CACHE ALIGNED DISTILLATION AND DRIFT CORRECTION FOR AUTOREGRESSIVE VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step autoregressive video world models must remain responsive to actions while repeatedly conditioning on their own generated history. Few-step distillation enables interactivity in generated worlds, but its errors can accumulate as lost motion, drifting latent statistics, and damaged scene structure. Recent work has improved teacher causality and reduced the training–inference mismatch; yet the ODE stage has not trained a causal student to use its KV cache directly when learning from a causal teacher's rollouts. Our ODE stage trains a four-step student through its inference-time KV-cache path against recorded teacher rollouts, using spatially local Gaussian KL supervision in place of pointwise regression. Autoregressive models are also commonly trained with token normalisation, channel normalisation, and spatial normalisation; the lack of temporal normalisation, however, causes latent values to explode or vanish as rollouts progress. We therefore introduce temporal normalisation to constrain latent statistics against an evolving real-scene reference. Finally, we observe causal autoregressive degradation in the form of highly structured, non-Gaussian, data-dependent noise, and we introduce an online Causal AutoRegressive Noise (CARN) network to model and correct this degradation. On 32 held-out contexts and nine distinct actions, we evaluate 30-second continuations against five open interactive world models. At six seconds, 70.8% of our rollouts remain clean across all six evaluation axes, compared with 24.7% on average externally; relocation is 6.3% versus 55.6%, while directional-control failure remains comparable at 21.1% versus 20.7%. Across longer rollouts, no model dominates every axis: our model preserves actions, high-frequency detail and comparatively low relocation, but accumulates style degradation and geometry drift rather than modes that are harder to recover from such as early collapse or relocation.

1 INTRODUCTION

Internet-scale video Diffusion Transformers (DiTs) provide powerful visual priors for world modelling, but their bidirectional generation and slow sampling make them unsuitable for interaction. Recent work has begun distilling them into causal autoregressive video models that are subsequently starting to evolve into world models. In this work we study how to better preserve the teacher's expressivity and power during distillation and how to better prevent accumulation and compounding of causal drift as the rollout progresses.

CausVid (Yin et al., 2025) established a two-stage route from a bidirectional teacher to a few-step causal student: ODE trajectory distillation followed by distribution matching distillation (DMD). Recent work since has aligned both the teacher and the distribution-matching stage with autoregressive generation but not the ODE stage, as Section 2 sets out; existing methods exclusively use clean ground truth latents as conditioning rather than train the student on sequentially generated latents that are subject to causal degradation.

Our teacher is Action Forcing (Sundar et al., 2026), a causal Wan2.1-1.3B world model trained on 88 hours of unlabelled real-world FrodoBots video. It recovers a signed, continuous throttle–yaw action space from tracked pixel motion and grounds these controls through an online latent critic, enabling reversal, scaling, and composition without recorded action labels. These capabilities make it an expressive world-model teacher, but its 40-step sampling is too slow for interaction. We therefore

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.

distil it into a four-step student, seeking to preserve not only its visual quality but also its learned action response under causal rollout.

The student is subsequently rolled on its own outputs during DMD, where the earlier cache-aligned initialisation does not prevent recurrent statistical drift. Standard token and channel normalisation acts within each denoising call, but does not control how the emitted latent distribution evolves across an autoregressive rollout. Related video methods address this instability by anchoring generated representations to seed or reference-frame statistics (Yang et al., 2023; Fan et al., 2024; Chong et al., 2025), or by retaining fixed visual anchors throughout generation (Blattmann et al., 2023; Ren et al., 2024; Jiang et al., 2024; Zhang et al., 2024; Yang et al., 2025). These mechanisms can prevent collapse, but they do not distinguish pathological drift from legitimate scene evolution: latent statistics should change with illumination, texture, viewpoint, and content, yet small shifts can compound until the latents explode or vanish. Fixing later chunks to the seed statistics can instead suppress valid changes, while predicting the appropriate future statistics introduces an additional forecasting problem (Luo et al., 2026; Devoto et al., 2025). We therefore treat temporal statistical evolution as a learning problem rather than impose a fixed inference-time bound. Our *Temporal Normalisation* objective supervises latent mean, variance, energy, and total variation against rolling statistics from the position-aligned real video. Although the recorded and generated futures need not depict identical content, this reference provides a seed-specific trajectory of plausible statistical change, teaching the generator to evolve its latent state without either runaway drift or artificial rigidity.

Temporal Normalisation controls the aggregate evolution of the latent field, but cannot remove the spatially structured, scene-dependent errors introduced by repeated self-conditioning. Supervising a corrector against the recorded future would conflate this degradation with legitimate differences in action realisation and generated content, effectively asking for semantic edits. We therefore introduce a network to model Causal AutoRegressive Noise (CARN): a lightweight, level-conditioned residual 3D ConvNet trained online from aligned continuations of the current student. The network is then applied to each student generated latent to correct it so later chunks condition on a less degraded history. Because CARN is trained alongside the changing student, it tracks the evolving error process and allows progressively longer training rollouts without imposing a ground-truth semantic trajectory.

When running ODE trajectory distillation on causal teacher rollouts, we observe a breakdown under traditional MSE supervision: ambiguous endpoints collapse towards blurred averages, leaving only 3.5% of rollouts clean at six seconds. Our variance-preserving local Gaussian KL raises this to 51.4%, matching minWM ODE's 47.6%. These similar clean fractions mask a large control gap: ours fails on 3.9% of six-second rollouts versus minWM's 37.5%. Weak or static responses avoid the viewpoint changes that expose visual degradation and can therefore appear stable without satisfying the action, as shown by minWM. We reach this performance with a quarter of minWM's optimisation steps (Zhao et al., 2026a). A matched ablation isolates the cache-aligned context itself: retraining the same student on teacher-forced clean prefixes, with every other setting held fixed, drops the six-second clean fraction from 51.4% to 8.7% and raises control failure from 3.9% to 43.4%.

During DMD, removing CARN's cache correction raises six-second control failure from 21.1% to 43.4%, geometry failure from 7.3% to 46.9%, and relocation from 6.3% to 27.1%. Restricting Temporal Normalisation to only mean and energy or variance and total variation raises geometry failure further, to 86.1% and 98.3%, respectively. Ablating the temporal normalisation entirely results in collapse.

Against five open interactive world models, ours degrades more slowly and more gracefully while remaining within the conditioned scene. At six seconds, 70.8% of its rollouts remain clean on all six axes, compared with 24.7% on average externally, and relocation is 6.3% versus 55.6%, while directional-control failure is comparable at 21.1% versus 20.7%. No model remains reliable at 30 seconds: ours retains lower HF and relocation failure than the external averages, but accumulates greater style and geometry failure.

Our contributions are:

1. **Cache-aligned distributional ODE transfer.** We transfer the teacher's action response, visual detail, and causal rollout behaviour by supervising all four student denoising rungs against endpoints from complete action-conditioned teacher rollouts, while executing the

student through its inference-time sequential KV-cache path. Spatially local Gaussian KL preserves the teacher's regional latent mean *and variance*, preventing ambiguous degraded histories from collapsing into blurred averages under pointwise regression.

2. **Temporal statistical control during DMD.** We introduce Temporal Normalisation, a training-only objective that teaches the autoregressive student how its latent statistics should evolve over time. By matching mean, variance, energy, and total variation to rolling, position-aligned real-video references, it prevents compounding explosion or vanishing without rigidly anchoring the rollout to its seed.
3. **Adaptive correction of causal drift.** We introduce CARN, a lightweight online corrector that estimates structured, model- and scene-dependent degradation from aligned student continuations with different amounts of real context. Applying its residual before each KV-cache commit limits error propagation, stabilising the self-rolled histories used for progressively longer training rollouts and the inference output.

2 Literature Review

CausVid (Yin et al., 2025) established the two-stage route that this literature follows: ODE initialisation of a few-step student, followed by distribution-matching distillation against a frozen teacher score. CausVid and Self Forcing (Huang et al., 2026) initialise from stored trajectories of a bidirectional teacher and train the student in a packed block-causal pass; Self Forcing changes the subsequent distribution-matching stage by training on student-generated histories. Causal Forcing (Zhu et al., 2026) instead generates each ODE endpoint with an autoregressive teacher, but conditions both teacher and student on the same teacher forced clean real-video prefix. minWM carries this design into an action conditioned world model pipeline (Zhao et al., 2026a). Causal Forcing++ (Zhao et al., 2026b) replaces stored causal ODE trajectories with online causal consistency distillation, while retaining the clean-prefix conditioning. Action Forcing (Sundar et al., 2026) instead recovers a signed throttle–yaw interface from unlabelled video and trains the action-conditioned causal model that serves as our many-step teacher but does not perform ODE distillation. These works align the teacher architecture or the later distribution-matching stage with autoregressive generation, but do not align the execution of few-step initialisation itself, which we do in this work.

The open interactive models in our evaluation reach streaming generation by different routes rather than sharing one distillation recipe. Matrix-Game 2.0 (He et al., 2025) combines bidirectional-teacher ODE initialisation with Self-Forcing-style distribution matching, whereas minWM uses causal ODE or causal consistency initialisation followed by asymmetric DMD. DreamX-World combines Causal Forcing, long-rollout DMD, retrieved scene memory, and post-distillation alignment (Team et al., 2026). LingBot-World-V2 starts from a natively causal action-conditioned model and follows consistency distillation with DMD over long self-rollouts (Gao et al., 2026). YUME-5B instead combines compressed historical context with adversarial distillation and denoising-feature caching (Mao et al., 2025). All expose autoregressive generation to online controls, which places them in our comparison, but their different teachers, context mechanisms, and post-training objectives are material to the failures they exhibit.

3 Preliminaries

3.1 Few-step ODE distillation

Let h denote the autoregressive context, $a \in \mathcal{A}$ an action, and $x_1 \sim p_{\mathrm{data}}(x \mid h, a)$ the next latent video chunk. A conditional flow-matching teacher (Lipman et al., 2022) maps noise ϵ to the endpoint $y = \Phi_T^{0\to 1}(\epsilon; h, a)$ through many numerical updates. The context and action select the conditional transport, while a distilled student approximates the same endpoint using only $K = 4$ rungs. At every rung the student predicts a clean endpoint and constructs its next input with a restart sampler:

$$\begin{aligned} x^S_{t_0} &= \epsilon, \qquad t_0 = 0, \\ \hat{y}_r &= G_\theta(x^S_{t_r}, h, a, t_r), \\ x^S_{t_{r+1}} &= \alpha_{t_{r+1}} \hat{y}_r + \sigma_{t_{r+1}} \xi_r, \qquad \xi_r \sim \mathcal{N}(0, I), \end{aligned} \tag{1}$$

where (α_t, σ_t) is the noise schedule and the last update applies only for $r < K - 1$. Teacher intermediate states are not used as student inputs. Each $\hat{y}_r$ is supervised against the same endpoint y; Section 4.1 specifies how those endpoints and their causal histories are constructed.

3.2 Distribution-matching distillation

We apply DMD2 (Yin et al., 2024) with a frozen causal teacher and an online fake score model. Unlike ODE transfer, DMD trains the student on its own generated history: every generated chunk is committed as context for the next. We first supervise seven-chunk rollouts from real context in a stationary phase. We then sample progressively greater rollout depths, generate each prefix without gradients, and supervise its terminal seven chunks. The DMD normalisation denominator is floored at 0.05 to limit gradient amplification. A frozen ego-motion critic additionally penalises disagreement between the generated motion code and the commanded action.

4 Method

4.1 ODE distillation

4.1.1 Cache-aligned causal rollout transfer

We first record the causal teacher under the same chunkwise execution pattern used for interactive generation. Starting from three real seed chunks, the 20-step teacher generates six three-frame chunks sequentially and commits each endpoint to its KV cache before generating the next:

$$\begin{aligned} y_j^T &= \Phi_T^{0\to 1}(\epsilon_j; h_j^T, a), \\ h_{j+1}^T &= \mathcal{C}(h_j^T, y_j^T), \qquad h_0^T = h, \end{aligned} \tag{2}$$

where $\mathcal{C}$ commits the generated chunk to the cache. The action a is held constant over the generated chunks; each chunk has a fresh noise draw ϵ_j. The recording therefore preserves each endpoint together with the action, initial noise, and generated causal history that produced it.

We then replay the sequence with the four-rung student. At chunk j, the student starts from the same ϵ_j, follows its own restart trajectory from Equation 1, and is supervised at every rung against the recorded endpoint y_j^T:

$$\mathcal{L}_{\text{rollout}} = \mathbb{E}\left[\sum_{j=0}^{J-1}\sum_{r=0}^{3} \ell_{\text{local-KL}}(\hat{y}_{j,r}^S, y_j^T)\right]. \tag{3}$$

After each chunk, we commit y_j^T, rather than the student's prediction, to the student's cache. The student therefore learns through the same sequential cache path used at inference, while every target remains consistent with the history on which the teacher generated it. Teacher commits also prevent weak early student predictions from corrupting later ODE conditioning; exposure to self-generated student history is introduced subsequently during DMD.

4.1.2 Spatially local Gaussian KL

Our rollout transfer deliberately trains the student on teacher-generated histories that already contain causal degradation. This makes endpoint regression ambiguous: similar degraded states can correspond to multiple clean teacher continuations. For a fixed distribution of regression pairs, the unrestricted optimum under pointwise squared error is their conditional mean:

$$G_{\text{MSE}}^*(x, t, h, a) = \mathbb{E}[y \mid x_t^S = x, t, h, a]. \tag{4}$$

Target variation that the four-rung student cannot resolve is therefore averaged. In our matched experiment this contracts local latent variance, produces blurred decoded chunks, and then feeds that contraction back through subsequent cache commits.

We therefore replace pointwise regression with a spatially local Gaussian KL objective that preserves the teacher endpoint's mean and variance. For each latent frame and channel, adaptive average

pooling defines a 3×4 grid of spatial regions in which we estimate the student and teacher statistics. The variance term makes the under-dispersed average costly, while the spatial grid prevents the student from satisfying the objective by placing the right amount of variation in the wrong part of the frame. Empirically, this allows the student to maintain a coherent continuation rather than repeatedly contracting towards an average trajectory. The objective does not uniquely determine fine texture or guarantee correct action response; it directly targets the variance collapse observed in the matched MSE student.

Writing these statistics as (μ_S, v_S) and (μ_T, v_T), respectively, we minimise

$$\ell_{\text{local-KL}}(S,T) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} D_{\text{KL}}(\mathcal{N}(\mu_{S,i}, v_{S,i}) \,\|\, \mathcal{N}(\mu_{T,i}, v_{T,i}))\,, \tag{5}$$

where $\mathcal{I}$ indexes frames, channels, and pooled spatial regions. For both prediction and teacher, we floor each regional variance for numerical stability:

$$v_i = \max\left(\mathbb{E}_i[x^2] - \mathbb{E}_i[x]^2,\ 10^{-6}\right). \tag{6}$$

The per-region divergence is

$$D_{\text{KL}}(S\|T) = \frac{1}{2}\left[\log \frac{v_T}{v_S} + \frac{v_S + (\mu_T - \mu_S)^2}{v_T} - 1\right]. \tag{7}$$

The positive teacher variance therefore makes a vanishing student variance explicitly costly.

4.2 DMD-STAGE STABILISATION

Self-conditioned DMD exposes two recurrent failure modes left unresolved by the ODE initialisation: statistical instability and structured drift in the autoregressive state. We address them with Temporal Normalisation and CARN, respectively.

4.2.1 TEMPORAL NORMALISATION

Autoregressive generation repeatedly feeds our three-frame outputs back as context without any normalisation in the time dimension. We observe small perturbations compounding over the rollout into large shifts that destabilise generation. To ensure that we preserve the semantic meaning in these statistic shifts, we supervise their temporal evolution using rolling real-video statistics at the corresponding sequence position.

Let $P = 3$ denote the chunk length and p_j the first latent-frame position of generated chunk j. For each statistic S, the reference is

$$A_S(j) = \frac{1}{|\mathcal{W}_j|} \sum_{f \in \mathcal{W}_j} S(x_f^{\text{GT}}), \qquad \mathcal{W}_j = [p_j - 2P,\ p_j + 3P), \tag{8}$$

covering the aligned chunk and two neighbours on either side. We penalise deviations of each generated frame from this reference:

$$\mathcal{L}_{\text{stat}} = \sum_S w_S\, \mathbb{E}_{j, f \in j}\left[\|S(\hat{x}_f) - A_S(j)\|_2^2\right]. \tag{9}$$

The four statistics are

$$\begin{aligned} &\mu_d(x) = \text{mean}_{H,W}\, x_d, \qquad V(x) = \sum_d \text{Var}_{H,W}(x_d), \\ &E_d(x) = \sum_{H,W} x_d^2, \qquad\qquad \text{TV}(x) = \sum_d \left(\text{mean}\,|\Delta_w x_d| + \text{mean}\,|\Delta_h x_d|\right). \end{aligned} \tag{10}$$

We use $(w_\mu, w_V, w_{\text{TV}}, w_E) = (0.9, 0.1, 0.1, 6 \times 10^{-8})$, with no tolerance band. Per-channel means expose signed shifts that cancel under pooling, variance and energy constrain contraction and growth, and total variation constrains the spatial detail lost to progressive smoothing. The moving reference allows statistics to evolve along the sequence while penalising the small persistent deviations that accumulate under self-conditioning.

4.2.2 CORRECTING RESIDUAL DRIFT (CARN)

Temporal Normalisation constrains latent statistics, but longer rollouts create causal autoregressive drift. This error is neither uniformly distributed nor is it Gaussian: it is highly structured error that is unique to the generating model and the scene content being generated. Direct supervision of these errors becomes increasingly difficult with rollout depth: the generated scene diverges from the recorded sequence, so its difference from ground truth contains both autoregressive degradation and legitimate changes in content. Because the error distribution changes as the model trains, neither a fixed positive nor negative target is available.

Thus, we use an online model to approximate this drift using paired student rollouts of the same sequence, with one seeded by two fewer real chunks than the other, thereby obtaining an online estimate for relative drift rather than a fixed offline estimate for absolute drift. At an aligned position p, let r_p^1 denote the continuation with more real context and r_p^2 the continuation with two additional autoregressive steps. This construction reduces the content mismatch of ground-truth supervision while exposing differences associated with rollout depth. The pairs remain approximate: neither continuation is guaranteed clean, and their difference need not consist exclusively of drift.

We formulate correction as generative restoration rather than exact inversion. When degradation discards information, distinct states u and v can produce the same degraded state, $D_\ell(u) = D_\ell(v) = z$. Any deterministic corrector must then satisfy

$$\frac{1}{2}\left(\|G_\xi(z,\ell) - u\|_1 + \|G_\xi(z,\ell) - v\|_1\right) \geq \frac{1}{2}\|u - v\|_1 > 0. \tag{11}$$

Thus, exact recovery of both states is impossible from the degraded input alone. Here D_ℓ denotes an abstract information-losing degradation, not a learned component. CARN instead learns a plausible correction towards the less degraded continuation, without assuming that lost detail can be recovered exactly.

The dedrifter is a level-conditioned residual network, $G_\xi(z,\ell) = z + \Delta_\xi(z,\ell)$, with a zero-initialised output projection. The level ℓ counts generated chunks from the last clean chunk and labels the more degraded member of each pair. We train it using

$$\mathcal{L}_{\text{CARN}} = \mathbb{E}_{p,\ell}\left[\|G_\xi(\text{sg}[r_p^2],\ell) - \text{sg}[r_p^1]\|_1\right]. \tag{12}$$

The correction is applied when the state is committed to the autoregressive cache. For $z' = G_\xi(z,\ell)$,

$$z_{\text{commit}} = z + \alpha(s)\big(\Pi_z(z') - z\big), \tag{13}$$

where Π_z restores the per-channel mean and overall centred mean absolute deviation of z. This separates structural correction from the statistical control supplied by Temporal Normalisation. Because the corrector is online and applied every chunk, its own biases could otherwise compound. We therefore keep $\alpha(s) = 0$ through step 150, ramp it to 0.125 over the following 100 steps, and cap the applied correction at a relative shift of 0.015.

5 EVALUATION

5.1 EVALUATION PROTOCOL

We evaluate each reported model by rolling out 32 held-out outdoor real-video seed contexts from a variety of datasets: eight each from Sekai, SpatialVID, FrodoBots, and Ego4D. A mix of video datasets are offered in the interests of fairness; each of the models has been trained on one or more of these datasets. Each autoregressive rollout lasts for 30 seconds and is evaluated under eight directional commands and one no-op command. The ODE and DMD stages are evaluated separately, in Sections 5.2 and 5.3.

Video world models' performance is reliant on three different axes: control response, visual quality, and consistency. We evaluate visual quality with two axes: geometry which measures the 3D warping errors in objects that co-occur with ego movement and subsequent viewpoint change; and high-frequency (HF) degradation, which measures loss of spatial detail as the rollout progresses. We score consistency by measuring scene relocation, style shifts, and object conjuring - some world

models avoid control and quality errors by teleporting to a place or a style they are more familiar with or by apparating a subject to take up the focus. Conjuration and relocation are onset events and are carried forward after detection; the other axes are scored independently in each window. We measure these on fixed six-second windows ending at 6, 12, 18, 24, and 30 seconds, most with reference to the original context seed and some (relocation and HF degradation) with reference to other rollouts. Every axis is a binary decision, a video may fail several, and lower is better throughout.

We follow the evaluation protocol of Action Forcing (Sundar et al., 2026), with the modifications detailed in Appendix D. Control is measured by projecting CoTracker motion onto the action basis; failure requires either an angular error above 45° or insufficient motion, using thresholds of 0.15 for cardinal actions and $0.15/\sqrt{2}$ for diagonals. No-op fails above 0.15. Style and geometry failures are respectively defined by DINOv2 distance above 0.72 and Qwen3-VL inconsistency probability above 0.5. High-frequency failure measures excessive loss of Laplacian variance relative to the family-balanced model panel. Conjuration candidates proposed by dense RT-DETR tracking are accepted only after temporal evidence review. Six-second relocation fails below 50 verified ORB–RANSAC correspondences and this decision is carried forward to later endpoints.

5.2 ODE DISTILLATION

The ODE comparison isolates the initial transfer objective before DMD. We vary two factors independently: the context the student is trained on, being either the teacher's own cache-aligned rollout or a teacher-forced clean prefix, and the transfer loss, being either spatially local Gaussian KL or pointwise MSE. All four arms otherwise share the same teacher recordings, denoising rungs, optimisation budget, and inference path, as shown in Figure 1. Among the evaluated systems, minWM is the only world model with a released ODE-stage checkpoint and therefore provides our external causal-ODE reference.

The CausVid-style ODE stage obtains whole-clip targets directly from a bidirectional teacher (Yin et al., 2025). minWM uses a causal teacher which cannot do the same, so it instead teacher-forces each block with clean preceding latents. Real trajectories may contain action expressivity outside the teacher's learned repertoire; from this imposed history, the teacher may respond weakly or with no motion, and the student then distils that response. We instead roll the teacher autoregressively under a sampled action and commit its own outputs, allowing the student to inherit the action response the teacher can express. This distinction is clearest in control: at six seconds minWM fails on 37.5% of rollouts, compared with 3.9% for our local-KL student. Their clean fractions are initially similar (47.6% versus 51.4%), but at 12 and 18 seconds minWM retains 3.5% and 0%, compared with 9.7% and 8.3% for local KL. By 30 seconds, minWM's control and HF failure rates reach 50.8% and 60.4%, compared with 25.0% and 17.0% for local KL.

The matched ablation isolates the context directly. Holding the loss at local Gaussian KL and replacing the teacher's own rollout with a teacher-forced clean prefix costs most of the transfer: the six-second clean fraction falls from 51.4% to 8.7%, control failure rises from 3.9% to 43.4%, and relocation from 24.7% to 89.9%. This is the comparison the ODE stage has not previously been given: the same student, budget and inference path with only the teacher execution path at ODE trajectory generation changed.

Cache alignment introduces a separate regression problem: degraded inputs can be compatible with several plausible clean futures. Pointwise MSE favours their conditional average and blurs the output; accordingly, the matched MSE student retains only 3.5% clean rollouts at six seconds, compared with 51.4% under local Gaussian KL. By preserving regional latent dispersion, local KL instead maintains a coherent continuation. Its six-second clean fraction of 51.4% is comparable to minWM's 47.6%. It achieves this with 500 optimisation steps, compared with minWM's 2,000 (Zhao et al., 2026a). At 30 seconds, its stacked failure burden is 256 percentage points, about four-fifths of minWM's 328.

5.3 DMD AND AUTOREGRESSIVE ROLLOUT

Figure 2 compares the complete DMD model with five external distilled models on the balanced four-dataset panel. Ablations of the DMD-stage stabilisers are reported in Appendix F.

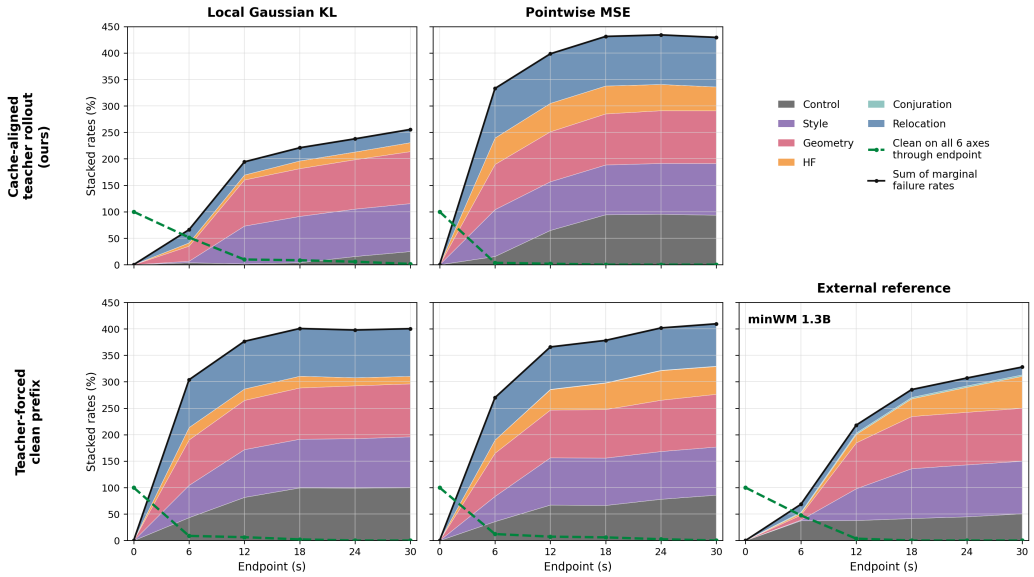

Figure 1: ODE-stage failure accumulation on the balanced four-dataset panel. Rows are the context the student is trained on and columns the transfer loss; all four arms are otherwise matched. The coloured bands stack directional control, style, geometry, HF, conjuration, and relocation marginal rates, so their height may exceed 100% when one rollout fails several axes. Relocation and conjuring bands persist after onset. The dashed green curve is the cumulative fraction of all 288 rollouts clean on all six axes through the indicated endpoint.

Our model exhibits the most graceful accumulation of failure, with the lowest stacked marginal burden at every endpoint: 34.6 percentage points versus 64.4 for the next-best model, minWM, at six seconds, to 199.1 versus 232.8 for the next-best model, LingBot, at 30 seconds. At six seconds it also ties minWM for the lowest relocation rate (6.3% each), compared with 44.8% for the next-lowest model, LingBot. Several apparently strong control scores instead coincide with early relocation: DreamX records 1.2% control failure versus our 21.1%, but relocates in 70.1% of rollouts versus our 6.3%; Matrix-Game records 8.6% versus 21.1% control failure but 87.8% versus 6.3% relocation. YUME likewise relocates in 68.8% of rollouts versus our 6.3%. minWM is the closest competitor, with 49.0% clean versus our 70.8% and the same 6.3% relocation rate, but its 42.2% control failure versus our 21.1% is highly directional: forward and backward fail at 96.9% and 87.5%, whereas pure left and right turns both fail at 0%. Its strong appearance scores should therefore be read alongside limited translational response, which avoids many of the viewpoint changes that expose visual degradation.

By 30 seconds every model has taken its own route to failure: the largest all-six-axis clean fraction is only 4.5% for YUME, followed by our 4.2%, while the other systems retain between 0% and 2.4%. Ours continues to respond within the original scene, with 14.1% control, 3.8% HF, and 6.3% relocation failure versus external averages of 30.5%, 20.2%, and 55.6%, but pays for this continuity in style and geometry failure (79.2% and 95.8% versus 69.2% and 68.9% externally). minWM also keeps relocation to 6.3%, but reaches 46.9% control and 95.1% geometry failure. DreamX retains the best control score at 5.9% but relocates in 70.1% of rollouts; Matrix-Game retains low HF failure at 6.3% but relocates in 87.8%. YUME has the lowest geometry failure at 44.1%, but combines it with 43.0% control failure and 68.8% relocation.

6 Discussion & Limitations

The evaluation axes are coupled: failing one task can make another easier. A model that does not move avoids the viewpoint changes that expose geometric and appearance errors. Relocation provides a second shortcut: replacing the conditioned scene can produce a sharp, locally coherent video while ceasing to continue the world supplied by the user. Visual quality should therefore be

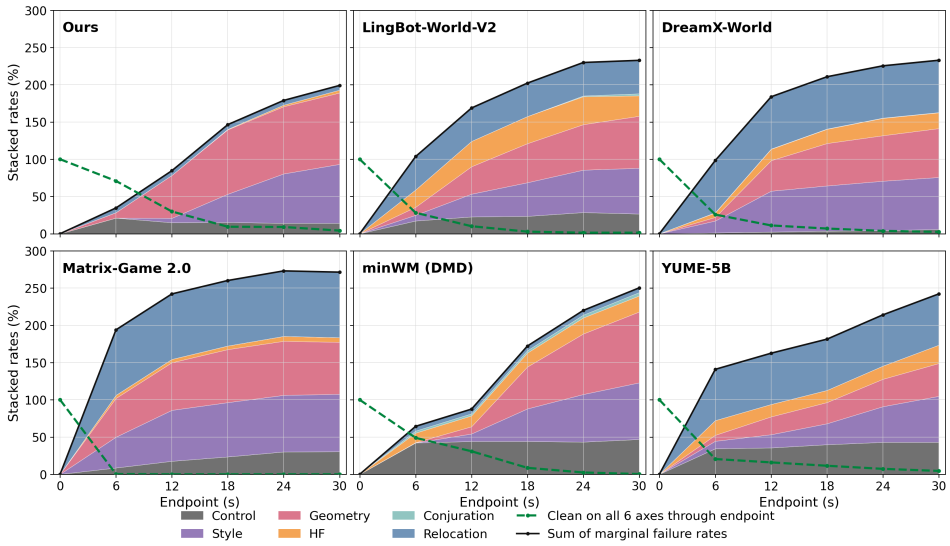

Figure 2: Failure accumulation for our complete DMD model and five external systems on the balanced four-dataset panel. Coloured bands stack directional control, style, geometry, HF, conjuration, and relocation marginal rates; because failures overlap, their sum is a burden index rather than a percentage of failed rollouts. Relocation and conjuration bands persist after onset. The dashed green curve is the cumulative fraction of all 288 rollouts clean on all six axes through each endpoint.

read conditional on both action success and scene preservation, rather than as evidence of world modelling in isolation.

Our results show that the model continues to act and generate within the conditioned scene without catastrophic collapse or scene replacement. Its remaining failure is predominantly gradual geometric warping and style drift. This graceful degradation leaves a continuous, recognisable world whose errors can be studied and repaired.

After multiple small differences in action realisation, a generated rollout and its ground truth counterpart may no longer share exact spatial or semantic correspondence but they may still remain plausible and high quality. GT based losses encourage generative models to move towards arbitrary recorded scenes rather than to repair their own; this subsequently appears as relocation errors. Long-horizon objectives must instead reward plausible generation distributionally rather than match a specific recorded scene, motivating learned objectives that distinguish degradation from valid change.

While we do not make use of such semantic loss models, we instead highlight that generative self repair is indeed possible with intelligent models through Temporal Normalisation and our CARN network. They supervise evolving statistics and structured degradation while leaving generated semantic content free to develop, which helps explain why the complete model retains its scene while the ablations and external models relocate or deform much earlier.

7 Conclusion & Future Work

This work shows that a few-step interactive world model can continue responding within the same world for long horizons and degrade gracefully rather than collapse, become static, or relocate. We have also shown through this work that generative models need appropriately designed generative losses rather than simple heuristic based losses that cannot untangle errors from creative generation.

The remaining challenge is consequently more precise: to repair progressive 3D warping in the generated worlds without forcing a generated world towards the wrong recorded future or imposing arbitrary rigid constraints upon it. Addressing it will require augmenting these world models with some sort of memory and creative model based losses that reward the model for using this memory to generate within distribution content.

AI USE STATEMENT

The research questions, method design, and interpretation of the results were developed by the authors. Generative AI tools were used to a limited extent for manuscript editing, extensively as coding assistants, and to assist experiment orchestration but not experiment design. They were not used to originate the research idea or to analyse the experimental results. All AI-assisted text, code, and commands were reviewed and validated by the authors, who take full responsibility for the contents of this work.

ETHICS STATEMENT

The model is a research artefact and has not been deployed to control a physical system. All training performed by us uses only FrodoBots video, under the dataset's published terms; the Sekai, SpatialVID, and Ego4D clips in our panel are held out for evaluation and do not update the model. The FrodoBots provider states that faces are blurred before release. We relied on this preprocessing and performed no additional face detection or removal. As with other generative video systems, the model may reproduce dataset biases or generate realistic but incorrect content, and its outputs should not be treated as reliable forecasts of real-world outcomes. Any deployment would require additional privacy, bias, and safety assessment.

REPRODUCIBILITY STATEMENT

Upon acceptance, we will release the training and inference code, trained checkpoints, configuration files, and the complete evaluation harness used to produce the reported per-video decisions and aggregate figures. The paper and appendix specify the training schedule, model variants, evaluation panel, detector thresholds, and aggregation rules. Third-party datasets and baseline checkpoints must be obtained from their respective providers under their original terms.

REFERENCES

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575. IEEE, 2023.

Zheng Chong, Wenqing Zhang, Shiyue Zhang, Jun Zheng, Xiao Dong, Haoxiang Li, Yiling Wu, Dongmei Jiang, and Xiaodan Liang. Catv2ton: Taming diffusion transformers for vision-based virtual try-on with temporal concatenation. *arXiv preprint arXiv:2501.11325*, 2025.

Alessio Devoto, Maximilian Jeblick, and Simon Jégou. Expected attention: Kv cache compression by estimating attention from future queries distribution. *arXiv preprint arXiv:2510.00636*, 2025.

Xiang Fan, Anand Bhattad, and Ranjay Krishna. Videoshop: Localized semantic video editing with noise-extrapolated diffusion inversion. In *European conference on computer vision*, pp. 232–250. Springer, 2024.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, et al. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

Yuming Jiang, Tianxing Wu, Shuai Yang, Chenyang Si, Dahua Lin, Yu Qiao, Chen Change Loy, and Ziwei Liu. Videobooth: Diffusion-based video generation with image prompts. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6689–6700. IEEE, 2024.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jiayi Luo, Qiyan Liu, Tengyang Wang, JunHao Liu, Jiayu Chen, Cong Wang, Hanxin Zhu, Chen Gao, Xiaobin Hu, Qingyun Sun, et al. Future forcing: Future-aware training-free kv cache policy for autoregressive video generation. *arXiv preprint arXiv:2605.30083*, 2026.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024.

Ashish Sundar, Tiankuo Hou, Zhong Fan, Chunbo Luo, and Xiaoyang Wang. Action forcing: Training world models on unsupervised video by recovering underlying egomotion bases. *arXiv preprint*, 2026.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026.

Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Zhongwei Zhang, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Ting Yao, Yang Cao, and Tao Mei. Trip: Temporal residual learning with image noise prior for image-to-video diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8671–8681. IEEE, 2024.

Min Zhao, Hongzhou Zhu, Bokai Yan, Zihan Zhou, Yimin Chen, Wenqiang Sun, Kaiwen Zheng, Guande He, Xiao Yang, Chongxuan Li, et al. minwm: A full-stack open-source framework for real-time interactive video world models. *arXiv preprint arXiv:2605.30263*, 2026a.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026b.

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.

A Implementation details

Model and data. The teacher is the Action Forcing Wan2.1-1.3B model described in Section 4.1, and the student retains the same backbone and latent representation while reducing sampling to four rungs. All training performed in this work uses FrodoBots data only. The additional datasets in the evaluation panel are held out and are never used for optimisation. Training uses bfloat16 arithmetic.

ODE initialisation. We construct 10,000 recorded causal-teacher rollout samples and train the four-rung student for 500 optimiser steps. The local-KL and pointwise-MSE arms use identical recordings, action sampling, initialisation, restart schedule, and cache execution; only the endpoint loss changes. We use AdamW with learning rate 2×10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 0.01. Training is distributed over 32 GH200 GPUs with one sample per rank, giving an effective batch size of 32.

DMD schedule. The reported model is initialised from the step-500 local-KL ODE checkpoint. Its stationary phase trains for 200 steps on seven generated chunks following real context. We then resume the same run for 800 additional rolling steps, reaching 1,000 DMD steps in total. In the rolling phase, the prefix is generated without gradients at a sampled depth and supervision is applied to the terminal seven chunks, with the maximum depth increasing over training. The generator and online fake-score model use learning rate 10^{-5} with 8-bit AdamW. Both DMD phases use four GH200 GPUs. The flagship configuration disables the adversarial/GAN objective; its additional losses are the frozen ego-motion guidance, Temporal Normalisation, and CARN objectives specified in Sections 4.2.1 and 4.2.2.

CARN architecture. CARN operates on the 16-channel VAE latent. It is a level-conditioned residual 3D ConvNet with width 512 and four factorised spatial–temporal residual blocks. Each block uses sinusoidal level embeddings and FiLM modulation; the output projection is zero-initialised so the network starts as the identity. The corrector contains 23.1M parameters and is trained online alongside DMD.

B Conditioning spans

The models retain their native conditioning spans: LingBot-World-V2, DreamX-World, Matrix-Game 2.0, and YUME-5B receive only real frame 32; minWM DMD receives two native seed blocks (eight latent frames, decoded as real-video pixel frames 4–32); minWM ODE receives frames 20–32; and our checkpoints receive nine latent frames, decoded as frames 0–32. Every model therefore begins generation immediately after the same real frame. YUME's packaged rollout contains that single conditioning frame followed by 480 generated frames at 16 fps. Matrix-Game's native 25-fps output is sampled by time for the 16-fps control reader while retaining its native spatial resolution.

C Matched-seed rollout examples

Figures 3–8 show forward and right continuations from held-out FrodoBots, Sekai, and SpatialVID contexts in the balanced panel. Every row starts from the same source-video frame 32, which is reproduced directly in the conditioning column; the models otherwise retain their native context lengths. Generated frames are sampled every two seconds through 14 seconds by time rather than by frame number, aligning the 16-fps and 25-fps systems. These examples are not used to set any metric threshold.

D Evaluation protocol

The six instruments follow those introduced by Action Forcing (Sundar et al., 2026). Appendix E summarises their human validation; here we specify the deployed configuration.

Panel and timing. Every decision is binary, and one video may fail several criteria at once. The balanced panel contains eight held-out outdoor contexts from each of Sekai, SpatialVID, FrodoBots,

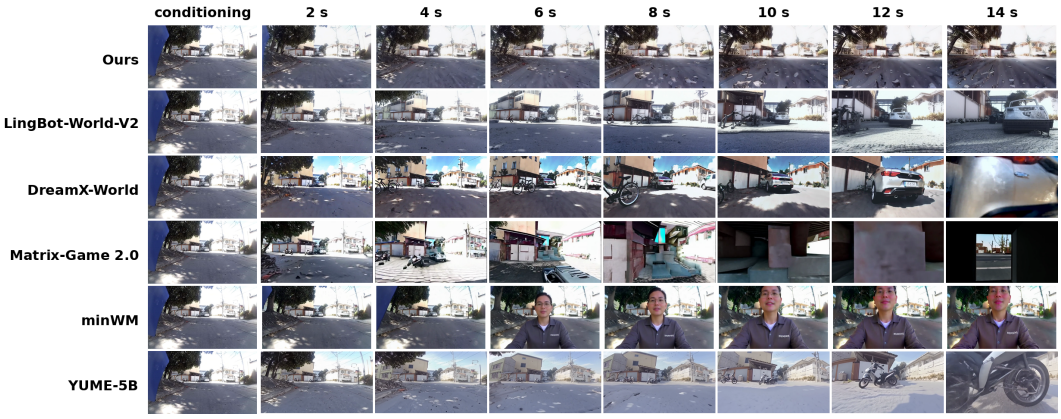

Figure 3: Matched forward-command continuations for FrodoBots context `u31`. Columns show the common conditioning frame and generated frames at two-second intervals.



Figure 4: Matched right-command continuations for FrodoBots context `u31`.

and Ego4D; every context is paired with eight directional commands and one no-op command. Control, style, geometry, and HF are evaluated independently per window; conjuration and relocation are onset events carried forward in the cumulative six-axis graphs. Let c be the number of real conditioning frames a model receives and f its native frame rate, so generated frame zero is index c and a window starting at s seconds spans $c + \mathrm{round}(sf)$ to $c + \mathrm{round}((s+6)f) - 1$. We score the five non-overlapping windows $s \in \{0, 6, 12, 18, 24\}$; the producers also compute an overlapping window at $s = 9$ for a 15-second endpoint, which is excluded from every trajectory reported here. Frames are decoded at native resolution before each instrument applies its stated internal resize; Matrix-Game 2.0's 25-fps output is selected by timestamp.

Directional control. Control is measured from motion rather than appearance. A frozen CoTracker3 tracker receives 12 seed frames and 96 generated frames at 16 Hz. The seed is the final 12 real conditioning frames in the first window and the 12 preceding frames thereafter; models with shorter contexts repeat their earliest available view. We project the tracked displacement field onto the frozen PCA basis of the teacher's action interface and average it over the window. The resulting vector g has the longitudinal and lateral components in which commands are issued. The eight directional commands are the sign pairs $(\pm 1, 0)$, $(0, \pm 1)$ and $(\pm 1, \pm 1)$. Each directional rollout is judged from its own vector. It fails when $\cos(g, \text{command}) < \cos(45°)$, or when $\|g\|_2 < 0.15$ for a cardinal command and $\|g\|_2 < 0.15/\sqrt{2}$ for a diagonal command. The diagonal cutoff preserves the corresponding per-axis scale without using separate cardinal rollouts as preconditions. We report directional control on the 256 directional videos per row. The no-op command, having no direction to compare against, instead fails when $\|g\|_2 \geq 0.15$ and is included in the all-axis clean fraction.



Figure 5: Matched forward-command continuations for Sekai context `cheongju`.



Figure 6: Matched right-command continuations for Sekai context `cheongju`.

Style drift. Style compares generated frames with the real frames supplied to the model. Frames are embedded with DINOv2 ViT-S/14 at 224×224 under ImageNet normalisation, ℓ_2-normalised and averaged. The reference is the final $\min(16, c)$ conditioning frames. The generated set comprises the 16 frames ending the first window, and 16 frames sampled uniformly from the final second for later windows. Style drift is one minus the cosine between the two and fails above 0.72, a fixed descriptive level that is not tuned on our evaluation set. A rolling variant, one minus the cosine between consecutive window vectors, separates continuing change from a continuation that has settled into a stable but different appearance; it carries no validated threshold and appears only as the dashed curves in the style figures.

Geometry. Qwen3-VL-8B-Instruct receives four real frames from the candidate's conditioning clip and 16 generated frames sampled uniformly across the window, all labelled and resized to 640×352. It is asked to identify significant reality-breaking failures—including impossible geometry, dissolving surfaces, and large corrupted regions—while ignoring minor blur and small local artifacts. All conditioning spans end at the same underlying real-video frame 32. The probe therefore uses source frames $\{20, 24, 28, 32\}$ for every multi-frame model and repeats source frame 32 for the one-frame systems; for minWM DMD these source frames occupy local conditioning positions $\{16, 20, 24, 28\}$. The reply is force-prefixed with `{"answer": "` so that no free text is decoded, the score is the softmax of *Yes* against *No* at that position, and a window fails above 0.5.

Conjuration. Conjuration is temporal: it asks whether an object was visible before its detected birth. Legitimate objects may enter from a border, approach from below the detector's resolution limit, or emerge from behind an occluder. RT-DETR R50 runs densely over every frame to 30



Figure 7: Matched forward-command continuations for SpatialVID context `sample04`.

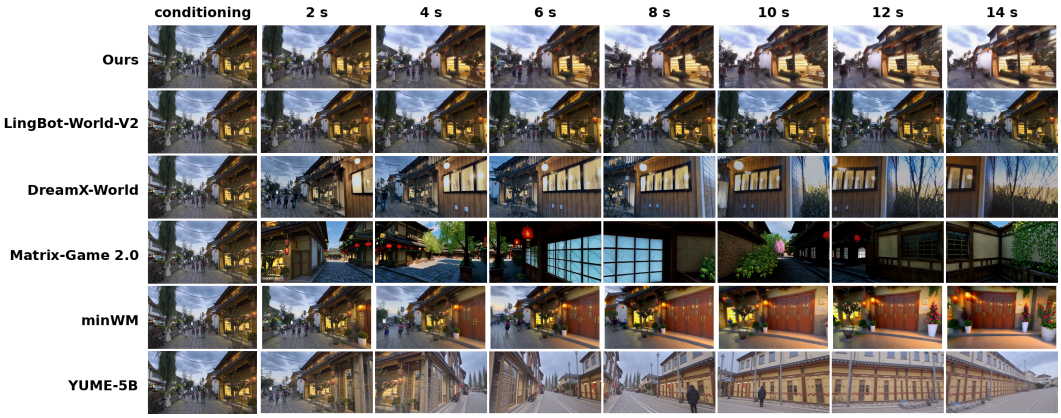

Figure 8: Matched right-command continuations for SpatialVID context `sample04`.

seconds, cached at confidence 0.10 so that the weak-detection tail remains available as evidence of prior presence, with tracks linked from detections scoring at least 0.30. A candidate must belong to a high-recall set of people, animals, vehicles, and portable or sports-object classes. It must appear after the context boundary, reach confidence 0.70, occupy at least 0.0015 of the frame at birth, and originate within the central 70% of the image width. It must also clear both lateral borders by 0.06 of the width during its first half second, survive through 75% of the clip, and appear in at least 55% of frames since birth. Every track passing these proposal gates is reviewed using a seven-frame temporal evidence sheet spanning one second before to two seconds after its detected birth; the earlier onset, zoom, and back-match scores are shown as evidence but are not automatic acceptance rules. A candidate is accepted only when the object is absent beforehand, appears abruptly in the interior, and persists. Boundary entries, egocentric hands or body parts, occlusion reveals, and detector false positives are rejected. Frame counts are quoted at 16 fps and rescaled to each clip's own rateThe crystallisation branch is disabled because it is not separable from an occlusion reveal with the evidence available. A confirmed birth is assigned to the first scored window containing its onset and is carried forward thereafter; later frames establish persistence of the detected track rather than supplying repeated trajectory endpoints.

High-frequency degradation. We resize each frame to 832×448, convert it to grayscale, and measure its Laplacian variance. The base window contains four frames one second after the context boundary; the end window contains eight stride-two frames spanning the final 15 frames of the scored window. Their difference is $d = L_{\text{end}} - L_{\text{base}}$. Because absolute sharpness confounds scene difficulty with model behaviour, d is referenced to a panel of six seats, one per model family, frozen per scene and endpoint. The scored video replaces its own family's seat. The statistic is

$\mathcal{B} = \text{median} - d$, and a window fails when $\mathcal{B} > 150$. The rate is therefore peer-relative, stating how often a model degrades more than its contemporaries on the same content rather than how often degradation is visible in absolute terms.

Six-second scene relocation. Six-second scene relocation uses the released Reloc-Cal-v2 feature instrument. Frames are resized to 640×352, converted to grayscale, and described with `ORB_create(3000)`. Descriptors are matched by brute-force Hamming distance with mutual cross-check and verified with a single homography using RANSAC at a five-pixel reprojection tolerance. For each candidate, the panel contains the endpoint from one representative of every other model family and real frames from the conditioning span that candidate actually received. The candidate replaces its own family seat before that seat is excluded, so an ablation cannot be supported by another checkpoint from our family and minWM DMD cannot be supported by minWM ODE. The score is the largest verified-inlier count over the eligible panel and fails below 50. All 288 videos are scored, including no-op.

Long-horizon relocation proposals. For longer rollouts, direct comparison with those fixed references is not a valid relocation test: continuous camera travel can remove all overlap with the starting view. We therefore sample the rollout at 4 Hz and search for a persistent break between the one-second banks immediately before and after each candidate boundary. A boundary is proposed only when the largest cross-bank ORB–RANSAC match has fewer than 20 inliers, the median adjacent match within each bank has at least 35, and each bank has a median of at least 100 detected keypoints. The within-bank requirements reject featureless collapse, while the cross-bank requirement allows gradual travel to remain connected through intermediate views.

Long-horizon relocation adjudication. Because sparse sampling can still turn a fast camera motion or occlusion into an apparent break, we decode the native frames around the boundary. We compare the HSV-histogram change across the exact adjacent-frame boundary with the median adjacent-frame change in the surrounding two seconds and retain only changes larger than three times that local baseline. Finally, every retained sequence is inspected in temporal order; a positive event requires persistent replacement of physical place identity and excludes turns, occlusions, lighting changes, blur, progressive corruption, and collapse. The complete scan generated 411 sparse correspondence proposals, of which 39 survived the native-frame gate. Inspection found every one to be continuous, so no new relocation event was added after the six-second anchor. Accepted onsets are carried forward in the cumulative failure graphs rather than reported as separate endpoint measurements.

E VALIDATION OF THE EVALUATION INSTRUMENTS

The appearance instruments are inherited from Action Forcing (Sundar et al., 2026), where candidate detectors were selected against human judgements rather than against model identity or the results reported here. We reproduce the validation evidence because the six axes are intentionally non-interchangeable: a relocated video can have clean geometry, and a geometrically corrupted video can remain sharp. Table 1 gives the labelled sets and the deployed detector's held-out discrimination. The validation videos are distinct from the ICLR fleet. Figures 9–13 are reproduced unchanged from Action Forcing (Sundar et al., 2026): they depict Action Forcing's validation videos and its detector decisions. We include them so that the instruments this paper depends on can be inspected.

Geometry and style. The geometric-corruption probe was compared with PAL4VST, two crop-based melt prompts, and depth-field curvature. Qwen3-VL obtains the highest AUC, while the same prompt remains consistent across InternVL3 and Cosmos-Reason1 (AUC 0.75 and 0.74). For style, DINOv2 outperforms CLIP, three direct VLM style judgements, and VGG Gram distance; the direct VLMs range from 0.55 to 0.67 AUC. Figures 9 and 10 expose representative positives and negatives, including a clean geometry judgement for a relocated rollout. That example is why relocation is not folded into geometry.

Relocation. The verified ORB score was compared with AKAZE–RANSAC (AUC 0.63), SIFT–RANSAC (0.62), and raw ORB matching without geometric verification (0.54). Figure 11 shows

Table 1: Human validation of the five appearance instruments inherited from Action Forcing. The family-balanced high-frequency row reports the exact deployed reference construction; the stronger dark-channel result came from the earlier candidate comparison and is not a matched rescore under that reference panel. ORB–RANSAC is used for relocation despite DINOv2's slightly higher AUC because its absolute verified-match count is interpretable and the same primitive supports the long-horizon continuity audit.

Instrument	Human-labelled set	Validated signal	Validation	Strongest alternative
Geometry	100 videos; 36 positives	Qwen3-VL inconsistency	AUC 0.78	PAL4VST, AUC 0.77
Style	100 videos; 20 positives	DINOv2 drift	AUC 0.72	CLIP drift, AUC 0.70
Relocation	100 videos; 19 positives	ORB–RANSAC inliers	AUC 0.70	DINOv2 drift, AUC 0.72
Conjuration	37 videos; 15 positives	RT-DETR + temporal evidence	$P = 1.00$, $R = 0.80$	Faster R-CNN, same P/R
High frequency	112 videos; 13 positives	family-balanced Laplacian loss	AUC 0.68	dark-channel drift, AUC 0.82

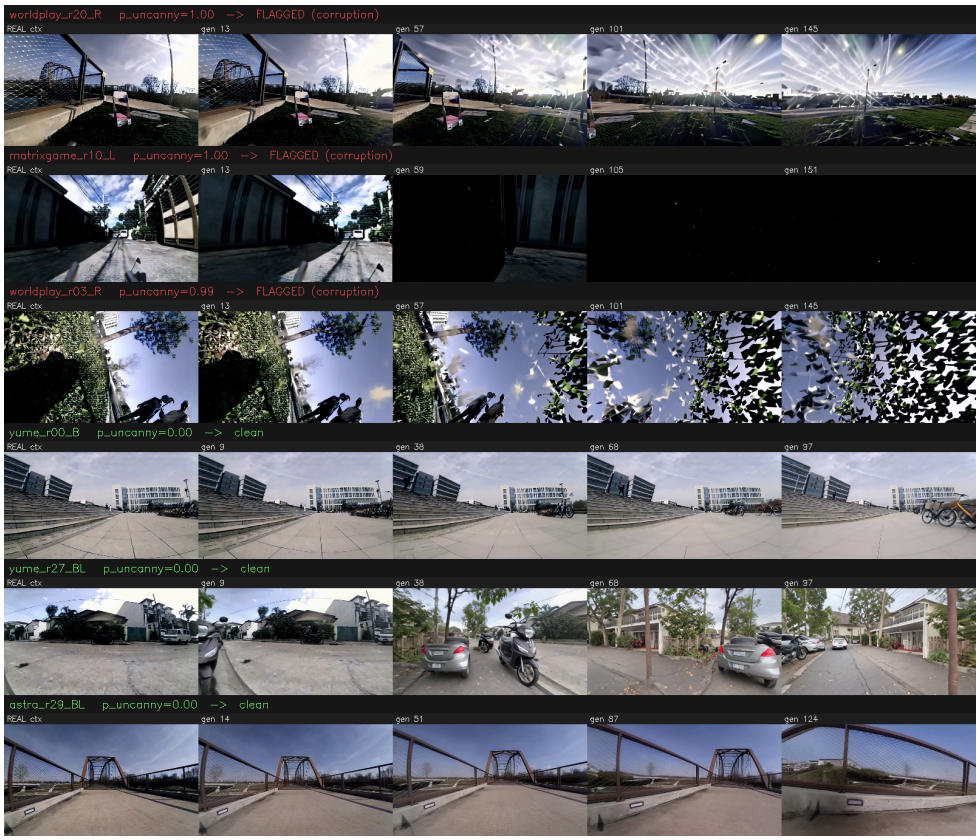

Figure 9: Representative geometry-probe decisions from the human-labelled validation set. The top rows contain visible structural corruption; the clean rows include a scene relocation that is correctly outside this instrument's definition. Reproduced from Action Forcing (Sundar et al., 2026).

representative cases and two known confounds: severe occlusion and post-collision content can remove correspondences without replacing the underlying scene. We retain these cases rather than tune the threshold to the evaluated models. The long-horizon extension is validated separately:



Figure 10: Representative DINOv2 style-drift decisions from the human-labelled validation set. The instrument responds to rendering changes rather than to camera displacement or object layout. Reproduced from Action Forcing (Sundar et al., 2026).

its native-frame gate rejects all 13 continuous real transitions and retains all 65 synthetic scene replacements in the construction pilot. Human inspection is then applied to the small surviving proposal set, as described in Appendix D.

Conjuration. Because spontaneous object birth is rare, the validation set is enriched with 15 confirmed births and 22 hard negatives. RT-DETR and Faster R-CNN produce identical decisions under the same prior-appearance, zoom, and back-match temporal evidence criteria ($P = 1.00$, $R = 0.80$, $F_1 = 0.89$); RT-DETR is retained because the end-to-end pipeline runs at 94.8 fps rather than 56.3 fps. RetinaNet obtains $P = 1.00$, $R = 0.73$. Figure 12 shows both true births and legitimate entries rejected by the temporal tests.

High-frequency degradation. The family-balanced Laplacian score is deliberately conservative: on 112 ablation rollouts it obtains AUC 0.68, compared with 0.82 for an earlier sibling-relative dark-channel measure and 0.64 for contrast drift. We use the Laplacian construction because it measures the stated loss of spatial detail and permits a fixed family-balanced reference, but its lower AUC means that this axis should be read as a noisy diagnostic rather than ground truth. Figure 13 shows representative high- and low-score cases. The ICLR panel has six family seats rather than the seven used in the original validation, so its absolute failure rates are not compared across papers.

Control. Control is not a learned binary detector selected on these rollout labels. CoTracker displacement is projected through the frozen PCA basis recovered from the teacher action interface, whose leading components encode longitudinal motion and yaw. The resulting decisions are deterministic tests on the frozen motion representation, not classifiers fitted to an evaluated model.



Figure 11: Representative scene-relocation decisions. Verified feature correspondence detects loss of place identity, but occlusion and collision are documented failure cases and motivate the subsequent temporal audit. Reproduced from Action Forcing (Sundar et al., 2026).

F DMD ABLATIONS

The ablations distinguish the roles of the DMD-stage stabilisers. Figure 14 follows the complete model and three selected variants across the six windowed axes and summarises the contraction of the jointly clean set. Removing the CARN cache commit raises six-second directional-control failure from 21.1% to 43.4%, and raises style from 0.0% to 7.3%, geometry from 7.3% to 46.9%, and relocation from 6.3% to 27.1%. Removing CARN entirely raises directional-control failure to 43.8% at six seconds and 37.5% at 30 seconds, compared with 21.1% and 14.1% for the complete model. Restricting Temporal Normalisation to mean and energy or to variance and total variation raises six-second geometry failure to 86.1% and 98.3%, respectively. Conjuration is 0.0% for all five rows at six seconds; relocation is 6.3%, 10.4%, 27.1%, 55.6%, and 82.6% for the complete, No CARN, No CARN commit, mean-and-energy, and variance-and-TV rows, respectively. The overall summary shows the full No CARN arm and omits the separate no-commit arm.

G RUNTIME ANALYSIS

Table 2 reports runner-level timing for a 30-second rollout. The runners time slightly different spans, so these values establish the operating regime rather than a tightly controlled latency ranking. The first five rows use one RTX 5090. YUME-5B instead requires its released four-rank distributed path in our environment and was run on four GH200 GPUs.



Figure 12: Conjuration validation. Flagged rows show empty ground before a persistent object appears; hard negatives show objects that approach, enter from a boundary, or were already weakly visible. Reproduced from Action Forcing (Sundar et al., 2026).

Table 2: Runner-reported time for a 30-second rollout and the nominal generated-seconds-per-wall-clock-second factor. The first five rows use one RTX 5090; YUME-5B uses four GH200 GPUs. Our value is the median of 18 runs. The YUME value is the median per-rollout batch latency over nine 32-rollout action jobs, each executed as eight batches of four. Both values include decoding and video output.

MODEL	REPORTED TIME (S)	REAL-TIME FACTOR
Ours	91.5	0.33
LingBot-World-V2 1.3B	119.6	0.25
DreamX-World 5B	73.8	0.41
Matrix-Game 2.0	49.9	0.60
minWM 1.3B (DMD)	119.1	0.25
YUME-5B	233.8	0.13



Figure 13: High-frequency validation examples. High-score rollouts lose texture and contrast over time; low-score examples preserve spatial detail. Reproduced from Action Forcing (Sundar et al., 2026).

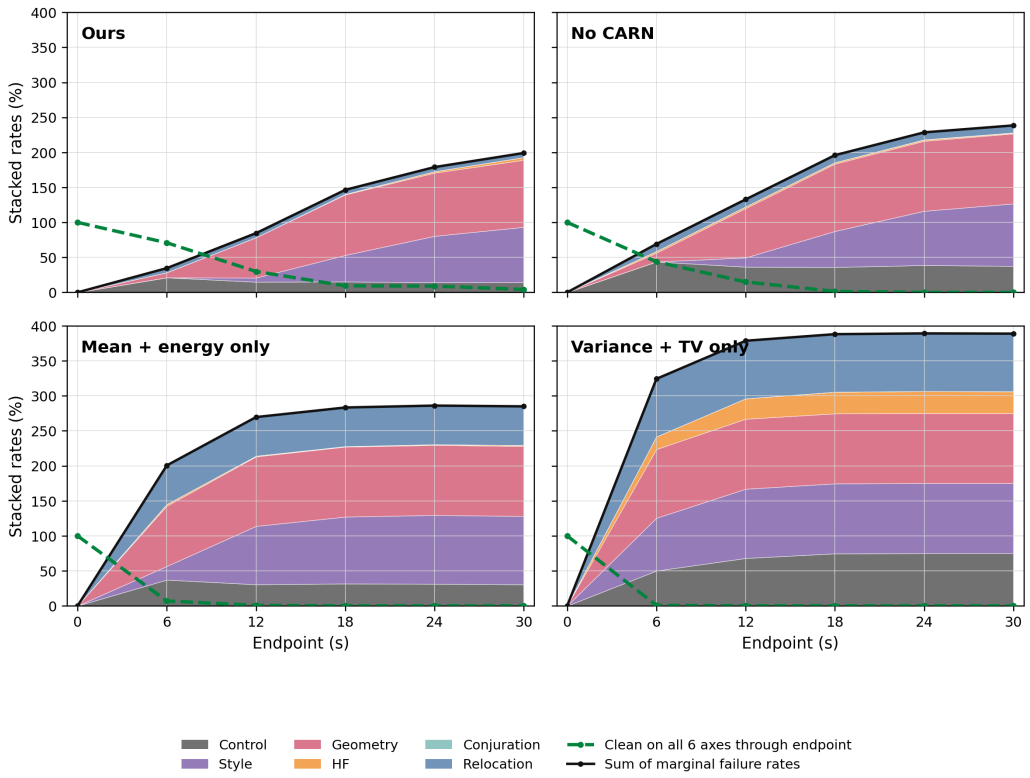

Figure 14: Failure accumulation for the complete model, the full No CARN arm, and the two Temporal Normalisation ablations on the balanced four-dataset panel. Coloured bands stack directional control, style, geometry, HF, conjuration, and relocation marginal rates and may therefore exceed 100%. Event bands persist after onset. The dashed green curve is the cumulative fraction of all 288 rollouts clean on all six axes through each endpoint.