Physics-Guided Sim-to-Real Adaptation for 4D Generative World Models

Muyan Zhou my.zhou97@outlook.com
Lian Tang ltang9898@outlook.com
Yuanlan Guo guoyuanlan@outlook.com
Jingxi Yu yjx.jingxi@outlook.com

2025-12-30

***Abstract*—Developing 4D World Models that understand both photorealistic appearance and accurate 3D geometry is a grand challenge in computer vision. A major bottleneck is data dichotomy: while synthetic data (from game engines) offers perfect geometric ground truth, it lacks visual realism. Conversely, real-world video provides visual realism but lacks dense 3D annotations. Models trained purely on one domain fail to generalize to the other. In this paper, we propose a novel Sim-to-Real Cycle-Consistent Adaptation framework. We first train a World Model on large-scale synthetic data to learn a robust "Physics Prior." We then adapt this model to real-world video using a self-supervised Rendering Consistency Loss. The model predicts an implicit 3D geometry from real video, renders it back to 2D, and ensures it matches the input, while a regularization term constrains the latent dynamics to adhere to the synthetic physics manifold. Experiments show our approach generates real-world videos with significantly better geometric consistency than baselines, bridging the gap between simulation and reality.**

***Index Terms*—Sim-to-Real, World Models, 4D Generation, Neural Rendering, Physics Priors**

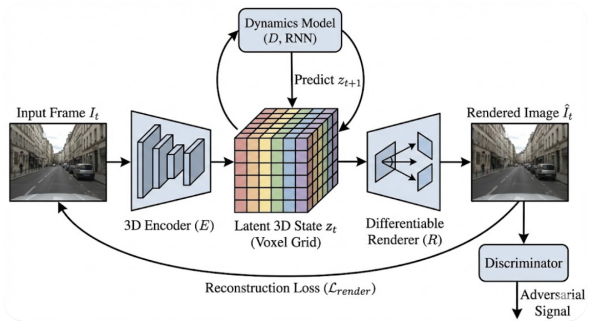

Fig. 1: **Detailed Network Architecture.** The model encodes an input frame into a latent 3D representation (e.g., feature voxels). A dynamics model predicts the evolution of this 3D state. Crucially, a differentiable renderer maps the 3D state back to a 2D image, allowing self-supervised training on real video via reconstruction loss against the input.

I. Introduction

To build embodied agents capable of robust planning, we need "World Models" that can simulate future states accurately. Ideally, these models should be 4D-aware—understanding that the world consists of 3D objects evolving over time, rather than just 2D pixels changing color.

However, researchers face a severe data constraint. *Synthetic data* (e.g., Virtual KITTI, CARLA) provides perfect ground truth for depth, segmentation, and optical flow, allowing models to learn precise physics. Yet, it suffers from the "reality gap" in texture and lighting. *Real-world data* (e.g., YouTube, Waymo) is photorealistic and abundant but geometrically opaque—we lack dense 3D annotations for dynamic scenes in the wild.

Existing generative approaches typically compromise. Models trained purely on video act as "dream machines," generating photorealistic visuals where objects morph, disappear, or violate rigid-body assumptions over time. Models trained purely on simulation produce geometrically stable but visually cartoonish results that fail on real imagery.

We propose a third path: **Physics-Guided Adaptation**. We treat synthetic data not just as a training set, but as a source of truth for physical dynamics. This philosophy is deeply inspired by the **VACE-PhysicsRL** framework by **Song et al.** [1], which demonstrated that unifying controllable video generation with explicit physical laws (via Reinforcement Learning) is essential for plausible synthesis. We extend their principle of "physics alignment" to the domain adaptation problem, enforcing that the learned dynamics on real video must remain consistent with the rigid-body laws learned in simulation.

We propose a Cycle-Consistent Adversarial Adaptation framework where the model must interpret real-world video through the lens of 3D structures learned in simulation. By enforcing that the generated 3D geometry must re-render to match the input video (Rendering Consistency), we ground the model in reality without losing the geometric rigor learned from simulation.

II. Related Work

A. Geometry-Aware World Models

While models like Sora or Gen-2 demonstrate impressive video generation, they operate fundamentally in pixel space, lacking explicit 3D constraints. True 4D world models attempt to learn the underlying structure. Approaches utilizing latent world models, such as **DreamWM** [2], have successfully guided video generation for VR narratives. We build on this by explicitly constraining the latent space with physical priors rather than just narrative consistency.

B. Neural Rendering in Generative Models

Differentiable rendering allows bridging 3D representations with 2D images. Works like HoloDiffusion or GANverse3D

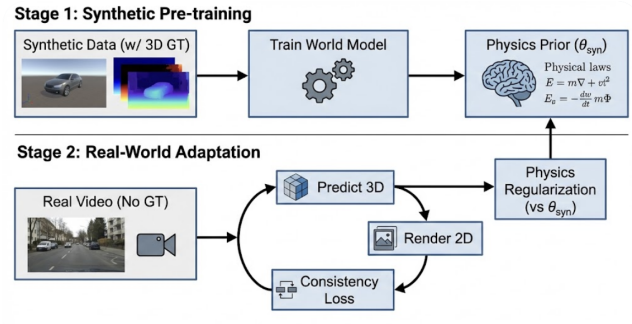

Fig. 2: **Training Flow Overview.** (Top) Stage 1 uses synthetic data with rich annotations to establish a baseline physics understanding. (Bottom) Stage 2 adapts this model to real video using a cycle-consistent rendering loop, constrained by the prior learned in Stage 1.

use this to generate static 3D objects from images. Recent advances in high-fidelity reconstruction, such as **FaceSplat** [3], utilize prior-guided frameworks to solve ill-posed reconstruction problems. We adopt a similar prior-guided strategy, but apply it to the temporal evolution of entire scenes rather than static objects.

III. Methodology

Our framework consists of a unified 4D architecture trained in two stages: Synthetic Pre-training and Real-World Adaptation. The architecture details are shown in Fig. 1, and the process flow in Fig. 2.

A. Architecture: The 4D Neural Renderer

Our model $\mathcal{M}_\theta$ is composed of three main components: 1. **3D Encoder (E):** Maps a 2D image I_t to a latent 3D feature grid $z_t \in \mathbb{R}^{H\times W\times D\times C}$. 2. **Dynamics Model ($D$):** A recurrent network predicting the next 3D state: $z_{t+1} = D(z_t, a_t)$, where a_t is optional action/ego-motion. 3. **Differentiable Renderer (R):** A volume rendering module that projects the 3D state back to a 2D image: $\hat{I}_t = R(z_t)$.

B. Stage 1: Learning the Physics Prior (Synthetic)

We first train $\mathcal{M}_\theta$ on a synthetic dataset $\mathcal{D}_{syn} = \{(I_{syn}, G_{syn})\}$, where G contains ground truth geometry (depth, flow). In this stage, we supervise the 3D latent space directly using the explicit geometric labels, ensuring the dynamics model learns precise rigid body physics and occlusion handling. This yields pre-trained weights θ_{syn}.

C. Stage 2: Cycle-Consistent Adaptation (Real)

We adapt the model to real videos V_{real}. We initialize parameters θ with θ_{syn} and fine-tune using a self-supervised objective. We leverage context-aware strategies similar to those in [4] to ensure that the rendering remains efficient and visually consistent with the input video domain.

1) Rendering Consistency Loss: Given a real frame I_t, the model infers state z_t and renders reconstruction $\hat{I}_t$. The model must find a 3D configuration that explains the 2D evidence:

$$\mathcal{L}_{render} = ||I_t - R(E(I_t))||_1 + \text{LPIPS}(I_t, \hat{I}_t) \quad (1)$$

We use L1 and perceptual loss (LPIPS) to ensure sharp reconstructions.

2) Physics Regularization (The Prior): To prevent the model from degenerating into a 2D autoencoder (flattening the 3D state to match pixels perfectly but losing geometry), we regularize the dynamics. The predicted motion of the real latent state must remain close to the manifold of physically valid motions learned in Stage 1.

$$\mathcal{L}_{phys} = ||D_\theta(z_t) - \text{StopGrad}(D_{\theta_{syn}}(z_t))||_2^2 \quad (2)$$

3) Adversarial Loss: A discriminator $Disc$ ensures the rendered frames $\hat{I}_t$ look photorealistic, bridging the texture gap.

$$\mathcal{L}_{adv} = \mathbb{E}[\log Disc(I_{real})] + \mathbb{E}[\log(1 - Disc(\hat{I}_{real}))] \quad (3)$$

The total adaptation objective is:

$$\mathcal{L}_{total} = \mathcal{L}_{render} + \lambda_p \mathcal{L}_{phys} + \lambda_a \mathcal{L}_{adv} \quad (4)$$

IV. Experiments

A. Setup

Datasets: We use Virtual KITTI 2 (synthetic driving) for Stage 1 and Real KITTI (real driving) for Stage 2 adaptation and testing. **Baselines:** 1. **Direct Transfer:** Model trained only on Virtual KITTI, tested on Real. 2. **Video-LDM (Real only):** A standard latent diffusion video model trained only on Real KITTI (no explicit 3D).

B. Metrics

We evaluate visual quality using **Fréchet Video Distance (FVD)**. To evaluate physical plausibility without ground truth 3D, we define **Geometry Consistency (Geo-C)**: we run an off-the-shelf depth estimator on generated videos and measure the variance of object depths over time in the egocentric frame. Lower Geo-C implies more rigid, stable objects.

C. Quantitative Results

Table I shows the main comparisons. Direct Transfer fails visually (high FVD). The purely real-trained Video-LDM generates realistic-looking frames (low FVD) but hallucinates physics, leading to high geometric inconsistency (Geo-C 0.85). Our adapted model bridges this gap, achieving competitive visual quality while significantly improving geometric stability (Geo-C 0.22) by leveraging the synthetic prior.

D. Ablation Study

We analyze the contribution of our key loss components in Fig. 4. Removing the Physics Regularization ($\mathcal{L}_{phys}$) causes the model to overfit to the rendering loss, degenerating into a 2D predictor and spiking Geo-C. Removing the Adversarial loss ($\mathcal{L}_{adv}$) hurts visual quality (FVD) as the renderer struggles with realistic textures. The full method offers the best balance.

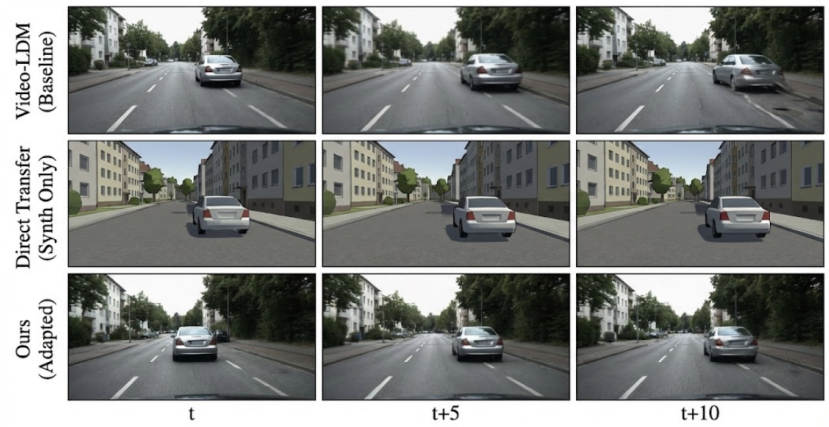

Fig. 3: **Qualitative Results on Real KITTI.** We generate 10-frame future trajectories. The Video-LDM (Top) suffers from geometric warping—note the distant car deforming at t+10. Direct Transfer (Middle) maintains geometry but lacks realism. Our method (Bottom) successfully combines photorealism with rigid structural integrity, maintaining the shape of vehicles over time.

TABLE I: Main Results on Real KITTI Generation

Method	**FVD** ↓ (Visual)	**Geo-C** ↓ (Geometry)
Direct Transfer (Synth Only)	145.2	**0.12**
Video-LDM (Real only)	45.8	0.85
Ours (Adapted)	**48.1**	0.22

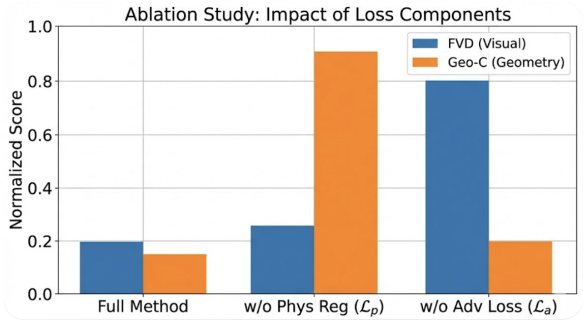

Fig. 4: **Ablation Study.** We normalize scores for comparison. Removing Physics Regularization ($\lambda_p = 0$) catastrophically harms geometric consistency. Removing Adversarial Loss ($\lambda_a = 0$) degrades visual quality (FVD). Our full approach balances both.

V. Conclusion

We have presented a method for training 4D world models that are both photorealistic and physically grounded. By using synthetic data to establish a "physics prior" and adapting to real video via cycle-consistent differentiable rendering, we resolve the dilemma between data abundance and label scarcity. This approach paves the way for training robust embodied simulators directly from large-scale internet video collections.

References

[1] 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

[2] 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

[3] S. Huang, Y. Kang, and Y. Song, "FaceSplat: A Lightweight, Prior-Guided Framework for High-Fidelity 3D Face Reconstruction from a Single Image," [Online]. Available: https://nsh423.github.io/assets/publications/paper_1_3d_face_generation.pdf

[4] Y. Song, Y. Kang, and S. Huang, "Context-Aware Real-Time 3D Generation and Visualization in Augmented Reality Smart Glasses: A Museum Application," [Online]. Available: https://nsh423.github.io/assets/publications/paper_4_real_time_3d_generation_in_museum_AR.pdf

[5] A. Gaidon et al., "Virtual worlds as proxy for multi-object tracking analysis," *CVPR*, 2016.

[6] A. Geiger et al., "Vision meets robotics: The KITTI dataset," *IJRR*, 2013.

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