DECOUPLING APPEARANCE AND GEOMETRY IN DIFFUSION WORLD MODELS

PREPRINT

Rajit Rajpal
University of Edinburgh
s2592586@ed.ac.uk

Tadgh Kelly
Qubit Analytics

Maxime Vandegar
Qubit Analytics

August 18, 2026

ABSTRACT

We study the problem of learning diffusion-based world models that jointly capture scene geometry and visual appearance in complex first-person environments. Existing approaches typically train a single action-conditioned model that predicts future RGB frames, but they either ignore geometry or require action supervision for all modalities, limiting scalability to large unlabeled video datasets. In this work, we introduce a two-stage diffusion world model that separates geometric and appearance prediction while maintaining joint temporal coherence. Our key idea is a staggered RGB–depth formulation: depth at time $t + 1$ is predicted first using an action-conditioned diffusion transformer, and the RGB frame at the same timestep is then generated conditioned on this predicted depth using a second diffusion model without action inputs. We show that this decoupled design faithfully approximates a hypothetical joint RGBD diffusion model trained with diffusion forcing, enabling the second stage to be trained or fine-tuned on additional RGB–depth data without requiring action labels. Experiments on the CS:GO Dust 2 dataset demonstrate that our approach improves both depth accuracy and RGB fidelity compared to fully joint models and RGB-only baselines, while significantly enhancing scalability to augmented or unlabeled datasets. Our results highlight the effectiveness of geometry-guided diffusion for world modeling and open a path toward large-scale action-free video model training. Code to be released upon acceptance of paper.

Keywords Diffusion Models, World Models, Video Generation, 3D

1 Introduction

Modeling complex 3D environments from high-dimensional sensory observations is a central challenge in computer vision and embodied AI. Recent advances in diffusion models—have dramatically improved our ability to synthesize realistic images, videos, and 3D structures. At the same time, world models have emerged as a powerful paradigm for learning predictive representations of environments, enabling agents to forecast future observations, reason about dynamics, and act effectively in interactive settings. However, building world models that are both temporally coherent and geometrically grounded remains an open problem. A key difficulty lies in jointly capturing visual appearance and scene geometry. Structural guidance (primarily from depth maps) has been shown to be very beneficial in video world modeling [23], [21], [20][5] [3]. While RGB observations provide rich photorealistic detail, they are inherently ambiguous: many distinct 3D configurations can produce similar images. Depth, on the other hand, offers strong geometric constraints but lacks texture and semantic information. There are also works that train Diffusion Models in 2 stages in various modalities (e.g Optical Flow, Depth, Camera Parameters etc.) [11] [6][22] [12]. A 2 stage pipeline can help to improve controllability as well. However, there does not exist a principled method to train both stages jointly.

Prior works [1] [10] [19] [7][4] take important steps toward learning diffusion world models from large-scale gameplay data, demonstrating that videos from first-person environments like CS:GO Dust2 and Minecraft offer an ideal testbed for studying long-horizon dynamics. Yet these approaches still rely heavily on action-conditioned modeling or unified representations that must simultaneously encode geometry, appearance, and dynamics—making training fragile, data-

hungry, or difficult to scale. A persistent challenge in this space is balancing structural fidelity and visual realism without entangling the two. Methods that treat RGB and depth jointly often struggle to enforce geometric consistency, while methods that train separate models typically require explicit action supervision to maintain temporal coherence. Moreover, action labels are rarely available outside interactive simulators, limiting the applicability of such models to real-world or web-scale video data.

These limitations raise an important question: Can we construct a world model that leverages depth to provide stable geometry, while using generative RGB modeling to capture appearance—without requiring action-label supervision throughout? A solution to this would significantly expand the scalability and practicality of diffusion-based world models. It would allow RGB models to be trained on far larger and more diverse datasets, including those without action annotations—while retaining the strong geometric grounding needed for coherent long-horizon rollouts. It would also open the door to training at higher frame rates or using synthetic depth derived from off-the-shelf estimators, thereby bypassing the need for simulator access. In this work, we explore this direction and develop a two-stage diffusion-based world modeling framework that decouples geometry from appearance while maintaining joint predictive capability. Our approach leverages the unique properties of depth and RGB modalities, and is evaluated on the widely used CS:GO Dust2 environment to ensure comparability with prior state-of-the-art systems.

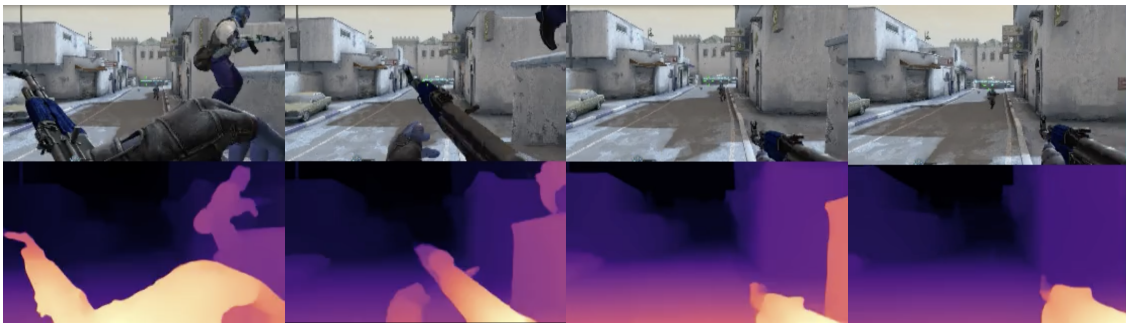

Figure 1: A video sequence of RGB (first row) from the CSGO Dust 2 dataset [13] and Depth maps (second row) extracted with MiDaS DPT-Hybrid [15].

2 Background

Diffusion Models. Diffusion models [9] have emerged as a powerful class of generative models capable of synthesizing high-fidelity images, videos, and 3D signals. They operate by learning to reverse a fixed Markovian noising process applied to a data sample $\mathbf{z}_0$. The forward diffusion process adds Gaussian noise in T steps:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

where $\{\bar{\alpha}_t\}$ denotes the cumulative product of noise schedule coefficients. The generative model learns a denoising function $f_\theta(\mathbf{z}_t, t)$ that estimates either the noise $\boldsymbol{\epsilon}$, the clean signal $\mathbf{z}_0$, or—following [17]—a blended latent variable known as the *velocity* $\mathbf{v}$. The v-prediction parameterization offers improved training stability and sampling efficiency. It reparameterizes Eq (1) in terms of velocity as:

$$\mathbf{v}_t = \sqrt{\bar{\alpha}_t}\,\boldsymbol{\epsilon} - \sqrt{1-\bar{\alpha}_t}\,\mathbf{z}_0.$$

This transformation is invertible, and the pair $(\mathbf{z}_t, t)$ uniquely determines $(\mathbf{z}_0, \boldsymbol{\epsilon}, \mathbf{v}_t)$. A neural network then predicts $\hat{\mathbf{v}}_\theta$ in place of $\hat{\boldsymbol{\epsilon}}_\theta$ or $\hat{\mathbf{z}}_{0,\theta}$ with optional conditioning $\mathbf{c}$ (e.g., actions, class), the denoiser becomes $\hat{\mathbf{v}}_\theta(\mathbf{z}_t, t \mid \mathbf{c})$ with a simple MSE training objective:

$$\mathcal{L}_{\text{v-pred}} = \mathbb{E}_{\mathbf{z}_0,t,\boldsymbol{\epsilon}}\Big[\|\mathbf{v}_t(\mathbf{z}_0, t) - \hat{\mathbf{v}}_\theta(\mathbf{z}_t, t \mid \mathbf{c})\|^2\Big]. \tag{1}$$

At sampling time, the predicted velocity is mapped back to the noise or clean-signal estimate. Because the network learns a nearly scale-invariant representation, $\mathbf{v}$-prediction is widely used in high-resolution video diffusion models.

Diffusion forcing for autoregressive prediction. To extend diffusion models to sequential or dynamical settings, the model must predict future observations conditioned on past ones. Diffusion forcing [2, 18] provides a simple and effective mechanism: instead of predicting a clean future frame $\mathbf{x}_{t+1}$ directly, the model predicts a *noisy* version of it at a randomly sampled diffusion timestep τ. Formally, given a history $\mathcal{H}_t = \{\mathbf{x}_{t-k}, \ldots, \mathbf{x}_t\}$,

$$(\mathcal{H}_t, \tilde{\mathbf{x}}_{t+1}), \qquad \tilde{\mathbf{x}}_{t+1} = q(\mathbf{x}_{t+1}, \tau),$$

is used as a training pair, where q is the forward noising operator. Because the target lies on the diffusion manifold, the model remains close to the support of the forward process even when unrolled autoregressively. This substantially mitigates compounding error and improves long-horizon stability while maintaining parallel training across frames in a given sequence.

Action conditioning in world models. World models generalize generative modeling to interactive environments in which future observations depend on both the environment state and agent actions. Diffusion world models incorporate actions by conditioning the denoiser $\hat{\mathbf{v}}_\theta(\mathbf{z}_\tau, \tau \mid \mathcal{H}_t, \mathbf{a}_t)$, which implicitly models $p(\mathbf{x}_{t+1} \mid \mathbf{x}_{1:t}, \mathbf{a}_t)$. Due to the requirement of action-conditioning at each frame, diffusion world models must be temporally autoregressive to see and control the rollout accordingly. Compared to unconditional diffusion models, whose goal is to model a static data distribution, action-conditioned models must learn transition dynamics consistent with control inputs. While effective, this introduces a dependence on action labels, which are often expensive to obtain or unavailable outside simulators. VLAs are also used to predict actions from unlabelled videos but they are often not accurate and require lots of training and supervision. This limitation motivates architectures that either weaken or remove this requirement, enabling training on broader datasets such as augmented videos or web-scale RGB–depth pairs.

3 Methodology

We begin by considering a hypothetical *joint* diffusion transformer (DiT) that models RGB and depth simultaneously. Let $\mathbf{x}_t^R$ denote the RGB latent at time t and $\mathbf{x}_t^D$ the corresponding depth latent. A naïve joint formulation concatenates the two modalities channel-wise, producing a latent tensor of shape $(C_R + C_D, H, W)$, and learns a denoising function

$$f_\theta(\mathbf{z}_\tau^{R,D}, \tau \mid \mathbf{z}_{\leq t}^{R,D}),$$

trained with diffusion forcing so that the next RGB latent is conditioned on noisy RGB and depth histories.

However, this joint setting obscures an asymmetry between appearance and geometry: RGB should be *guided* by geometry, whereas depth should be *driven* by dynamics and actions. To expose this structure, we introduce a temporal **staggering** of modalities. Instead of aligning RGB and depth at the same timestep, we offset depth forward by one frame:

$$\{\mathbf{x}_0^R, \ldots, \mathbf{x}_t^R\}, \qquad \{\mathbf{x}_1^D, \ldots, \mathbf{x}_{t+1}^D\}.$$

This induces two asymmetric conditioning relationships:

$$\mathbf{x}_{t+1}^D \leftarrow (\mathbf{z}_{\leq t}^D, \mathbf{z}_{\leq t-1}^R, \mathbf{a}_t), \tag{2}$$

$$\mathbf{x}_{t+1}^R \leftarrow (\mathbf{z}_{\leq t}^R, \mathbf{z}_{t+1}^D). \tag{3}$$

Thus the geometric future (depth at $t{+}1$) depends on past geometry, past appearance, and the agent action; but appearance at $t{+}1$ may directly depend on the *current* depth $\mathbf{x}_{t+1}^D$, serving as a geometric guide.

Decoupling into two diffusion stages. Instead of training one large joint model, we factor the problem into two specialized DiTs:

1. **Stage 1: Depth DiT (action-conditioned).**
   Learns $p(\mathbf{x}_{t+1}^D \mid \mathbf{x}_{\leq t}^D, \mathbf{x}_{\leq t-1}^R, \mathbf{a}_t)$. This model captures scene geometry and dynamics.
2. **Stage 2: RGB DiT (depth-guided, action-free).**
   Learns $p(\mathbf{x}_{t+1}^R \mid \mathbf{x}_{\leq t}^R, \mathbf{x}_{t+1}^D)$, enabling RGB synthesis that leverages the depth predicted in Stage 1.

To preserve the training dynamics of a joint model, each DiT receives inputs constructed by concatenating modality channels at matched diffusion noise levels. Specifically, for each training sample we generate noisy latents

$$\mathbf{z}_\tau^R = q(\mathbf{x}^R, \tau), \qquad \mathbf{z}_\tau^D = q(\mathbf{x}^D, \tau),$$

and feed each DiT the same noisy (RGB, depth) pairing it would have seen in the joint model. Diffusion forcing is applied identically in both stages, ensuring that both networks learn to denoise autoregressively on the same manifold of noisy inputs.

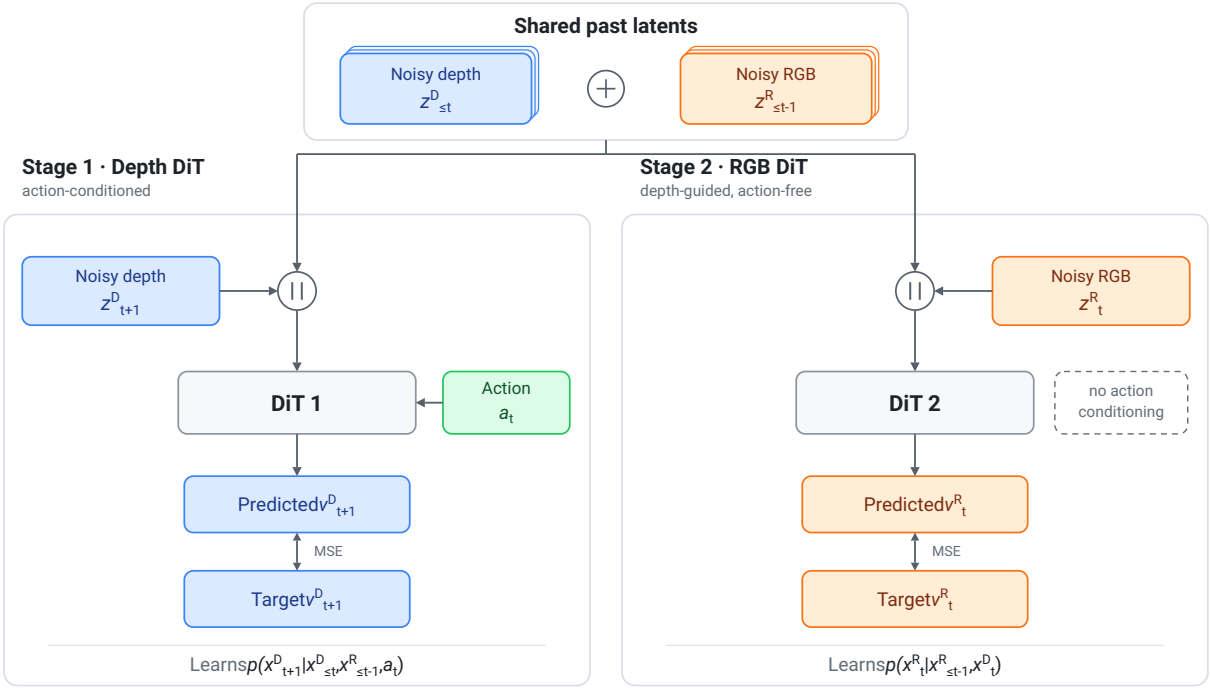

Figure 2: **Staggered Joint Diffusion Training Strategy.** We illustrate the decoupled training procedure which factorizes the joint modeling of video into two specialized Diffusion Transformers (DiTs). **Stage 1 (Left):** The Action-Conditioned Depth DiT learns scene dynamics by denoising the depth latent $\mathbf{z}^D_{t+1}$, conditioned on past depth $\mathbf{z}^D_{\leq t}$, staggered past RGB $\mathbf{z}^R_{\leq t-1}$, and the agent action $\mathbf{a}_t$. **Stage 2 (Right):** The Depth-Guided RGB DiT learns appearance by denoising $\mathbf{z}^R_t$. Crucially, it receives structural guidance via the *noisy ground-truth* depth $\mathbf{z}^D_t$ (channel-wise concatenated), alongside the full history $\mathbf{z}^{R,D}_{\leq t-1}$. Both models utilize v-prediction with a shared noise schedule τ, ensuring the decoupled stages approximate a valid joint distribution.

To formally justify our staggered training strategy, we start with the standard autoregressive factorization of the joint distribution for the next frame's RGB ($\mathbf{x}^R_{t+1}$) and depth ($\mathbf{x}^D_{t+1}$). Conditioned on the full history of both modalities and the agent's action $\mathbf{a}_t$, the joint probability decomposes via the chain rule as:

$$p(\mathbf{x}^R_{t+1}, \mathbf{x}^D_{t+1} \mid \mathbf{x}^R_{1:t}, \mathbf{x}^D_{1:t}, \mathbf{a}_t) = \underbrace{p(\mathbf{x}^D_{t+1} \mid \mathbf{x}^R_{1:t}, \mathbf{x}^D_{1:t}, \mathbf{a}_t)}_{\text{Dynamics Term}} \cdot \underbrace{p(\mathbf{x}^R_{t+1} \mid \mathbf{x}^R_{1:t}, \mathbf{x}^D_{1:t}, \mathbf{x}^D_{t+1}, \mathbf{a}_t)}_{\text{Appearance Term}}. \tag{4}$$

Equation 4 represents a fully connected joint model. To decouple the training into the specialized stages illustrated in Figure 2, we introduce two conditional independence assumptions that reflect the causal asymmetry of video generation.

We posit that the physical evolution of the scene (depth at $t+1$) is driven by the geometric history and agent actions, rather than the synchronous appearance at time t. We therefore assume $\mathbf{x}^D_{t+1}$ is conditionally independent of $\mathbf{x}^R_t$:

$$\mathbf{x}^D_{t+1} \perp \mathbf{x}^R_t \mid \mathbf{x}^R_{1:t-1}, \mathbf{x}^D_{1:t}, \mathbf{a}_t. \tag{5}$$

This simplifies the dynamics term to $p(\mathbf{x}^D_{t+1} \mid \mathbf{x}^R_{1:t-1}, \mathbf{x}^D_{1:t}, \mathbf{a}_t)$, where the appearance input is staggered by one frame.

We assume that appearance generation is primarily a function of texture history and the target surface geometry. Once the target geometry $\mathbf{x}^D_{t+1}$ is known, the specific action $\mathbf{a}_t$ and past geometry $\mathbf{x}^D_{1:t}$ become redundant. Formally,

$$\mathbf{x}^R_{t+1} \perp \{\mathbf{x}^D_{1:t}, \mathbf{a}_t\} \mid \mathbf{x}^R_{1:t}, \mathbf{x}^D_{t+1}. \tag{6}$$

This simplifies the appearance term to $p(\mathbf{x}^R_{t+1} \mid \mathbf{x}^R_{1:t}, \mathbf{x}^D_{t+1})$, where the RGB generation is guided by the current depth.

Substituting these simplified terms back into Equation 4 yields our final staggered factorization:

$$p(\mathbf{x}^R_{t+1}, \mathbf{x}^D_{t+1} \mid \mathbf{x}^R_{1:t}, \mathbf{x}^D_{1:t}, \mathbf{a}_t) = \underbrace{p(\mathbf{x}^D_{t+1} \mid \mathbf{x}^R_{1:t-1}, \mathbf{x}^D_{1:t}, \mathbf{a}_t)}_{\text{Stage 1: Staggered Depth DiT}} \cdot \underbrace{p(\mathbf{x}^R_{t+1} \mid \mathbf{x}^R_{1:t}, \mathbf{x}^D_{t+1})}_{\text{Stage 2: Depth-Guided RGB DiT}}. \tag{7}$$

Depth-guided RGB generation. During inference, the two stages are executed sequentially. Stage 1 predicts $\mathbf{x}_{t+1}^D$ using action-conditioned rollout. Stage 2 then takes this predicted depth—at the *current* timestep—as an input for RGB generation:

$$\mathbf{x}_{t+1}^R = f_{\theta_R}(\mathbf{z}_\tau^R, \tau \mid \mathbf{x}_{t+1}^D, \mathbf{x}_{\leq t}^R).$$

Thus RGB can "peek" at the depth of the same future frame, yielding a depth-guided appearance model without requiring explicit action-conditioning.

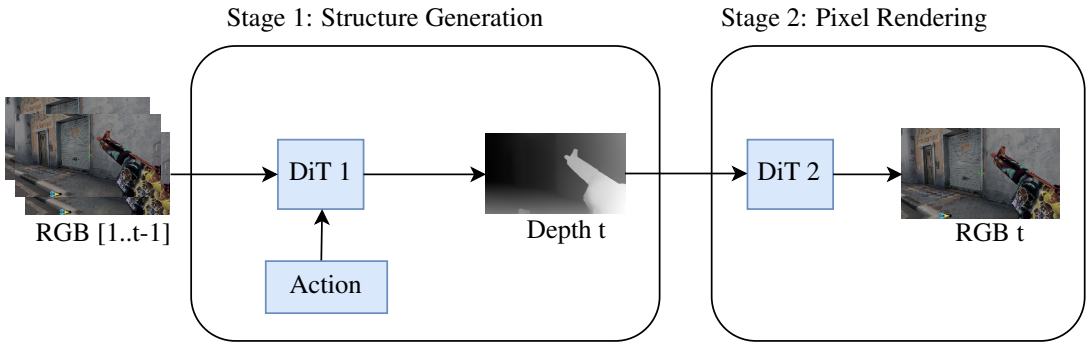

Figure 3: **Two-Stage Video Synthesis Pipeline**. Stage 1 generates depth maps from the current frame, action and history. Stage 2 renders the next frame conditioned on these history and depth with channel concatenation.

Action-free scalability. A key implication of this factorization is that only Stage 1 requires action labels. Stage 2 can be trained or fine-tuned on any dataset with RGB–depth pairs, including augmented data or higher frame rates, with depth obtained from an off-the-shelf estimator. This makes Stage 2 scalable beyond simulator data and enables significantly larger appearance models without requiring interactive trajectories.

4 Experiments

4.1 Dataset

We conduct all experiments on the CS:GO Dust 2 dataset [13], which provides large-scale first-person gameplay trajectories suitable for evaluating world models. The dataset consists of approximately 5.5M frames (∼96 hours) recorded at 16 FPS with image resolution 150×280. We apply reflection padding to obtain a uniform resolution of 160×280. We use 5M frames for training and 0.5M frames for validation. Each frame is accompanied by 51-dimensional one-hot encoded actions. As the dataset does not contain ground-truth geometry, we estimate depth using MiDaS DPT-Hybrid [15].

4.2 Model Training Setup

Our models follow the Oasis architecture [4] for both the VAE and DiT components, but are retrained to support the Dust 2 resolution and our latent tokenization. All diffusion models operate entirely in the VAE latent space of shape $(16, 28, 16)$.

VAE

The VAE is a ViT-based [8] image autoencoder with no temporal compression. Using a patch size of 10, each frame is encoded into a $16 \times 28 \times 16$ latent tensor. The VAE is trained for 800k iterations with a batch size of 24 (six sequences of four frames) using the following composite objective:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{recon}} + \beta \mathcal{L}_{\text{LPIPS}} + \gamma \, d_{\text{weight}} \mathcal{L}_{\text{GAN}} + \lambda \mathcal{L}_{\text{KL}}.$$

Here d_{weight} is the dynamic GAN weighting factor,

$$d_{\text{weight}} = \max\left(\frac{\|\nabla \mathcal{L}_{\text{GAN}}\|}{\|\nabla \mathcal{L}_{\text{NLL}}\| + 10^{-4}}, \; 10^5 \right),$$

and we adopt the same balancing scheme as [16]. Training is performed in bfloat16 and completes in eight days on a single H100 GPU.

Table 1: Comparison of validation metrics across all model variants. Evaluated on 2400 videos. RGB metrics evaluate appearance fidelity, while depth metrics evaluate geometric accuracy.

Model	RGB Metrics					Depth Metrics	
	FVD↓	FID↓	PSNR↑	LPIPS↓	SSIM↑	AbsRel↓	$\delta_{1.25}$ ↑
Diamond [1]	368.5	87.2	18.1	0.51	0.44	–	–
Vid2World [10]	106.6	17.5	18.7	0.404	0.48	–	–
RGB-Only	52.8	10.1	18.7	0.35	0.49	–	–
Joint RGBD	54.2	11.5	18.5	0.38	0.49	340.3	44.2
2-Stage	53.1	11.1	18.5	0.36	0.48	318.1	44.9

Diffusion Transformer (DiT)

Architecture. The diffusion model is a standard DiT [14] adapted for video generation. Attention is causal along the temporal dimension, and conditioning is injected via AdaLN-Zero. Actions are one-hot encoded and added to the sinusoidal timestep embedding for action-conditioned models. We use noisy ground-truth previous frames as context tokens, following common diffusion-forcing-style conditioning.

Training Procedure. Models are trained on sequences of 16 frames with a sigmoid beta noise schedule. Table 2 summarizes the main training configurations.

Inference. At inference time, we use DDIM sampling. Given initial context frames $\{\mathbf{x}_t\}_{t=0}^{t_0}$ and a future action sequence $\{\mathbf{a}_t\}_{t=t_0}^{T}$, we encode the context frames into VAE latents, and the DiT autoregressively generates future latents $\{\mathbf{z}_t\}_{t=t_0}^{T}$ consistent with the provided action sequence.

4.3 Models Evaluated

To assess the impact of joint RGB–depth modeling and our proposed two-stage formulation, we evaluate five model variants trained under comparable settings. All models use the same VAE tokenizer and DiT backbone described in Section 4. Unless otherwise stated, training uses action-conditioned diffusion forcing and identical optimization schedules.

RGB-Only Baseline (Action-Conditioned). A standard DiT trained purely on RGB latents with one-hot action conditioning. This corresponds to a conventional video diffusion world model without any geometric information.

Joint RGBD Model. A single DiT trained on concatenated RGB and depth latents. Both modalities are provided at matching timesteps and are denoised jointly. Actions are used as in the baseline model. This model tests whether depth improves predictive fidelity when fused at the token level.

Two-Stage RGB–Depth Model. Our proposed approach factorizes the model into two DiTs trained end-to-end:

- **Stage 1 (Depth DiT).** Predicts future depth conditioned on past depth, past RGB, and actions.
- **Stage 2 (RGB DiT).** Predicts RGB conditioned on past RGB and the *current* predicted depth. No action conditioning is used in Stage 2.

We simulate joint-channel diffusion forcing by providing both DiTs with matched noisy RGB and depth latents during training. We find that our 2-stage pipeline achieves competitive performance with RGB-only, and improved performance over Joint RGBD (as it uses double the number of parameters). With this, however, we are able to have coupled depth and RGB while keeping them decoupled sequentially during generation. We also display samples comparing 2-stage to joint RGBD and Vanilla in Fig.

4.4 Finetuning RGB DiT on unlabeled data

We tried take the pretrained two-stage model above, freeze the Stage 1 Depth DiT, and fine-tune only the Stage 2 RGB DiT on additional unlabeled augmented RGB data. We initially settled for augmented data (temporal, flipping, random cropping) because we did not have access to more RGB data. However, we found that this did not improve performance by much. Rather than collecting more RGB data from a game engine, which would have been plausible, we opted

instead to simulate our desired setting with a more academic point of view by simply pretraining on a % of the data and finetuning on the remainder. In Fig, we show how reducing the dataset size during training affects the validation metrics similar to the findings from GameNGen. In Fig, we show how finetuning only the RGB DiT on unlabeled RGB data can improve the quality despite the Depth DiT only trained on some % of the data. This corroborates our hypothesis (also CITE) that learning depth is more sample efficient for the model. Additionally, the training loss is also an order of magnitude lower.

5 Limitations and Future Work

Some of the limitations of our work that need to be addressed are:

- More ablations needed while varying model sizes.
- While we need less action labels, we require estimating (imperfect) depth maps or acess to ground truth depth maps which can be a big ask. Imperfect depth learnt renders the Depth DiT useless.
- We have 2 separate models and that might result in more accumulative error. There is also a distribution mismatch between the two DiTs due to the frame offset.

Some extensions of our work may include:

- Incorporating other modalities such as camera parameters, occlusion maps, and optical flow maps.
- Being able to generalize to arbitrary OOD datasets which would require a much larger scale model with a more diverse dataset.
- Scaling depth prediction using multi-view or self-supervised priors, extending our approach to real-world video with estimated camera poses, and exploring hierarchical or multi-rate variants of two-stage diffusion for long-horizon planning and embodied agents.

6 Conclusion

We present a two-stage diffusion world model that decouples geometric and appearance prediction through a staggered RGB–depth formulation. By training a depth-predictive DiT with action conditioning and an RGB-predictive DiT that is guided by the current depth but free of action inputs, we showed that the factorized model retains the expressiveness of a joint RGBD diffusion model while enabling practical advantages in scalability and data efficiency. Our formulation is theoretically grounded through its correspondence to the conditionals learned by a fully joint diffusion model under diffusion forcing, and empirically validated on the CS:GO Dust 2 dataset, where it improves RGB fidelity, depth accuracy, and long-horizon rollout stability relative to traditional baselines. A key benefit of our approach is that the RGB model can be fine-tuned on arbitrary RGB–depth data, including augmented or unlabeled sources, since it requires no action supervision. This opens the door to substantially larger training corpora and richer generative models of appearance, while maintaining geometric consistency through the depth stage.

Acknowledgements

We thank Esmeralda Whitammer for her helpful discussions on VAE training. Cameron Barker for his useful suggestions on attention masking. Danier Duolikun for his helpful advice on video diffusion modeling papers and guidance.

References

[1] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *Thirty-eighth Conference on Neural Information Processing Systems*.

[2] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2025.

[3] Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, and Tong He. Deepverse: 4d autoregressive video generation as a world model, 2025.

[4] Decart and Spruce Campbell Xinlei Chen Robert Wachen Julian Quevedo, Quinn McIntyre. Oasis: A universe in a transformer. 2024.

[5] Jiacheng Lv Jianqi Ma Jun Zhang Junyan Lv Junyou Li Minwen Deng Mingyu Yang Qiang Fu Wei Yang Wenkai Lv Yangbin Yu Yewen Wang Yonghang Guan Zhihao Hu Zhongbin Fang Zhongqian Sun Deheng Ye, Fangyun Zhou. Yan: Foundational interactive video generation. *arXiv preprint arXiv:2508.08601*, 2025.

[6] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7346–7356, October 2023.

[7] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models, 2025.

[8] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[10] Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv: 2505.14357*, 2025.

[11] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis, 2025.

[12] Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc Van Gool, and Rakesh Ranjan. Movideo: Motion-aware video generation with diffusion model. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLIV*, page 56–74, Berlin, Heidelberg, 2024. Springer-Verlag.

[13] Tim Pearce and Jun Zhu. Counter-strike deathmatch with large-scale behavioural cloning. In *2022 IEEE Conference on Games (CoG)*, page 104–111. IEEE Press, 2022.

[14] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023.

[15] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ICCV*, 2021.

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[17] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[18] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion, 2025.

[19] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *The Thirteenth International Conference on Learning Representations*, 2025.

[20] Andrii Zahorodnii. Improving world models using deep supervision with linear probes, 2025.

[21] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Learning 4d embodied world models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5337–5347, October 2025.

[22] Zhide Zhong, Haodong Yan, Junfeng Li, Xiangchen Liu, Xin Gong, Tianran Zhang, Wenxuan Song, Jiayi Chen, Xinhu Zheng, Hesheng Wang, and Haoang Li. Flowvla: Visual chain of thought-based motion reasoning for vision-language-action models, 2025.

[23] Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8535–8546, October 2025.

A Model Configuration

Table 2: Model configuration for both DiTs.

Component	**Specification**
Number of parameters	600M
Batch size	24
Sequence length	16
EMA decay	0.999
Optimizer	Adafactor (lr = 1×10^{-4})
Precision	bfloat16
Architecture	DiT
Patch size	2
GPU	NVIDIA GH200
Training time	$\sim$2 days