A LOW-DIMENSIONAL DETERMINISTIC DYNAMICS CORE FOR CONTROLLABLE LONG-HORIZON LATENT WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive world models must generate plausible observations while preserving the consequences of actions over long autoregressive rollouts for game agents, model-based planning, and embodied control. Yet commonly used one-step and teacher-forced metrics can make flat and factorized transition models appear similar under matched capacity and exposure-bias noise, even when their free-running behavior differs. We study this discrepancy in frozen pretrained latent spaces and introduce WORLDBRIDGE, a factorized transition model that predicts a compact dynamics code before rendering the remaining appearance coordinates deterministically. Both components retain access to the previous latent and the action, so the compact code guides prediction without imposing a strict information bottleneck. Under capacity-, training-budget-, and exposure-noise-matched comparisons, the factorized model has comparable short-range fidelity but produces more controllable and radially stable long-horizon rollouts than a flat predictor, with the clearest separation at the single-evaluation operating point. Comparisons with stronger memory-based predictors further show that the training objective and the factorization contribute distinct parts of the observed improvement, and that the advantage is not explained by generative stochasticity alone. A larger-scale validation retains the controllability advantage on temporally held-out trajectories and improves action identification. Finally, component-wise distillation yields a real-time student using only a few network evaluations per frame. These results support evaluating latent world models by their closed-loop behavior, not by short-range fidelity alone.

1 INTRODUCTION

A world model predicts how an environment evolves under an agent's actions, and learning one directly in pixels or latent video has quickly become a unifying objective across game engines, model-based planning, and embodied control (Ha & Schmidhuber, 2018; Hafner et al., 2023; Bruce et al., 2024). The most capable recent instances are continuous generative models, diffusion or flow networks that denoise the next frame conditioned on past frames and an action, and they already drive playable, interactive experiences such as GAMENGEN, DIAMOND, and Oasis (Valevski et al., 2024; Alonso et al., 2024; Decart et al., 2024). A usable world model requires an autoregressive loop: it must roll out for hundreds of steps without drifting, remain controllable, and run fast enough to be played, which has made real-time single or few-step generation a central research target (Yin et al., 2025; Huang et al., 2025).

Progress on these models is reported almost entirely with short-range or static criteria: one-step or teacher-forced distances, per-frame PSNR and LPIPS, and clip-level Fréchet R3D-18 distance. But the operative regime is the autoregressive loop, and we show the two regimes can disagree sharply. Under matched capacity and exposure-bias noise, flat and factored predictors are essentially tied on every tested one-step, teacher-forced, PSNR, and R3D-feature distance, yet they diverge over long rollouts: a flat predictor silently contracts toward an over-smoothed, less controllable attractor while a factored one better preserves latent scale and action response. Standard metrics do not see this gap, which recent position and benchmark work also flags and which we make concrete in the frozen generative-latent setting.

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.

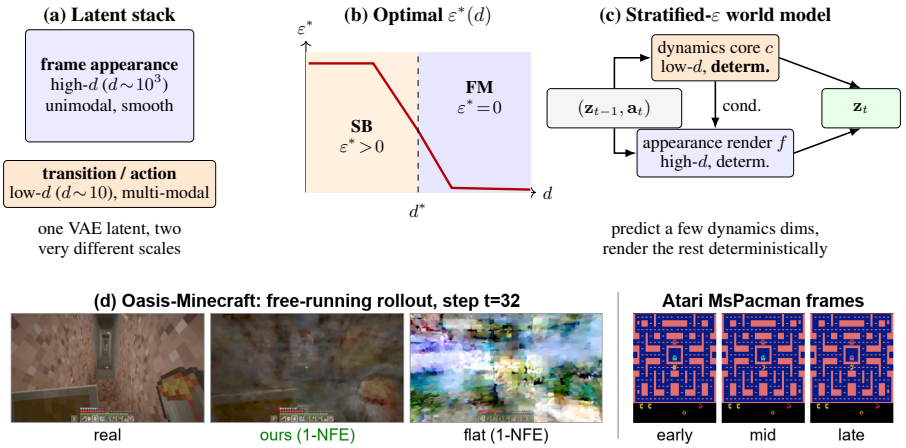

Figure 1: A low-dimensional deterministic dynamics core for world models. (a) A world-model VAE latent is a multi-scale stack: high-dimensional frame appearance and a low-dimensional transition/action code. (b) Motivation: the optimal interpolation stochasticity ε^* falls toward zero as the modeled dimension grows, so the high-d appearance prefers $\varepsilon^*{=}0$; under fair comparison ε acts as a fidelity–diversity dial. (c) Our model, WORLDBRIDGE, predicts a low-d dynamics core c (a 2–4 dimensional coarse code already saturates tested one-step reconstruction) and renders the high-d appearance f deterministically from it into one next-latent $\mathbf{z}_t$, conditioned on the previous latent and action; this improves long-horizon controllability and radial stability relative to flat predictors. (d) The two frozen pretrained domains: on Oasis-Minecraft, at free-running step $t{=}32$ the distilled factored core (1-NFE) keeps scene geometry and HUD while a flat student of equal budget collapses (one shared brightness boost for print; Fig. 18); right, decoded Atari MsPacman frames.

We study, on frozen pretrained world-model latents, which design choice controls this long-horizon behaviour, and find that it is architectural rather than the amount of generative stochasticity. In the tested PCA-width sweep, a deterministic next-latent predictor that models a low-dimensional coarse code and renders the rest from it saturates one-step reconstruction with a 2–4 dimensional coarse code (Fig. 1b,c), far below the ambient dimension of 4096 to 9216 or even the participation-ratio dimension of about 20. Building the world model, WORLDBRIDGE, around this dynamics core, by predicting the low-dimensional core and rendering the high-dimensional appearance deterministically from it (Fig. 1c), yields at matched parameter and training budget rollouts that over 1000 steps stay markedly more action-controllable, drift less from the real marginal, and preserve latent scale, while a flat predictor of equal capacity loses all three. A second result is negative: along the interpolant from Flow-Matching ($\varepsilon{=}0$) to a Schrödinger bridge ($\varepsilon{=}1$) (Albergo et al., 2023; De Bortoli et al., 2021; Lipman et al., 2023), generative stochasticity is, under fair matching, only a fidelity/diversity dial and does not improve accuracy, stability, or controllability. Finally, the core distills to one or two network evaluations for real-time interaction.

On two pretrained world-model latent spaces, Oasis-Minecraft and Atari, we make three contributions:

- **A factorized transition model with a low-dimensional dynamics core.** We introduce WORLDBRIDGE, which predicts a compact coarse code first and then renders the remaining latent coordinates conditionally. Across the tested frozen latent spaces, a 2–4 dimensional coarse code is sufficient to match the tested one-step reconstruction criterion, while the factorized model yields more stable and more action-responsive long-horizon rollouts than a parameter-matched flat predictor.
- **A long-horizon evaluation protocol for frozen generative latents.** We evaluate free-running rollouts with controllability, marginal drift, and latent scale in addition to one-step and teacher-forced fidelity. This comparison shows that standard short-range metrics can leave flat and factorized predictors nearly tied even when their long-horizon behavior differs substantially.
- **A controlled study of stochasticity and real-time distillation.** Under matched capacity, training budget, and exposure-noise settings, varying the interpolant stochasticity changes the fidelity–

diversity trade-off without improving the tested long-horizon outcomes. The factorized transition can also be distilled to a two-evaluation student that supports real-time interactive rollout.

2 Related Work

Diffusion and flow world models. World models learn environment dynamics for prediction, planning, and control (Ha & Schmidhuber, 2018; Hafner et al., 2021; 2023), and a recent line makes them generative and interactive: GameNGen renders playable DOOM (Valevski et al., 2024), DIAMOND captures the visual detail that matters for Atari agents (Alonso et al., 2024), Oasis brings real-time open-ended Minecraft generation with a ViT-VAE latent and a DiT backbone (Decart et al., 2024; Peebles & Xie, 2023), and Genie learns a latent action space for controllable environments (Bruce et al., 2024). These systems treat the frame latent as monolithic and are evaluated mostly by short-range or static criteria; we instead identify the factors governing their long-horizon rollout behaviour, using Oasis and an Atari SD-VAE only as sources of pretrained latent spaces, not baselines to beat.

Low-dimensional, parsimonious, and controllable latent dynamics. Action-controllable structure in an environment can be low-dimensional even when observations are not; this is a recurring observation: PLSM penalizes state-dependent action effects (Saanum et al., 2024), Sensorimotor World Models bias a JEPA encoder toward controllable degrees of freedom (Ivashkov et al., 2026), TC-WM learns a compact dynamical subspace on frozen foundation embeddings (Fu et al., 2026), and GRWM identifies latent geometry as the long-horizon bottleneck (Xia et al., 2025). We extend this view to frozen pretrained generative world-model latents: we change only the autoregression target, find that two to four coarse dimensions already saturate the tested one-step criterion (without claiming they span the intrinsic action dimension), and evaluate the consequence under a capacity- and exposure-noise-matched 1000-step protocol on which standard short-range metrics are blind, joining the call for interventional, closed-loop evaluation (Yu et al., 2026; Zhou et al., 2026; Liang et al., 2026; Tang et al., 2026).

Fast, causal rollout and few-step distillation. Real-time interactive generation has driven distilled causal video generators: CausVid turns slow bidirectional teachers into fast causal students (Yin et al., 2025), Self-Forcing rolls out on the model's own predictions to fight exposure bias (Huang et al., 2025), and causal-forcing variants distill an autoregressive consistency map (Zhu et al., 2026); one/few-step generators are likewise obtained by progressive distillation (Salimans & Ho, 2022), consistency models (Song et al., 2023; Lu & Song, 2025), distribution matching (Yin et al., 2024), and average-velocity/shortcut maps (Frans et al., 2025; Geng et al., 2025; Zhang et al., 2025a). These works treat the frame latent as monolithic; we adopt their exposure-bias remedy (context noising) and contribute the observation that the low-d core distills cleanly to one/few steps whereas an equal-capacity flat predictor becomes unstable in this regime.

Flow matching, stochastic interpolants, and bridges. Flow Matching and rectified flow learn deterministic transport (Lipman et al., 2023; Liu et al., 2023); stochastic interpolants unify flows and diffusions through a noisy interpolant (Albergo et al., 2023); and Schrödinger/Brownian bridges add an entropic coupling between endpoints (De Bortoli et al., 2021; Shi et al., 2023; Zhou et al., 2024). Our ε-interpolant uses these public primitives as a controlled probe of generative stochasticity. Prior work already observes deterministic predictors matching stochastic ones in accuracy (Zhang et al., 2025b; Yoon et al., 2024) and context noise stabilizing autoregressive rollouts (Chen et al., 2024); a dimension-dependent scaling argument (restated in App. A) motivates deterministic appearance rendering.

3 Method

We first define the one-parameter generative family we use for each module (Sec. 3.1) and note a secondary observation, that its optimum drifts toward determinism as dimension grows (Sec. 3.2), which motivates rendering appearance deterministically. The core of the method is the low-dimensional deterministic dynamics core world model (Sec. 3.3) and its real-time distillation (Sec. 3.4). Under the fair comparison of Sec. 4.2, the stochasticity level ε is not the lever for long-horizon quality; the factorization and the training objective are.

3.1 THE ε-INTERPOLANT FAMILY

Let $\mathbf{x}_0$ denote a data point (a latent we wish to model) and $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I}_d)$ a noise endpoint. For a stochasticity level $\varepsilon \geq 0$ we define the ε-interpolant between them, for $\tau \in [0, 1]$,

$$\mathbf{x}_\tau \;=\; (1-\tau)\,\mathbf{x}_0 + \tau\,\mathbf{x}_1 + \sqrt{\varepsilon\,\tau(1-\tau)}\,\mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_d), \tag{1}$$

so $\mathbf{x}_\tau \sim \mathcal{N}(\mu_\tau, \sigma_\tau^2 \mathbf{I})$ with $\mu_\tau = (1-\tau)\mathbf{x}_0 + \tau\mathbf{x}_1$ and $\sigma_\tau^2 = \varepsilon\,\tau(1-\tau)$. At $\varepsilon = 0$, Eq. (1) is the straight, deterministic transport of Flow Matching (Lipman et al., 2023; Liu et al., 2023); at $\varepsilon = 1$ it is a Brownian/ Schrödinger-bridge interpolant with an entropic stochastic coupling (Albergo et al., 2023; De Bortoli et al., 2021). The conditional velocity that transports Eq. (1) is

$$\mathbf{v}_\tau \;=\; \underbrace{(\mathbf{x}_1 - \mathbf{x}_0)}_{\text{deterministic drift}} \;+\; \underbrace{\tfrac{1}{2}\,\frac{1-2\tau}{\tau(1-\tau)}\,(\mathbf{x}_\tau - \mu_\tau)}_{\text{stochastic correction } \mathbf{s}_\tau}, \tag{2}$$

and we learn a field $\mathbf{v}_\theta(\mathbf{x}, \tau) \approx \mathbb{E}[\mathbf{v}_\tau \mid \mathbf{x}_\tau = \mathbf{x}]$ by regression. Sampling is identical for every ε: we integrate the ODE $\dot{\mathbf{x}} = \mathbf{v}_\theta(\mathbf{x}, \tau)$ from a noise sample at $\tau{=}1$ to $\tau{=}0$ with K Euler steps; only the training target in Eq. (2) depends on ε, through $\mathbf{s}_\tau$. This lets us treat ε as a pure modeling knob at fixed sampling cost.

Per-coordinate stochasticity. Crucially, ε may be a vector $\boldsymbol{\varepsilon} \in \mathbb{R}^d_{\geq 0}$ applied coordinate-wise: coordinates with $\varepsilon_i{=}0$ contribute only the deterministic drift $(\mathbf{x}_1 - \mathbf{x}_0)_i$, while those with $\varepsilon_i{>}0$ carry bridge noise. A multi-scale latent can thus be assigned different stochasticity per block, the mechanism behind stratification below.

3.2 DETERMINISTIC RENDERING OF HIGH-DIMENSIONAL APPEARANCE

A simple scaling argument makes the optimal interpolation stochasticity dimension-dependent: the conditional target of Eq. (2) carries an irreducible stochastic term whose expected squared norm grows as $O(\varepsilon\, d)$, while the curvature-smoothing benefit of bridge noise is $O(1)$ in d (and shrinks with K), so above a threshold dimension $d^*(K)$ the deterministic choice $\varepsilon^*{=}0$ is preferred. We recap this argument in App. A and measure the transition on our own data as a sanity check in App. C and App. D. The consequence we rely on is that the high-dimensional appearance code prefers $\varepsilon^*{=}0$ and can be rendered deterministically, while Sec. 4.2 shows that adding stochasticity to the low-dimensional core does not help long-horizon quality either. The whole model can therefore be deterministic, and ε is best read as a fidelity/diversity dial; the win comes from the factorization, not the noise level.

3.3 THE LOW-DIMENSIONAL DYNAMICS-CORE WORLD MODEL

We model action-conditioned dynamics in the latent space of a frozen pretrained encoder. Frames are encoded and projected onto the top-d_z principal directions of the latent marginal (natural scale; see Sec. 5 and App. D), giving an ordered code $\mathbf{z}_t \in \mathbb{R}^{d_z}$ in which early coordinates carry the most variance. A transition is the conditional $p(\mathbf{z}_t \mid \mathbf{z}_{t-1}, \mathbf{a}_t)$ for previous latent $\mathbf{z}_{t-1}$ and action $\mathbf{a}_t$.

Factorization into a dynamics core and a render. We split $\mathbf{z}_t = [\,c\,; f\,]$ into a low-d dynamics core $c = \mathbf{z}_t[{:}d_c] \in \mathbb{R}^{d_c}$ (with $d_c = 2$–4 sufficient for the tested one-step reconstruction sweep; see Sec. 4.2) and a high-d appearance residual $f = \mathbf{z}_t[d_c{:}] \in \mathbb{R}^{d_f}$, $d_f = d_z - d_c$, and model

$$\underbrace{p(c \mid \mathbf{z}_{t-1}, \mathbf{a}_t)}_{\text{core flow}} \quad \text{and} \quad \underbrace{p(f \mid \mathbf{z}_{t-1}, \mathbf{a}_t, c)}_{\text{render flow}}, \tag{3}$$

with two conditional velocity networks trained by Eq. (2): a small core flow that carries the action-conditioned dynamics, and a render flow that deterministically reconstructs appearance given the sampled core. Both default to $\varepsilon{=}0$ (deterministic); an optional core stochasticity ε_c is available but, under fair comparison, is not needed (Sec. 4.2). Sampling draws c (K steps), then f conditioned on c (K steps), and decodes $[\,c; f\,]$ (Alg. 1). The flat baseline is a single flow over all of $\mathbf{z}_t$ at matched parameter budget. This is a factorized predictor, not a strict information bottleneck: the render also receives $\mathbf{z}_{t-1}$ and $\mathbf{a}_t$, so action information can bypass c. The experiments test whether the factorization improves rollout stability; they do not identify a unique causal mechanism or an intrinsic action-subspace dimension.

Algorithm 1 Dynamics-core transition (one autoregressive step)

Require: previous latent $\mathbf{z}_{t-1}$, action $\mathbf{a}_t$; core flow $\mathbf{v}_c$, render flow $\mathbf{v}_f$ (both $\varepsilon{=}0$ by default); steps K

$c \leftarrow \text{ODESOLVE}(\mathbf{v}_c\,;\,(\mathbf{z}_{t-1}, \mathbf{a}_t),\, K)$ // low-d dynamics core

$f \leftarrow \text{ODESOLVE}(\mathbf{v}_f\,;\,(\mathbf{z}_{t-1}, \mathbf{a}_t, c),\, K)$ // deterministic appearance render

$\mathbf{z}_t \leftarrow [\,c\,;\,f\,]$; **return** $\mathbf{z}_t$ (decode with frozen VAE for pixels)

3.4 REAL-TIME 1-NFE DISTILLATION

For interaction we distill each flow to a single function evaluation with an average-velocity flow map $F_\theta(\mathbf{x}, t, s \mid \text{cond})$ that predicts the mean velocity from time t to s (Frans et al., 2025; Geng et al., 2025). We keep an always-on Flow-Matching anchor and warm in a one-step consistency target,

$$\mathcal{L} = \underbrace{\left\|F_\theta(\mathbf{x}_\tau, t, t) - \mathbf{v}_t\right\|^2}_{\text{FM anchor}} + \lambda_c(\text{step})\ \underbrace{\left\|F_\theta(\mathbf{x}_\tau, t, 0) - \left(\mathbf{v}_t - t\,\tfrac{d}{dt}F_\theta\right)\right\|^2}_{\text{consistency (stop-grad target)}}, \tag{4}$$

where the total derivative $\frac{d}{dt}F_\theta$ is estimated by finite differences and λ_c ramps from 0; gradients are clipped. (The anchor is essential: bootstrapping the map from scratch diverges, since these recipes were designed to fine-tune a pretrained field.) One-step generation is then $\mathbf{z}_t = \mathbf{x}_1 - F_\theta(\mathbf{x}_1, t{=}1, s{=}0 \mid \text{cond})$, $\mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$: the flat model costs 1 evaluation per frame, the stratified model 2 (coarse then fine).

Exposure-bias context noise. A one-step student rolled out on its own predictions accumulates error and diverges. We adopt the standard remedy (Valevski et al., 2024; Huang et al., 2025): during distillation we perturb only the conditioning past latent (and its coarse dimensions) with Gaussian noise of std σ, leaving the action untouched, so the student learns to correct perturbed rollout contexts. As shown in Sec. 4.2, this makes 1-NFE rollout stable, with a mild stability–sharpness trade-off in σ.

4 EXPERIMENTS

The main result is the long-horizon rollout study (Sec. 4.2): under a fair, capacity-matched comparison a low-dimensional deterministic dynamics core is more controllable and less drifting over 1000 steps than a flat predictor, while standard short-range and static metrics see no difference. Sec. 4.1 motivates deterministic appearance rendering; Secs. 4.4 and 4.6 cover controllability and real-time distillation. Unless noted the metric is Sliced-Wasserstein-2 between generated and held-out latents (protocols in App. B).

4.1 MOTIVATION: HIGH-DIMENSIONAL APPEARANCE PREFERS DETERMINISTIC RENDERING

The scaling argument of Sec. 3.2 (restated in App. A) predicts that the optimal interpolation stochasticity ε^* falls to zero above a threshold dimension, because the smoothing benefit of bridge noise is $O(1)$ in dimension while its variance cost grows as $O(\varepsilon d)$. As a sanity check that our regime is the deterministic one, we measure this transition on a controlled d-scaling Gaussian mixture and on both real world-model latent spaces (Oasis-Minecraft, Atari; participation ratio ≈ 20): ε^* drops from 1 to 0 across $d \approx 256$ to 512 (App. Fig. 4). The high-dimensional frame appearance therefore prefers $\varepsilon^*{=}0$ and can be rendered deterministically (full sweeps in App. C and App. D). Sec. 4.2 then shows that, once capacity and exposure-bias noise are matched, stochasticity is not the lever for long-horizon quality.

4.2 FACTORS GOVERNING LONG-HORIZON ROLLOUT AND MATCHED CONTROLS

This is the core of the paper. We train action-conditioned dynamics $\mathbf{v}(\mathbf{z}_t, \tau \mid \mathbf{z}_{t-1}, \mathbf{a}_t)$ on ordered sequences and roll them out for up to $H{=}1000$ self-conditioned steps, on both Oasis-Minecraft (strides 1 and 3; 3000 and 1200 frames) and Atari (6000 frames). Every variant is matched on parameter count, training budget, sampling budget K, and exposure-bias context noise $\sigma{=}0.1$: a flat predictor over all of $\mathbf{z}_t$ ($\varepsilon{=}0$ and $\varepsilon{=}1$) versus our dynamics core (d_c-d coarse, deterministic; render

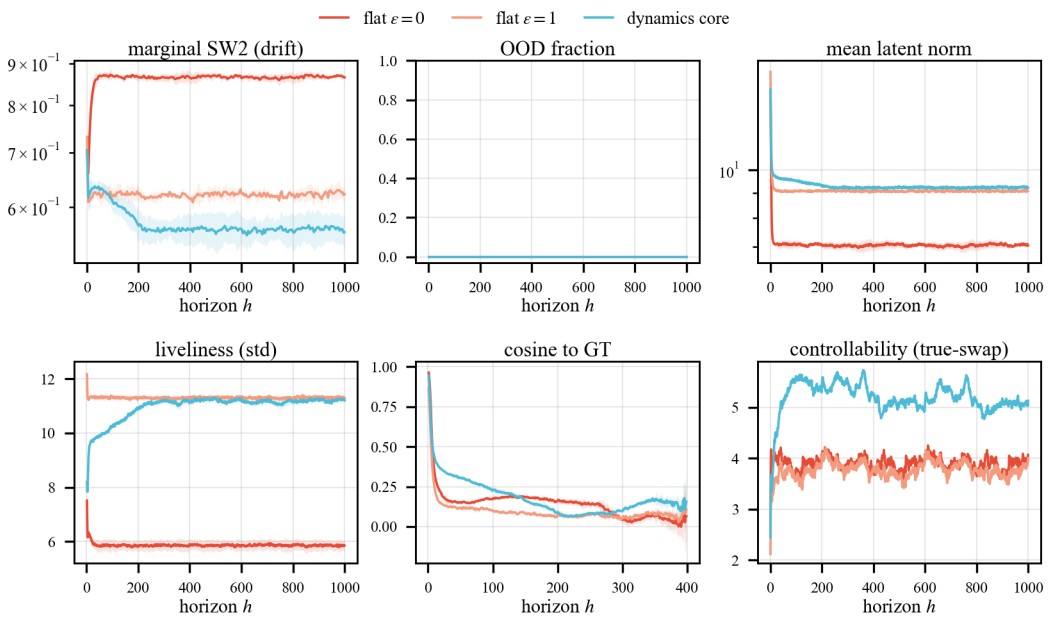

Figure 2: Long-horizon rollout (Oasis stride-3, $H{=}1000$, capacity and σ-matched, 3 seeds). With fair exposure-bias noise no model diverges (OOD$=0$), so short-range metrics tie; over the horizon flat $\varepsilon{=}0$ (red) drifts up in marginal-W2 and contracts in latent norm, while the dynamics core (blue) holds the lowest drift, preserves scale, and is most action-controllable (bottom-right).

$\varepsilon{=}0$). Per horizon we report distribution drift (sliced-W2 against the real next-latent marginal), the fraction exceeding the maximum training-latent norm, mean latent norm, cosine to ground truth, and counterfactual controllability (the ℓ_2 rollout change when the action stream is swapped at fixed sampling noise; higher $=$ more action-driven).

Standard short-range/static metrics do not separate the models. Under the matched setup, the flat and core predictors are statistically comparable on every metric the field usually reports (App. E, Table 14): one-step SW2 ($0.31/0.26/0.28$ for flat $\varepsilon{=}0$ / flat $\varepsilon{=}1$ / core), teacher-forced decoded PSNR ($25.8/23.6/24.8$ dB), and Fréchet R3D-18 distance ($155/200/157$; the core ties flat $\varepsilon{=}0$). A practitioner reading only these numbers would see no meaningful difference. With matched exposure-bias noise, no model exceeds the training-norm threshold to $H{=}1000$.

Over long rollouts the core improves markedly over a flat predictor. The differences appear only in free-running rollout (App. E, Table 15 and Fig. 2), and they are consistent across all three domains. Against a flat predictor of equal capacity, the dynamics core has lower long-horizon drift (marginal-W2@1000 $0.56/0.67$ on Oasis-s3 / Atari vs. $0.87/0.78$ for flat $\varepsilon{=}0$), preserves the real latent scale (flat $\varepsilon{=}0$ contracts: norm@1000 ≈ 7 vs. the core's 9), and, most importantly for a world model, is markedly more action-controllable: its counterfactual response at $H{=}1000$ is about 26 to 40% above flat $\varepsilon{=}0$ in the reported small-scale domains. A flat predictor of equal capacity silently relaxes toward an over-smoothed, weakly-controllable attractor, whereas the factorized predictor better preserves radial scale and response, an advantage invisible to one-step, teacher-forced, PSNR, and R3D-feature scores. We next test whether stronger architectures than a flat MLP also recover this.

Dissecting the gain: objective and factorization each contribute a factor of about two. We add stronger, capacity-matched ($\sim$2.1M) baselines at the same budget and exposure-bias noise: a Dreamer-style RSSM (Hafner et al., 2021; 2023), a history+attention model, a history-conditioned flat model, and a flat model trained with the same multi-step rollout loss as the core (flat-ms). Two findings (App. E, Table 12). (i) The gain decomposes: the multi-step objective lifts history-conditioned flat from ctrl@1000 ≈ 1.2 to 2.5 up to 8.0 (flat-ms), and on Atari the factorization doubles it again to 16.3 under the same objective. (ii) Against the memory baselines the core leads on Atari (ctrl 16.3

Table 1: Scaled-up comparison ($\sim$21M params, 400k frames, 4 Atari games, 3 seeds). The dynamics core doubles controllability over the flat baseline while maintaining comparable tracking and radial stability.

Model	CTRL@$H\uparrow$	cos@50$\uparrow$	W2@$H\downarrow$	OOD@$H\downarrow$
flat (ε=0)	0.694	0.911	1.203	0.000
PCA-core	1.843	0.917	1.220	0.000
learned core (ms)	**1.375**	**0.921**	1.251	**0.000**

Table 2: Held-out evaluation (80/20 temporal split; 100k frames/game; $\sim$21M params; 3 seeds). The dynamics core maintains high cosine tracking on unseen trajectory regions through 500 steps.

Model	CTRL@$H\uparrow$	cos@50	cos@200	cos@500	OOD@H
flat (ε=0)	0.336$\pm$0.036	0.922	0.909	0.912	0.000
PCA-core	1.295	0.921	0.898	0.894	0.000
learned core (ms)	**2.098$\pm$0.418**	**0.930**	0.892	0.889	**0.000**

vs. RSSM 9.9, attention 8.6); on Oasis history+attention matches it (14.4 vs. 14.7) and beats it on drift (0.51 vs. 0.89), so the multi-step advantage there is shared. The factorization's distinct property is robustness: it alone stays controllable without exposure noise (σ=0: ctrl $\approx$ 21 while flat and PCA-core diverge), at lower architectural complexity. Long-horizon controllability is thus governed by training and architecture choices that standard metrics cannot see, and the factored core combines both.

A few coarse dimensions suffice for tested reconstruction. Sweeping the coarse-code width $d_c \in \{2, 4, 8, 16, 32, 64\}$, one-step accuracy saturates almost immediately: cosine to the next latent is 0.94 already at d_c=2 and rises only to 0.96 at d_c=64, and long-horizon drift is flat across this range. Thus 2–4 coarse dimensions suffice for the tested one-step reconstruction criterion.

The core can be learned, not just PCA'd. Replacing the fixed top-d_c PCA slice with a learned d_c-dimensional bottleneck (jointly training an encoder $E\colon \mathbf{z}_t \to c$, a core flow $p(c \mid \mathbf{z}_{t-1}, \mathbf{a}_t)$, and a deterministic render D at matched budget) sharply increases long-horizon action-controllability on both domains, and a multi-step training objective increases it further.

Scaling to 21M parameters and 400k frames. We scale the comparison from 2.1M to $\sim$21M parameters (hidden = 1280, depth = 8) and from 6k to 100k frames per game across four Atari games (MsPacman, Breakout, SpaceInvaders, Qbert; 400k transitions total). Under the same capacity-matched protocol, the learned multi-step core achieves CTRL@H=1.375 versus flat's 0.694 while maintaining comparable cosine tracking (cos@50=0.921 vs. 0.911) and zero rounded norm-threshold exceedance. The PCA core further achieves CTRL@H=1.843, cos@50=0.917 (Table 1).

Held-out generalization requires data scale. With 6k frames per game, held-out ground-truth tracking collapses (cosine-to-GT ≈ -0.02 at h=200; App. Table 13). At 100k frames per game (temporal 80/20 split, PCA fit on training frames only) the learned core reaches cos@50=0.930, cos@200=0.892, cos@500=0.889 on unseen trajectory regions, and its held-out controllability advantage widens to 6.2$\times$ (CTRL@H=2.098$\pm$0.418 vs. flat's 0.336$\pm$0.036; Table 2). Flat attains slightly higher cos@500 (0.912 vs. 0.889); at this scale the core's advantage is controllability rather than tracking.

Stochasticity is not the lever. Replacing the deterministic core with a stochastic one (ε_c>0, the original "stratified" design) changes none of the above within noise: at matched capacity it ties the deterministic core on drift, controllability, and one-step SW2, and is slightly worse on Fréchet R3D-18 distance; across the FM-to-SB axis the only consistent effect of ε is a fidelity/diversity trade-off (App. F). We therefore report a fully deterministic model and treat ε as a dial.

Table 3: Single-step (K=1) operating point (strict held-out; σ=0.1; H=1000). CTRL is counterfactual response. NX/F is the fraction exceeding the maximum training-latent norm, with a non-finite rollout counted as failure; a radial stability check, not manifold membership. Atari averages three seeds. Oasis weights all six seed–fold units equally; flat and PCA-core each have one non-finite unit (finite CTRL uses $5/6$ units).

	Atari (3 seeds)		**Oasis** (2 seeds × 3 folds)	
arch	CTRL↑	NX↓	CTRL↑	NX/F↓
flat	1.55	0.00	2.47±0.50	0.17
PCA-core	1.43	0.00	1.29±0.44	0.17
RSSM	9.21	0.98	14.34±2.79	0.50
hist+attn	9.86	0.28	10.52±2.61	0.16
core (ours)	**16.49**	**0.00**	**12.67±5.49**	**0.01**

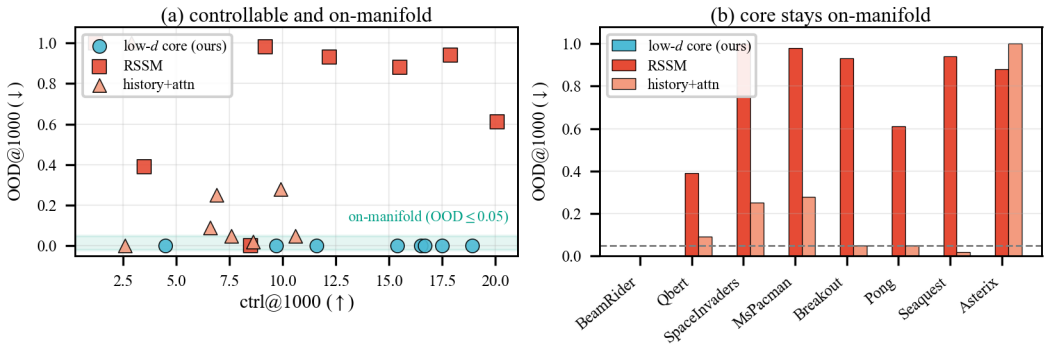

Figure 3: The K=1 separation across eight Atari games. (a) ctrl@1000 vs. norm-threshold exceedance@1000: the core occupies the controllable, radially stable corner in $7/8$ games. (b) Per-game exceedance; the core's exceedance rounds to zero in every game while the memory baselines often exceed the 0.05 band.

4.3 THE SINGLE-STEP (K=1) OPERATING POINT

The cross-architecture comparison above uses a multi-step sampling budget (K=4), where strong memory baselines match the core (Table 12). We evaluate the same training recipe at K=1: one network evaluation per frame for flat, two for the factored models (coarse then render); throughput is reported in Table 26. Each operating point is trained under the same recipe, so this is a recipe-level rather than fixed-weight comparison.

Atari. Compressing to one step leaves the core's controllability essentially unchanged (CTRL@1000 $\approx$ 16.5, vs. 16.3 at K=4), with no norm-threshold exceedance in the three seeds. RSSM is radially unstable (NX= 0.98) and the attention model partially diverges (NX= 0.28); the core gives the best joint response/stability point in this experiment.

Oasis. Equal weighting of the six LOSO seed–fold units gives 12.67±5.49 for the core and 10.52±2.61 for attention; the variance is too large to claim a clean response advantage. The clearer difference is radial stability: the core's norm-exceedance rate is 0.008, while RSSM reaches 0.50 and attention 0.16. Flat and PCA-core each become non-finite on one unit. These results support a stability advantage for the factored recipe, not an isolated controllability win on Oasis.

The separation is domain-general (eight Atari games). Replicating the same K=1 protocol on seven additional Atari games (eight in total; fresh 12k-frame collections, participation ratio 9.5–91.4; Fig. 3; App. E.1, Table 18), the low-d core is the best joint response/radial-stability model in $7/8$ games, with zero rounded norm-threshold exceedance in every domain (mean: core 0.00, RSSM 0.72, attention 0.22). The exception is the lowest-dimensional latent (BeamRider, PR=9.5), where controllable dynamics already fill the latent and a fixed PCA core suffices. The full free-running suite

(App. E.1, Table 19) corroborates the radial-instability reading; the core's drift advantage is clearest on higher-dimensional latents and reverses on the two lowest.

4.4 Action obedience

Two static probes on Oasis (stride-6) corroborate the controllability gap of Sec. 4.2 (Fig. 15; App. G): the 16-dim core carries $\approx 51\%$ of per-dimension action sensitivity despite being 6% of dimensions, and a linear action probe reaches macro-AUC 0.912 (real 0.927) versus 0.892 for flat $\varepsilon{=}0$: the factored model encodes the action effect; flat averages it away.

4.5 Held-out action ranking

Does higher controllability help a downstream decision-maker? Given a held-out transition (z_t, z_{t+1}) with true action a_t, we score each candidate action by predicting $\hat{z} = \text{stepfn}(z_t, a_i)$ and picking $\arg\min_i \|\hat{z} - z_{t+1}\|$. On the 400k-frame multi-game setup ($V{=}18$ discrete actions, 3 seeds), the core reaches top-5 accuracy $0.372{\pm}0.002$ versus flat's $0.343{\pm}0.003$, an 8.5% relative gain (chance $5/18{=}0.278$); top-1 is similar across models ($\sim$0.074, chance 0.056). The difference concerns action identifiability, not closed-loop planning.

4.6 Distilling each component to one network evaluation

Beyond the recipe-level $K{=}1$ evaluation of Sec. 4.3, we distill each flow to one or few evaluations per component (Sec. 3.4; quality, speed, and σ sweep in App. H, Tables 25–26). With context noise $\sigma{=}0.1$ the distilled core student reaches 24.1 dB rollout PSNR, within 0.9 dB of its $K{=}4$ teacher, whereas the distilled flat student collapses ($22.1 \to 12.7$): in this aggressive regime the flat model becomes unstable while the core stays stable and controllable. On one L40 at batch 1, 1-NFE accelerates the latent step 3.5–$3.6\times$; end-to-end is decoder-bound, and the factored student (two evaluations per frame) runs at 75 FPS (teacher 69), both real time.

5 Conclusion

Standard short-range metrics miss long-horizon behavior. On two frozen pretrained latents, a a low-dimensional coarse code suffices for the tested one-step reconstruction criterion, the factorized predictor preserves controllability and latent scale during long free-running rollouts, and the larger-data validation retains its controllability advantage on unseen trajectories and improves held-out action ranking.

Limitations. These findings are a controlled architectural result rather than a claim about all world models. We study two frozen pretrained latents only; at multi-step budgets strong memory baselines match the core, and the clearest single-step advantage is joint controllability and radial stability (Sec. 4.3; App. K).

Reproducibility statement

The study uses two frozen public pretrained encoders (Oasis-500M's ViT-VAE and a Stable-Diffusion VAE) and public Minecraft and Atari recordings. Assets, fixed protocol choices, common architectures, and per-setting result tables are documented in Apps. L and B.

AI use statement

Generative AI tools were used for manuscript polishing and typesetting, organizing code and experimental results, and code standardization. The authors reviewed all AI-assisted work and take responsibility for the final manuscript.

References

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in Atari. *NeurIPS*, 2024.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *NeurIPS*, 2024.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. *NeurIPS*, 2021.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. oasis-model.github.io, 2024.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. *ICLR*, 2025.

Minghao Fu, Fan Feng, Nicklas Hansen, and Biwei Huang. Back to parsimonious latents: Learning task-centric world models from visual foundations. *arXiv preprint arXiv:2605.25620*, 2026.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *NeurIPS*, 2018.

Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *ICLR*, 2021.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Petr Ivashkov, Randall Balestriero, and Bernhard Schölkopf. Sensorimotor world models: Perception for action via inverse dynamics. *arXiv preprint arXiv:2606.20104*, 2026.

Ao Liang, Lingdong Kong, Tianyi Yan, Hongsi Liu, Yu Yang, Ziqi Huang, et al. Worldlens: Full-spectrum evaluations of driving world models in real world. In *CVPR*, 2026.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ICLR*, 2023.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. *ICLR*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. *ICCV*, 2023.

Tankred Saanum, Peter Dayan, and Eric Schulz. Simplifying latent dynamics with softly state-invariant world models. In *NeurIPS*, 2024.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion Schrödinger bridge matching. *NeurIPS*, 2023.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *ICML*, 2023.

Zhijiang Tang, Jiaxin Qi, Bing Zhao, and Jianqiang Huang. Long-code: Isolating pure long-context as an orthogonal dimension in video evaluation. *arXiv preprint arXiv:2604.17428*, 2026.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

Zaishuo Xia, Yukuan Lu, Xinyi Li, Yifan Xu, and Yubei Chen. Cloning deterministic worlds: The critical role of latent geometry in long-horizon world models. *arXiv preprint arXiv:2510.26782*, 2025.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. *CVPR*, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. *CVPR*, 2025.

Donggeun Yoon, Minseok Seo, Doyi Kim, Yeji Choi, and Donghyeon Cho. Probabilistic weather forecasting with deterministic guidance-based diffusion model. In *ECCV*, 2024.

Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? a decision-making-centric position. *arXiv preprint arXiv:2606.15032*, 2026.

Huijie Zhang, Aliaksandr Siarohin, Willi Menapace, Michael Vasilkovsky, Sergey Tulyakov, Qing Qu, and Ivan Skorokhodov. α-Flow: Understanding and improving MeanFlow models. *arXiv preprint arXiv:2510.20771*, 2025a.

Yu Zhang, Xingzhuo Guo, Haoran Xu, Jialong Wu, and Mingsheng Long. Consistent world models via foresight diffusion. *arXiv preprint arXiv:2505.16474*, 2025b.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *ICLR*, 2024.

Yang Zhou, Hao Shao, Letian Wang, Zhuofan Zong, Hongsheng Li, and Steven L. Waslander. Drivinggen: A comprehensive benchmark for generative video world models in autonomous driving. In *ICLR*, 2026.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

APPENDIX

The appendix is organized as follows. App. A recaps the dimension-dependent optimal-stochasticity criterion (a known result restated for completeness). App. B collects all common protocols and hyperparameters. App. C details the synthetic study; App. D the two real domains and their full sweeps; App. E the rollout study; App. F the ablations; App. G controllability; App. H 1-NFE distillation. App. K expands every limitation, and App. L states reproducibility details and scope.

Study scales and common protocol. The appendix reports results at two scales: a controlled small-scale study (2.1M params, 6k frames) that isolates architectural choices, and a scaled-up validation ($\sim$21M params, 400k frames, 4 Atari games) that tests held-out generalization. Throughout the appendix, "stratified" denotes the factored architecture of the main text; with $\varepsilon_c{=}0$ it is exactly the deterministic dynamics core, and Sec. 4.2 shows the stochastic variant ($\varepsilon_c{>}0$) ties it within noise. Two small-scale observations hold only without exposure-bias noise and are therefore not part of the main-text claims: the apparent divergence of flat $\varepsilon{=}0$ and the large PSNR gap. Under the σ-matched protocol the measured advantage is long-horizon controllability and radial stability, evaluated by latent norm, and it is due to the factorization rather than ε. The scale-up also reports held-out tracking (cos@500$=$0.89) and action ranking. Small-scale numbers are kept below for completeness.

A Dimension-dependent optimal stochasticity: the scaling argument

For completeness we restate the scaling argument that motivates rendering the high-dimensional appearance deterministically (Sec. 3.2). We give only the qualitative scaling needed to justify the deterministic appearance render; none of the paper's claims depend on a closed form for the threshold.

The criterion, qualitatively. For the ε-interpolant of Eq. (1), the conditional regression target retains an irreducible stochastic component whose expected squared norm scales linearly in the modeled dimension, $\Theta(\varepsilon\, d)$, because each of the d coordinates contributes an independent $\varepsilon\,\tau(1-\tau)$ of target variance. The curvature-smoothing benefit of bridge noise, by contrast, is governed by the scalar smoothing scale relative to the data radius and is therefore $O(1)$ in d (and shrinks with the sampling budget K). Balancing an $O(1)$ benefit against an $O(d)$ cost yields a threshold dimension $d^*(K)$ above which the deterministic choice $\varepsilon^*{=}0$ is preferred, with ε^* decreasing in both d and K. We do not derive a closed form for $d^*(K)$: the qualitative direction (that deterministic rendering is preferred in the high-dimensional appearance subspace) is all the architecture uses, and Sec. 4.1 verifies it empirically in both latent spaces.

Role of the criterion. We invoke this criterion only as motivation: the high-dimensional frame appearance ($d{\sim}10^3$) sits well above any plausible threshold, so it prefers $\varepsilon^*{=}0$ and can be rendered deterministically. We additionally measure the transition on our own synthetic and two real latent spaces (App. C, App. D) purely as a sanity check that the regime we operate in is the deterministic one; these measurements serve as motivation, not as a theoretical contribution. Crucially, the main text (Sec. 4.2) further shows that stochastic variants of the low-dimensional core leave long-horizon quality unchanged, so the whole model can be deterministic and ε is best read as a fidelity–diversity dial.

Remark 1 (Two distinct roles of "few steps"). The same criterion separates two senses of "few steps" that should not be conflated. As a sampling budget, smaller K leaves more discretization error for bridge noise to smooth and so raises ε^*. As a distillation target, a one-step consistency/flow-map student no longer integrates an ODE, so that smoothing benefit vanishes and the student prefers determinism ($\varepsilon^* \to 0$) even in low dimensions. The stochastic teacher's multi-step ceiling can nevertheless retain multimodality (App. H). We keep these two roles distinct throughout.

B Common protocols and hyperparameters

Metric. Sliced-Wasserstein-2 (SW2) uses 100 random unit projections; we report the root mean over projections of the 1D W_2^2 between sorted samples, on a held-out 20% split. Lower is better. SW2 is cheap, unbiased under projection, and avoids fitting a Gaussian, which matters for multi-modal transition codes.

ε-interpolant training. Velocity networks are SiLU MLPs with width/depth scaled by dimension (128–640 wide, 4–5 layers), Adam (10^{-3}) with cosine decay, $\tau \sim \mathcal{U}[0.05, 0.95]$, batch 512. The conditional target is Eq. (2) with per-coordinate ε (coordinates with $\varepsilon_i{=}0$ contribute only the deterministic drift, exactly). Sampling integrates $\dot{\mathbf{x}} = \mathbf{v}_\theta$ with K backward Euler steps over $[\tau_{\min}, \tau_{\max}]$; the ODE is ε-independent.

Hardware. Synthetic and pilot runs: a single RTX 4080S (16 GB). Real-latent extraction, rollout, and timing: 4× NVIDIA L40 (46 GB). Scaled-up experiments (∼21M params, 400k frames): 2× RTX 4090 (49 GB). Teacher latents come from a frozen Oasis-500M ViT-VAE (Minecraft) and a frozen Stable-Diffusion VAE (Atari); we never fine-tune the encoders.

Architectures. Table 4 lists the networks. All are residual-free SiLU MLPs; conditional flows concatenate $(\mathbf{x}, \tau, \text{cond})$, where the context is $(\mathbf{z}_{t-1}, \mathbf{a}_t)$ for the coarse/flat flows and $(\mathbf{z}_{t-1}, \mathbf{a}_t, c)$ for the fine flow (it is conditioned on the true coarse code during training and on the sampled coarse code at inference). The flow map additionally takes the target time s. The small-scale controlled study uses ∼2.1M parameters to isolate architectural choices; the scaled-up validation (Sec. 4.2) uses ∼21M parameters with gradient clipping and reduced learning rate to confirm the findings generalize.

Table 4: Network and optimization settings. "cap." indicates width/depth scaled with dimension (App. B).

component	input	width × depth	steps / opt.
synthetic velocity MLP	$(\mathbf{x}, \tau)$	128–640 × 4–5	6k–100k, Adam 10^{-3}, cos
conditional flow (coarse/flat)	$(\mathbf{x}, \tau, \mathbf{z}_{t-1}, \mathbf{a}_t)$	512 × 5	10k–12k, Adam 10^{-3}, cos
conditional flow (fine)	$(\mathbf{x}, \tau, \mathbf{z}_{t-1}, \mathbf{a}_t, c)$	512 × 5	10k–12k, Adam 10^{-3}, cos
flow map (distill)	$(\mathbf{x}, t, s, \text{cond})$	384–512 × 5	12k, Adam 10^{-3}, clip 1.0
Scaled-up setting (∼21M params, 400k frames, 4 Atari games)			
conditional flow (coarse/flat)	$(\mathbf{x}, \tau, \mathbf{z}_{t-1}, \mathbf{a}_t)$	1280 × 8	50k, Adam 3×10^{-4}, cos, clip 1.0
conditional flow (fine/render)	$(\mathbf{x}, \tau, \mathbf{z}_{t-1}, \mathbf{a}_t, c)$	1280 × 8	50k, Adam 3×10^{-4}, cos, clip 1.0

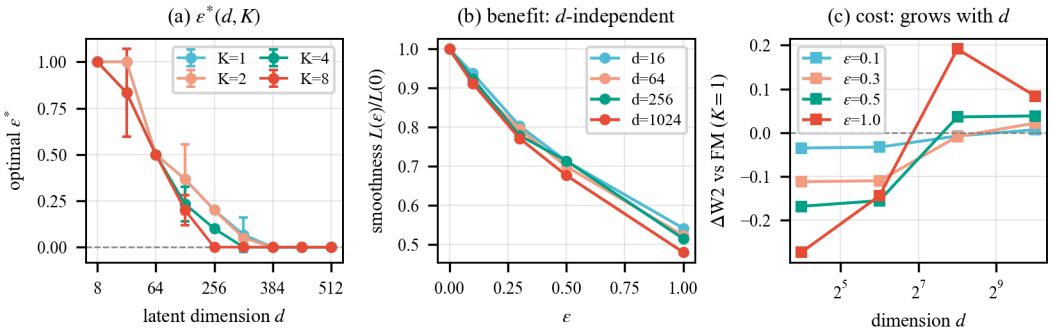

Figure 4: Synthetic dimension-dependent transition (a) and its mechanism (b,c). (a) $\varepsilon^*(d, K)\to 0$ above $d^* \approx 256$ to 320. (b,c) the ε benefit is d-independent while ΔSW2 develops a positive, cost-dominated slope as d grows, an $O(1)$ benefit losing to an $O(d)$ cost.

C Synthetic Study (Full)

Data and sweep. d-dimensional Gaussian mixture, 8 modes uniform on a sphere of radius 3, per-mode σ=0.3, 50k samples. Sweep $d \in \{8, 32, 64, 128, 256, 320, 384, 448, 512\}$, $\varepsilon \in \{0, 0.05, 0.1, 0.2, 0.3, 0.5, 1.0\}$, $K \in \{1, 2, 4, 8\}$, 3 seeds. One velocity MLP is trained per $(d, \varepsilon, \text{seed})$ and evaluated by SW2 at each K; $\varepsilon^*(d, K) = \arg\min_\varepsilon$. Figure 5 shows the per-ε SW2 curves and the (d, ε) heatmap underlying Table 5.

Stratified: modular vs. monolithic. We form a two-block stack: a 512-d "frame" block (ε^*=0) and a 32-d "action" block (ε^*>0). With separate modules (one network per block) the stratified assignment beats the best single global ε by $+22.5\%/12.1\%/4.7\%/2.1\%$ at K=1/2/4/8 (Fig. 6, Table 6). With a monolithic shared network the stratified assignment instead loses to flat ε=0, because the 32-d block is only $\sim 6\%$ of the loss and is drowned out. The negative result is useful: stratification requires modular capacity, which motivates the separate coarse/fine flows in the world model.

D Two Real Domains (Full)

D.1 Oasis-Minecraft (ViT-VAE)

Data. 1800 frames from three Minecraft recordings (open-oasis sample_data) are encoded by the Oasis ViT-VAE to $[1800, 9216]$ (576 tokens × 16), with one-hot 25-d actions (20 non-constant). The latent is strongly anisotropic: participation ratio ≈ 21, $90/95\%$ variance at d=461/683.

Frame transition (G1). Sweeping a top-d PCA (whitened to make the dimension axis comparable to synthetic), ε^* is 1 up to d=128, 0.3 at d=256, and 0 for $d \geq 512$ (Table 8, Fig. 8): $d^* \approx 256$–512, matching synthetic.

Table 5: Synthetic $\varepsilon^*(d, K)$. $\varepsilon^* \downarrow$ in d and K; $d^* \approx 256$ to 320.

d	K=1	K=2	K=4	K=8
8	1.0	1.0	1.0	1.0
64	0.5	0.5	0.5	0.5
128	0.5	0.5	0.1	0.1
256	0.2	0.2	0.1	0.0
320	0.0	0.05	0.0	0.0
512	0.0	0.0	0.0	0.0

Table 6: Synthetic stratified (d_{hi}=512, ε=0 | d_{lo}=32, ε^*) beats the best flat ε; gain largest at few steps.

	K=1	K=2	K=4	K=8
gain over best flat	+22.5%	+12.1%	+4.7%	+2.1%

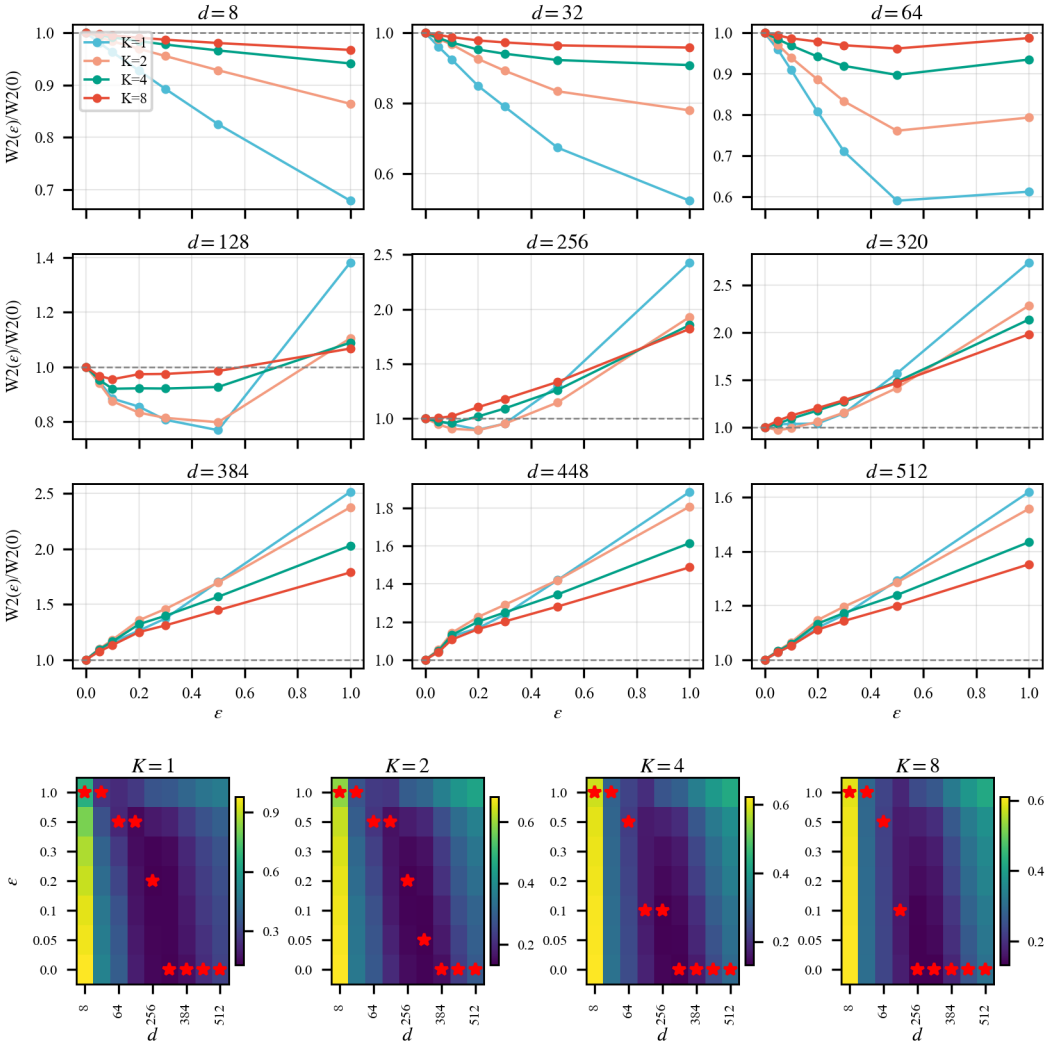

Figure 5: Synthetic study. (Top) relative SW2 vs. ε at several d (the minimizer ε^* migrates toward ε=0 as d grows). (Bottom) SW2 heatmaps over (ε, d) for $K \in \{1, 2, 4, 8\}$; the red star marks ε^*, which drops to ε=0 above d^*.

Action latent (G2). Standardizing the 20 non-constant action dimensions, SW2 falls monotonically in ε (0.61 $\rightarrow$ 0.45) so ε^*=1 at all K: a low-d code wants a Schrödinger bridge.

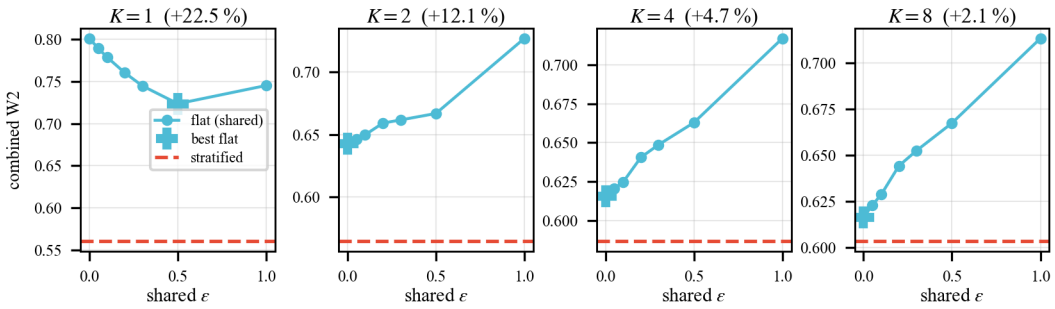

Figure 6: Synthetic stratified (modular). Combined cost of frame+action blocks vs. a shared flat ε; the dashed line (stratified, frame $\varepsilon{=}0$ | action ε^*) lies below every flat choice, with the largest margin at $K{=}1$.

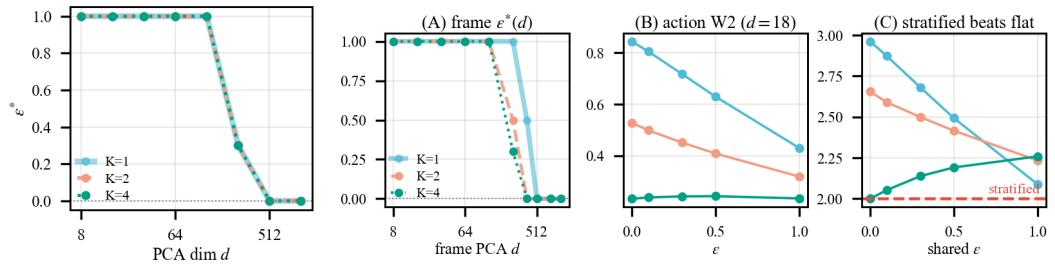

Figure 7: The dimension-dependent transition on two real pretrained latent spaces. Left: Oasis-Minecraft frame latents, $\varepsilon^*{=}1$ up to $d{=}128$, transitional at $d{=}256$, and $\varepsilon^*{=}0$ for $d \geq 512$ ($d^* \approx 256$ to 512). Right: Atari (SD-VAE), the frame latent shows the same drop of ε^* near $d \approx 256$ to 512 (A), the low-d action latent prefers $\varepsilon^*{>}0$ (B), and stratified beats the best flat ε (C); numbers in Table 7.

Stratified stack (G3). Frame module (PCA $d{=}1024$, $\varepsilon{=}0$) plus action module ($d{=}20$, $\varepsilon{=}1$), normalized combined cost: stratified beats the best flat ε by $+9.9\%/6.6\%/3.2\%$ at $K{=}1/2/4$ (Fig. 9).

D.2 Atari MsPacman (Stable-Diffusion VAE)

Data. 6000 frames resized to 256, encoded by an SD-VAE to ambient 4096, with 18 discrete actions (roughly balanced). Participation ratio ≈ 19.6, $90/95/99\%$ variance at $d{=}142/258/643$.

Frame transition (E7). On the natural-scale PCA (see the whitening caveat, App. K), the best ε is 1 up to $d{=}128$, transitional at $d{=}256$–384 ($\varepsilon^*_{K=4}{=}0.3$ at 256, 0 at 384), and 0 for $d \geq 512$ (Table 9), i.e. $d^* \approx 256$–512, consistent with Oasis and synthetic.

Action latent and stratified. The 18-d action latent has SW2 decreasing in ε, with $\varepsilon^*{=}1$ at $K{\leq}2$ (and 0 at $K{=}4$, where even the action code's discretization error is negligible). Stratified (frame $d{=}512$, $\varepsilon{=}0$ | action, ε^*) beats the best flat by $+4.2\%/10.5\%$ at $K{=}1/2$ (Fig. 7, right C; Fig. 10).

Full Atari frame sweep (real numbers). Table 10 reports the complete one-step ($K{=}1$) SW2 for the Atari frame latent over (d, ε) on natural scale (mean over 2 seeds). The row-wise minimum (bold) migrates from $\varepsilon{=}1$ at small d to $\varepsilon{=}0$ at $d \geq 512$, with the crossover at $d \approx 384$: the dimension-dependent transition in raw measurements.

Atari action sweep (real numbers). Table 11 gives the 18-d action-latent SW2 vs. ε for $K{=}1, 2, 4$ (mean over 3 seeds). At $K{\leq}2$ the optimum is $\varepsilon{=}1$ (Schrödinger bridge); at $K{=}4$, $\varepsilon{=}0$ and $\varepsilon{=}1$ are statistically tied (0.233 vs 0.234), consistent with $\varepsilon^* \downarrow$ in K.

Table 7: Two real world-model latent spaces both exhibit the transition. PR: participation ratio; d_{95}: 95%-variance dimension; ε_a^*: action-latent optimum; stratified gain over the best flat ε.

Domain	Encoder	ambient	PR	d_{95}	frame d^*	action (d, ε_a^*)	stratified gain
Synthetic GMM	–	–	–	–	256 to 320	(32, >0)	+22.5%/12.1%/4.7%
Oasis-Minecraft	ViT-VAE	9216	21.0	683	256 to 512	(20, 1.0)	+9.9%/6.6%/3.2%
Atari MsPacman	SD-VAE	4096	19.6	258	256 to 512	(18, 1.0)	+4.2%/10.5%

Table 8: Oasis frame-latent ε^* vs. retained PCA dimension (whitened).

d	8	32	128	256	512	1024
ε^* (K=1–4)	1.0	1.0	1.0	0.3	0.0	0.0

E AUTOREGRESSIVE ROLLOUT (FULL)

Sequences. Oasis recordings are sub-sampled at stride $\in \{1, 3, 6, 9\}$, encoded, and stored as ordered $[T, 576, 16]$ latents plus 25-d actions. We learn $\mathbf{v}(\mathbf{z}_t, \tau \mid \mathbf{z}_{t-1}, \mathbf{a}_t)$ in a top-d_z PCA basis (natural scale, single global normalization). Larger stride = larger inter-frame jumps = more stochastic/multi-modal transitions.

Sequence statistics. Table 16 lists the frame and transition counts per stride and controllability split. The natural-frame-count protocol uses all available frames per stride (so larger strides have fewer frames); the controlled protocol caps every stride to N=134 frames to remove the data-quantity confound. Test transitions for the stochastic strides are deliberately small, so we emphasize trends and AUC over single SW2 digits.

Stride sweep: two protocols. The natural-frame-count version reproduces existing numbers and the monotone trend (Table 14; σ=0 rollout curves in Fig. 11); the controlled N=134 version fixes frames per stride to isolate stochasticity from data quantity. In the controlled version the stratified model wins at every stride including stride 1; e.g. free-running rollout W2 drops from 3.83 (flat) to 0.41 (stratified) at stride 1, where the 256-d flat flow is hardest to fit from few samples. This reveals a second source of advantage (low-d sample efficiency) beyond transition stochasticity, which we state as a confound (App. K); the natural-count protocol, which has abundant stride-1 data, isolates the stochasticity axis and is the one used in the main text.

Exposure bias and context noise (E13). Free-running flat ε=0 rollout diverges: the latent norm explodes (unclipped stride-1 SW2 reaches $\sim 10^8$). Adding Gaussian noise of std σ to the conditioning past latent during training (Table 17, Fig. 12) monotonically raises the long-horizon cosine asymptote ($0.03 \to 0.31$ at H=50) and removes the distribution blow-up, with a distribution-matching sweet spot at $\sigma \approx 0.05$. This isolates exposure bias as the cause of flat divergence and a simple cure.

Decoded video (E12). Rollout latents are inverse-PCA'd to $[576, 16]$ and decoded by the frozen VAE (d_z=512, H=32, stride 1). The PCA-512 ceiling is PSNR 33.4 / LPIPS 0.172. Full numbers (Table 15): free-running PSNR 24.7 (strat) vs. 14.8/14.6 (flat 0/1); teacher-forced 29.1/28.9/27.4. LPIPS uses AlexNet; FVD is omitted (App. K). Figure 13 shows the decoded montage.

E.1 MULTI-DOMAIN K=1 FRONTIER (EIGHT ATARI GAMES)

We replicate the exact K=1 protocol of Table 3 on seven additional Atari games, eight environments in total, checking that the real-time separation is not specific to one domain. Each game is collected fresh (12,000 frames, 2× the original, uniform-random full-18-action ALE rollouts, same frozen SD-VAE 4096-d latent), and run through the identical capacity-matched K=1 pipeline (strict temporal 80/20 held-out, σ=0.1, H=1000, 3 seeds). The eight latents span a wide range of effective dimension (participation ratio 9.5 to 91.4), so this is a genuine generality test rather than eight near-identical copies.

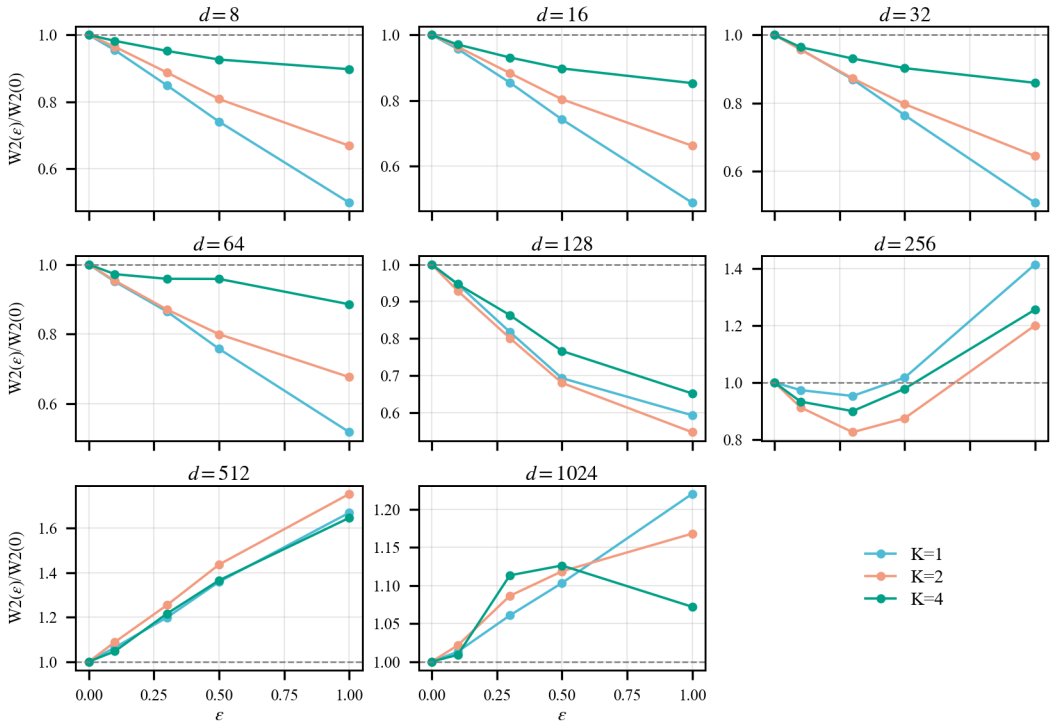

Figure 8: Oasis-Minecraft frame latent: relative SW2 vs. ε at several retained PCA dimensions (<1 means $\varepsilon>0$ beats FM). The minimizer sits at $\varepsilon=1$ for small d, forms a U at $d=256$, and moves to $\varepsilon=0$ for $d\geq 512$.

Table 9: Atari frame-latent best ε vs. retained PCA dimension (natural scale).

d	8	64	128	256	384	512	768	1024
best ε ($K=4$)	1.0	1.0	1.0	0.3	0.0	0.0	0.0	0.0

Across the eight games, the joint response/radial-stability pattern holds in $7/8$ games across a $10\times$ range of latent dimensionality. The lone negative is BeamRider (PR$=9.5$): when the controllable dynamics already fill almost the entire low-d latent, the learned bottleneck has little to gain and a fixed PCA core is better, which we report. Controllability variance is non-trivial on SpaceInvaders (± 4.8) and Qbert (± 5.0), so there the win rests partly on the radial-stability separation rather than raw controllability alone.

Full rollout suite across the eight games (drift and scale). For completeness we report the remaining free-running metrics at $H=1000$ for the same runs (Table 19): marginal drift (sliced-W2 to the real next-latent marginal) and the mean latent norm (scale). The picture is more mixed than the ctrl/norm-threshold separation. (i) Drift. The core attains the lowest drift among the radially stable models on the higher-dimensional latents (Pong 0.39, Seaquest 0.46, Asterix 0.50, Breakout 0.59) and a lower mean drift overall (0.63 vs. flat 0.80), but it is over-energetic on the two lowest-PR latents, where its drift is worse than the flat baseline (BeamRider 0.74 vs. 0.32; MsPacman 0.85 vs. 0.65), the same over-energetic free-running behaviour flagged in Sec. 5. (ii) Scale corroborates divergence. The diverging memory baselines inflate the latent norm to 20–30 (RSSM) versus the core's stable 12–15, so the OOD and norm signals agree where a model's scale blows up. The clean reported pattern is therefore the joint controllability/radial-stability separation; drift and scale are reported here in full and read as corroborating, not as an additional clean win.

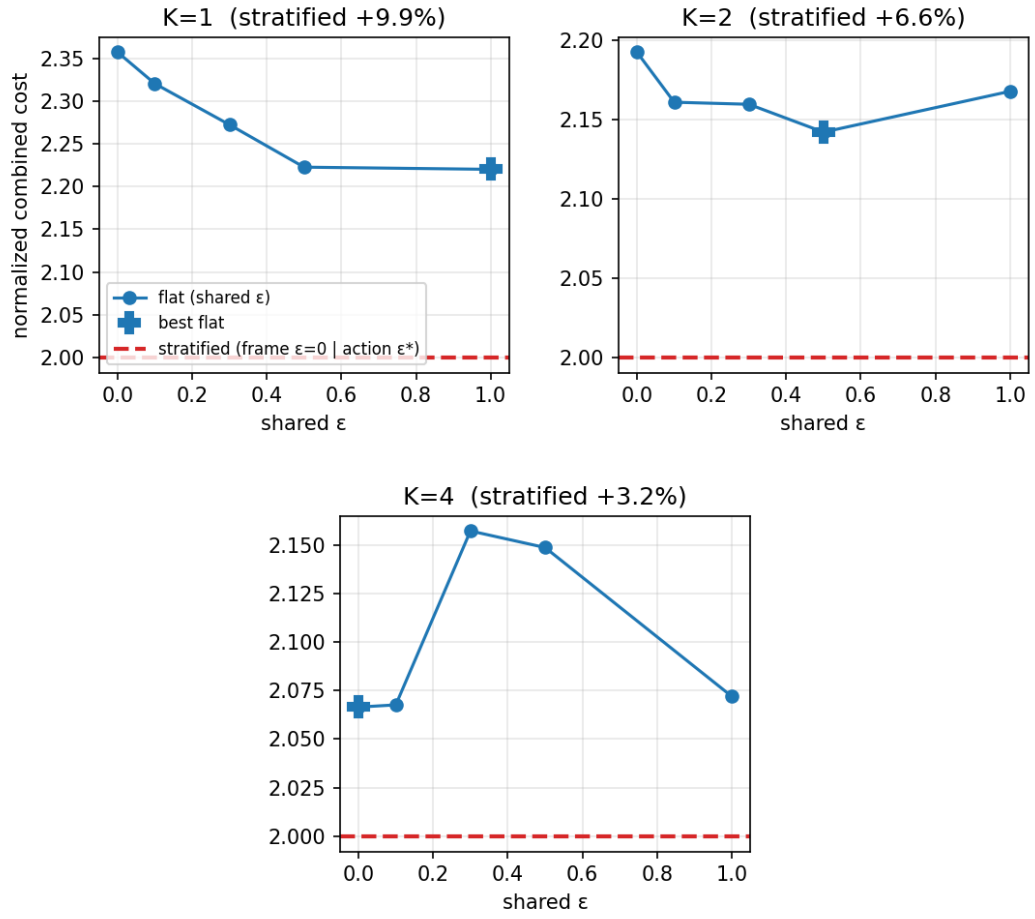

Figure 9: Oasis-Minecraft stratified stack: normalized combined cost of the stratified stack (frame ε=0 | action ε=1, red dashed) vs. the best flat ε at each K; stratified lies below every flat choice.

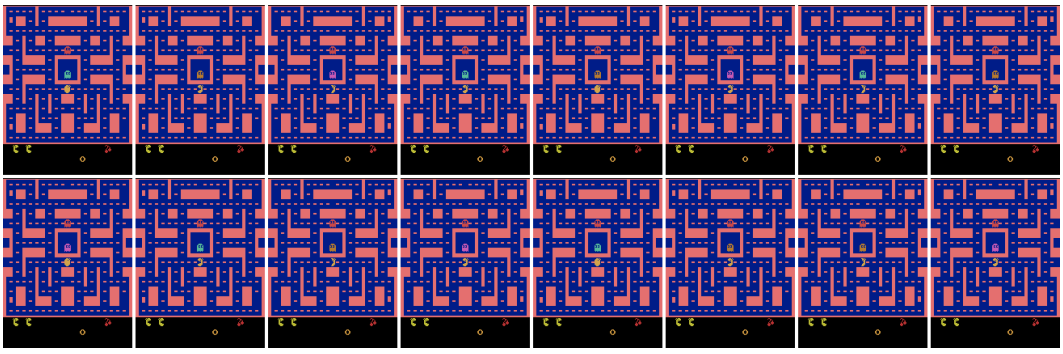

Figure 10: Second-domain sanity check. The Stable-Diffusion VAE latent we study in the Atari phase-transition experiment (Sec. 4.1) decodes to genuine MsPacman frames (maze walls, pellets, the agent, coloured ghosts, and the lives/score row), confirming we operate on a real pretrained world-model latent space, not a synthetic toy. 16 frames sampled along one episode.

F Rollout ablations (full)

Table 20 is the complete stride-6 ablation (K=4 unless noted; rollout W2 clipped at 3× the real norm; real diversity scale ≈0.27), summarized by Fig. 14. Reading it: flat ε=0 diverges in rollout (W2

Table 10: Atari frame latent, one-step (K=1) SW2 vs. (d, ε), natural scale. Row minimum in bold; the optimum crosses from ε=1 to ε=0 near $d \approx 384$.

$d \setminus \varepsilon$	0.0	0.1	0.3	0.5	1.0
8	0.812	0.774	0.689	0.596	**0.372**
16	0.787	0.754	0.675	0.586	**0.379**
32	0.697	0.674	0.612	0.543	**0.347**
64	0.659	0.631	0.565	0.496	**0.326**
128	0.522	0.502	0.453	0.394	**0.255**
256	0.291	0.282	0.256	0.222	**0.177**
384	0.203	0.196	0.180	**0.172**	0.177
512	**0.188**	0.188	0.190	0.193	0.204
768	**0.190**	0.192	0.194	0.196	0.201
1024	**0.201**	0.202	0.203	0.204	0.207

Table 11: Atari action latent (d=18), SW2 vs. ε at each budget K. Row minimum in bold.

$K \setminus \varepsilon$	0.0	0.1	0.2	0.3	0.5	1.0
1	0.844	0.805	0.762	0.718	0.630	**0.430**
2	0.528	0.499	0.475	0.451	0.409	**0.319**
4	**0.233**	0.238	0.241	0.242	0.243	0.234

1.29, 23% OOD); flat ε=1 minimizes one-step SW2 but over-disperses (diversity 0.526); widening d_c monotonically stabilizes rollout (W2 $0.89 \to 0.60 \to 0.47$); ε_c is mild; and K=4 is necessary (K=1 diverges for both flat and stratified). The factored ("stratified") configuration is balanced across one-step accuracy, rollout stability, diversity scale, and decode sharpness; as the capacity-matched analysis of Sec. 4.2 shows, this balance comes from the low-d factorization, and the deterministic core (ε_c=0) matches it, so ε_c is a dial, not the source of the advantage.

Conditional diversity calibration. A subtle but important axis is whether a model's per-state future diversity matches the real data, rather than being merely large. We measure conditional diversity as the mean per-dimension standard deviation of M sampled futures for a fixed $(\mathbf{z}_{t-1}, \mathbf{a}_t)$, and compare to the real transition's scale (≈ 0.27 on stride-6). Table 20 shows the calibration story: flat ε=0 is slightly under-dispersed (0.284) and rolls out unstably; flat ε=1 is badly over-dispersed ($0.526 \approx 2\times$ real), trading fidelity for noise; the stratified model is the closest to real (0.270–0.280) while remaining stable. Therefore, "beating any flat ε" must be read on the full set of axes (one-step accuracy, rollout stability, diversity calibration, perceptual quality), not on one-step SW2 alone: flat ε=1 can match stratified on the latter while failing diversity calibration and video quality. The stride-sweep diversity panel (Fig. 2, center) shows the same behavior across strides: stratified tracks the real diversity curve, flat ε=0 runs below it.

The two K-axes, restated. Two ablation rows make the distinction in App. A concrete. As a sampling budget, smaller K both raises ε^* (Table 5) and destabilizes rollout (K=1 diverges, Table 20); these are different failure modes of too-coarse integration. As a distillation target, the one-step student prefers determinism ($\varepsilon^* \to 0$): the stochastic coarse teacher supplies multi-modality, and the student merely reproduces its multi-step rollout in one evaluation. Conflating the two would wrongly predict that one-step models should use large ε.

G ACTION CONTROLLABILITY (FULL)

Protocol. d_z=256, d_c=16, ε_c=1, K=4, 12k steps, natural-scale PCA. Two regimes: stride-1 (near-deterministic, N=2997) and stride-6 (stochastic, N=597). Actions are 25-d (20 non-constant), with $\sim$14–19 key classes by argmax.

(a) Counterfactual divergence. Fixing $\mathbf{z}_{t-1}$ and the sampling noise, we compare the prediction under the true action vs. an alternative real action (action effect), against a noise floor (same action,

Table 12: Cross-architecture baselines, strictly held-out (Atari temporal $80/20$; Oasis leave-one-sequence-out), capacity-matched ($\sim$2.1M), σ=0.1, H=1000. CTRL=counterfactual controllability (↑); W2=marginal drift (↓). Memory baselines (RSSM, attention) close most of the gap to the core and surpass it on Oasis; the multi-step objective drives the gain (a history-conditioned flat model without it does not improve, see text).

	flat	PCA-core	**core (ours)**	flat-ms	RSSM	hist+attn
Atari CTRL↑	2.4	2.7	**16.3**	8.0	9.9	8.6
Atari W2↓	0.70	**0.64**	0.83	0.66	0.80	0.79
Oasis CTRL↑	4.6	3.9	14.7	7.9	13.6	**14.4**
Oasis W2↓	0.96	1.19	0.89	0.87	0.62	**0.51**

Table 13: Strict held-out validation. Rollout start points lie in trajectory regions never seen in training (Atari: temporal 80/20 split of the single 6k-frame sequence; Oasis: leave-one-sequence-out), and the PCA basis is fit on training frames only. Counterfactual controllability (CTRL@H, H=1000) and ground-truth tracking (COS@h) as mean±std over 3 seeds (Atari) / 3 folds×3 seeds (Oasis), beside the original non-held-out values where the eval overlapped training. All models capacity-matched ($\sim$2.1M params).

		flat	PCA-core	learned-core (1-step)	learned-core (MS)
Atari	CTRL@H (held)	2.5±0.4	2.7±0.3	10.4±0.7	**16.3±0.1**
	COS@50 (held / orig)	0.09/0.02	0.08/0.04	0.19/0.49	0.29/0.99
	COS@200 (held / orig)	0.10/0.02	0.09/0.02	0.01/0.13	−0.02/0.95
Oasis	CTRL@H (held)	3.5±1.2	3.3±1.4	7.7±4.1	**14.5±5.3**
	COS@50 (held / orig)	0.03/0.18	0.02/0.31	0.06/0.56	0.13/1.00
	COS@200 (held / orig)	0.04/0.16	−0.01/0.09	0.01/0.33	0.05/1.00

resampled noise). All models respond (action effect is 30–42% of a real single-step change); Table 21. The stratified coarse code (16-d) carries ≈51% (stride-6) / 53% (stride-1) of the per-dimension action sensitivity despite being 6% of the dimensions. The action/noise ratio favors flat ε=0 (narrower conditional spectrum); since the ratio conflates responsiveness with conditional sharpness, we rely on action-effect magnitude and the identifiability probe below.

(b) Action identifiability probe. A ridge probe predicts the action from $(\mathbf{z}_{t-1}, \mathbf{z}_t^{\text{pred}})$; the $\mathbf{z}_{t-1}$-only baseline controls scene→action correlation and the real transition $(\mathbf{z}_{t-1}, \mathbf{z}_t^{\text{real}})$ is the upper bound. Stride-1 is saturated (baseline AUC already 0.953, low discriminability). Stride-6 (the decisive stochastic regime) separates the models: stratified AUC 0.912 (+0.020 over baseline), flat ε=0 0.892 (= baseline, no gain), approaching real 0.927 (Table 23, Fig. 15). We also report, as a caveat, a delta probe (input $\mathbf{z}_t^{\text{pred}} - \mathbf{z}_{t-1}$) where flat can exceed real because a model whose output is a cleaner deterministic function of the action is more linearly decodable than the noisy real transition; the delta probe therefore does not rank realism and is not used as a headline (Table 22). On the delta probe the real transition even scores below chance ($\text{AUC} = 0.31$) because its action-conditional displacement is the noisiest, confirming the probe measures decodability, not fidelity; the context probe of Table 23 (with real as a proper upper bound) is the correct identifiability metric.

Saturation in the near-deterministic regime. Figure 16 shows the stride-1 counterpart. Here transitions are nearly deterministic, so the action is already almost fully predictable from the scene alone: the $\mathbf{z}_{t-1}$-only baseline probe reaches AUC 0.953, leaving little room for any model to add. All models cluster near this ceiling (flat ε=0 0.950, factored 0.953, real 0.967), so stride-1 cannot discriminate controllability. The more stochastic stride-6 regime (Table 23) is therefore the decisive test, and it is consistent with the main-text finding that the factored model's controllability advantage is largest where transitions are most multi-modal (and, per Sec. 4.2, stems from the factorization rather than from core stochasticity).

Table 14: Short-range/static metrics tie (Oasis, capacity and σ-matched, K=4): one-step SW2, teacher-forced PSNR, and Fréchet R3D-18 distance.

	SW2↓	TF-PSNR↑	Fréchet R3D↓
flat ε=0	0.306	**25.8**	**155**
flat ε=1	**0.258**	23.6	200
core	0.282	24.8	157

Table 15: Long-horizon (H=1000) separates them: drift W2 (vs. real marginal), controllability (counterfactual response), norm (real≈ core).

dom.	model	drift↓	ctrl↑	norm
Oasis	flat ε=0	0.87	4.1	7.0
	core	**0.56**	**5.1**	9.2
Atari	flat ε=0	0.78	2.1	7.2
	core	**0.67**	**2.9**	9.2

Table 16: Sequence/transition counts (Oasis). "nat." = natural-frame-count protocol; "ctrl." = controlled N=134.

stride	1	3	6	9
frames (nat.)	1000	400	200	134
frames (ctrl.)	134	134	134	134
controllability N (transitions)	2997	–	597	–

H 1-NFE DISTILLATION (FULL)

Distiller. A single flow map $F_\theta(\mathbf{x}, t, s \mid \mathbf{z}_{t-1}, \mathbf{a}_t)$ learns the average velocity, trained with the FM anchor plus warmed-in consistency of Eq. (4), finite-difference total derivative (δ=10^{-3}), λ_c=0.1 ramped over the first 40% of training, gradient clip 1.0, 12k steps. One-step: $\mathbf{z}_t = \mathbf{x}_1 - F_\theta(\mathbf{x}_1, 1, 0 \mid \mathbf{z}_{t-1}, \mathbf{a}_t)$. Flat = 1 evaluation/frame; stratified = 2 (coarse then fine). The always-on FM anchor is essential: bootstrapping the map from scratch diverges, since these recipes were designed to fine-tune a pretrained field.

Static-latent distillation (pilot). On static frame/action latents, a naive regression distiller is worse than the raw one step on the high-d frame ($0.362 > 0.289$) and is hurt by whitening, whereas the flow-map self-distiller on natural scale gives the numbers in Table 24: action 0.310 (better than its own multi-step) and frame 0.265 (better than raw one step, +0.047 from the multi-step ceiling). This confirms that the dimension-aware ε choice matters more than distillation tricks: a stochastic action teacher's raw one step is already near its ceiling, while the deterministic frame must be distilled.

Rollout distillation. Config d_z=512, d_c=32, stride-1, H=32, teacher K=4. A one-step student diverges without context noise (OOD@H=100%, W2@$H \approx 4.4$); the σ sweep (Table 25) shows σ=0.1 gives student-strat PSNR 24.1 (≈ teacher 25.0), σ=0.2 erases the gap (student ≥ teacher) at the price of mild over-smoothing. Speed is in Table 26: 3.5–3.6× latent-step speedup, end-to-end decoder-bound at ∼75 FPS (decode-only 12.3 ms ≈81 FPS); full per-model throughput is in Table 26. Figure 17 plots quality and latency, and the montage of Fig. 18 confirms the one-step stratified student tracks the multi-step teacher while flat students collapse.

I EXTENDED METRIC DISCUSSION

SW2 for latent distributions. For the static and rollout-distribution studies we need a distance between empirical distributions of (possibly multi-modal) latents that (i) does not assume a parametric form, (ii) is cheap enough to evaluate inside sweeps, and (iii) is sensitive to mode structure rather than only first/second moments. The Sliced-Wasserstein-2 distance satisfies all three: it averages the exact one-dimensional W_2 over random projections, is an unbiased estimator of a metric, and, unlike a Gaussian-fit Fréchet distance, does not collapse the multi-modality of the transition code. We use 100 projections; variance across projection seeds is small relative to the effects we report.

Pixel metrics and Fréchet-feature coverage. For decoded rollout video we report PSNR (no learned weights) and LPIPS with an AlexNet backbone (the only perceptual network reachable in

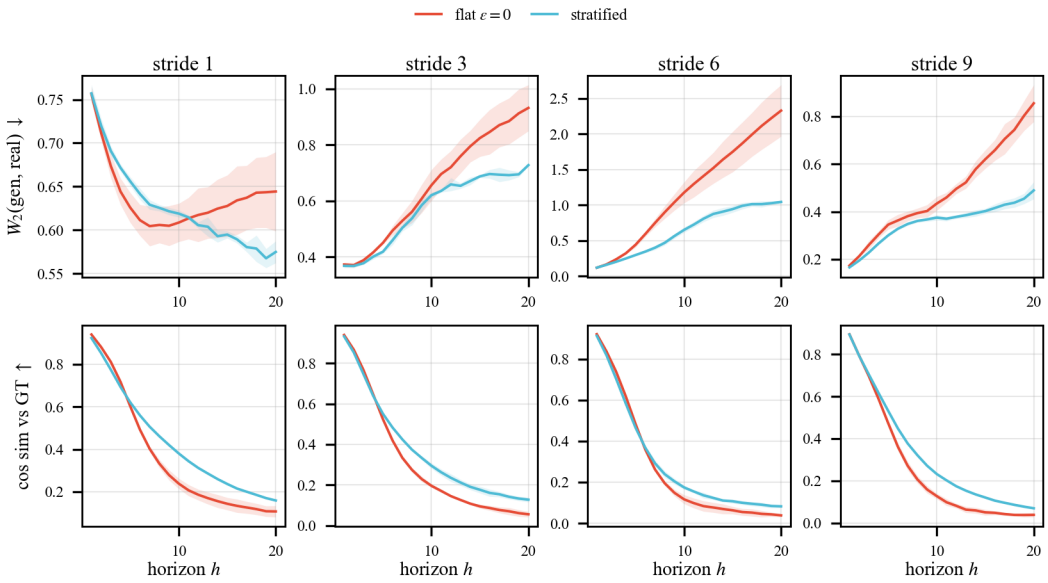

Figure 11: Free-running rollout vs. horizon at each stride, at σ=0 (no exposure-bias noise; superseded by the σ-matched protocol of the main text). Distribution distance to the real marginal (upper four panels) and trajectory drift (lower four panels). Without context noise, flat ε=0 (red) degrades and its latent norm grows (clipped at 3× the real norm to plot); with the fair σ of the main text no model diverges, and the core's durable advantage there is controllability, not averting this σ=0 blow-up.

Table 17: Exposure-bias context noising (Oasis stride-1, H=50).

ctx σ	0.0	0.05	0.1	0.2
cos@H=50 ↑	0.03	0.09	0.22	**0.31**
W2@H=50 ↓	∼1.0–1.5 (div.)	**0.76**	0.87	0.87

our environment). PSNR rewards mean-reverting blur: a model that regresses to the conditional mean scores well under teacher forcing, which is so the teacher-forced PSNR gaps between flat and stratified are small (Table 15) while the free-running gap, driven by stability rather than per-frame fidelity, is large (24.7 vs 14.7 dB). LPIPS, being perceptual, better reflects the visible collapse of the flat rollouts (0.39 vs 0.65). The short-range comparison in Table 14 uses Kinetics-pretrained R3D-18 features and is therefore a Fréchet R3D-18 distance, not standard I3D-FVD. No Fréchet video metric is reported for the decoded rollout experiment.

Per-frame profiles. The free-running PSNR is a per-frame mean over the H=32 horizon; the flat models start near the ceiling and decay monotonically as exposure-bias error accumulates, whereas the stratified model (and its 1-NFE student) stays in the 20–30 dB band throughout (Table 27). This is the per-frame view of the free-running collapse: the decisive difference is accumulation over horizon, not single-step fidelity.

Latent spectra. Both real latents are strongly anisotropic, so a low-d coarse code can capture the transition while the leading appearance directions are far above d^* (Table 28). The participation ratio $\mathrm{PR} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ (λ_i the PCA eigenvalues) is ≈ 20 for both, an order of magnitude below the ambient and 95%-variance dimensions, so the effective dynamics live in a small subspace even though the appearance code is high-dimensional.

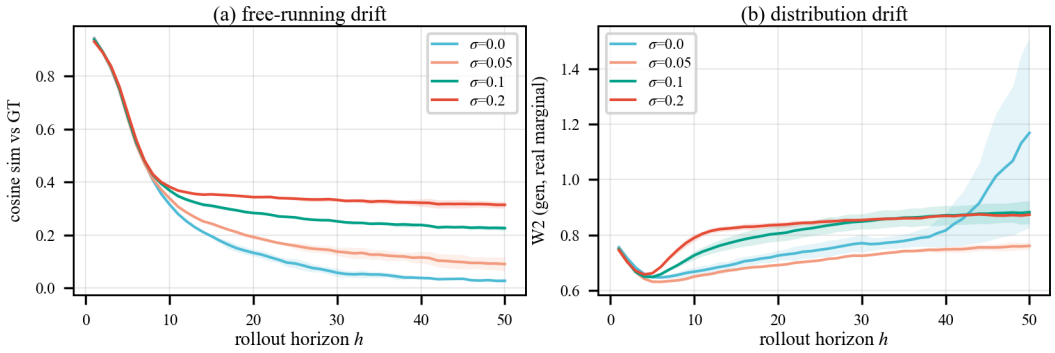

Figure 12: Exposure-bias context noising. Left: free-running trajectory drift (cosine vs. ground truth) over horizon; larger σ raises the asymptote. Right: distribution distance over horizon; $\sigma{=}0$ blows up, $\sigma{\approx}0.05$ is best for distribution matching.

Table 18: Multi-domain single-step ($K{=}1$) frontier (ctrl@1000 $\uparrow$ / norm-threshold exceedance@1000 $\downarrow$; 3 seeds; games ordered by participation ratio PR). The low-d core is the best joint response/radial-stability model in $7/8$ games with zero rounded exceedance in every domain (mean: core 0.00, RSSM 0.72, attention 0.22). BeamRider (lowest PR) is the single exception, where the fixed PCA core suffices. ∞ marks a controllability blow-up.

game (PR)	flat	PCA-core	**core (ours)**	RSSM	hist+attn
BeamRider (9.5)	5.1/0.00	11.5/0.00	4.5/0.00	8.5/0.00	2.6/0.00
Qbert (10.8)	∞/0.67	∞/0.01	**11.6/0.00**	3.5/0.39	6.6/0.09
SpaceInvaders (15.2)	2.2/0.00	4.3/0.02	**9.7/0.00**	1.2/1.00	6.9/0.25
MsPacman (19.6)	1.6/0.00	1.4/0.00	**16.5/0.00**	9.2/0.98	9.9/0.28
Breakout (20.4)	6.3/0.00	9.8/0.00	**15.4/0.00**	12.2/0.93	10.6/0.05
Pong (22.0)	3.0/0.01	∞/0.00	**17.5/0.00**	20.1/0.61	7.6/0.05
Seaquest (44.4)	9.8/0.00	3.0/0.00	**18.9/0.00**	17.9/0.94	8.6/0.02
Asterix (91.4)	2.0/0.00	1.9/0.00	**16.7/0.00**	15.5/0.88	2.9/1.00

J DESIGN RULE AND PRACTITIONER RECIPE

We distill the study into a recipe for adding dimension-aware stochasticity to a continuous (diffusion/flow) world model.

1. **Locate the scales.** In the frozen encoder's latent, identify a low-d transition/action code (here, the leading PCA coordinates and the action) and the high-d appearance remainder. The effective dimension (participation ratio) and the $d^* \approx 256$–512 landmark tell you which side of the transition each block is on.

2. **Assign ε by dimension.** Set $\varepsilon{=}0$ (Flow Matching) on the high-d appearance block and $\varepsilon{>}0$ (toward a Schrödinger bridge) on the low-d transition block. A coarse code width d_c of a few tens and $\varepsilon_c \in [0.3, 1]$ work well; wider d_c buys rollout stability at little one-step cost.

3. **Use modular capacity.** Realize the two blocks with separate conditional flows (coarse first, fine conditioned on the sampled coarse code); a single shared network erases the benefit.

4. **Use $K \geq 4$ Euler steps** for the multi-step teacher; $K{=}1$ rollouts diverge.

5. **For real time, distill to a flow map** with an always-on FM anchor and warmed-in consistency, and add exposure-bias context noise ($\sigma \approx 0.1$) to the conditioning latent only. Expect the latent step to drop to one (flat) or two (stratified) evaluations and the bottleneck to move to the decoder.

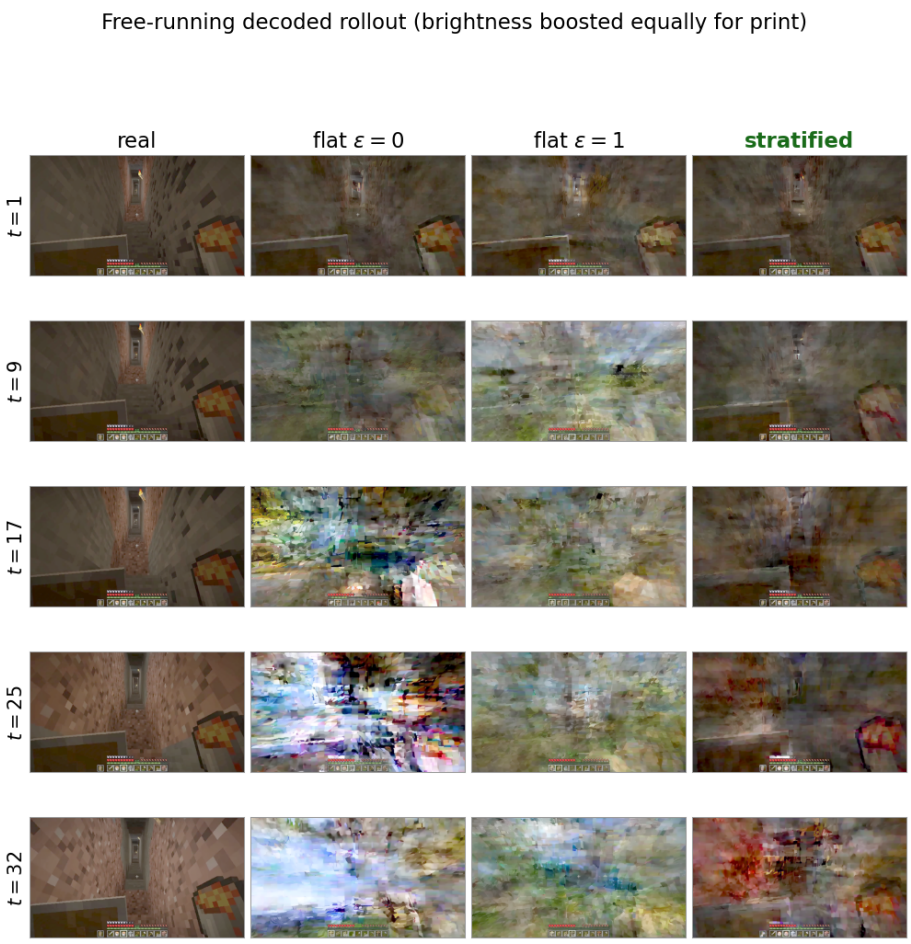

Figure 13: Free-running decoded rollout at σ=0 (rows: rollout step t; columns labelled above; superseded by the σ-matched protocol of the main text). Without exposure-bias noise both flat models accumulate error and dissolve, whereas the factored model keeps the cave geometry and HUD bar (24.7 vs. 14.7 dB PSNR); under the fair σ-matched setting of the main text this gap largely closes and decoded PSNR is comparable, so the figure illustrates the σ=0 failure mode, not a durable PSNR advantage. The Minecraft footage is genuinely dark (cave scenes), so for print we apply one identical brightness (γ=0.5) transform to all panels; this is a display-only change and alters no content.

K Additional limitations

1. **Fréchet video coverage is partial.** The short-range table uses Kinetics-pretrained R3D-18 features, not standard I3D-FVD. The rollout comparison uses PSNR, LPIPS, and latent-space drift and makes no FVD claim.

2. **PCA proxy / small test sets.** Real-latent studies operate on a top-d PCA of the frame latent with a finite number of frames; stride-6/9 test sets are $\sim$90–120 transitions, so we emphasize AUC and monotone trends over single SW2 digits, and rollout shares the train transition distribution identically for flat and stratified (a fair comparison).

3. **Whitening artifact and the natural-scale fix.** Early sweeps whitened the top-d PCA to align the dimension axis with synthetic data; whitening inflates low-variance tail directions, amplifying noise in distillation (the naive frame distiller failure) and shifting apparent transition locations. All rollout, controllability, and distillation results use natural-scale PCA (single global normalization, no per-dimension whitening), which we recommend in practice.

Table 19: Full-suite free-running drift and scale at $H=1000$ across the eight Atari games (marginal-W2 drift ↓ / latent norm; 3 seeds; games by PR). The core has the lowest drift on the higher-PR latents and the lowest mean drift (0.63 vs. flat 0.80), but is over-energetic on BeamRider/MsPacman; RSSM's inflated norm (20–30) corroborates radial instability (Table 18).

game (PR)	flat	PCA-core	core (ours)	RSSM	hist+attn
BeamRider (9.5)	0.32/14.2	0.30/14.6	0.74/7.1	0.76/20.1	0.74/19.3
Qbert (10.8)	2.09/∞	0.71/∞	0.73/13.2	1.15/24.3	0.76/21.0
SpaceInvaders (15.2)	0.75/12.3	0.46/∞	0.74/15.2	1.04/26.0	0.73/20.0
MsPacman (19.6)	0.65/9.0	0.52/10.7	0.85/15.5	1.26/26.8	0.75/20.2
Breakout (20.4)	0.71/11.1	0.61/13.0	0.59/12.1	1.37/26.6	0.76/19.0
Pong (22.0)	0.58/9.5	0.51/∞	0.39/13.7	0.85/23.6	0.46/15.6
Seaquest (44.4)	0.62/10.0	0.66/10.5	0.46/13.8	1.20/28.6	0.52/15.4
Asterix (91.4)	0.68/7.0	0.70/7.6	0.50/12.0	0.91/22.8	1.09/29.9

Table 20: Complete rollout ablation (Oasis stride-6). ↓/→real as indicated.

config	one-step SW2↓	rollout W2@H ↓	OOD frac@H ↓	diversity ($\to 0.27$)
flat ε=0 (K=4)	0.258	1.291	0.23	0.284
flat ε=1 (K=4)	**0.233**	0.480	0.04	0.526 (over)
stratified d_c=16, ε_c=1 (K=4)	0.236	0.604	0.06	0.280
stratified d_c=8	0.260	0.890	0.13	0.282
stratified d_c=32	0.234	**0.471**	**0.01**	0.270
stratified ε_c=0.3	0.238	0.647	0.07	0.263
stratified ε_c=0.5	0.237	0.622	0.07	0.268
stratified K=1	0.297	3.358	0.81	0.505
stratified K=2	0.243	0.996	0.15	0.319
flat ε=0 K=1	0.368	3.802	1.00	0.506
flat ε=0 K=2	0.274	2.106	0.57	0.308

4. **Data-efficiency confound.** The controlled equal-frame stride sweep shows a stratified advantage at stride 1 too, so the benefit has two sources: genuine transition stochasticity and the sample efficiency of a low-d code. We state both rather than attributing the entire gain to stochasticity.
5. **1-NFE needs context noise.** One-step rollout is stable only with exposure-bias context noise; σ=0.05 still diverges, σ=0.1 leaves a 0.9 dB gap, σ=0.2 closes it but over-smooths, producing a stability–sharpness trade-off.
6. **Decoder-bound end-to-end speedup.** The 3.5× acceleration is in the latent step; the frozen VAE decoder dominates end-to-end FPS. The contribution is that the world model ceases to be the bottleneck; further speedups need a lighter decoder or batched decoding.
7. **Residual static frame 1-NFE gap.** A +0.047 SW2 gap remains for the one-step high-d frame map vs. its multi-step ceiling; the present experiments do not test whether an auxiliary marginal-velocity alignment term would narrow it.
8. **Two K-axes.** "Few steps" means two different things: the sampling budget K (smaller ⇒ larger ε^*) and the distillation budget (one-step student ⇒ $\varepsilon^* \to 0$, Remark in App. A). We keep them distinct throughout.

L Reproducibility, assets, and scope

Assets. The study uses two frozen pretrained encoders as latent-space sources: the Oasis-500M ViT-VAE (Minecraft) and a Stable-Diffusion VAE (Atari MsPacman). We never fine-tune them; all trained components are the small velocity MLPs, conditional flows, and flow maps of Table 4. Data are public gameplay recordings (open-oasis sample data; Atari frames via a standard environment). The Sliced-Wasserstein-2 metric and the linear action probes are implemented from scratch with fixed projection/seed schedules for determinism.

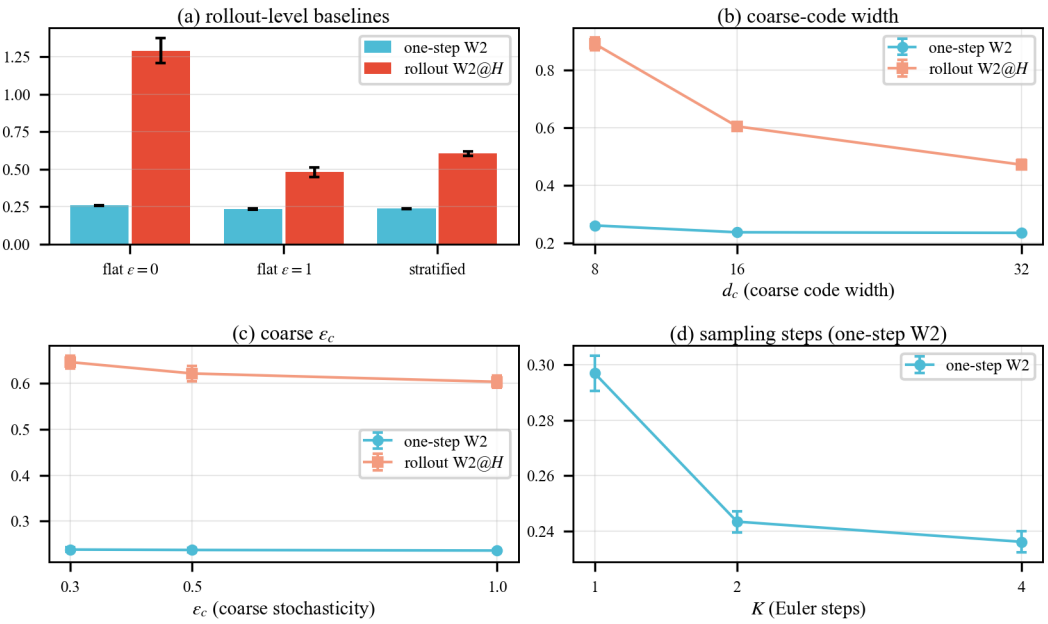

Figure 14: Rollout-level ablations on stochastic (stride-6) latents: baselines (a), coarse-code width d_c (b), coarse ε_c (c), and sampling steps K (d).

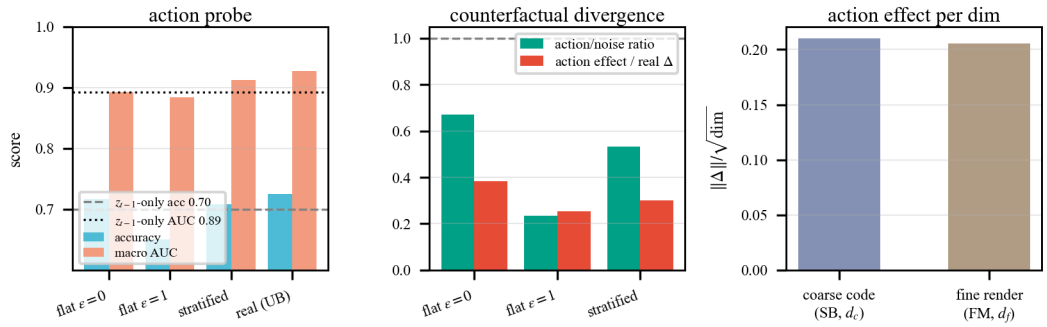

Figure 15: Controllability (stride-6). Left: action probe (accuracy, macro-AUC). Center: counterfactual divergence (action effect and action/noise ratio). Right: per-dimension action sensitivity is comparable for the coarse SB code and fine FM render: action information concentrates in the low-d code.

Fixed quantities in the epsilon sweep. Across the ε sweeps we hold network capacity, optimizer, schedule, batch size, and the train/test split fixed, varying only ε (and, where stated, d or K), so that differences are attributable to stochasticity rather than tuning. Each cell is averaged over 2–3 seeds; error bars in the figures are seed standard deviations.

We deliberately limit the claim to a controlled study at two scales: a small-scale setting ($\sim$2.1M params, 6k frames) that isolates architectural choices, and a scaled-up validation ($\sim$21M params, 400k frames, 4 Atari games) that confirms generalization and resolves the memorization concern. We do not claim a state-of-the-art interactive generator, do not benchmark against full video-generation leaderboards, and do not train at the scale of production world models. We study two domains and a PCA latent proxy; whether the precise d^* shifts under a learned (non-PCA) coarse code, under much larger encoders, or under longer horizons than $H{=}50$ is left open. The dimension-dependent-transition argument is a leading-order, finite-capacity/finite-step heuristic under a bounded-radius assumption, offered to explain measurements, not to provide tight constants or a population theorem. We expect the qualitative design rule (a separately modeled low-d transition code, with deterministic

Table 21: Counterfactual divergence (stride-1 / stride-6). Action effect as a fraction of a real single-step change (↑ more responsive).

model	action effect / real Δ (s1 / s6)	action/noise ratio (s1 / s6)
flat ε=0	0.42 / 0.38	0.35 / 0.67
flat ε=1	0.36 / 0.25	0.15 / 0.23
stratified	0.37 / 0.30	0.26 / 0.53

Table 22: Delta probe (input $\mathbf{z}_t^{\text{pred}} - \mathbf{z}_{t-1}$), Oasis stride-6. Reported as a caveat: flat/strat exceed real because cleaner deterministic outputs are more linearly decodable; this probe does not rank realism.

	baseline	flat ε=0	flat ε=1	stratified	real
macro-AUC	0.500	0.541	0.511	**0.617**	0.308
accuracy	0.375	0.392	0.308	**0.408**	0.325

rendering of the high-d appearance) to transfer, and the two-domain agreement on $d^* \approx 256$–512 is our main evidence for that.

Ethics. World models that generate interactive video inherit the dual-use and content-provenance concerns of generative video broadly; our study is small-scale and uses public gameplay data, but practitioners deploying such models should consider provenance, watermarking, and misuse. No human-subjects data are used.

Table 23: Action probe macro-AUC. Stride-1 is saturated; stride-6 is decisive.

regime	baseline ($\mathbf{z}_{t-1}$)	flat ε=0	flat ε=1	stratified	real (UB)
stride-1 (saturated)	0.953	0.950	0.939	0.953	**0.967**
stride-6 (stochastic)	0.892	0.892	0.884	**0.912**	**0.927**

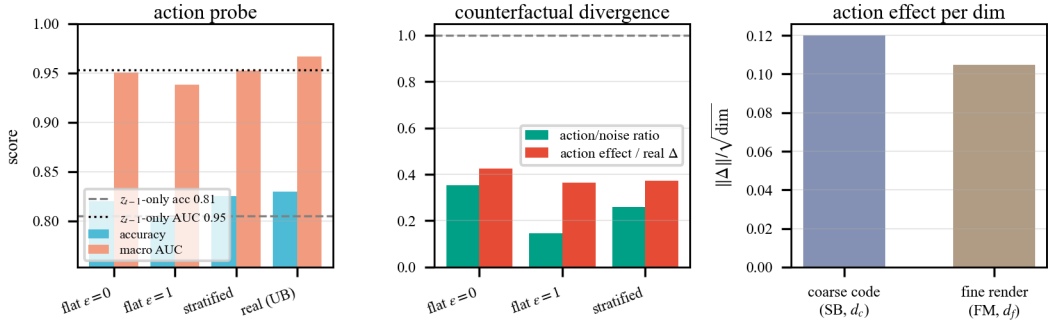

Figure 16: Controllability (stride-1, near-deterministic). The probe is saturated: the $\mathbf{z}_{t-1}$-only baseline already reaches AUC $\approx$0.95, so all models cluster near the ceiling and the regime cannot separate them, motivating the stochastic stride-6 evaluation in the main text.

Table 24: Static-latent flow-map distillation (SW2↓, 2 seeds). The low-d action map is best at 1-NFE; the high-d frame map beats raw one step and approaches its multi-step ceiling.

module	flow-map 1-NFE	flow-map K=16	teacher raw 1-step	teacher K=16
action (d=20)	**0.310**	0.404	0.631	0.481
frame (d=256, natural)	0.265	**0.217**	0.326	0.198

Table 25: 1-NFE quality vs. context noise σ (Oasis H=32). Teacher (strat, K=4) PSNR = 25.0.

σ	student PSNR↑	OOD@H ↓	W2@H ↓
0.05	12.6 (diverges)	0.97	3.98
0.10	**24.1**	0.15	1.17
0.20	26.5 ($\geq$ teacher)	**0.00**	**0.89**

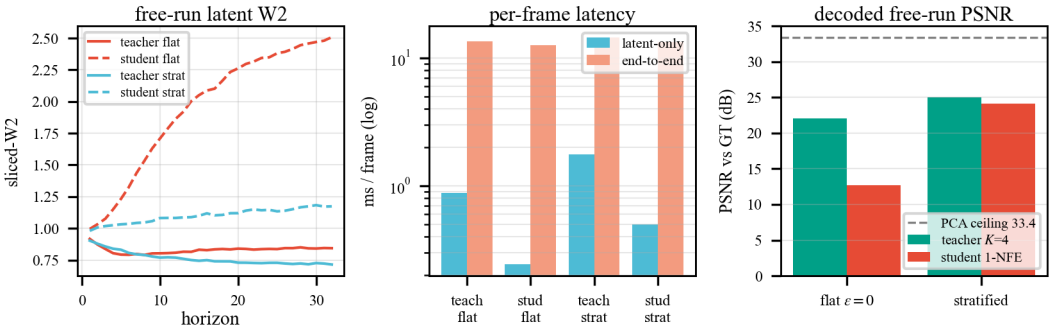

Figure 17: Real-time 1-NFE rollout. Free-run latent W2 vs. horizon (left), per-frame latency (center), and decoded PSNR (right): the 1-NFE stratified student preserves quality while accelerating the latent step 3.5×.

Table 26: Throughput (single L40, batch 1). 1-NFE gives a 3.5× latent-step speedup; end-to-end is decoder-bound but real time for all stratified settings.

model	NFE/frame	latent latency (ms)	latent FPS	end-to-end FPS
teacher flat (K=4)	4	0.88	1139	73.7
teacher strat (K=4)	8	1.76	568	68.7
student flat (1-NFE)	1	0.24	**4113**	**79.4**
student strat (1-NFE)	2	0.50	1996	75.1



Figure 18: Free-running 1-NFE decoded rollout (rows: rollout step t; columns labelled above). The one-step (1-NFE) stratified student keeps the cave geometry and HUD consistent with the multi-step teacher and with real (24.1 vs. 25.0 dB), whereas the flat student collapses into colour noise. As in Fig. 13 one identical brightness (γ=0.5) transform is applied to all panels for print (display-only; no content change).

Table 27: Per-frame free-running PSNR (dB) at selected horizons h (Oasis, H=32). Multi-step models (E12) and the 1-NFE students (E15). Flat models decay; stratified stays stable.

	h=1	h=8	h=16	h=24	h=32
flat ε=0 (multi-step)	30.4	22.5	13.8	7.8	7.3
flat ε=1 (multi-step)	26.3	11.5	11.8	13.2	12.9
stratified (multi-step)	30.2	25.5	26.1	22.9	20.6
student flat (1-NFE)	29.9	11.7	10.0	6.8	7.8
student strat (1-NFE)	31.0	23.7	23.5	23.1	20.0

Table 28: Latent spectra of the two domains. PR: participation ratio; d_p: dimension capturing $p\%$ of variance.

domain	ambient	PR	d_{90}	d_{95}	d_{99}
Oasis-Minecraft (ViT-VAE)	9216	21.0	461	683	–
Atari MsPacman (SD-VAE)	4096	19.6	142	258	643