DECOUPLED DRIFT-DIFFUSION WORLD MODELS FOR CONSISTENT FUTURE FRAME PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

World models of visual information predict future image frames conditioned on given action sequences. This predicted scene evolution is driven by deterministic changes such as controlled motion on a robot platform, as well as stochastic changes such as rendering of previously occluded regions. Our key insight is that the spatial location of these stochastic changes is itself learnable. We leverage this insight by jointly regressing both the deterministic next-state prediction and the spatial uncertainty map, then using the uncertainty map to guide a conditional diffusion process. We propose Decoupled Drift-Diffusion World Model (D3WM), which uses a transformer trained with a Gaussian negative-log-likelihood objective to produce both the deterministic next-state prediction and a spatial map of heteroscedastic aleatoric uncertainty. The learned uncertainty map then provides a structured spatial gate for a conditional diffusion model: deterministic regions are anchored to the predicted state, while uncertain regions, such as disocclusions or dynamic textures, are sampled by the diffusion process. We frame this decomposition under a variance-gated stochastic differential equation that separates deterministic drift from stochastic diffusion. D3WM shows improved results over competitive baselines in two distinct types of action spaces: 1) trajectory planning for robotic manipulation from generated images and 2) camera movement in long-horizon novel view synthesis.

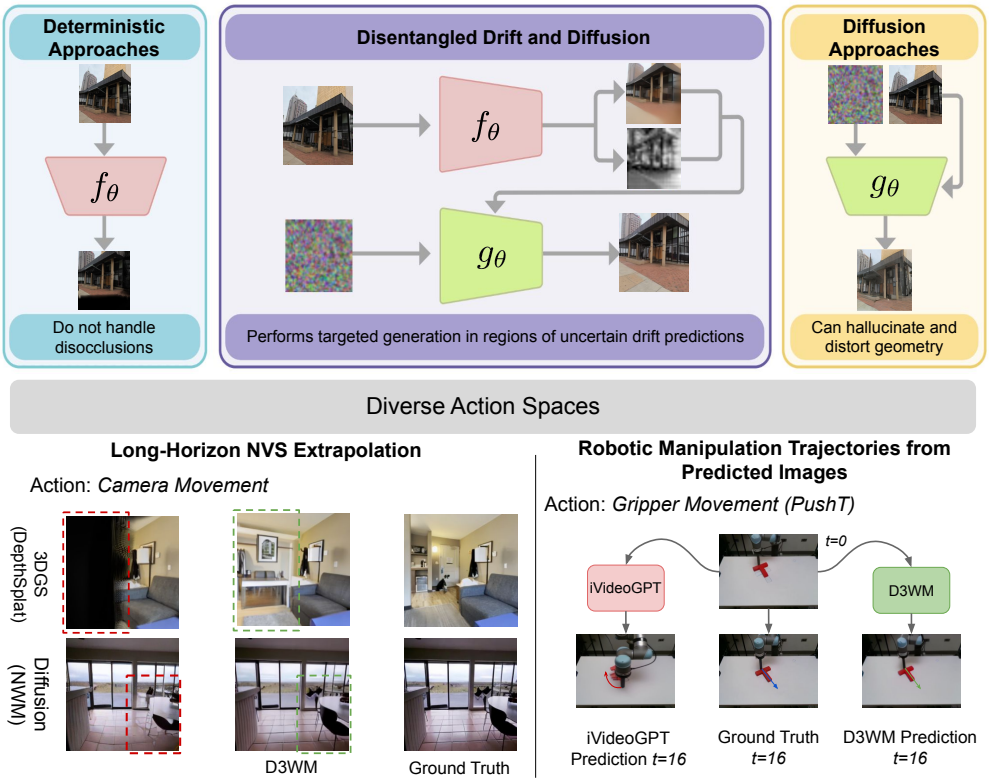
1 INTRODUCTION

Perception in humans is an active process in which the outcomes of actions are anticipated before they happen to inform decision making. From throwing a ball to jumping over a puddle, visually grounded outcomes are anticipated based on planned actions Von Holst & Mittelstaedt (1950). To advance artificial intelligence, computational world models of visual information mimic this biological action-conditioned visual prediction Zhu et al. (2025b); Zhou et al. (2025); Ali et al. (2025). A natural tension arises in predicting future image frames: how can a model consistently predict deterministic changes from given actions simultaneously with stochastic changes such as rendering of previously occluded regions?

Modern 3D modeling and generative methods address this task by optimizing either the deterministic or stochastic side of the problem at the expense of the other. Gaussian splats render 3D scenes from image frames, enabling automatic transformations from camera displacements Charatan et al. (2024); Xu et al. (2025); Chen et al. (2024b). However, projective transformations alone leave information-less gaps in previously occluded regions. Alternatively, conditional generative models leverage learned priors to synthesize plausible content for stochastic and deterministic changes Ren et al. (2025); Bar et al. (2025); Wu et al. (2024). However, they do not preserve geometric structure in deterministic transitions since these transitions are not explicitly identified.

Recent work bridges these gaps by refining Gaussian splatting novel view synthesis (NVS) results with diffusion to fill information gaps in disoccluded regions Dong et al. (2026); Zhang et al. (2025b). This conditional diffusion approach aligns with our own contributions, but these methods are directly designed to improve performance of NVS where the deterministic transformations are always perspective warps from camera motion.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



D3WM architecture handles disocclusions while preserving the structure in Long-Horizon NVS Extrapolation and provides physically consistent predictions for planning Robotic Manipulation Trajectories.

Figure 1: D3WM uses a learnable variance map to disentangle deterministic aspects (drift) from stochastic and texture generation (diffusion) achieving structurally-consistent and photorealistic action-conditioned observation prediction. In contrast, deterministic methods struggle with disocclusions in previously unseen regions, and approaches that rely solely on diffusion are prone to hallucinations and distortions of deterministic parts of the scene.

In our work, we consider predicting future image frames conditioned on arbitrary action sequences — a setting that arises whenever an agent's actions interact with a dynamic environment, as in robotic manipulation. This more general formulation must synthesize the "world model" capabilities of generative frameworks with the rigid geometric constraints of 3D modeling. Predicted states must be not only visually plausible but also physically and geometrically consistent with the underlying scene structure and the agent's intended trajectory.

Predicting future frames conditioned on arbitrary actions therefore requires learning the deterministic component of the action's effect, while identifying regions where this deterministic prediction is uncertain. Our key insight is that the spatial location of these uncertain regions is itself learnable: high aleatoric uncertainty in the deterministic prediction identifies areas requiring stochastic generation. We use this spatial uncertainty map to gate generation, retaining geometric structure where prediction is reliable while synthesizing new content where it is not.

Contributions. We introduce **Decoupled Drift-Diffusion World Model (D3WM)** to perform action-conditioned prediction of future image frames. D3WM uses a transformer to jointly regress the deterministic next-state prediction and a spatial map of heteroscedastic aleatoric uncertainty. This learned variance map then conditions a diffusion model, guiding the generation of new content toward regions of high aleatoric uncertainty while anchoring the prediction to the deterministic state elsewhere.

We evaluate D3WM n action-conditioned image prediction for robotic manipulation using the Push-T and BridgeData V2 datasets, the geometric consistency of D3WM's generated images yields improved downstream trajectory planning and candidate evaluation compared to baseline world model

architectures. We also validate D3WM on NVS extrapolation from 2-view inputs, extrapolating up to 12 frames into the future on the DL3DV benchmark with the performance gap over the 3DGS baseline widening at longer horizons and demonstrating strong out-of-distribution generalization to RealEstate10k without retraining.

2 Related Work

Conditional Generative Models. Conditional generative architectures enable controllable generation of high-dimensional visual data through explicit structural or physics control Zhang et al. (2023); Guen & Thome (2020). Many 3D-aware diffusion pipelines incorporate geometric inductive biases for view synthesis Chan et al. (2023). Explicit representations like 3D Gaussian Splatting (3DGS) impose geometric constraints in feed-forward NVS Charatan et al. (2024); Xu et al. (2025); Chen et al. (2024b;c). Recent work reconciles explicit geometry with generative hallucination by jointly optimizing splatting and diffusion for camera-displacement actions Dong et al. (2026); Zhang et al. (2025b). Similarly, we combine an estimate of the deterministic transformation with targeted diffusion-based refinement. Unlike these methods, we model arbitrary action sequences and predict the deterministic mean alongside a spatial estimate of aleatoric uncertainty, which conditions the diffusion process.

SDEs in Neural Networks. Neural Stochastic Differential Equations (SDEs) capture continuous-time dynamics by jointly learning deterministic drift and stochastic diffusion components Pedersen (1995); Tzen & Raginsky (2019); Liu et al. (2020); Li et al. (2020); Kidger et al. (2021). However, when applied to high-dimensional data challenges like latent image generation, these approaches experimentally underperform due to the computational intractability of joint pathwise variational inference Course & Nair (2023) and the failure of isotropic diffusion constraints in highly correlated spatial domains Vandersanden et al. (2024). Score-based generative models Song et al. (2021), flow matching Zhang et al. (2024), and stochastic interpolants Chen et al. (2024a) bypass joint learning entirely, relying on fixed forward noise processes reversed via score matching. In contrast, we learn a spatial variance map that controls the location and quantity of novel image features generated, yielding competitive results against large-scale diffusion baselines on standard action-conditioned prediction tasks.

World Models for Future Frame Prediction. Recent advances in generative modeling have sparked interest in action-conditioned visual world models that simulate complex physical environments. In the domain of robotics and embodied AI, models such as Navigation World Models (NWM) Bar et al. (2025), iVideoGPT Wu et al. (2024), Ctrl-World Guo et al. (2025), and Prophet-DWM Wang et al. (2025) have demonstrated that large-scale diffusion architectures can synthesize highly realistic future video frames conditioned on actions. These models act as neural simulators allowing agents to foresee visual consequences of actions in manipulation or navigation Hu et al. (2025); Guo et al. (2025); Finn & Levine (2017); Gao et al. (2024).

Physical planning in robotics requires geometric consistency to achieve desired outcomes. Efforts to enforce multi-view and spatial constraints Zhang et al. (2025a); Kim et al. (2025); Liang et al. (2025); Du et al. (2026) have steadily improved the physical plausibility of these generated environments. However, standard diffusion models map noise to pixel space without disentangling physical state, making them prone to hallucinating structural shifts or overriding action-conditioning in favor of visual priors.

We bridge this gap by explicitly separating deterministic transitions from stochastic generation. Unlike prior works that rely solely on diffusion Bar et al. (2025); Wu et al. (2024); Kim et al. (2025); Zhang et al. (2025a); Zhu et al. (2025a), our framework anchors generation to the deterministic outcome of the given actions, using aleatoric uncertainty of the deterministic prediction to gate targeted generation of new scene information. This produces the structural consistency required for reliable trajectory evaluation in downstream policy learning.

3 Decoupled Drift-Diffusion World Models

We mathematically formulate a visual prediction architecture built on several intuitive concepts. We first observe that actions yield learnable deterministic changes in a scene. For instance, shifting a

camera's viewpoint applies a strict geometric warping to already-visible background structures, just as a robotic actuator command predictably translates a rigid object. Critically, regions requiring hallucination outside of these expected changes can be identified from these deterministic transformations. In visual scenes, previously occluded areas for which there are no prior observations most commonly require this hallucination. Then, the generated information should be predicted using these targeted regions instead of broadly in the scene to preserve deterministic transformations.

We translate these observations into a mathematical formulation of latent future frame prediction. First, we use a deterministic transformer to approximate the predicted mean μ_t of the deterministic change and aleatoric variance σ_t^2 in latent space, of the next latent state z_t given a previous latent state z_0 and an action sequence $a_{[0,t)}$ for some time horizon t (Section 3.1). The predicted variance approximates the spatial heteroscedastic aleatoric uncertainty of the scene Kendall & Gal (2017), capturing regions where the deterministic future is inherently ambiguous, such as disocclusions or complex dynamic textures. This variance provides a spatial map for targeted hallucination, used both as conditioning for the diffusion model (Section 3.2) and as a spatial gate at inference time (Section 3.3).

3.1 Learning the Predicted Mean and Aleatoric Uncertainty

We model the temporal evolution of the latent state under an action sequence as a stochastic process whose deterministic component (the SDE drift coefficient in Eq. 4) we denote f_θ, and whose stochastic component is characterized by a spatial variance σ_t^2. Rather than train f_θ as an instantaneous rate and integrate it numerically, we directly regress its integrated effect over $[0, t]$, the conditional mean $\mu_t = \mathbb{E}[z_t \mid z_0, a_{[0,t)}]$, together with the aleatoric variance σ_t^2. The spatial uncertainty σ_t^2 conditions the diffusion-time refinement of Section 3.3. Throughout the paper we use "drift" to refer to the SDE drift coefficient f_θ in Eq. 4, while the network's outputs are the predicted mean μ_t and the variance map σ_t^2.

We approximate this predicted mean and aleatoric uncertainty using a deterministic transformer f_θ^μ. Given the previous latent state z_0 and action sequence $a_{[0,t)}$, the network outputs the mean and variance parameters of a spatial distribution

$$\mu_t, \sigma_t^2 = f_\theta^\mu(t, z_0, a_{[0,t)}). \tag{1}$$

Formally, we model the spatial heteroscedastic aleatoric uncertainty by assuming the ground-truth future state z_t is drawn from a Gaussian distribution parameterized by these outputs, such that we model the probability over spatial locations as a factorized Gaussian: $p(z_t \mid z_0, a_{[0,t)}) = \mathcal{N}(z_t; \mu_t, \sigma_t^2 I)$.

To ensure σ_t^2 captures this true aleatoric uncertainty, we train f_θ^μ via a negative log-likelihood (NLL) objective Kendall & Gal (2017). By treating the target latent state $z_t^{(i,j)}$ at a spatial location (i, j) as a sample from our predicted distribution, the loss becomes:

$$\mathcal{L}_{\text{drift}} = \frac{1}{2} \sum_{i,j} \left(\log \sigma_t^{2\,(i,j)} + \frac{(z_t^{(i,j)} - \mu_t^{(i,j)})^2}{\sigma_t^{2\,(i,j)}} \right). \tag{2}$$

This loss forces the network to predict high variance σ_t^2 explicitly in spatial regions where the deterministic prediction error is high. We use this variance prediction to create a learned spatial gate for the diffusion process in Section 3.2.

3.2 Conditional Diffusion Formulation

We construct a stochastic diffusion model g_θ using a ControlNet architecture Zhang et al. (2023) to sample from a μ_t- and σ_t^2-conditioned distribution over the next state $\hat{z}_t$. The deterministic prediction network f_θ^μ produces the mean μ_t and variance σ_t^2 that condition g_θ. To preserve fine textures from the input observation, g_θ is also conditioned on the reference image frame z_0 via IP-Adapter-Plus Ye et al. (2023). Texture refinement of the predicted mean is performed at inference time via the spatially-gated procedure of Section 3.3.

The diffusion model is trained to predict the noise residual ϵ via the standard conditional ϵ-prediction objective. To avoid notation collision with the temporal frame index t, let $\tau \in [0, T]$ denote the

diffusion denoising timestep, and let $z_t^{(\tau)}$ denote the noisy target latent at diffusion step τ. We train g_θ as:

$$\mathcal{L}_{\text{score}} = \mathbb{E}_{\tau, z_t, \epsilon \sim \mathcal{N}(0,I)} \left[\| \epsilon - g_\theta(z_t^{(\tau)}, \tau, \mu_t, \sigma_t^2, z_0) \|_2^2 \right]. \tag{3}$$

Minimizing this objective is equivalent to minimizing the Fisher Divergence between the data distribution and the model distribution Song et al. (2021), learning the conditional gradient field of the next-state distribution given the predicted-mean and variance conditioning.

The deterministic transformer f_θ^μ and the conditional diffusion model g_θ are trained sequentially. At inference, we sample from the learned conditional distribution using the variance-gated procedure detailed in Section 3.3, which uses σ_t^2 as a spatial gate to anchor the diffusion process to the deterministic prediction μ_t in regions of low aleatoric uncertainty.

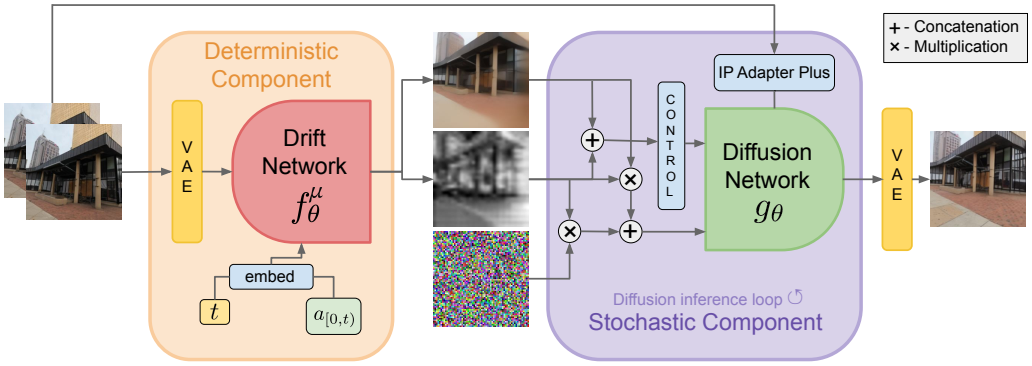


Figure 2: **Architecture of D3WM.** The input action sequence is denoted $a_{[0,t)}$, and t corresponds to the time shift into the future that the model aims to predict. For a single inference run, our deterministic prediction network f_θ^μ runs only once, while the diffusion network g_θ runs a standard diffusion inference loop using the variance-gated refinement procedure described in Eq. 6. Following Section 3.1 terminology, the deterministic prediction network (f_θ^μ) is shown as 'Drift Network' for figure compactness.

3.3 Variance-Gated Inference

We formulate the transition to the next future state as a continuous-time stochastic differential equation (SDE) over the observation temporal axis t. The dynamics of the scene can be modeled as:

$$\mathrm{d}z_t = f_\theta(t, z_0, a_{[0,t)})\mathrm{d}t + \sigma_t \odot \mathrm{d}w_t, \tag{4}$$

where f_θ is the deterministic drift coefficient, $\mathrm{d}w_t$ is a standard Wiener process representing physical stochasticity, and σ_t is the spatial standard deviation map that scales the uncertainty element-wise. Eq. 4 characterizes the data-generating process, and the network f_θ^μ from Section 3.1 regresses its integrated mean μ_t and aleatoric variance σ_t^2 directly rather than numerically integrating f_θ.

A direct numerical integration of this SDE via Euler-Maruyama would sample w_t from an isotropic Gaussian. However, isotropic Gaussian noise produces physically implausible artifacts (e.g., pixelated static in disoccluded regions) rather than realistic textures. We therefore replace the stochastic term with the trained conditional diffusion model g_θ from Section 3.2, which acts as a learned, image-manifold-aware sampler for the next-state distribution conditional on μ_t and σ_t^2.

Concretely, we sample the conditional next-state distribution $p(z_t \mid z_0, a_{[0,t)})$ via the standard Variance-Preserving DDIM reverse process, with a single modification: at each reverse diffusion step we replace the diffusion model's noisy latent with a spatially-weighted combination of the model's prediction and a forward-noised version of the deterministic prediction μ_t. Define the per-pixel latent trust map

$$\gamma_t = (\mathbf{1} - \sigma_t^2) \odot \lambda_\tau, \tag{5}$$

where $\lambda_\tau \in [0, 1]$ is a diffusion-time-dependent anchoring schedule. At reverse step τ, the conditional latent is

$$\tilde{z}_t^{(\tau)} = \gamma_t \odot \tilde{\mu}_t^{(\tau)} + (\mathbf{1} - \gamma_t) \odot z_t^{(\tau)}, \qquad \tilde{\mu}_t^{(\tau)} = \sqrt{\bar{\alpha}_\tau}\, \mu_t + \sqrt{1 - \bar{\alpha}_\tau}\, \xi_\tau, \tag{6}$$

where $\xi_\tau \sim \mathcal{N}(0, I)$ and $\bar{\alpha}_\tau$ is the cumulative noise schedule of the underlying VP diffusion process. Both terms are valid noisy latents at noise level τ, ensuring $\tilde{z}_t^{(\tau)}$ remains compatible with the next reverse step. We then proceed with the standard reverse diffusion update using $\tilde{z}_t^{(\tau)}$ in place of $z_t^{(\tau)}$. The full SDE-level derivation is provided in the supplementary material.

This formulation enforces our core objective in two limits. In regions of low aleatoric uncertainty ($\sigma_t^2 \to 0$), the trust map γ_t approaches λ_τ and the conditional latent is dominated by the noisy deterministic prediction $\tilde{\mu}_t^{(\tau)}$; the final $\hat{z}_t$ collapses toward μ_t, preserving deterministic geometric structure. In regions of high structural ambiguity ($\sigma_t^2 \to 1$), γ_t approaches zero and the conditional latent equals the diffusion model's prediction, allowing the diffusion process to synthesize plausible content. The final output of the reverse process at $\tau = 0$ yields our final predicted state $\hat{z}_t$.

4 Experiments

D3WM is designed to perform action-conditioned future image frame prediction of stochastic processes. To experimentally validate this capability, we consider two significant types of actions to model: 1) robotic manipulation (Sections 4.1) and 4.2) 2) camera movement (Section 4.3). In the case of camera movement, we perform 2-view novel view synthesis (NVS) on the DL3DV benchmark Ling et al. (2024) and show long horizon extrapolation capabilities of new frames. We also show learning geometric structure can improve generalization of this approach by testing on RealEstate10k Zhou et al. (2018) which we did not train on. For robotic manipulation, we predict image frames after sequences of manipulation actions and use those predicted frames as input to a trajectory generation policy trained on the Real PushT dataset Chi et al. (2023; 2024) and BridgeData V2 (Walke et al., 2023).

4.1 Robotic Manipulation Trajectories from Predicted Images on Real PushT

Experiment Design. We design an experiment with robotic manipulation data to evaluate the effectiveness of our method in producing generated images useful for robot control on Real PushT dataset. We use one fixed mounted camera view facing the tabletop and one wrist camera attached to the robot arm as the two input camera views to our network together with trajectories of action sequences to be executed by the robot arm. We predict a frame at the end of this trajectory from each camera view and use it to extend the manipulation planning horizon prior to executing actions for a trained policy by using model's generated frame as input to a trajectory generating policy to generate a new trajectory. We compare the Average Displacement Error (ADE) and Final Displacement Error (FDE) between generated trajectories using a generated image and the ground truth image.

Training on Real PushT. We train D3WM on Real PushT dataset Chi et al. (2023; 2024), with the deterministic network outputting 4 latents per inference: the predicted-mean and variance maps for both the side and wrist camera views. We allow up to 16 actions as conditioning input, where each action vector concatenates gripper world pose, arm rotation, and time shift.

Evaluation on Real PushT. When designing our prediction architecture, we claimed that the physical and geometric consistency our model enforces is critical to successful use of generated images in robotic control. To test this, we select 100 frames at random from each of the 14 episodes in the test split of the Real PushT dataset. For each of these frames, we run D3WM to predict 16 frames into the future conditioned on the corresponding actions. We then use our predicted frames as input to the pre-trained PushT Diffusion Policy Chi et al. (2023; 2024) to generate a trajectory. We do this setup on the same randomly selected frames using iVideoGPT Wu et al. (2024), which was trained on a subset of Open X-Embodiment Collaboration et al. (2024) dataset including the Real PushT dataset. We compute the Average Displacement Error (ADE) and Final Displacement Error (FDE) between trajectories and the trajectory generated by running PushT Diffusion Policy on the ground truth image.

Trajectory Generation Results. As shown in Table 1, we significantly outperform iVideoGPT on trajectory error on both ADE and FDE. This result highlights that our predictions are much more useful for robotic planning and control. The qualitative results in Figure 3 show that iVideoGPT distorts and hallucinates geometry, resulting in physically impossible gripper and object locations

as well as distortions of the rigid T object being pushed. In contrast, D3WM accurately models deterministic changes including physically consistent placements of the gripper and object.

Method	ADE ↓	FDE ↓
Ground Truth	1.552	1.319
iVideoGPT	45.134	65.228
D3WM	**12.864**	**17.550**

Table 1: Quantitative result comparing ADE and FDE of the manipulation trajectories generated from image predicted by D3WM and iVideoGPT. "Ground Truth" is the average AVE and FDE error from the stochastic generation of the trajectory when run twice in a row on the same ground truth frame.

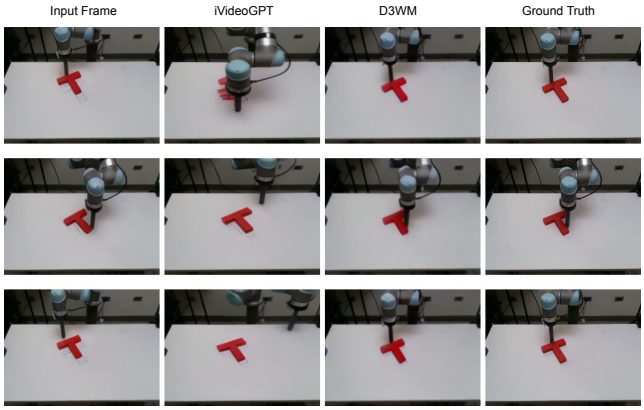


Figure 3: Qualitative comparison of predicting 16 frames into the future on Real PushT dataset. D3WM achieves high accuracy in terms of the gripper and "T"-object placement, while iVideoGPT suffers from hallucinated gripper movement and object deformations.

4.2 BridgeData V2

We evaluate D3WM on BridgeData V2 (Walke et al., 2023), which is a tabletop manipulation by a robotic arm, with fixed camera observations and 7-DoF actions. We compare against two generative world models for this dataset, IRASim (Zhu et al., 2025c) and WorldGym (Quevedo et al., 2026).

Open-loop rollouts. Table 2 shows both D3WM deterministic-only (μ_t) and full D3WM variants taking the top two positions among world models on PSNR, and the lead growing with the horizon, where D3WM μ_t is $+1.4$ to $+1.6$ dB over IRASim and up to $+3.6$ dB over the copy-frame control, with the full model second throughout. On LPIPS, D3WM μ_t leads every model from $g{=}3$ and is the only one to pass the copy control at $g{=}4$ (0.112 against 0.121). D3WM μ_t produces mean blur which is favored by PSNR, as shown in Fig. 4. Full D3WM model, in contrast, samples sharp outcomes. Whether the sharpness is worth what it costs in PSNR is not decidable from Table 2 alone.

Goal-image candidate verification. At each state, K candidate action chunks are ranked by predicted latent similarity to the true $t{+}4$ frame, and the selected chunk is scored against the demonstrated action. Random-of-K and oracle-of-K bracket every row, so a verifier is read by the fraction of that gap it captures (Table 3). D3WM is the best verifier in the table, beating IRASim at every pool size, 38–40% of the random-to-oracle gap closed against 23–26%, at half the compute per candidate. WorldGym gets worse as the pool grows ($0.044 \rightarrow 0.063$), and it under-commands, so candidates that command little motion are the ones it scores best. The D3WM μ_t row is weak because an averaged prediction is close to every candidate's outcome and discriminates between none of them, though at 4.9 ms per candidate it remains the only variant cheap enough to screen at control rate.

Policy consistency. We measure ADE and FDE exactly as in the real Push-T experiment (§4.1), using Octo policy (Octo Model Team, 2024). Table 4 shows that the stochastic component is worth a large margin over the deterministic alone (1.063 against 1.445 ADE), which is the same sharpness that cost it PSNR now being paid back by a policy that was never trained on blurred input. D3WM (full) is second behind IRASim, and the gap is structural rather than a deficit of fidelity, since IRASim diffuses the whole trajectory and can enforce consistency across the horizon, where D3WM, like WorldGym, predicts a single frame and has no mechanism tying independently predicted futures together, traded-off by its $2\times$ speedup compared to IRASim.

	PSNR↑				LPIPS↓			
gap g	1	2	3	4	1	2	3	4
copy-frame (control)	24.48	21.48	20.04	19.19	**.051**	**.084**	**.105**	.121
IRASim	23.34	22.37	21.64	21.12	.101	.116	.127	.136
WorldGym	23.80	21.10	19.70	18.86	.059	.093	.118	.137
D3WM μ_t	**24.73**	**23.54**	**23.09**	**22.76**	.086	.099	.107	**.112**
D3WM	24.29	22.80	22.32	21.91	.104	.118	.124	.130

Table 2: Quantitative on BridgeData V2. The two D3WM variants hold the top two model positions on PSNR at every gap, by a margin that grows with the horizon. Both metrics here are reference-based and therefore reward a hedged prediction, where the D3WM deterministic prediction leads partly because it blurs, and the full D3WM model, which samples a sharp completion, is penalized for missing the one future that occurred (Fig. 4). The copy-frame control wins on LPIPS because scenes are dominated by the mostly static background, passed only at $g{=}4$ by D3WM deterministic prediction.

Verifier	ms/cand	K@1s	selected-action MSE at K		
			4	16	32
random	—	—	.040	.038	.038
oracle	—	—	.006	.003	.003
IRASim	5478	0.2	.031	.029	.030
D3WM μ_t	**4.9**	**204**	.054	.043	.038
D3WM	2643	0.4	**.027**	**.024**	**.022**
WorldGym	1094	0.9	.044	.056	.063

Table 3: Goal-image candidate verification. Each world model ranks K candidate action chunks by predicted similarity to the true $t{+}4$ frame, and the selected chunk is scored against the demonstrated action. Random-of-K and oracle-of-K bracket every row. The D3WM deterministic prediction discriminates weakly dues to mean blur, and D3WM full method wins due to sharp and accurate sampling.

world model	ADE↓	FDE↓
copy-frame (control)	1.577	1.994
WorldGym	1.448	1.783
D3WM μ_t	1.445	1.649
D3WM	1.063	1.193
IRASim	**0.846**	**1.035**

Table 4: Policy consistency. Frozen Octo policy acts on the predicted and on the real future under paired sampling keys, so a perfect simulator scores 0. The D3WM full method is worth a large margin over D3WM deterministic-only prediction, but IRASim leads since it diffuses the whole trajectory and can enforce consistency across the horizon, where D3WM and WorldGym predict single frames.

4.3 NVS Extrapolation

DL3DV Benchmark Experiment Design. We perform a two-view long-horizon NVS extrapolation task, where we seek to predict an observation conditioned on the action of camera movement. We train our model with 2-view input, conditioned on the action of camera movement on DL3DV dataset Ling et al. (2024). We include camera instrinsics together with the input views, matching the formulation in NVS baseline methods. We evaluate on a subset of DL3DV Benchmark, predicting the observation 3, 6, 9 and 12 frames into the future. The DL3DV Benchmark contains 140 episodes, with hundreds of frames each. For our subset we randomly sample 20 frames and associated poses for each episode, for each time shift 3, 6, 9 and 12. We compare the performance of D3WM to the contemporary 3DGS state-of-the-art approach DepthSplat Xu et al. (2025), using the same two reference frames.

Ablations. We ablate the components of D3WM at $t = 3, 6, 9, 12$ frame extrapolation on the DL3DV Benchmark dataset (Figure 6).The *Deterministic Only* baseline (μ_t alone) achieves the highest PSNR/SSIM by producing locally-blurred predictions in disocclusion regions, but also the worst LPIPS. PSNR rewards averaging over plausible outcomes since per-pixel deviation remains bounded, while LPIPS penalizes the lack of high-frequency structural detail this averaging produces. Diffusion-based variants commit to specific high-frequency textures, improving perceptual realism at the cost of pixel-wise deviation when committing incorrectly. We evaluate on both metrics throughout. Variance gating ablations are discussed in App. §A.

Extrapol.	Method	PSNR ↑	SSIM ↑	LPIPS ↓
3 fr.	DepthSplat	20.8678	**0.7610**	**0.1997**
	D3WM	**21.4845**	0.6690	0.2218
6 fr.	DepthSplat	17.3864	**0.6431**	0.2975
	D3WM	**20.1228**	0.6092	**0.2829**
9 fr.	DepthSplat	15.3978	0.5565	0.3706
	D3WM	**19.1892**	**0.5629**	**0.3378**
12 fr.	DepthSplat	14.1356	0.4955	0.4208
	D3WM	**18.3521**	**0.5245**	**0.3824**

Table 5: Comparison to DepthSplat on DL3DV in-distribution Benchmark.

Extrapol.	Method	PSNR ↑	SSIM ↑	LPIPS ↓
3 fr.	NWM	18.4498	0.6318	0.2334
	D3WM	**21.5518**	**0.7901**	**0.1415**
6 fr.	NWM	15.8612	0.5541	0.3145
	D3WM	**19.2070**	**0.6191**	**0.1891**
9 fr.	NWM	14.4502	0.5086	0.3750
	D3WM	**17.6283**	**0.5561**	**0.2369**
12 fr.	NWM	13.5672	0.4795	0.4219
	D3WM	**16.7798**	**0.5223**	**0.2658**

Table 6: Comparison to NWM on out-of-distribution RealEstate10K.

DL3DV Benchmark Results. As shown in Table 5, D3WM outperforms DepthSplat across all extrapolation horizons. We outperform on PSNR at every horizon, on LPIPS for 6 frames or further, and on all three metrics (PSNR, SSIM, LPIPS) at 9 and 12 frames. The performance gap widens with horizon: DepthSplat's reliance on warping pixels from the input views breaks down as the camera moves further from the input viewpoint, leaving expanding "black holes" in disoccluded regions where no input pixels can be projected (Figure 5). D3WM's diffusion network handles these regions cleanly, while the variance-gated bridge anchors the prediction to the deterministic projection elsewhere. The remaining SSIM gap at short horizons (3 frames) reflects the same Gaussian-splatting advantage observed in our ablations: in regions where geometric warping is sufficient, splatting derives pixels directly from input frames, producing very high SSIM scores. D3WM's diffusion-synthesized textures look natural but deviate at the pixel level.

Out-of-Distribution Generalization to RealEstate10k. We also perform a 2-view long-horizon NVS extrapolation experiment on an out-of-distribution dataset. We use a subset of RealEstate10K dataset Zhou et al. (2018) for evaluation, from which we select 80 episodes and randomly sample 20 frames and associated poses for each episode. We compare our performance to that of Navigation World Models (NWM) Bar et al. (2025) on the same out-of-distribution evaluation dataset, using the same reference and target frames. Our method was only trained on data from DL3DV.

We outperform NWM on out-of-distribution NVS extrapolation in terms of PSNR, SSIM and LPIPS for all time shifts (3, 6, 9 and 12), as shown in Table 6. Although diffusion approaches like NWM are very good at high-resolution details, they often distort structure and hallucinate artifacts, as can be seen in our qualitative results in the supplement. The structural anchoring provided by D3WM's variance-gated diffusion process produces both higher PSNR and lower LPIPS than NWM's free-form diffusion approach, demonstrating that the architectural separation of deterministic dynamics from stochastic generation generalizes beyond the training distribution.

5 Conclusion

We introduce Decoupled Drift-Diffusion World Model (D3WM), a novel method for action-conditioned future frame prediction. D3WM jointly regresses the deterministic next-state prediction and a spatial map of heteroscedastic aleatoric uncertainty, then uses this uncertainty map to gate generation by a conditional diffusion model. We validate D3WM on novel view synthesis with camera-motion actions and on robotic manipulation prediction, outperforming strong baselines on both. The geometric consistency of D3WM's predictions translates to improved downstream trajectory planning and out-of-distribution generalization to RealEstate10k.

Limitations and Future Work. While D3WM can predict arbitrary horizons into the future, it produces only a single frame at each query. This is efficient when only the trajectory endpoint matters (e.g., the result of a single robotic manipulation), but limiting when assessing the viability of intermediate states. Our formulation also assumes scene changes are driven entirely by the input action sequence, and does not account for predictable agent-independent dynamics. Finally, while we input two reference views, the architecture is not restricted to two, hence we plan to experiment with varying view counts and to extend the targeted-generation framework beyond RGB images.

ETHICS STATEMENT

This work studies pretrained generative models and public datasets of robot manipulation and indoor scenes. It involves no human subjects, no personally identifying data and no new data collection. The audit we describe is intended to make the failure modes of action-conditioned generators easier to detect before deployment.

REPRODUCIBILITY STATEMENT

The Method section and Appendix give provide the full architectural and evaluation details. Core implementation of all training, inference and evaluation components is provided in the supplement.

AI USE STATEMENT

Generative AI tools were used for manuscript drafting and copy-editing, for producing explanatory diagrams from the authors' own figure specifications, and for searching and verifying related-work citations, as well as polishing code for the release. They were not used for research ideation, experimental design, or analysis; the probing harness, models, experiments and all reported numbers are the authors' own work. The authors reviewed all AI-assisted text and figures, verified every number against the experiment logs that produced it, and take responsibility for the final content of this work.

REFERENCES

Arslan Ali et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. *CVPR*, 2025.

Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3D-aware diffusion models. In *CVPR*, 2023.

David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, 2024.

Yifan Chen, Mark Goldstein, Mengjian Hua, Michael Samuel Albergo, Nicholas Matthew Boffi, and Eric Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and föllmer processes. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6728–6756. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/chen24n.html.

Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. *ECCV*, 2024b.

Yuedong Chen, Chuanxia Zheng, Haofei Xu, Bohan Zhuang, Andrea Vedaldi, Tat-Jen Cham, and Jianfei Cai. Mvsplat360: Feed-forward 360 scene synthesis from sparse views. 2024c.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2024.

Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. *ICRA*, 2024.

Kevin Course and Prasanth Nair. Amortized reparametrization: Efficient and scalable variational inference for latent sdes. *Advances in Neural Information Processing Systems*, 36:78296–78318, 2023.

Yitong Dong, Qi Zhang, Minchao Jiang, Zhiqiang Wu, Qingnan Fan, Ying Feng, Huaqi Zhang, Hujun Bao, and Guofeng Zhang. One-shot refiner: Boosting feed-forward novel view synthesis via one-step diffusion. *arXiv preprint arXiv:2601.14161*, 2026.

Hongyang Du, Junjie Ye, Xiaoyan Cong, others, and Yue Wang. Videogpa: Distilling geometry priors for 3d-consistent video generation. *arXiv preprint arXiv:2601.23286*, 2026.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2786–2793. IEEE, 2017.

Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability, 2024. URL https://arxiv.org/abs/2405.17398.

Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11474–11484, 2020.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *International Conference on Machine Learning (ICML)*, 2025.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, volume 30, 2017.

Patrick Kidger, James Foster, Xuechen Chen Li, and Terry Lyons. Efficient and accurate gradients for neural sdes. *Advances in Neural Information Processing Systems*, 34:18747–18761, 2021.

Byungjun Kim, Taeksoo Kim, Junyoung Lee, and Hanbyul Joo. Dexterous world models. *arXiv preprint arXiv:2512.17907*, 2025.

Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3870–3882. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/li20i.html.

Dingkang Liang, Dingyuan Zhang, Xin Zhou, Sifan Tu, Tianrui Feng, Xiaofan Li, Yumeng Zhang, Mingyang Du, Xiao Tan, and Xiang Bai. Unifuture: A 4d driving world model for future generation and perception. *arXiv preprint arXiv:2503.13587*, 2025.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, 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.

Yingru Liu, Yucheng Xing, Xuewen Yang, Xin Wang, Jing Shi, Di Jin, and Zhaoyue Chen. Learning continuous-time dynamics by stochastic differential networks. *arXiv preprint arXiv:2006.06145*, 2020.

Octo Model Team. Octo: An open-source generalist robot policy. In *Robotics: Science and Systems (RSS)*, 2024.

Asger Roer Pedersen. Consistency and asymptotic normality of an approximate maximum likelihood estimator for discretely observed diffusion processes. *Bernoulli*, pp. 257–279, 1995.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. In *ICLR*, 2026. URL https://arxiv.org/abs/2506.00613.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Belinda Tzen and Maxim Raginsky. Neural stochastic differential equations: Deep latent gaussian models in the diffusion limit. *arXiv preprint arXiv:1905.09883*, 2019.

Jente Vandersanden, Sascha Holl, Xingchang Huang, and Gurprit Singh. Edge-preserving noise for diffusion models, 2024.

Erich Von Holst and Horst Mittelstaedt. Das reafferenzprinzip: Wechselwirkungen zwischen zentralnervensystem und peripherie. *Naturwissenschaften*, 37(20):464–476, 1950.

Homer Walke, Kevin Black, Abraham Lee, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

Xiaofeng Wang et al. Prophetdwm: A driving world model for rolling out future actions and videos. *arXiv preprint arXiv:2505.18650*, 2025.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. In *Advances in Neural Information Processing Systems*, 2024.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. In *CVPR*, 2025.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Josh Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *CVPR*, 2025a. URL https://arxiv.org/abs/2412.01821.

Xi Nicole Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modelling. *Advances in Neural Information Processing Systems*, 37:107198–107224, 2024.

Xiang Zhang, Yang Zhang, Lukas Mehl, Markus Gross, and Christopher Schroers. High-fidelity novel view synthesis via splatting-guided diffusion. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025b.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. In *International Conference on Machine Learning (ICML)*, 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*, 2025a.

Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *Proceedings of Robotics: Science and Systems (RSS)*, 2025b.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *ICCV*, pp. 9834–9844, 2025c.

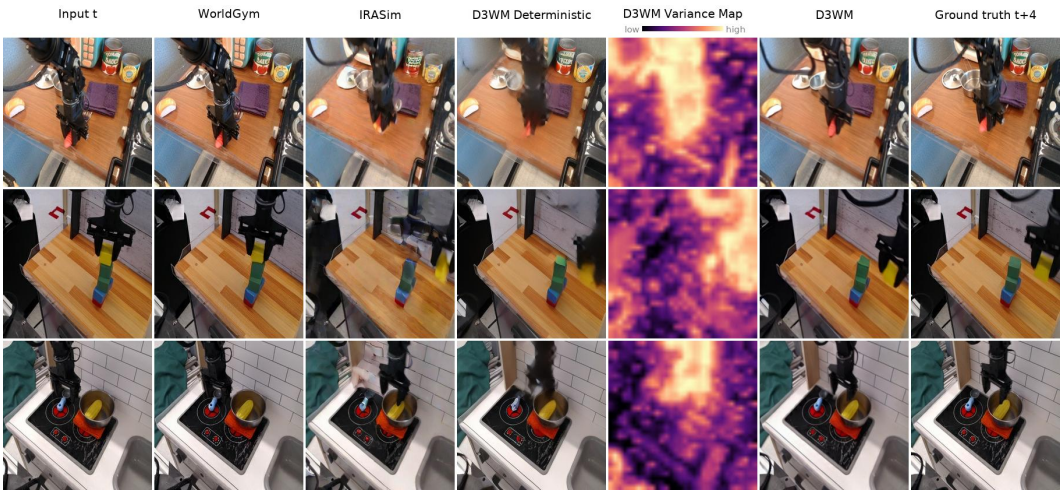


Figure 4: Qualitative results on BridgeData V2. The D3WM deterministic prediction hedges and blurs, which is what earns it the pixel metrics in Table 2, while the D3WM full model samples a sharp completion and is scored against the single future that model committed to. WorldGym undercommands and IRASim shows occasional hallucinations or wrong motion.

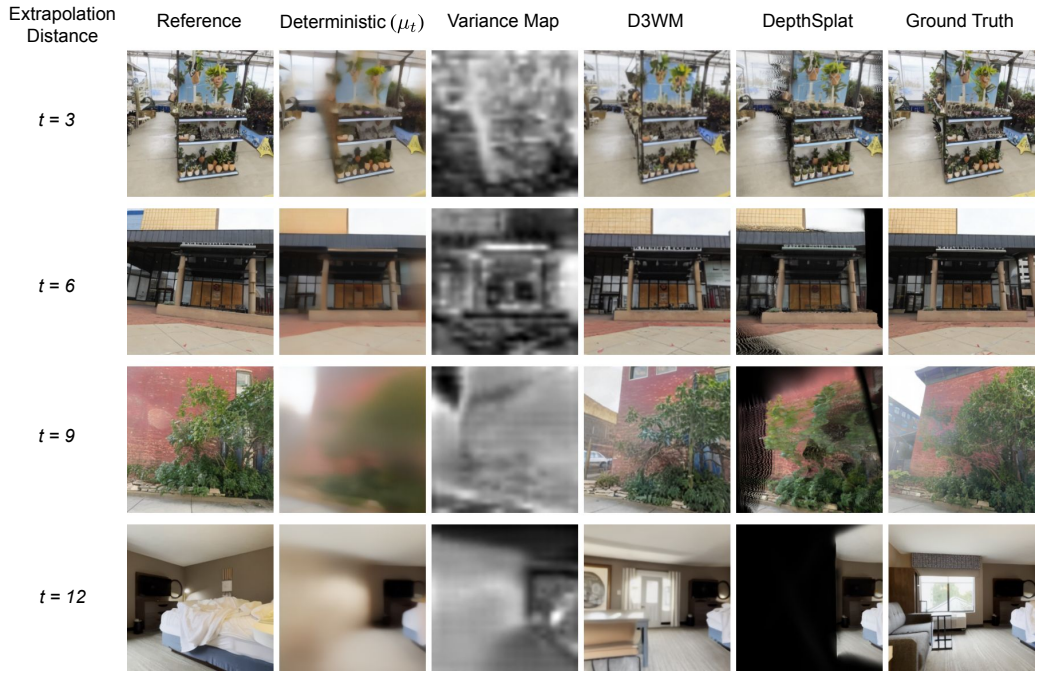


Figure 5: Qualitative comparison of 2-view NVS extrapolation between D3WM and DepthSplat, predicting 3, 6, 9, and 12 frames into the future. D3WM identifies regions of high variance (Variance Map) and uses the diffusion network to reconstruct textures and generate previously occluded regions, while preserving the geometric correctness provided by the deterministic prediction. DepthSplat produces "black holes" in regions that were previously occluded.

A Expanded Ablation Results

Variance Gate Ablations. Here we describe the full list of ablations shown in Figure 5 and Table 7. Disabling all variance-aware components (*No Variance*) recovers a standard ControlNet baseline that improves LPIPS but loses structural fidelity. Variance used only as ControlNet conditioning without inference-time gating (*Condition Only*) underperforms both, indicating that variance is

most useful as an inference-time spatial gate rather than as a training-time conditioning signal alone. The D3WM Full method, combining variance conditioning, attention bias, and the spatially-gated bridge of Eq. 6, achieves the best LPIPS and competitive structural metrics. Variance gating without ControlNet conditioning (*Gate Only*) recovers most of the structural benefit of the full method, suggesting the inference-time spatial gate is the dominant contribution and the conditioning + attention bias provide perceptual polish on top. Binarizing the variance map (*Binary Gate*) improves LPIPS at the cost of structural fidelity, suggesting the continuous variance values are calibrated meaningfully rather than just as a hard occlusion mask. Replacing the noisy-bridge gate of Eq. 6 with a clean-state posterior-mean formulation (*Clean State Blend*) drops PSNR by approximately 1.97 dB averaged across extrapolation horizons and reverts to the *Condition Only* regime, confirming that the specific choice of where to apply the variance gate during the reverse diffusion process is empirically consequential. The main observation is that the inference-time variance gate (*Gate Only*) recovers most of the structural benefit of the full method, while combining it with the variance-conditioned ControlNet and attention bias (*D3WM Full*) provides additional perceptual quality (LPIPS).

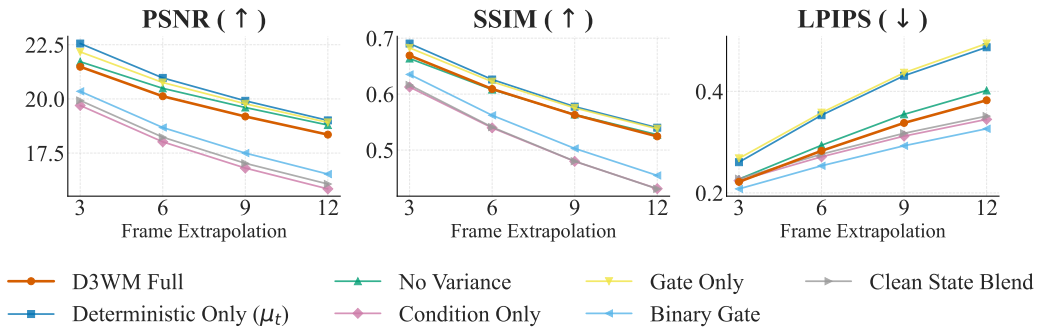


Figure 6: Ablations of D3WM method. The *D3WM Full* provides the best balance of structural similarity, signal fidelity and perceptual realism.

Schedule Variants. In our final configuration we use a step-function schedule for λ_τ in Eq. 12 that disables anchoring after the first 70% of denoising steps. We compared this against a quarter-cosine schedule (smooth decay over all steps), linear ramp, and constant anchoring across all steps. The stepfunction performed comparably to the cosine schedule and outperformed both the linear and constant variants. The constant schedule preserved structural fidelity (high SSIM) but produced blurry textures (worse LPIPS), indicating that the diffusion model needs unanchored final steps to develop coherent texture detail. The linear ramp was the worst of the four because it most strongly anchors at the first denoising step, where the diffusion model's contribution is most informative. We do not include these variants in the main paper as their effect was small relative to the design choices reported in Figure 3.

Variance Map Shape Variants. We also investigated whether modifying the dynamic range of the variance map changes downstream quality. Raising the variance map to power 0.5 before use widens high-uncertainty regions (favouring more diffusion-driven generation), while raising to power 2.0 shrinks them. Both modifications produced small changes to metrics consistent with the corresponding shift in the trust threshold, suggesting the learned variance values are calibrated meaningfully and robust to monotonic transformations.

B NVS Results Discussion

PSNR-LPIPS-SSIM Across Horizons. The DepthSplat comparison in Table 1 of the main paper shows that D3WM dominates on PSNR and LPIPS at 6-frame extrapolation and beyond, and on all three metrics at 9 and 12 frames. The persisting SSIM gap at the shortest 3-frame horizon reflects a fundamental property of the methods rather than a weakness of D3WM: Gaussian splatting is a rendering technique that derives output pixels directly from input frames where geometric warping is sufficient, so when the field of view overlap is large (small extrapolation distance), splatting produces near-perfect pixel reconstruction in non-disoccluded regions and very high SSIM scores. As extrapolation distance grows, disocclusion volume grows with it; DepthSplat then leaves expanding "black holes" (Figures 7, 8, 9, 10) where no input pixels can be projected. The squared-error penalty

Extrapolation	Ablation	PSNR ↑	SSIM ↑	LPIPS ↓
3 frames	**D3WM Full**	21.4845	0.6690	0.2218
	Deterministic Only (μ_t)	22.5551	0.6904	0.2605
	Condition Only	19.6972	0.6127	0.2242
	Gate Only	22.1707	0.6827	0.2680
	Binary Gate	20.3500	0.6352	0.2079
	Clean State Blend	19.9284	0.6162	0.2279
6 frames	**D3WM Full**	20.1228	0.6092	0.2829
	Deterministic Only (μ_t)	20.9693	0.6261	0.3530
	Condition Only	18.0184	0.5397	0.2711
	Gate Only	20.7563	0.6222	0.3580
	Binary Gate	18.6787	0.5621	0.2534
	Clean State Blend	18.2324	0.5410	0.2763
9 frames	**D3WM Full**	19.1892	0.5629	0.3378
	Deterministic Only (μ_t)	19.9124	0.5776	0.4307
	Condition Only	16.8081	0.4798	0.3118
	Gate Only	19.7741	0.5753	0.4366
	Binary Gate	17.5000	0.5029	0.2928
	Clean State Blend	17.0277	0.4801	0.3174
12 frames	**D3WM Full**	18.3521	0.5245	0.3824
	Deterministic Only (μ_t)	19.0078	0.5400	0.4868
	Condition Only	15.8471	0.4310	0.3448
	Gate Only	18.9141	0.5381	0.4941
	Binary Gate	16.5306	0.4546	0.3263
	Clean State Blend	16.0728	0.4306	0.3512

Table 7: Ablations of D3WM method. The *D3WM Full* provides the best balance of structural similarity, signal fidelity and perceptual realism.

in PSNR strictly punishes these massive deviations, which is why DepthSplat's PSNR drops sharply with horizon while D3WM's diffusion network handles disoccluded regions cleanly.

Deterministic Only vs Diffusion-Based Predictions. The ablation table (Section A) shows that the *Deterministic Only* baseline (drift transformer alone) achieves the highest PSNR/SSIM scores but the worst LPIPS. This is the same fundamental metric tension affecting the DepthSplat comparison, but expressed within our own architecture. The deterministic prediction μ_t averages over all plausible outcomes in regions of high aleatoric uncertainty, producing locally-blurred predictions that preserve average pixel intensity. PSNR's squared-error penalty rewards this averaging because the per-pixel deviation from any single plausible outcome remains bounded, while LPIPS heavily penalizes the lack of high-frequency structural detail this blur produces. Our full method intentionally trades a small amount of PSNR/SSIM for significantly improved perceptual realism (LPIPS), with the sharp-but-occasionally-incorrect textures synthesized by diffusion in high-variance regions. The visualizations (Figures 7, 8, 9, 10) make this tradeoff visible: drift-only outputs look identifiably blurry in disocclusion regions, while D3WM's outputs look natural.

Expanded Qualitative Comparison to NWM. D3WM produces a significant improvement in the visual consistency of generated frames compared to NWM. Figures 11, 12, 13, and 14 show additional examples of hallucinations and structure distortions produced by NWM compared to D3WM. These results support our motivation for grounding predictions by modeling deterministic processes first, and applying generative diffusion only in spatial areas of high stochasticity. The structural anchoring provided by the variance-gated bridge produces both higher PSNR and lower LPIPS than NWM's free-form diffusion approach across all four horizons, even though RealEstate10k was not used during training.

C PushT Results Discussion

In the context of autonomous robots, particularly for precise manipulation tasks such as PushT, accurate prediction of the deterministic outcomes of actions is critical. Although Real PushT scenes

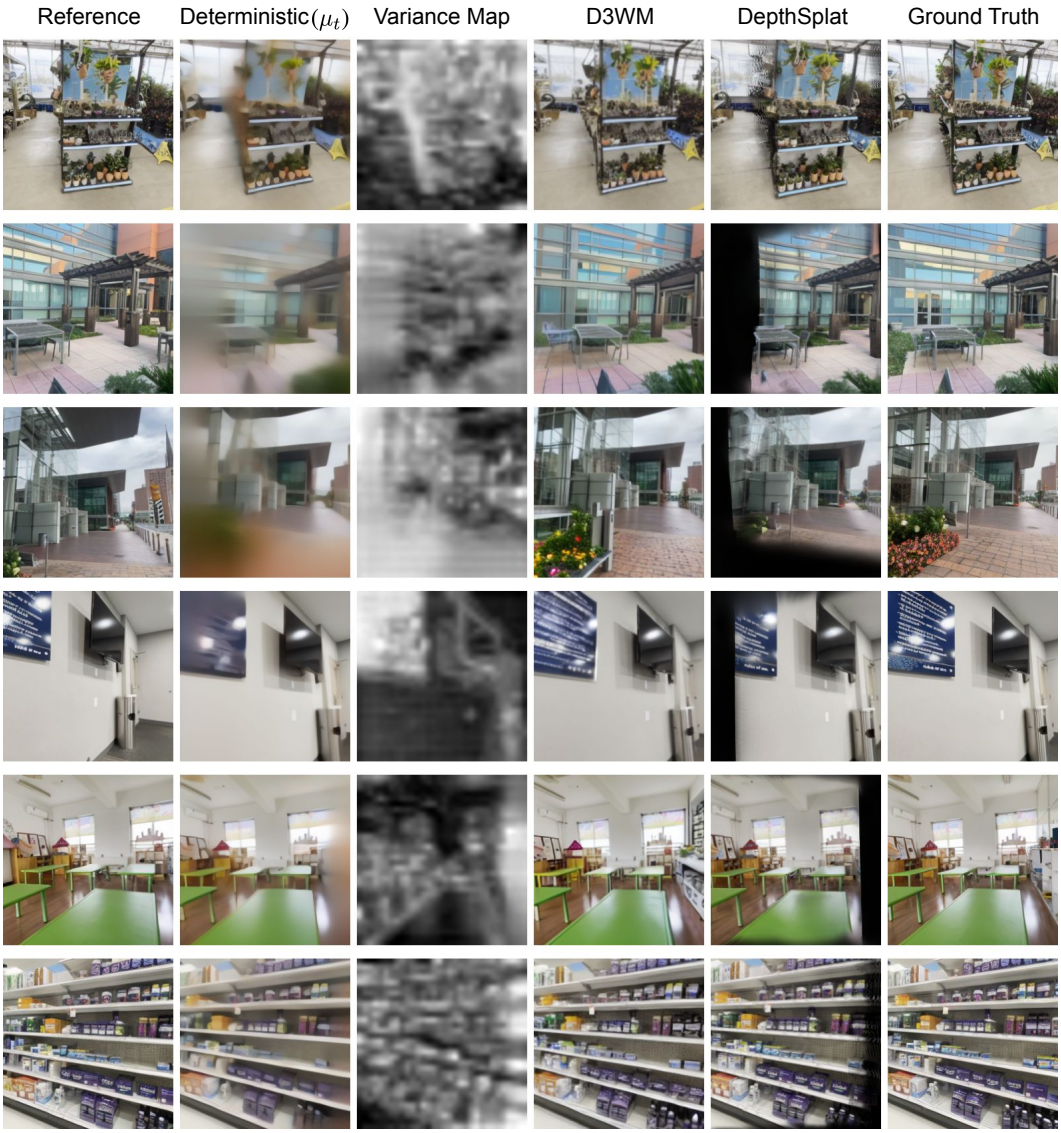


Figure 7: Additional results for qualitative comparison between D3WM and DepthSplat on DL3DV Benchmark dataset with extrapolation time shift $t = 3$ frames into the future. Highlighting how D3WM handles the "black hole" problem as compared to 3DGS approaches while retaining structural consistency of the rest of scene.

naturally exhibit partial occlusions and inherent stochasticity, robust robotic planning fundamentally depends on the deterministic consequences of taken actions. Consequently, approaches that rely exclusively on diffusion-based generation, such as iVideoGPT, frequently struggle in this domain. As demonstrated by the quantitative and qualitative results in Section 4.2 of the main paper, these purely generative models are highly susceptible to visual distortions and physically implausible hallucinations, which ultimately undermine downstream planning.

By contrast, as illustrated in Figures 15 and 16 , D3WM effectively overcomes these limitations by accurately modeling the deterministic outcomes of manipulation actions. This stability is largely driven by our deterministic image prediction (μ_t) and variance map, which anchors the generation process. Furthermore, the deterministic transformer within D3WM successfully anticipates the complex appearances and rigid-body movements of the "T" object from the wrist camera's perspective.

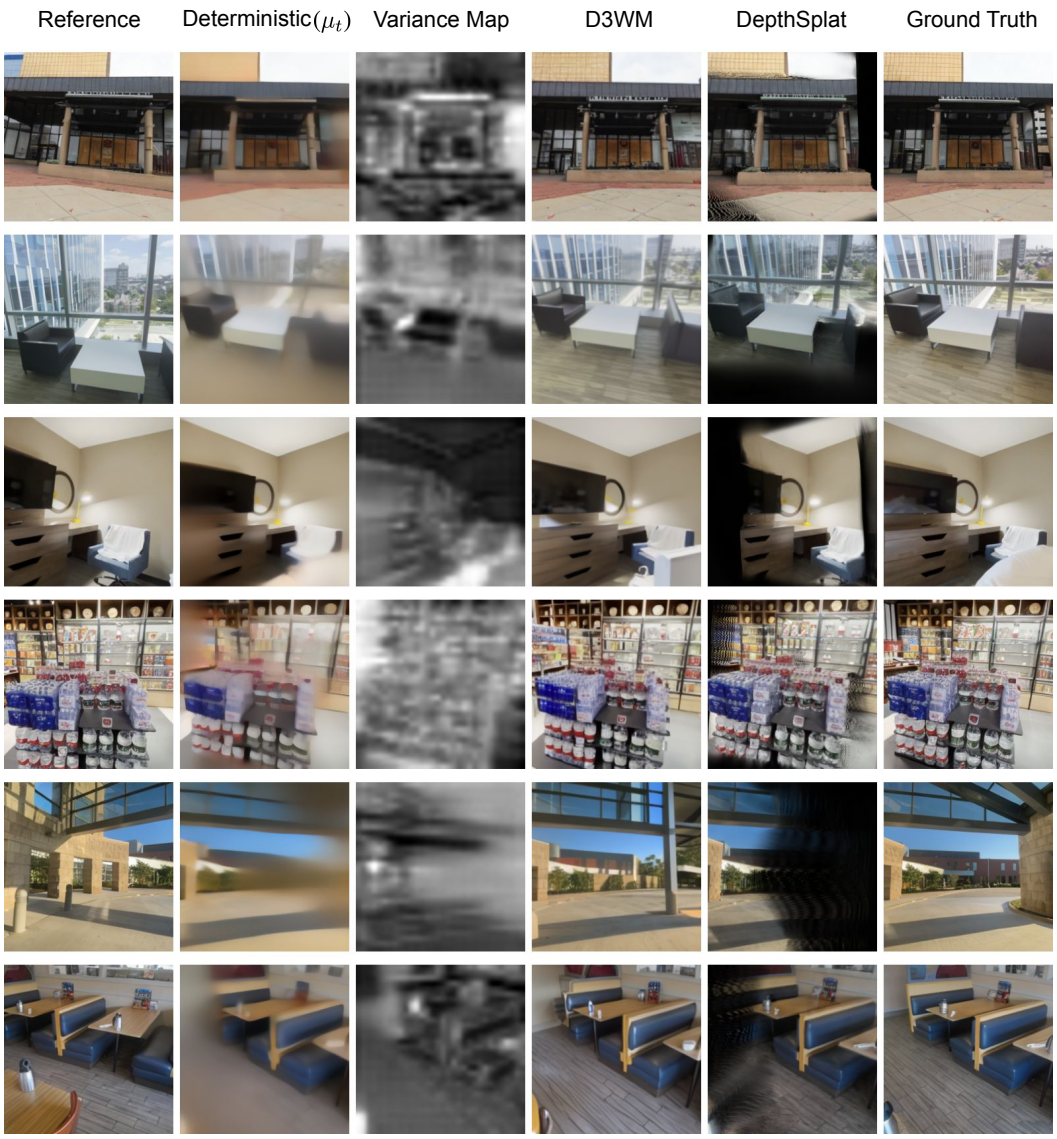


Figure 8: Additional results for qualitative comparison between D3WM and DepthSplat on DL3DV Benchmark dataset with extrapolation time shift $t = 6$ frames into the future. Highlighting how D3WM handles the "black hole" problem as compared to 3DGS approaches while retaining structural consistency of the rest of scene.

By predicting this local wrist view concurrently with the side camera view, our model ensures that the resulting action-conditioned changes are actively informed by and highly consistent with the broader spatial context of the scene from different camera views.

D BridgeData V2 Results Discussion

One small difference in the architecture of the stochastic component for BridgeData V2 is worth mentioning. Dues to the mostly static background in the dataset, we found that injecting reference frame into the ControlNet worked better via reference-attention rather than IP-Adapter. Additional qualitative results on BridgeData V2 dataset can be found in Figures 17 and 17. Consistent with our observations, D3WM sometimes samples not the most precise stochastic outcomes in terms of

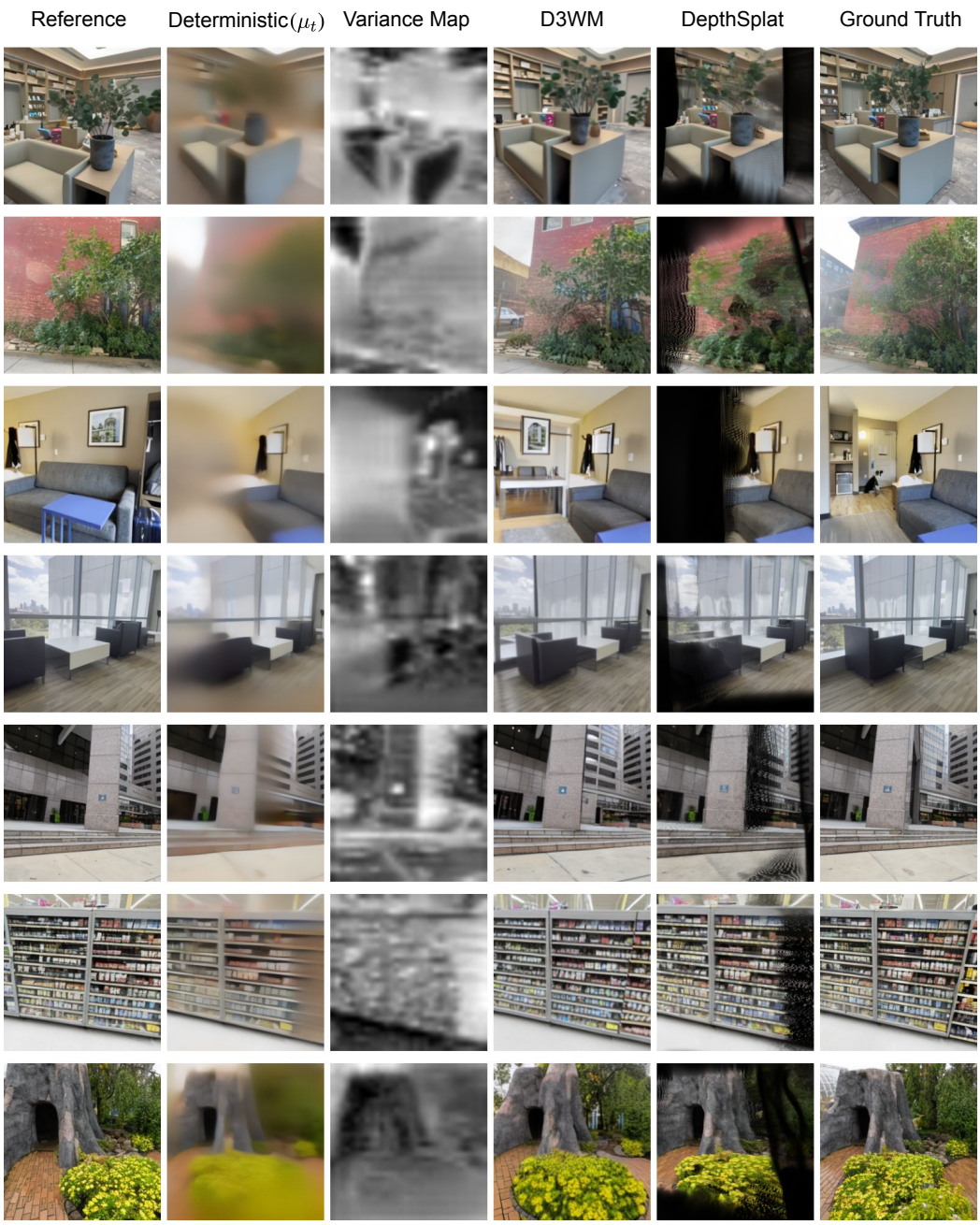


Figure 9: Additional results for qualitative comparison between D3WM and DepthSplat on DL3DV Benchmark dataset with extrapolation time shift $t = 9$ frames into the future. Highlighting how D3WM handles the "black hole" problem as compared to 3DGS approaches while retaining structural consistency of the rest of scene.

end-effector shape and object colors and textures, but preserves the generally correct motion dues to the deterministic prediction grounding. IRASim, on the other hand, frequently shows incorrect motion or distorts the objects' shapes and the background. WorldGym does generally much better in terms of consistency and sharpness, but realizes only a fraction of the commanded motion, and sometimes shows incorrect motion as well.

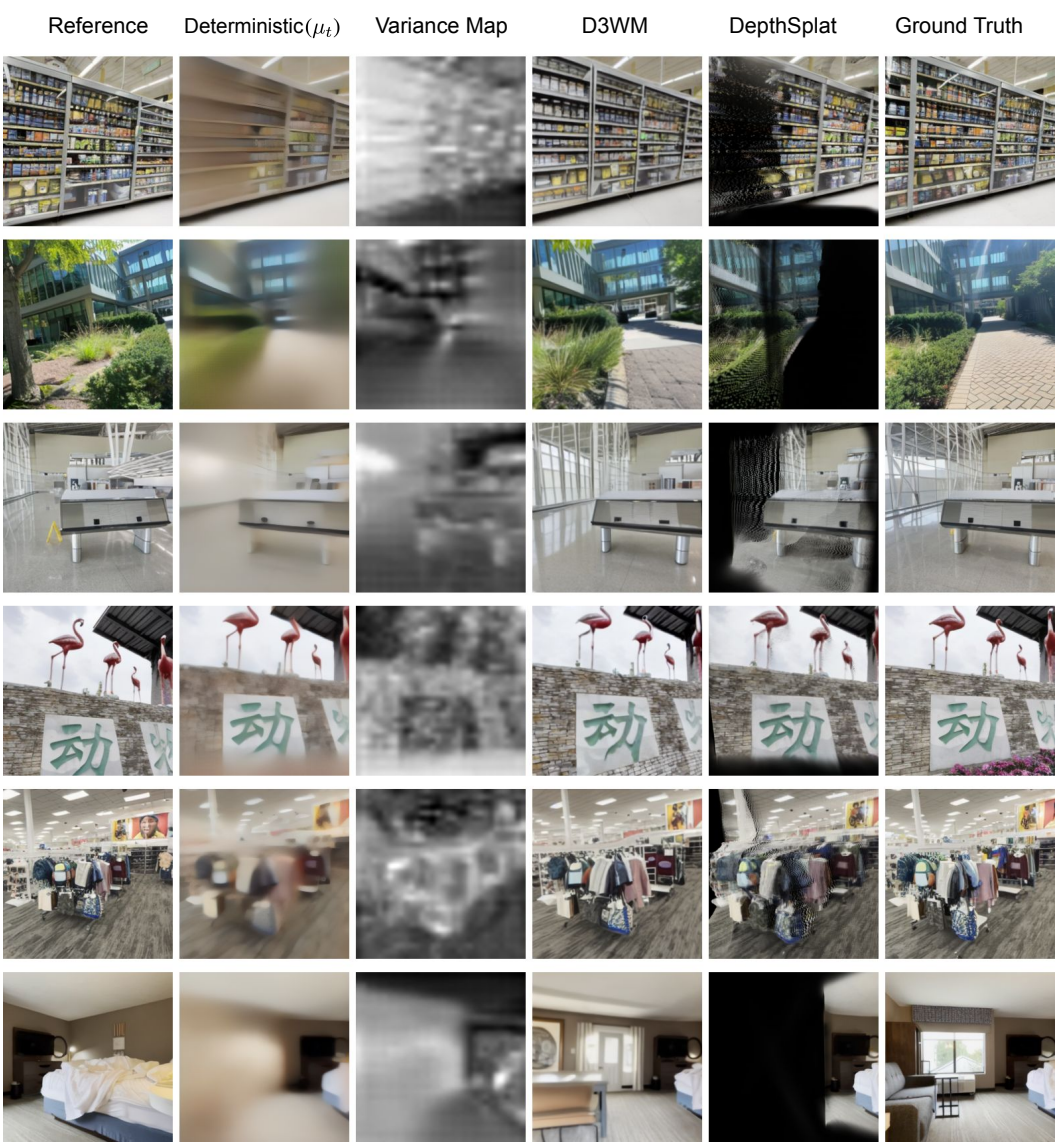


Figure 10: Additional results for qualitative comparison between D3WM and DepthSplat on DL3DV Benchmark dataset with extrapolation time shift $t = 12$ frames into the future. Highlighting how D3WM handles the "black hole" problem as compared to 3DGS approaches while retaining structural consistency of the rest of scene.

E Derivation of the Variance-Gated Inference Update

In this section, we derive the variance-gated bridge inference of Section 3.3 of the main paper. We work entirely in the Variance-Preserving (VP) regime that matches our underlying diffusion backbone (Stable Diffusion 1.5). We first present the noisy-bridge formulation used by D3WM, then contrast it with a clean-state posterior-mean alternative which we ablate in the main paper as *Clean State Blend*.

Let t denote the temporal frame index of the world model, and $\tau \in [0, T]$ denote the diffusion noise level. We sample the next state z_t given the deterministic prediction μ_t and aleatoric variance σ_t^2 produced by our transformer.



Figure 11: Additional results for qualitative comparison between D3WM and NWM on out of distribution RealEstate10K dataset subset with extrapolation time shift $t = 3$ frames into the future. Highlighting how D3WM controls and handles the hallucinations that are common for diffusion-only approaches.

E.1 Variance-Preserving Forward and Reverse SDE

The VP forward diffusion process over τ is

$$\mathrm{d}z_t^{(\tau)} = -\tfrac{1}{2}\beta(\tau)\, z_t^{(\tau)}\,\mathrm{d}\tau + \sqrt{\beta(\tau)}\,\mathrm{d}w_\tau, \tag{7}$$

with closed-form marginal distribution

$$z_t^{(\tau)} = \sqrt{\bar{\alpha}_\tau}\, z_t^{(0)} + \sqrt{1-\bar{\alpha}_\tau}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{8}$$

where $\bar{\alpha}_\tau = \exp\big(-\int_0^\tau \beta(s)\,\mathrm{d}s\big)$ is the cumulative noise schedule. The corresponding reverse-time SDE that generates samples from noise is

$$\mathrm{d}z_t^{(\tau)} = \Big[-\tfrac{1}{2}\beta(\tau)\, z_t^{(\tau)} - \beta(\tau)\,\nabla_z \log p_\tau(z_t^{(\tau)})\Big]\mathrm{d}\tau + \sqrt{\beta(\tau)}\,\mathrm{d}\bar{w}_\tau. \tag{9}$$

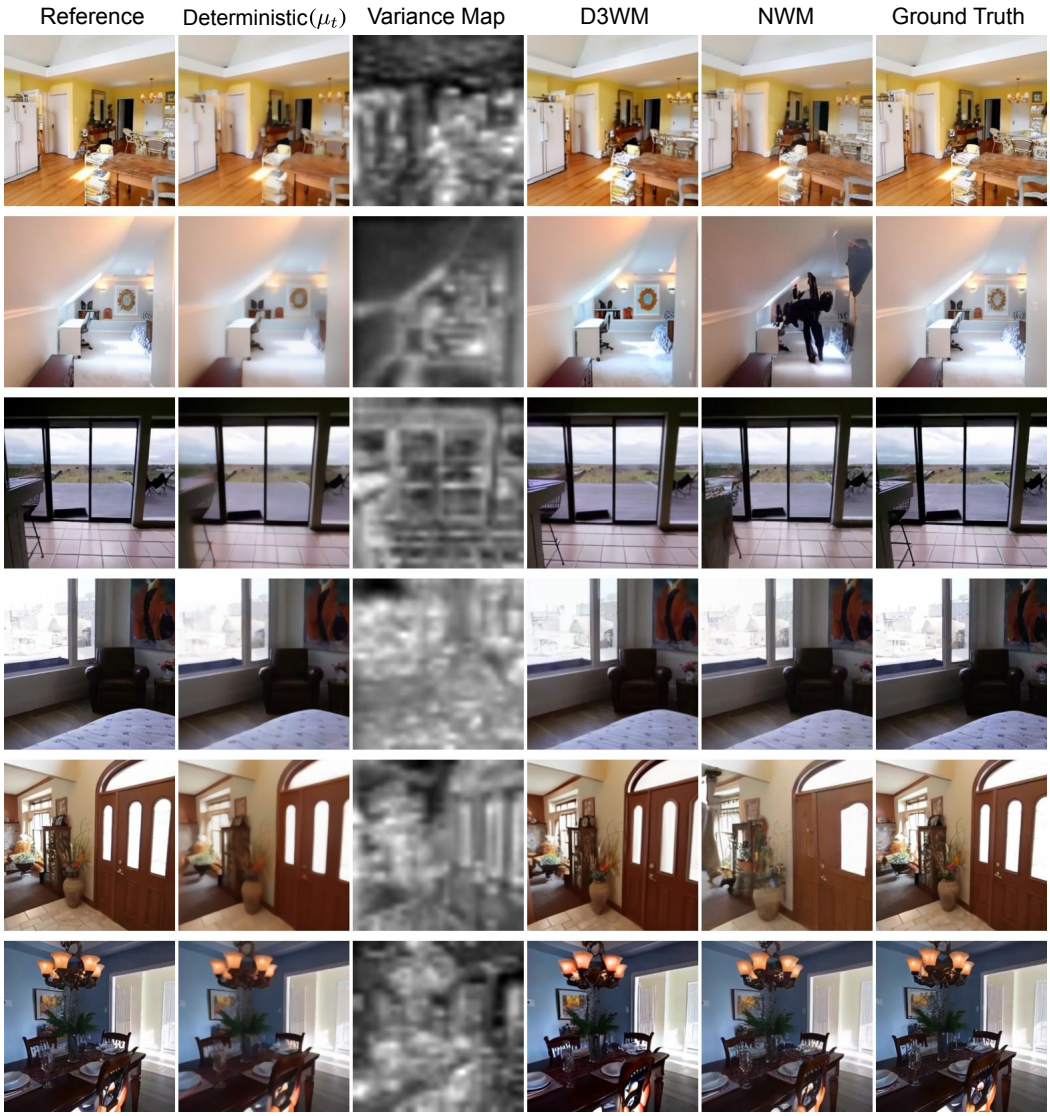


Figure 12: Additional results for qualitative comparison between D3WM and NWM on out of distribution RealEstate10K dataset subset with extrapolation time shift $t = 6$ frames into the future. Highlighting how D3WM controls and handles the hallucinations that are common for diffusion-only approaches.

Tweedie's formula for the VP marginal gives the score in terms of the conditional clean-state expectation,

$$\nabla_z \log p_\tau(z_t^{(\tau)}) = \frac{\sqrt{\bar{\alpha}_\tau}\,\mathbb{E}[z_t^{(0)} \mid z_t^{(\tau)}] - z_t^{(\tau)}}{1 - \bar{\alpha}_\tau}. \tag{10}$$

We omit the t-subscript on the score function for clarity; all quantities pertain to the world-model timestep t being predicted.

E.2 Deterministic Prior from the Transformer Network

The deterministic prediction network f_θ^μ provides a Gaussian prior over the clean target state given the input observation z_0 and action sequence $a_{[0,t)}$:

$$p(z_t^{(0)} \mid z_0, a_{[0,t)}) = \mathcal{N}\big(z_t^{(0)};\, \mu_t,\, \sigma_t^2 \odot I\big), \tag{11}$$

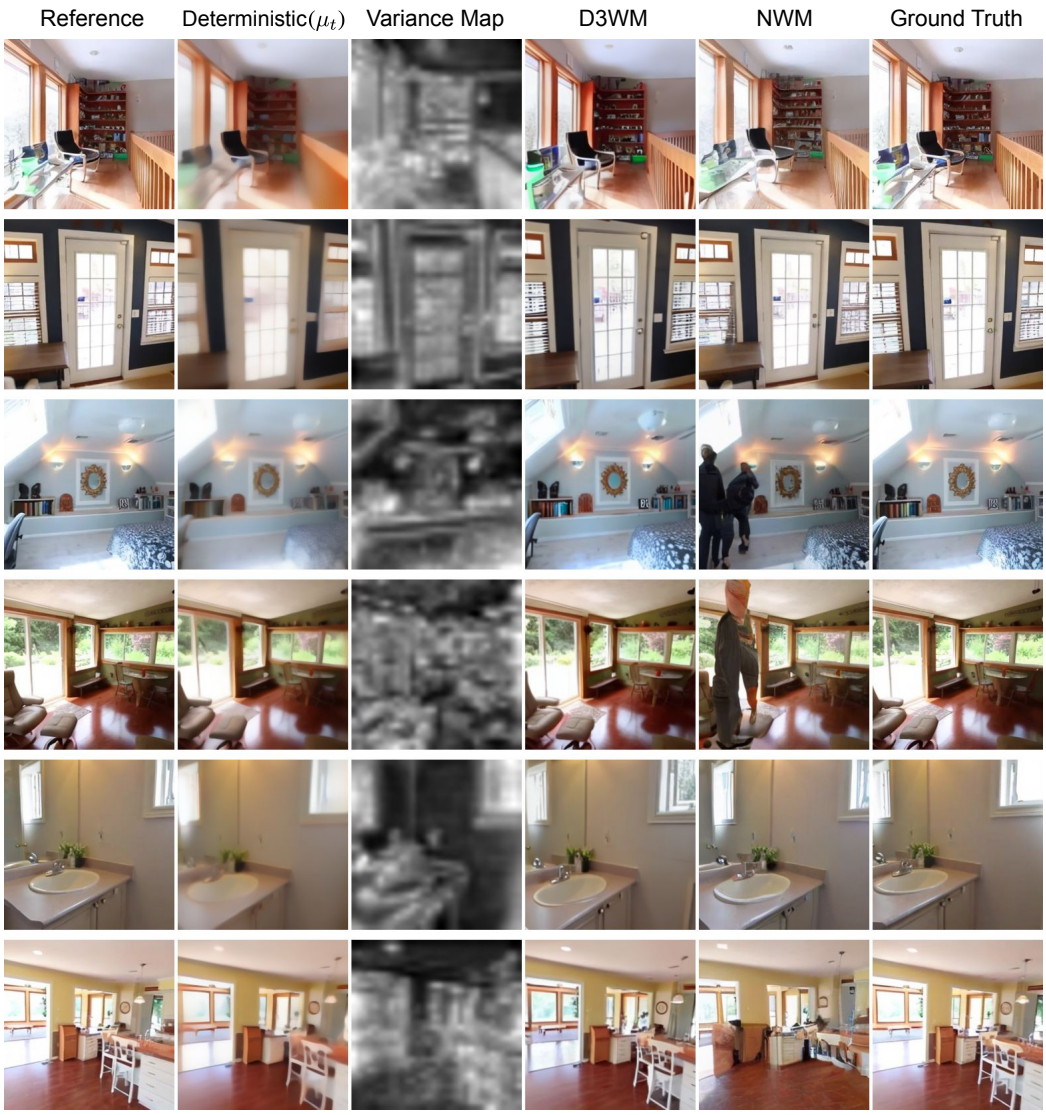


Figure 13: Additional results for qualitative comparison between D3WM and NWM on out of distribution RealEstate10K dataset subset with extrapolation time shift $t = 9$ frames into the future. Highlighting how D3WM controls and handles the hallucinations that are common for diffusion-only approaches.

with spatial diagonal variance $\sigma_t^2 \in (0, 1]^{C \times h \times w}$. This prior captures the heteroscedastic aleatoric uncertainty learned by f_θ^μ: low variance in regions where the action's effect is deterministic (e.g., visible structures undergoing predictable warping) and high variance in regions where it is stochastic (e.g., disocclusions).

E.3 Noisy-Bridge Inference (D3WM Main Method)

Pure unconditional diffusion samples from the unconditional latent prior $p(z_t^{(0)})$. Pure deterministic prediction returns μ_t. We construct an intermediate process that respects the deterministic prior strongly where σ_t^2 is small, while leaving the diffusion model free to act where σ_t^2 is large.

Spatially-gated forward bridge. We define the per-pixel trust map

$$\gamma_t \;=\; (\mathbf{1} - \sigma_t^2) \odot \lambda_\tau, \tag{12}$$

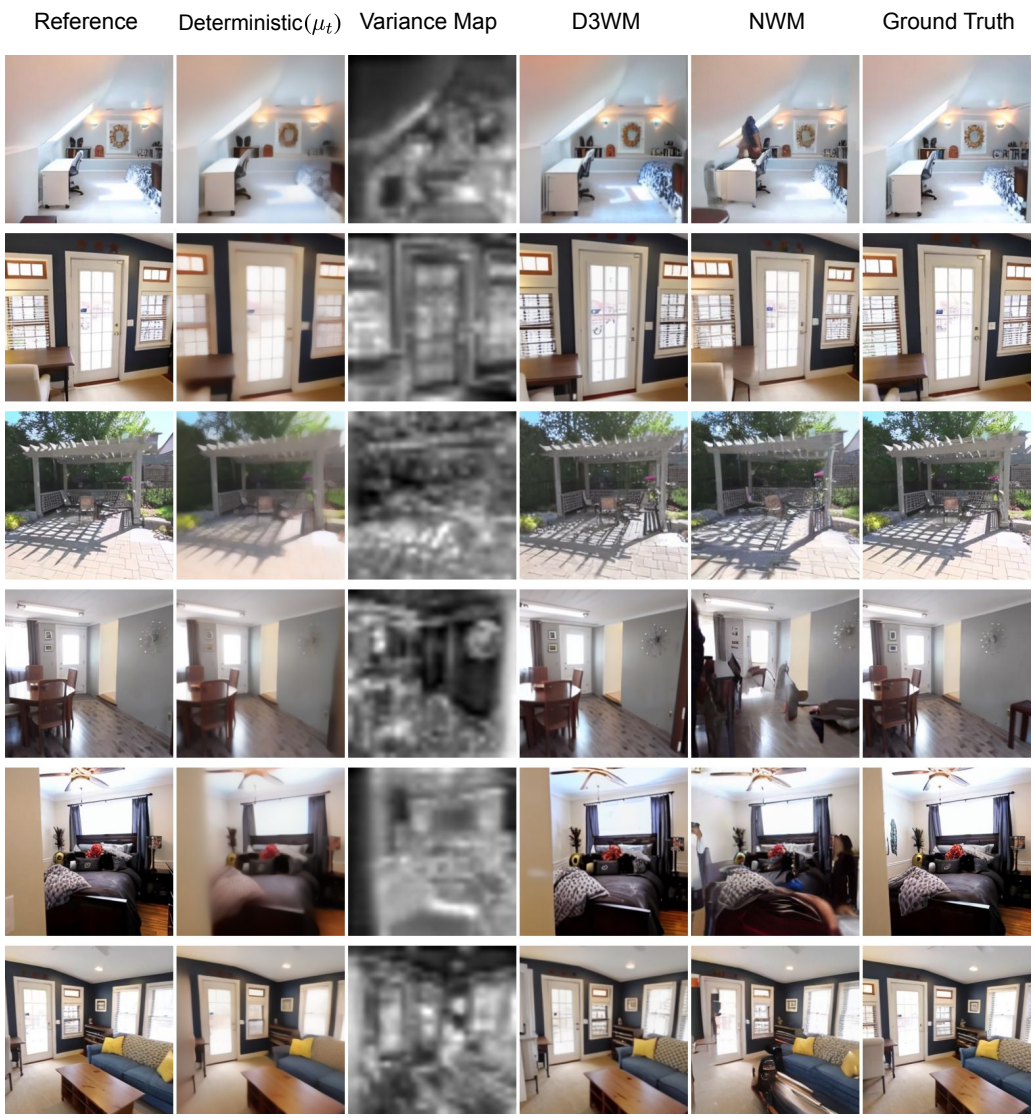


Figure 14: Additional results for qualitative comparison between D3WM and NWM on out of distribution RealEstate10K dataset subset with extrapolation time shift $t = 12$ frames into the future. Highlighting how D3WM controls and handles the hallucinations that are common for diffusion-only approaches.

where $\lambda_\tau \in [0, 1]$ is a diffusion-time-dependent anchoring schedule controlling the overall strength of deterministic anchoring. At each reverse step, we replace the diffusion model's noisy latent $z_t^{(\tau)}$ with a per-pixel convex combination

$$\tilde{z}_t^{(\tau)} \;=\; \gamma_t \odot \tilde{\mu}_t^{(\tau)} + (\mathbf{1} - \gamma_t) \odot z_t^{(\tau)}, \tag{13}$$

where $\tilde{\mu}_t^{(\tau)}$ is the deterministic prior mean transported to the noise level of step τ via the same VP forward marginal as the data,

$$\tilde{\mu}_t^{(\tau)} \;=\; \sqrt{\bar{\alpha}_\tau}\, \mu_t + \sqrt{1 - \bar{\alpha}_\tau}\, \xi_\tau, \qquad \xi_\tau \sim \mathcal{N}(0, I). \tag{14}$$

Because both terms in the convex combination are valid noisy latents at level τ (Eq. 8), the blended latent $\tilde{z}_t^{(\tau)}$ also lies on the noise-level manifold expected by the next reverse step. We then proceed with the standard VP reverse update (Eq. 9) using $\tilde{z}_t^{(\tau)}$ in place of $z_t^{(\tau)}$.

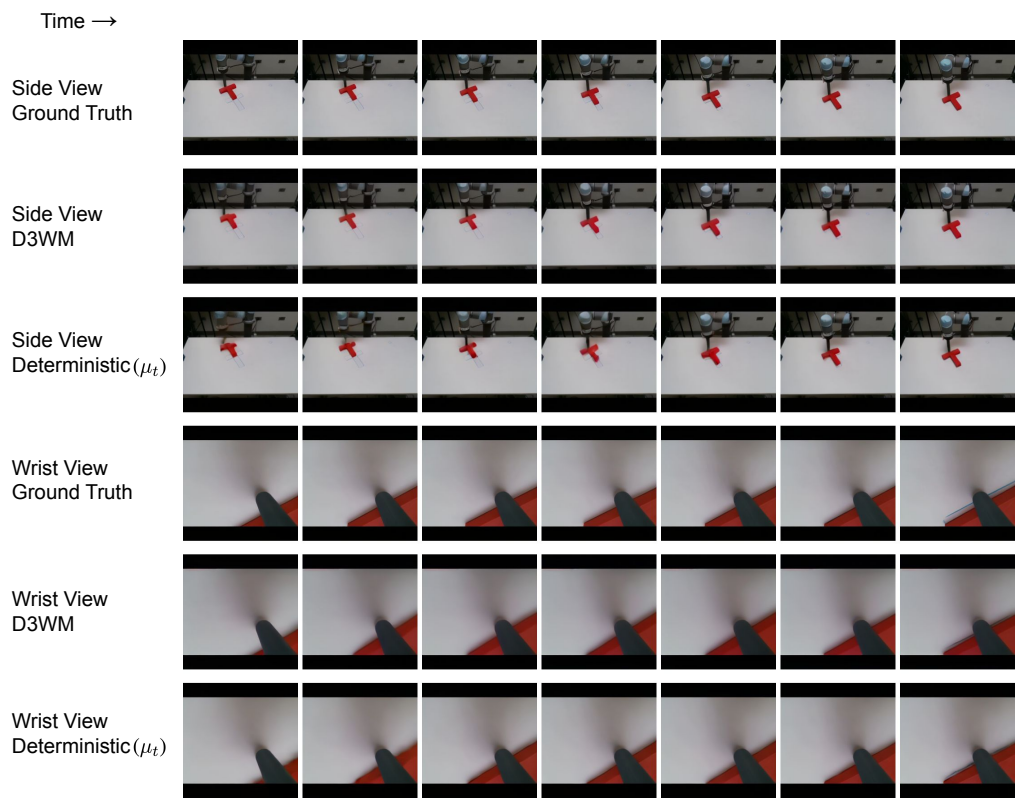


Figure 15: Qualitative results on Real PushT test dataset. D3WM for PushT predicts both side view and wrist view, learning the correct correlation between the two and accurate point of contact with the object. In this example, we can see every 5th frame of a predictions made in a row from ground truth frames, showing that we correctly capture the deterministic dynamics of the scene, as evident from the Drift Only predictions. These example frames show the object being pushed straight into the target location.

Limit behavior. Eq. 13 realizes the qualitative behavior described in the main paper:

- **Low aleatoric uncertainty** ($\sigma_t^2 \to 0$): $\gamma_t \to \lambda_\tau$. With λ_τ near 1, the blended latent is dominated by the noisy deterministic prior $\tilde{\mu}_t^{(\tau)}$, and the final $\hat{z}_t$ collapses toward μ_t. Geometric structure is preserved.

- **High aleatoric uncertainty** ($\sigma_t^2 \to 1$): $\gamma_t \to 0$. The blended latent equals $z_t^{(\tau)}$, and the diffusion model freely synthesizes content.

- **Schedule effect** (small λ_τ): in the second half of denoising the schedule reduces the deterministic anchoring even in low-variance regions, allowing the diffusion model to refine fine textures. In our default configuration we use a step-function schedule that disables anchoring after a fixed fraction of denoising steps.

E.4 Alternative: Clean-State Posterior-Mean Blending

For comparison and motivation of our design choice, we derive the alternative *clean-state* formulation that we ablate in the main paper. In this alternative, the deterministic prior is incorporated as a posterior correction to the diffusion model's clean-state estimate at each step, rather than as a noisy-latent replacement.

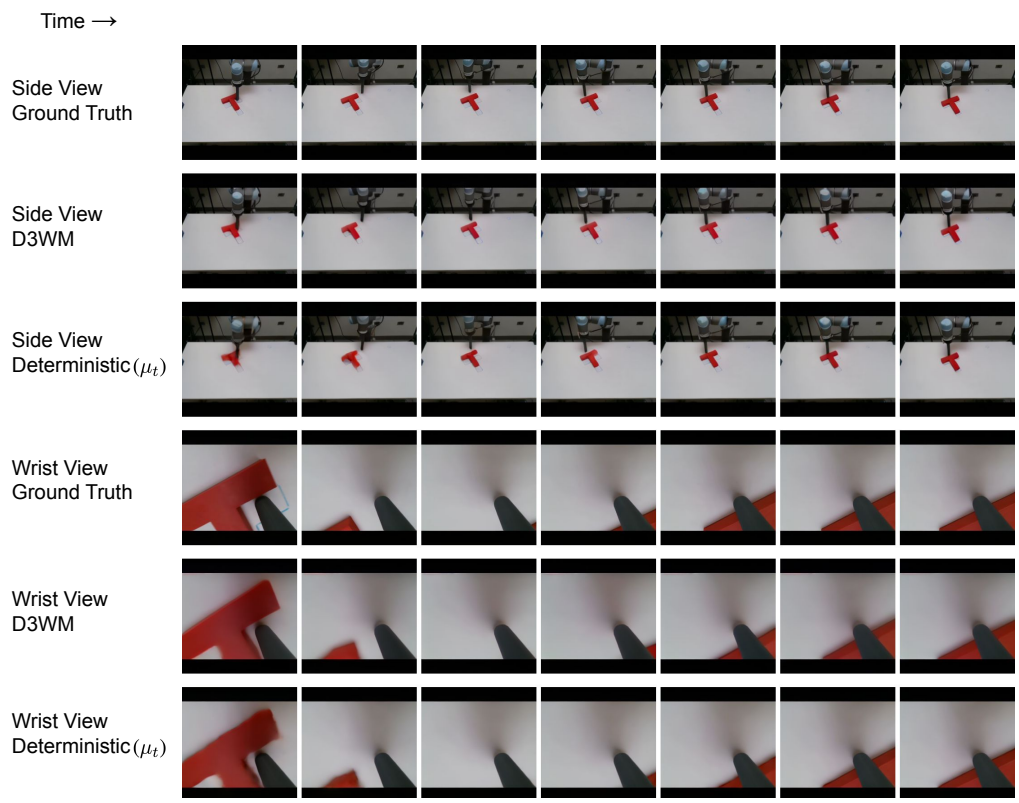


Figure 16: Additional qualitative results on Real PushT test dataset. D3WM for PushT predicts both side view and wrist view, learning the correct correlation between the two and accurate point of contact with the object. In this example we can see every 5th frame of a predictions made in a row from ground truth frames, showing that we correctly capture the deterministic dynamics of the scene, as evident from the Drift Only predictions. The sample frames show robotic gripper go around the object, and then pushing it into the the target location, with D3WM correctly capturing appearance and disappearance of the object in the wrist view.

At each step τ, the diffusion model produces a clean-state estimate $\hat{z}_{t,\tau}^{(0)}$ via the standard VP ϵ-prediction

$$\hat{z}_{t,\tau}^{(0)} = \frac{z_t^{(\tau)} - \sqrt{1-\bar{\alpha}_\tau}\,\hat{\epsilon}}{\sqrt{\bar{\alpha}_\tau}}, \tag{15}$$

with implicit variance

$$\rho_\tau^2 = \frac{1-\bar{\alpha}_\tau}{\bar{\alpha}_\tau}, \tag{16}$$

which is the variance of $\hat{z}_{t,\tau}^{(0)}$ around the true $z_t^{(0)}$ under the noise schedule.

We have two Gaussian sources of information about $z_t^{(0)}$:

- The temporal prior $\mathcal{N}(\mu_t, \sigma_t^2 I)$ from the deterministic network.
- The diffusion observation $\mathcal{N}(\hat{z}_{t,\tau}^{(0)}, \rho_\tau^2 I)$ from the score model.

The Gaussian posterior over $z_t^{(0)}$ has a precision-weighted mean

$$\tilde{z}_{t,\tau}^{(0)} = \frac{\sigma_t^{-2} \odot \mu_t + \rho_\tau^{-2} \odot \hat{z}_{t,\tau}^{(0)}}{\sigma_t^{-2} + \rho_\tau^{-2}}. \tag{17}$$

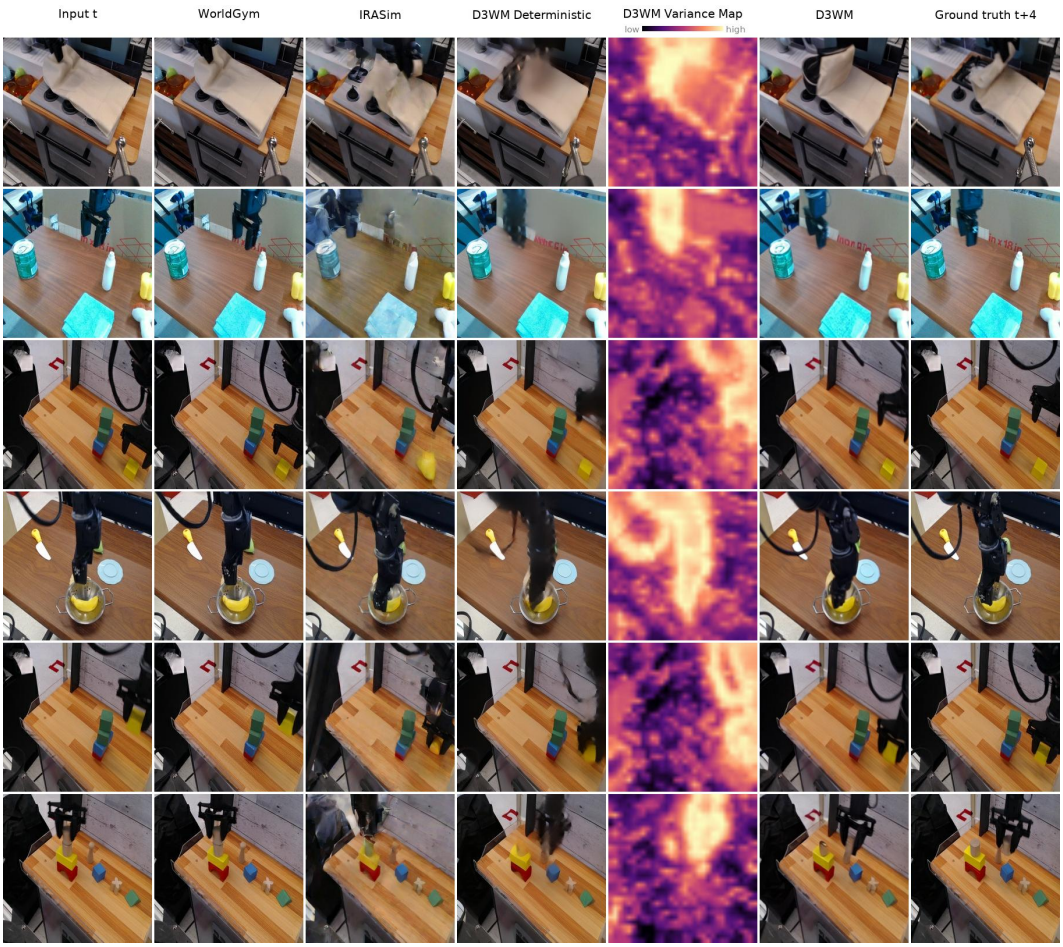


Figure 17: Additional qualitative results on BridgeData V2 dataset. D3WM Deterministic prediction correctly realizes the commanded motion, while blurring in the high-variance regions. Variance Map, plotted for visualization purposes only from its stretched latent representation, is used to sample a committed outcome using the stochastic component, resulting in a sharp D3WM prediction.

Multiplying numerator and denominator by $\sigma_t^2 \odot \rho_\tau^2$,

$$\tilde{z}_{t,\tau}^{(0)} \;=\; \frac{\rho_\tau^2}{\rho_\tau^2 + \sigma_t^2} \odot \mu_t + \frac{\sigma_t^2}{\rho_\tau^2 + \sigma_t^2} \odot \hat{z}_{t,\tau}^{(0)}. \tag{18}$$

Defining $\lambda_\tau^{\text{cs}} = 1/(\rho_\tau^2 + \sigma_t^2)$ gives the same convex-combination form as the noisy-bridge blend, but with the diffusion clean-state estimate $\hat{z}_{t,\tau}^{(0)}$ in place of the noisy-bridge term $\tilde{\mu}_t^{(\tau)}$, and the clean-state posterior variance ρ_τ^2 in place of the noise-level marginal variance $1 - \bar{\alpha}_\tau$.

Why the noisy-bridge outperforms. The clean-state formulation in Eq. 18 has a structural failure mode that explains its inferior empirical performance (see *Clean State Blend* ablation in main paper Section 4.1). At early diffusion steps τ, the implicit variance $\rho_\tau^2 = (1 - \bar{\alpha}_\tau)/\bar{\alpha}_\tau$ of $\hat{z}_{t,\tau}^{(0)}$ becomes very large because $\bar{\alpha}_\tau$ approaches zero. Consequently, the precision-weighted average gives disproportionate weight to μ_t in early denoising steps everywhere in the image, regardless of the spatial variance map. This over-anchors the trajectory to the blurry deterministic prediction in early steps, leaving the diffusion model insufficient room across the remaining steps to recover sharp textures even in high-uncertainty regions.

The noisy-bridge formulation (Section E.3) avoids this failure because the schedule λ_τ that controls overall anchoring strength is independent of the diffusion noise level — we can keep λ_τ near 1 for the first portion of denoising (where the diffusion model benefits from the deterministic prior's

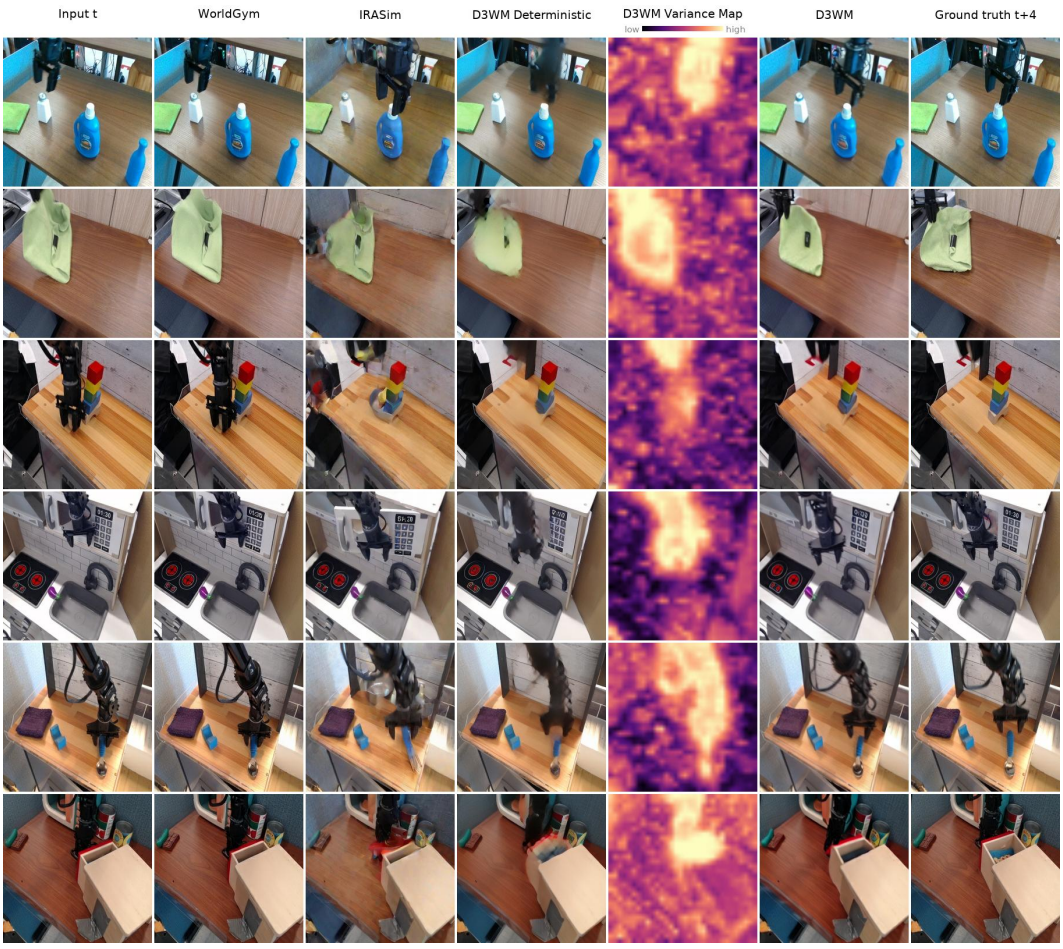


Figure 18: Additional qualitative results on BridgeData V2 dataset. D3WM Deterministic prediction correctly realizes the commanded motion, while blurring in the high-variance regions. Variance Map, plotted for visualization purposes only from its stretched latent representation, is used to sample a committed outcome using the stochastic component, resulting in a sharp D3WM prediction.

structure) and reduce it to 0 for the second portion (where the diffusion model produces texture detail). The spatial gate $(\mathbf{1} - \sigma_t^2)$ and the temporal schedule λ_τ are decoupled.

E.5 Relationship to the Temporal SDE

The world-model temporal SDE in Eq. 4 of the main paper models the *data-generating process* along the trajectory axis t. The deterministic prediction network regresses the integrated effect of the temporal drift coefficient over $[0, t]$ (i.e., the conditional mean $\mu_t = \mathbb{E}[z_t \mid z_0, a_{[0,t)}]$) together with the spatial variance σ_t^2 of the unobserved residual. We do not numerically integrate the temporal SDE, instead, the trained network f_θ^μ produces the integrated mean and variance directly. The reverse-time diffusion described in this section operates on a separate axis τ (diffusion noise level) and provides image-manifold-aware sampling from the conditional distribution $p(z_t \mid z_0, a_{[0,t)})$. The two SDEs share only the spatial variance map σ_t^2, which couples them via the trust map γ_t.

F Model Architecture Details

We built our drift network using a vision transformer architecture with about 450 million parameters, also using AdaLN for conditioning. We used an MLP to embed the action conditioning so that it fits up to 7*16 size, the size requirement dictated by our PushT experiment. This size can be easily

changed for action spaces with different dimensions. Our drift network trained for NVS outputs two images in latent space, one corresponding to the predicted drift image and the other being the variance map, both in latent space. Our PushT network outputs four images in latent space: the latent drift image prediction for the side camera view, the side view variance map, the latent drift image for the wrist camera view, and the wrist view variance map. For the NVS task, our drift network takes two reference frames as input. For the PushT task, our drift network takes one wrist view and one side view as input.

We used the Stable Diffusion Canny ControlNet architecture to build the diffusion component of D3WM. We used Stable Diffusion 1.5 together with the IP Adapter Plus for Stable Diffusion 1.5 as the backbone. We changed the control input layer of our ControlNet to condition on latent representations of the drift image and variance map. During inference, all of the gating and selective noising of the intermediate frames during each diffusion step are done in latent space using the latent deterministic prediction image (μ_t) and the latent variance map (σ_t^2) output from the deterministic network f_θ^μ.

G Compute Resources

We trained all models on a workstation with two NVIDIA RTX PRO 6000 Blackwell GPUs (96 GB GDDR7 each). Training the deterministic prediction network for NVS required approximately one week of wall-clock time across both GPUs (1M training steps at global batch size 256). The PushT deterministic prediction network trained considerably faster, in approximately 12 hours, due to a smaller batch size (32) and 5x fewer training steps (200k) for the simpler scene distribution. The conditional diffusion ControlNet required approximately 2 days for NVS (100k steps at batch size 16, with conditioning images at 512x512 resolution) and approximately 10 hours for PushT (20k steps at batch size 16). We note that comparable performance was achieved by the NVS deterministic network at 300k training steps with additional LPIPS and SSIM auxiliary loss terms, and the additional 700k steps yielded diminishing returns.