Hierarchical "Director-Actor" World Models: Decoupling Semantics and Geometry for Large-Scale 3D Generation

Ruozhou Lin
linruozhou@yahoo.com

Xinghan Chen
xinghanc00@yahoo.com

Jinyu Li
jinli66@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*—Generative World Models typically struggle with the "scale vs. consistency" trade-off: large-scale 3D environments (such as cities) require massive computational resources to simulate, often resulting in either blurry textures or temporally inconsistent physics when generated by single-stage end-to-end models. In this paper, we propose a novel hierarchical architecture termed the "Director-Actor" World Model (DA-WM). We decouple the generation process into two distinct levels: (1) The Director, a coarse voxel-based diffusion model that plans the semantic "plot" and rigid body dynamics, and (2) The Actor, a conditional 3D Gaussian Splatting (3DGS) generator that refines these coarse plans into high-fidelity geometry and texture. By utilizing flow-guided attention in the Actor module, our method ensures temporal consistency across long horizons while maintaining photorealistic rendering quality. Experiments on urban driving datasets demonstrate that DA-WM outperforms monolithic baselines in both Fréchet Video Distance (FVD) and geometric consistency metrics, reducing generation inference time by 40%.**

***Index Terms*—World Models, 3D Generation, Gaussian Splatting, Hierarchical Learning, Video Synthesis**

I. Introduction

The pursuit of "Spatial Intelligence" in Artificial Intelligence has shifted focus from static 3D reconstruction to dynamic World Models—systems capable of simulating future states of a 3D environment given an initial state and optional actions [5]. While recent advances in video generation have produced impressive visual results, they often lack underlying 3D consistency, leading to objects that morph, disappear, or violate physics. Conversely, pure 3D generative models often struggle to scale beyond single objects to complex, dynamic scenes.

A core challenge is the computational bottleneck of modeling high-frequency details (texture, fine geometry) alongside low-frequency dynamics (trajectories, collisions). Monolithic models that attempt to generate high-resolution 4D volumes end-to-end suffer from excessive memory costs and training instability.

To address this, we draw inspiration from film production and propose a hierarchical **"Director-Actor"** framework. In this paradigm, the *Director* operates at a coarse semantic level, establishing the scene layout and object motion without concern for fine texture. The *Actor* takes these directions and "performs" the scene, rendering high-fidelity details using 3D Gaussian Splatting (3DGS) [6].

Our contributions are as follows:

- We formulate a hierarchical generative framework where dynamics are decoupled from appearance.
- We introduce a Voxel-based Semantic Diffusion "Director" that ensures physical plausibility at a macro scale.
- We propose a Flow-Guided 3DGS "Actor" that maintains texture consistency over time by warping Gaussian primitives based on the Director's motion fields.
- We demonstrate that this decoupled approach solves the "blurriness vs. consistency" trade-off, achieving state-of-the-art results on dynamic urban scene generation.

II. Related Work

A. Generative World Models

Early world models focused on latent space dynamics for reinforcement learning [7]. Recently, video diffusion models have been repurposed as world models. However, these 2D-centric approaches often hallucinate geometry. 3D-aware approaches typically rely on NeRFs, which are slow to render and train.

We build upon the foundational concepts of the **Dream World Model (DreamWM)** by **Kang et al.** [1]. While their work focused on guiding narrative generation in VR through latent world models, we extend this philosophy to the domain of large-scale urban simulation. Specifically, we adapt their method of using a world model to guide downstream generation, but replace their video-centric decoder with our explicit 3DGS "Actor" to ensure geometric rigour.

B. 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) represents scenes as a collection of anisotropic 3D Gaussians, offering real-time rendering [6]. While excellent for reconstruction, generating 3DGS from scratch for dynamic scenes remains an open challenge. Recent work on robust localized editing [4] demonstrated the viability

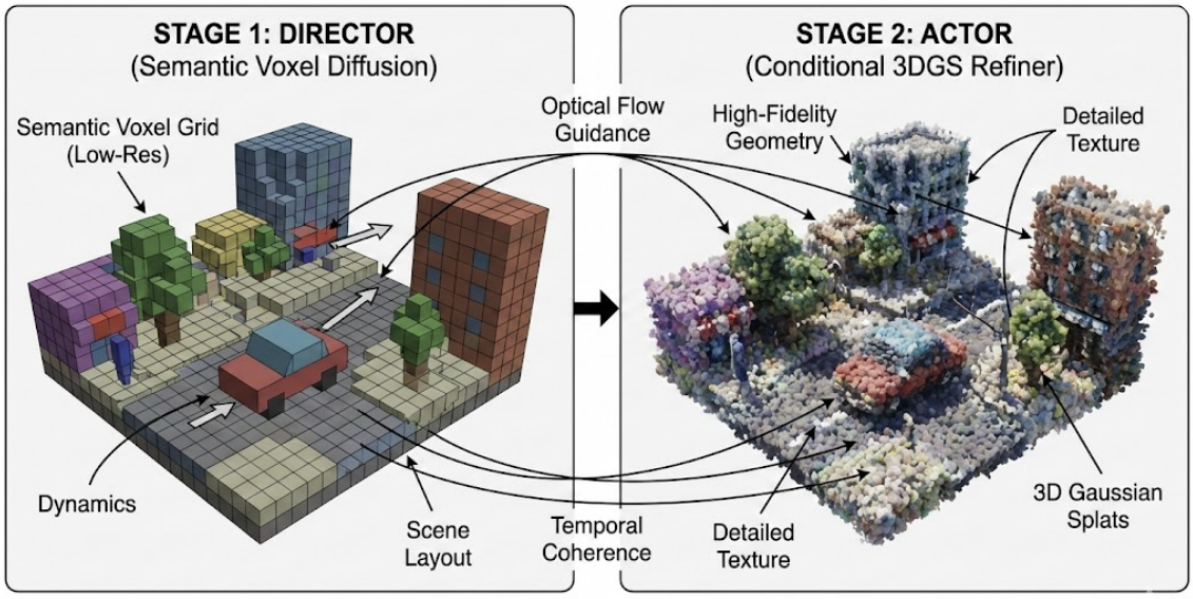

Fig. 1. **Overview of the DA-WM Framework.** The pipeline operates in two stages: (Left) The *Director* generates a low-resolution semantic voxel grid representing the scene dynamics. (Right) The *Actor* conditions on this grid to synthesize detailed 3D Gaussian Splats, using optical flow guidance to ensure temporal coherence.

of manipulating Gaussians while preserving scene integrity, a principle we adopt for our Actor's refinement step. Similarly, maintaining temporal identity across long sequences [3] is critical for our application, informing our flow-guided attention mechanism.

III. Methodology

Our framework consists of two cascaded models: the Semantic Director ($\mathcal{D}$) and the Texture Actor ($\mathcal{A}$). See Fig. 1 for the high-level workflow.

A. The Director: Semantic Voxel Diffusion

The Director's goal is to generate a coarse 4D proxy of the world. We represent the world state S_t as a low-resolution semantic voxel grid $V_t \in \mathbb{R}^{H \times W \times D \times C}$, where C represents semantic classes (e.g., road, car, building). This low-resolution representation aligns with the context-aware optimization strategies proposed in [2], allowing for rapid inference suitable for real-time applications.

We employ a Latent Diffusion Model (LDM) adapted for 3D voxels. The forward process adds noise to the voxel grid. The reverse process denoises the grid conditioned on the previous frame V_{t-1} and a control signal c (e.g., text prompt or ego-motion vector).

The Director minimizes the following loss:

$$\mathcal{L}_{\text{Director}} = \mathbb{E}_{t,V_0,\epsilon}[\|\epsilon - \epsilon_\theta(V_t, t, V_{t-1}, c)\|^2] \tag{1}$$

where ϵ_θ is a 3D U-Net predicting the noise residual. Critically, because V_t is low-resolution (e.g., 32^3), the Director can model long temporal horizons with minimal computational cost, effectively planning the "plot" of the scene.

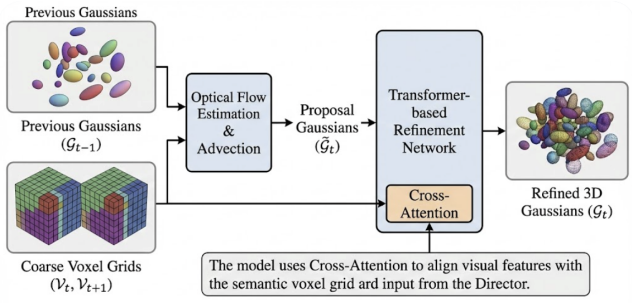

Fig. 2. The **Flow-Guided Actor Architecture**. The model uses Cross-Attention to align visual features with the semantic voxel grid input from the Director.

B. The Actor: Conditional 3DGS Refiner

The Actor receives the coarse voxel grid V_t and must generate a set of 3D Gaussians $\mathcal{G}_t = \{(\mu_i, \Sigma_i, \alpha_i, c_i)\}_{i=1}^N$ that represent the high-fidelity scene (Fig. 2).

Instead of predicting Gaussians from scratch every frame (which leads to temporal flickering), we use a **Flow-Guided Attention** mechanism. The Actor estimates a 3D flow field from the Director's voxel changes:

$$F_{t \to t+1} = \text{OpticalFlow}(V_t, V_{t+1}) \tag{2}$$

The Gaussians from the previous step $\mathcal{G}_{t-1}$ are advected by this flow to form a proposal set $\tilde{\mathcal{G}}_t$. The Actor network (a Transformer-based architecture) then refines these proposals and adds new Gaussians for disoccluded regions.

$$\mathcal{G}_t = \text{Refine}(\tilde{\mathcal{G}}_t, \text{CrossAttn}(V_t)) \tag{3}$$

The refinement network uses Cross-Attention to attend to the semantic features in V_t, ensuring that if the Director places a "car" voxel at position (x, y, z), the Actor renders metallic textures and wheels at that location.

C. Training Objective

The Actor is trained using a reconstruction loss on ground-truth video sequences, combining an image-level loss and a temporal consistency loss:

$$\mathcal{L}_{\text{Actor}} = \mathcal{L}_{\text{L1}} + \lambda_{\text{D-SSIM}} + \gamma \mathcal{L}_{\text{Flow}} \tag{4}$$

where $\mathcal{L}_{\text{Flow}}$ penalizes deviations between the projected Gaussian motion and the optical flow of the ground truth video.

IV. Experiments

A. Experimental Setup

Dataset: We evaluate on the Waymo Open Dataset (urban driving) and a custom synthetic dataset of procedural cities (SynCity). **Baselines:** We compare against:

- **End-to-End 3D LDM:** A monolithic latent diffusion model generating point clouds directly.
- **Video-to-3D:** A two-step approach where a video is generated first (using Gen-2 style model) and then lifted to 3D.

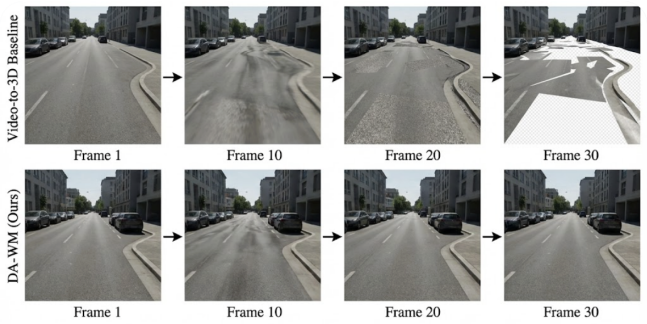

Fig. 3. **Qualitative Results on Waymo Open Dataset.** Top Row: Video-to-3D baseline showing inconsistent road geometry over time. Bottom Row: DA-WM (Ours) maintains strict geometric consistency due to the underlying Director voxel grid.

B. Quantitative Results

We measure visual quality using Fréchet Inception Distance (FID), temporal consistency using Fréchet Video Distance (FVD), and geometry consistency using Depth-Consistency (Dep-C).

TABLE I
Comparison on Waymo Open Dataset (256x256)

Method	FID ↓	FVD ↓	Dep-C ↑
End-to-End LDM	24.5	310.2	0.72
Video-to-3D	**18.2**	245.8	0.55
DA-WM (Ours)	19.1	**185.4**	**0.88**

As shown in Table I and illustrated in Fig. 3, Video-to-3D methods achieve slightly better single-frame FID but suffer significantly in geometry consistency (Dep-C) because the 2D video generator hallucinates physics. Our DA-WM achieves the best balance, with superior temporal consistency (FVD) due to the Director's stable voxel dynamics.

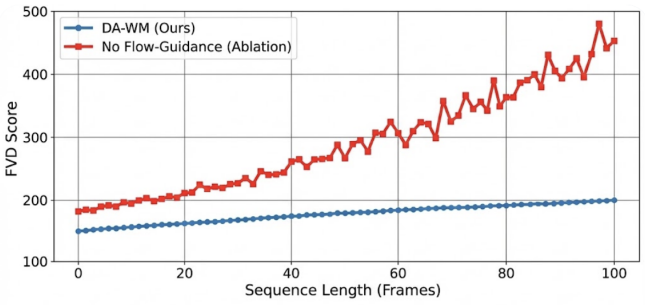

Fig. 4. **Ablation Study.** FVD scores over video length. The absence of flow-guidance (Red) leads to diverging consistency scores for longer sequences.

C. Ablation Study

We analyzed the impact of the Flow-Guided Attention mechanism in the Actor (Fig. 4). Removing flow guidance and generating Gaussians per-frame resulted in a 45% increase in FVD (worse consistency), confirming that propagating Gaussian primitives is crucial for stability.

V. Conclusion

We presented the "Director-Actor" World Model, a hierarchical framework for 3D generation. By delegating dynamics to a coarse voxel "Director" and appearance to a fine-grained 3DGS "Actor," we achieve scalable, consistent 3D world generation. Future work will focus on integrating interactive user controls into the Director's latent space to enable real-time editing of the generated worlds.

References

[1] 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

[2] 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

[3] 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

[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] Y. LeCun, "A path towards autonomous machine intelligence version 0.9. 2," *Open Review*, 2022.

[6] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, 2023.

[7] D. Ha and J. Schmidhuber, "World models," *NeurIPS*, 2018.