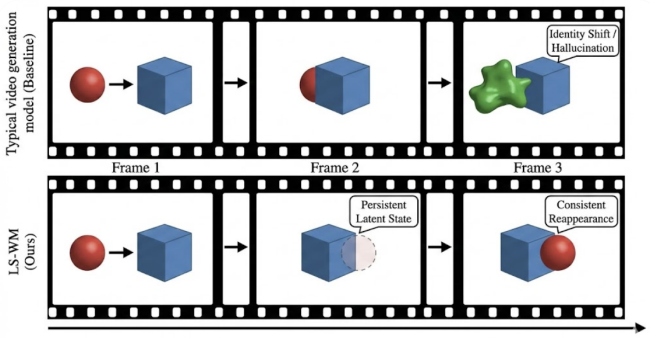

Fig. 1: **The Object Permanence Problem in Generative Models.** Top Row: Typical video generation models lose track of objects when occluded (center frame), leading to identity shifts upon reappearance (right frame). Bottom Row: Our proposed LS-WM maintains a persistent latent representation of the occluded object, ensuring consistent identity and physics when it emerges.

The core problem lies in the underlying scene representation. Grid-based (voxel) representations entangle geometry with visibility and suffer from cubic memory scaling, limiting their ability to track objects outside a narrow field of view. To address this, we propose the **Latent Set-Based World Model (LS-WM)**.

Our approach shifts the paradigm from *spatially-anchored* grids to *object-anchored* sets. By representing a scene as a dynamic set of latent descriptors $\{z_1, z_2, ..., z_n\}$, we decouple the *simulation* of an object from its *rendering*. A dynamics transformer updates the state of hidden objects, ensuring they continue to move and interact even when occluded from the camera's view (Fig. 1).

Our primary contributions are:

- **Decoupled Architecture:** A novel framework separating

physical dynamics (performed on latent sets) from visual rendering (performed via compositional Gaussian Splatting), enforcing object permanence as an architectural bias.
- **"Blind" Dynamics Modeling:** A transformer-based dynamics model that updates object states purely based on their latent interactions, agnostic to camera visibility.
- **Superior Occlusion Reasoning:** State-of-the-art performance on the complex Kubric MOVi-E benchmark, demonstrating robust identity retention over long occlusion horizons compared to Voxel and 2D baselines.

II. Related Work

A. Video Generation as World Models

Recent advances in video diffusion have led to models capable of generating realistic short clips. Some approaches frame this as "world modeling," aiming to simulate physics. However, these models operate fundamentally in 2D pixel space. They lack an explicit 3D representation, meaning occlusions are handled merely as pixel in-painting tasks rather than reasoning about depth relationships. This leads to poor long-term consistency.

B. 3D-Aware Dynamic Scene Representations

To address 3D consistency, researchers have extended Neural Radiance Fields (NeRFs) to dynamic scenes. Approaches like D-NeRF or HyperNeRF use deformation fields to map time-variant states to a canonical space. While effective for playback, these methods generally require complete observations of the scene from multiple angles during training and do not easily generalize to generating *new*, unobserved future states with complex interactions.

Voxel-based approaches, such as Block-NeRF or recent generative voxel diffusion models, offer explicit 3D structure. However, they are constrained by memory. Objects that move out of the voxel grid boundary or are heavily occluded by foreground voxels are often "forgotten" by the update mechanism to save memory, re-introducing the permanence problem.

C. Object-Centric Learning

Our work is heavily inspired by unsupervised object-centric learning. Models like SLOT Attention [6] learn to decompose images into a set of abstract "slots." However, standard slot attention is typically applied to static 2D images or simple sprite-based videos. Our work extends this concept to complex, physically rich 4D environments. This aligns with recent efforts to unify controllable generation through physical laws [2], using the set representation as the substrate for a generative dynamics model rather than just a reconstruction tool.

III. Methodology

Our LS-WM framework is composed of three main stages: Initialization, Persistent Dynamics, and Compositional Rendering. The overall architecture is illustrated in Fig. 2.

A. Scene Initialization

To begin the simulation, we must map visual observations to our latent set representation. Given k initial context frames $I_{0:k}$, we use a convolutional backbone followed by a slot-attention mechanism to extract an initial set of N latent vectors:

$$\mathcal{S}_k = \text{SetEncoder}(I_{0:k}) = \{z_{i,k} \mid z_{i,k} \in \mathbb{R}^d\}_{i=1}^{N} \quad (1)$$

We use a fixed number of slots N (e.g., 16) to handle a variable number of objects. Empty slots are handled via a "presence" probability incorporated into the latent vector.

B. Latent Set Representation

Each vector $z_{i,t}$ in the set $\mathcal{S}_t$ is a compact, disentangled representation of an object i at time t. Conceptually, it encodes:

- **Appearance code** (z^{app}): Texture, color, material properties.
- **Shape code** (z^{shape}): Geometric structure.
- **State code** (z^{state}): 3D position, rotation, and linear/angular velocity.

This unordered set representation is crucial because it does not bake in spatial position into the representation structure itself, unlike a voxel grid. We draw on the latent world modeling principles established in *DreamWM* [3], ensuring that these latent codes can effectively drive downstream generation tasks.

C. Persistent Dynamics: The "Blind" Updater

The core of our proposal is the separation of dynamics from rendering. The evolution of the scene is modeled by a Transformer architecture that operates on the set $\mathcal{S}_t$.

Crucially, this module is "blind"—it does not receive the camera pose π_t as input. It predicts the next state based solely on interactions between object latent vectors (simulating physics):

$$\mathcal{S}_{t+1} = \text{TransformerDynamics}(\mathcal{S}_t) \quad (2)$$

The self-attention mechanism in the Transformer allows each object to attend to every other object to resolve collisions and forces. Because the dynamics model does not know which objects are visible to the camera, it is forced to update $z_{i,t}$ for **all** existing objects i. This enforces the persistence of state through occlusion.

D. Renderer: Set-to-3DGS

To generate an image I_t from the latent set $\mathcal{S}_t$ given a camera pose π_t, we need a renderer that can handle occlusion explicitly. We uphold the object-centric nature by using 3D Gaussian Splatting (3DGS).

A decoding network maps each latent object vector $z_{i,t}$ to a collection of 3D Gaussians parameterized by position μ, covariance Σ, color c, and opacity α:

$$\{\mu_j, \Sigma_j, c_j, \alpha_j\}_{j=1}^{M_i} = \text{Decoder}(z_{i,t}) \quad (3)$$

where M_i is the number of Gaussians used to represent object i. This decoding step benefits from robust editing techniques

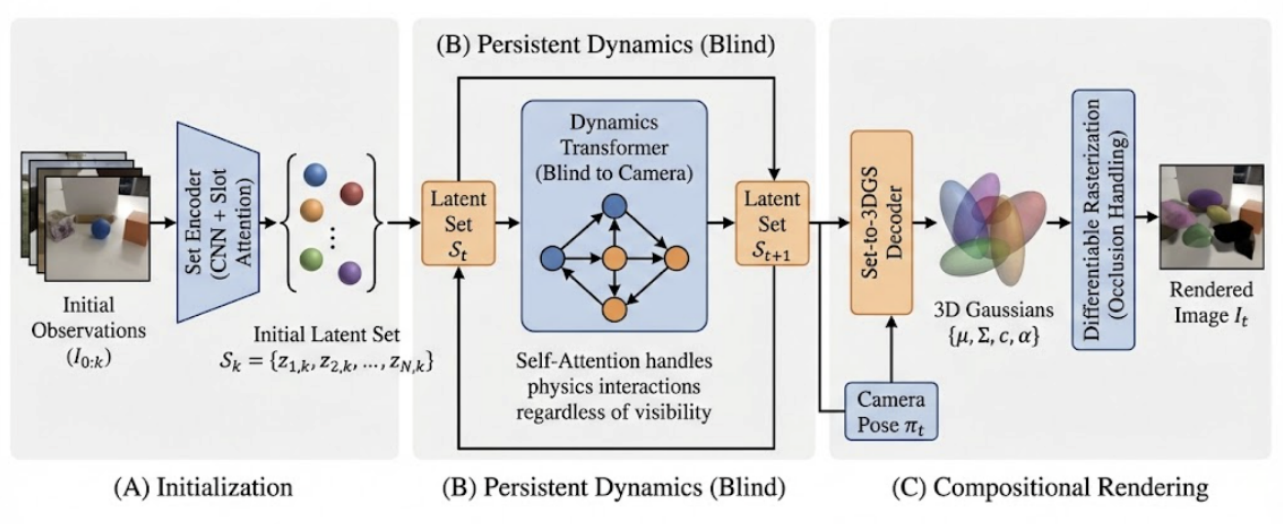

Fig. 2: **The LS-WM System Architecture.** (A) **Initialization:** An encoder initializes a set of object-centric latent vectors $\mathcal{S}_t$ from observation frames. (B) **Persistent Dynamics:** A permutation-equivariant Transformer updates the latent set from t to $t+1$ based on physics interactions. Crucially, this module is "blind" to the camera, updating all objects regardless of visibility. (C) **Compositional Rendering:** A separate decoder converts each latent vector $z_{i,t}$ into a set of 3D Gaussians. These are combined via alpha-blending based on the target camera pose π_t to handle occlusion at the rendering stage.

[4], ensuring that the geometry remains consistent even as the latent vector evolves over time.

To render the complete scene, we concatenate all Gaussians from all N objects into a global set. We then use standard differentiable Gaussian rasterization, which sorts Gaussians by depth relative to the camera π_t and performs α-blending.

$$I_t = \text{Rasterize}(\bigcup_{i=1}^{N} \text{Decoder}(z_{i,t}), \pi_t) \tag{4}$$

During rasterization, front-to-back compositing determines visibility. While the *pixels* of object i might be occluded by object k, the latent vector $z_{i,t}$ remains intact in $\mathcal{S}_t$, ready for the next dynamics step.

E. Training Objectives

We train the model end-to-end on synthetic data (Kubric) where ground truth future frames I_{gt} and object segmentation masks M_{gt} are available. The total loss is a weighted combination of three terms:

1. RGB Reconstruction Loss: Standard L_2 and perceptual loss (LPIPS) between rendered and ground truth images.

$$\mathcal{L}_{rgb} = \|I_t - I_{gt,t}\|_2^2 + \text{LPIPS}(I_t, I_{gt,t}) \tag{5}$$

2. Mask Consistency Loss: To ensure the slots decompose objects correctly, we supervise the aggregated alpha maps of each slot against GT masks.

$$\mathcal{L}_{mask} = \text{CrossEntropy}(\text{RasterAlpha}(\mathcal{S}_t), M_{gt,t}) \tag{6}$$

3. Latent Dynamics Loss: We encourage the dynamics model to make accurate future predictions in the latent space itself, vital for long-term stability.

$$\mathcal{L}_{dyn} = \|\text{StopGrad}(\text{SetEnc}(I_{t+1})) - \text{Dynamics}(\mathcal{S}_t)\|_2^2 \tag{7}$$

IV. Experiments

A. Experimental Setup

Dataset: We utilize the Kubric MOVi-E dataset. This synthetic dataset features complex rigid body dynamics, 3D objects of varying shapes and textures, and crucially, frequent, long-duration inter-object occlusions. We generate sequences of 50 frames.

Baselines: We compare LS-WM against two representative categories of world models:

- **Video-LDM (2D Baseline):** A standard latent video diffusion model conditioned on past frames, predicting future frames in pixel space.
- **Voxel-WM (3D Baseline):** A grid-based world model where the state is a 64^3 feature voxel grid updated by a 3D CNN U-Net. This represents state-of-the-art explicit 3D approaches.

B. Metrics

We focus on metrics that specifically evaluate persistence through occlusion:

1) **Re-ID Accuracy (%):** We identify occlusion events lasting > 10 frames. We use a pre-trained classifier to determine if the object reappearing after occlusion has the same semantic identity (shape/color) as the object that went behind the occluder.
2) **PSNR (Occluded):** Peak Signal-to-Noise Ratio calculated *only* in image regions where disocclusion occurs, measuring geometric and textural accuracy upon reappearance.

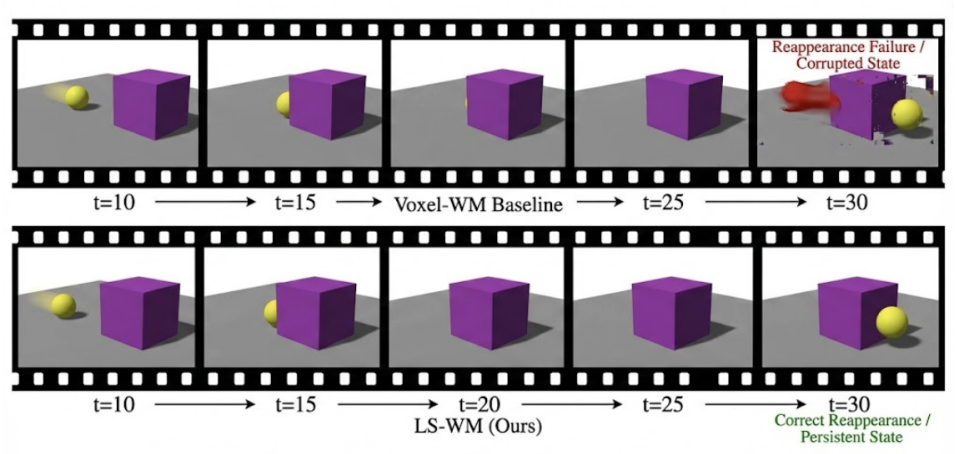

Fig. 3: **Qualitative Results on Kubric MOVi-E Long-Term Occlusion.** We visualize a sequence where a small yellow friction-less object slides behind a large static purple cube (frames $t = 15$ to $t = 25$). Top Row (Voxel-WM Baseline): The model loses track of the yellow object whilst occluded. Upon disocclusion at $t = 30$, it either fails to regenerate it or hallucinates a different object (red smear). Bottom Row (LS-WM): Our model successfully tracks the hidden state, resulting in the correct object reappearing at the correct time and location.

C. Quantitative Results

Table I presents the comparative results. The Video-LDM performs poorly on Re-ID, often morphing object identities after they disappear, confirming the lack of object permanence in pure 2D methods. The Voxel-WM performs better but still struggles. Analysis reveals that Voxel-WM tends to fail when a hidden object moves significantly, exiting the localized memory buffer it occupied before occlusion. Our LS-WM achieves superior performance, with a 94.3% Re-ID accuracy, demonstrating that the latent set representation effectively maintains object identity regardless of visibility duration or motion magnitude while hidden.

TABLE I: Occlusion Resilience on Kubric MOVi-E (50 frame sequences)

Method	Re-ID Acc ↑	PSNR (Occ Region) ↑
Video-LDM (2D)	42.5%	18.4 dB
Voxel-WM (3D Grid)	68.1%	22.1 dB
LS-WM (Ours)	**94.3%**	**28.5 dB**

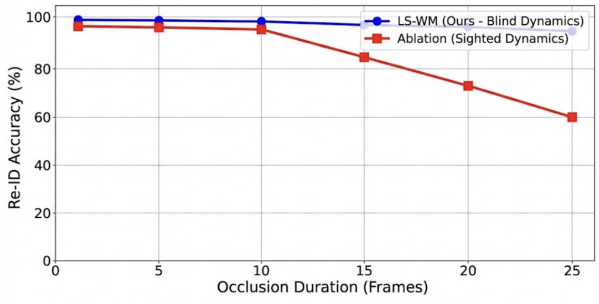

Fig. 4: **Ablation Study: The importance of "Blind" Dynamics.** We compare our full model against an ablated version where the dynamics transformer receives the camera pose as input ("Sighted Dynamics"). The ablated model learns to neglect occluded objects to minimize immediate rendering loss, causing Re-ID accuracy to plummet as occlusion duration increases. Our blind dynamics forces the model to track all objects, maintaining high accuracy even for long occlusions.

D. Qualitative Results

Figure 3 visualizes a challenging long-term occlusion scenario. The baseline Voxel-WM fails to track the small yellow object once it is fully hidden by the large purple cube. When it should reappear, the voxel grid has "forgotten" its features, resulting in a corrupted hallucination. Our LS-WM, by tracking the latent vector z_{yellow} throughout the simulation, correctly predicts its re-emergence with consistent appearance and trajectory.

E. Ablation Study: Importance of Decoupling

To validate our core hypothesis that decoupling dynamics from rendering is essential, we performed an ablation study. We modified our dynamics transformer to accept the target camera pose π_{t+1} as an additional input ("Sighted Dynamics").

As shown in Fig. 4, the "Sighted" model's performance degrades rapidly as occlusion duration increases. Because it knows where the camera is looking, it learns a shortcut: it stops updating the states of objects it knows will not be visible in the next frame to minimize training effort. By making the dynamics model "blind," we force it to learn consistent physics for all objects as a necessary condition for handling arbitrary future camera views.

V. Conclusion

We have presented LS-WM, a generative world model architecture designed to enforce object permanence. By representing scenes as unordered sets of object-centric latent vectors and decoupling "blind" physical dynamics from occlusion-aware rendering, we achieve robust state tracking even during extended periods of invisibility. This addresses a fundamental cognitive deficit in current generative models, paving the way for more reliable and interactive 4D simulations in AR/VR and robotics. Future work will focus on scaling this approach to real-world datasets using unsupervised slot discovery methods.

Acknowledgment

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

References

[1] Y. Song, S. Huang, and Y. Kang, "Temporal-ID: Robust Identity Preservation in Long-Form Video Generation via Adaptive Memory Banks," [Online]. Available: https://nsh423.github.io/assets/publications/paper_2_video_gen_consistency.pdf

[2] Y. Song, Y. Kang, and S. Huang, "VACE-PhysicsRL: Unified Controllable Video Generation through Physical Laws and Reinforcement Learning Alignment," [Online]. Available: https://nsh423.github.io/assets/publications/paper_5_VACE.pdf

[3] Y. Kang, Y. Song, and S. Huang, "Dream World Model (DreamWM): A World-Model-Guided 3D-to-Video Framework for Immersive Narrative Generation in VR," [Online]. Available: https://nsh423.github.io/assets/publications/paper_3_dream.pdf

[4] Y. Kang, S. Huang, and Y. Song, "Robust and Interactive Localized 3D Gaussian Editing with Geometry-Consistent Attention Prior," [Online]. Available: https://nsh423.github.io/assets/publications/paper_6_RoMaP.pdf

[5] R. Baillargeon, "Object permanence in $3^1/_2$-and $4^1/_2$-month-old infants," *Developmental psychology*, vol. 23, no. 5, p. 655, 1987.

[6] F. Locatello et al., "Object-centric learning with slot attention," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 11525–11538, 2020.

[7] B. Mildenhall et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *ECCV*, 2020.

[8] K. Greff et al., "Kubric: A scalable dataset generator," in *CVPR*, 2022.

[9] B. Kerbl et al., "3D Gaussian Splatting for Real-Time Radiance Field Rendering," *ACM Trans. Graph.*, vol. 42, no. 4, 2023.