# EVIDENCE-GATED STATE ANCHORING FOR PERSISTENT VIDEO WORLD MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Pretrained video world models change regions that nothing has touched, and they fail to restore a view after the camera looks away and comes back. Both behaviors come from the same gap. Self-attention can read any token in the window, while the training loss never asks whether a location was licensed to change. Recent memory-augmented world models decide which history stays readable. They leave the local write unconstrained.

Evidence-Gated State Anchoring (EGSA) is a frozen-backbone retrofit that turns correspondence and propagated-change evidence into write authorization. Where that evidence is absent, a residual hold mixes the clean-state prediction toward a warped copy of the model's own previous estimate, inside every denoising step. Training touches fewer than 0.4 percent of parameters. A warmup multiplier and zeroed LoRA output projections make the retrofit identical to the base model at step zero, and the evidence detector sees only the noised model input. A no-bias ablation matches the full model to within 0.5 percent on both backbones, which places the gain on the anchoring pathway.

On a 0.5B latent flow-matching model and the 0.46B pixel-space DFoT RealEstate10K model, EGSA improves real-capture palindromic revisits against a matched LoRA twin. On DFoT, look-away error drops by 67 percent and near-origin return error by 12 to 14 percent (86 to 93 percent of held-out scenes, scene-level $p < 10^{-4}$), across training seeds, rollout noise seeds, and a monotone inference dose. On the flow-matching model, all three training seeds lower return error on 128 stronger-departure held-out scenes, by up to 12 percent. Revisit targets are captured frames, fidelity is reported together with a view-change tracking ratio, and all revisit scores are computed in decoded pixels.

## 1 INTRODUCTION

Action- and pose-conditioned video generators now roll out long sequences as candidate simulators (Ha & Schmidhuber, 2018; Hafner et al., 2023; Brooks et al., 2024; Bruce et al., 2024; Agarwal et al., 2025; Zhao et al., 2026). The same diffusion and flow-matching stacks used for video synthesis (Ho et al., 2020; Liu et al., 2023; Peebles & Xie, 2023) have been turned into interactive worlds (Valevski et al., 2024; Parker-Holder et al., 2024; Chen et al., 2024; Song et al., 2025). A planner that looks away and comes back needs the room it left.

Two violations show up once rollouts leave the training horizon. Regions that received no contact change on their own, and a view seen before a look-away is a different view when the camera returns. Calling this temporal inconsistency misses the state claim underneath: a location that received no influence should still hold its last observation when the ray comes back.

In a DiT or U-ViT (Peebles & Xie, 2023; Bao et al., 2023; Vaswani et al., 2017), every token can read every other token in the window. The flow or denoising loss scores the frame, and it does not record whether that location was allowed to change. Small unauthorized writes then accumulate. Giving the model a longer context, a retrieved keyframe, or a 3D cache makes old state readable (Xiao et al., 2025; Wu et al., 2025b; Ren et al., 2025; Yu et al., 2025; Yi et al., 2026; Wu et al., 2026; Wang et al., 2026; Garcin et al., 2026). Read access still leaves the denoiser free to overwrite what it reads.



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.

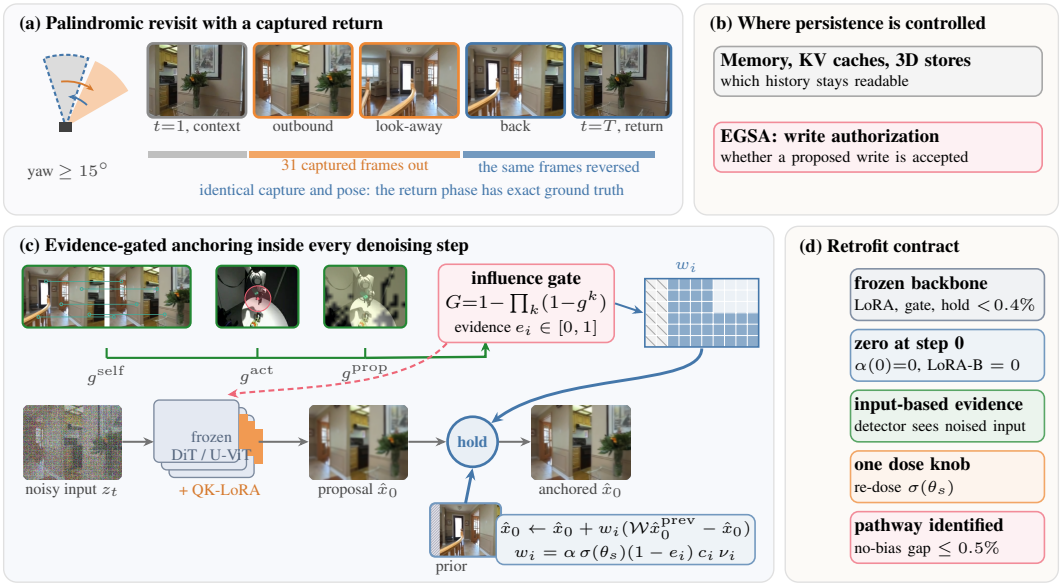



Figure 1: **Overview of EGSA.** (a) Palindromic revisit with a captured return target. (b) Read control versus write control. (c) Schematic of one denoising step: evidence sets the write weight $w_i$ that anchors the proposal to the warped previous estimate. (d) Retrofit properties.

Evidence-Gated State Anchoring (EGSA; Figure 1) puts a license on the write. An influence gate scores per-token support from correspondence, the continuation of a scene ray, from the reach of an action, and from controlled propagation of detected change. The quantity that enters the sampler is a per-token weight. It mixes the current clean-state estimate with a detached, correspondence-warped copy of the model's own previous estimate. The same evidence can also drive an additive attention-logit bias (Qiu et al., 2025; Cong et al., 2024; Ma et al., 2026); on both backbones, removing that bias changes revisit error by at most 0.5 percent, so the method is named for the anchoring pathway that carries the gain.

Three constraints keep the comparison interpretable. The backbone stays frozen. Trainable parameters are fused-QK LoRA (Hu et al., 2022), the gate, and the hold, under 0.4 percent of the model, and every comparison uses a LoRA-only twin with the same rank, data order, and seeds. A warmup multiplier and zeroed LoRA output projections reproduce the base model at step zero, in the same spirit as ControlNet's zero convolutions (Zhang et al., 2023). The evidence detector reads the noised input the network itself sees, so training and rollout use the same evidence.

We measure persistence with real-capture palindromic revisits (Ye et al., 2026; Gu et al., 2026). Outbound and return frames are photographs, so the return target is exact. The two backbones are MiniWorld-0.5B, a latent flow-matching DiT (Zhao et al., 2026), and DFoT, a 0.46B pixel-space continuous-diffusion U-ViT (Song et al., 2025). On DFoT, EGSA lowers look-away error by 67 percent and near-origin return error by 12 to 14 percent against its LoRA twin, across training seeds, rollout noise seeds, and a monotone dose. On MiniWorld, every training seed lowers outbound, away, and return error on 128 stronger-departure held-out scenes, with return gains of up to 12 percent.

Our contributions are:

- **Write authorization.** A zero-influence, dose-controllable retrofit that treats model-input influence evidence as permission to change persistent state (Section 3).
- **Pathway identification.** The anchoring pathway accounts for the measured gain. Disabling the attention-logit bias changes results by at most 0.5 percent on both backbones (Section 4.3).

Table 1: Where persistence methods act.

| Method | Persistent store | Controlled quantity |
|---|---|---|
| WorldMem | pose-keyed frames | which history is readable |
| Spatial memory | 3D point map + keyframes | which history is readable |
| Context as Memory | retrieved RGB frames | which history is readable |
| WorldKV / WorldTrace | evicted or summarized KV | which cache is readable |
| GEN3C / Mirage / PERSIST | 3D or latent 3D cache | rendering / query of a store |
| Closing the Loop | retrieved chunks + depth bias | attention routing at revisit |
| EGSA (this work) | warped previous estimate | whether a local write is accepted |

- **Transfer.** The same recipe, signals, and protocol improve revisit fidelity on a latent flow-matching DiT and a pixel-space continuous-diffusion U-ViT, trained and evaluated independently (Section 4.2).
- **Revisit measurement.** Captured-frame palindromes with exact return targets, a view-change tracking ratio reported alongside fidelity, and decoded-pixel scores. Raw residuals in a temporally compressing codec mix motion direction into the content error, so we score pixels (Section 4.1).

## 2 Related work

**From world models to video simulators.** The classical line models latent dynamics for planning (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2023; Micheli et al., 2023) and has recently been rebuilt on diffusion (Alonso et al., 2024; Valevski et al., 2024). In parallel, video generators scaled from U-Nets to DiTs and flow-matching transformers (Ho et al., 2022; Blattmann et al., 2023b;a; Peebles & Xie, 2023; Liu et al., 2023; Esser et al., 2024; Yang et al., 2025; Wan Team, 2025), and a subset of that machinery is now sold as a world simulator (Brooks et al., 2024; Bruce et al., 2024; Parker-Holder et al., 2024; Agarwal et al., 2025). Streaming rollout typically uses block-causal attention and per-frame noise, following diffusion forcing and its transformer descendants (Chen et al., 2024; Song et al., 2025; Zhao et al., 2026). Sampling a plausible next frame still leaves unobserved state free to drift.

**Memory as a read channel.** The dominant answer to look-away forgetting is to keep old state addressable: a geometry-grounded point map plus episodic keyframes (Wu et al., 2025b), a pose-and-time keyed frame bank (Xiao et al., 2025), raw context frames retrieved by field-of-view overlap (Yu et al., 2025), evicted KV chunks reinstated by camera or action correspondence (Yi et al., 2026), summary slots with in-distribution virtual positions (Wu et al., 2026), a latent 3D cache queried by warping (Wang et al., 2026), or an evolving latent 3D scene with an explicit camera and renderer (Garcin et al., 2026). GEN3C conditions generation on renderings of a 3D point cache (Ren et al., 2025); Geometry Forcing aligns diffusion features with a geometric foundation model (Wu et al., 2025a). Table 1 places these methods by the quantity they control. The rows above decide which history the model can read. EGSA decides whether a local write is accepted. A retrieved or rendered prior can supply the incumbent that our one-frame warp currently provides, and the same authorization weight can protect that prior from being overwritten.

**Revisit protocols.** Symmetric forward/reverse paths already underpin memory and action-control benchmarks (Ye et al., 2026); reversible cycles provide self-verifying training signals (Gu et al., 2026); LoopBench probes recall after a detour (Wu et al., 2026); WorldModelBench scores physics adherence at scale (Li et al., 2025). We instantiate the cycle on real static-scene video, so every return frame is a captured photograph with a captured pose, and we report fidelity together with a view-change tracking ratio.

**Gating, correspondence, and parameter-efficient control.** Convex interpolation of old and new state is as old as LSTMs and GRUs (Hochreiter & Schmidhuber, 1997; Cho et al., 2014; Srivastava et al., 2015), and head-specific output gates now improve large language models (Qiu et al., 2025).

Optical-flow trajectories have guided video attention (Cong et al., 2024; Teed & Deng, 2020); Closing the Loop retrieves pose-matched history and adds a depth-reprojection bias (Ma et al., 2026). The gate in this paper acts on the clean-state write inside each denoising step, and it is driven by cross-frame physical evidence that a second system can check. The no-bias ablation then shows that this write, with the attention bias removed, accounts for the revisit gain on both backbones. Adapters and LoRA (Houlsby et al., 2019; Hu et al., 2022) let us modify a frozen generator without paying for a second pretraining run; zero-initialized control paths (Zhang et al., 2023) let us start from exact equivalence. That combination is what makes a matched-twin comparison possible.

# 3 METHOD

Figure 1c shows the retrofit. The frozen world model still proposes a clean state at every sampler step. An influence gate turns physical evidence into a per-token weight. A zero weight leaves the proposal unchanged. A larger weight mixes it more strongly toward a warped copy of the model's own previous estimate. The same evidence also feeds an optional attention-logit bias, which Section 4.3 separates from the write.

## 3.1 DESIGN PRINCIPLES

The design follows four principles.

**Conservative authorization.** A channel is enabled when its physical source can be checked independently; otherwise the token keeps the backbone's own prediction. Action reach is tied to an effector calibration, and on angular-departure windows the correspondence field uses rotation, the dominant term in that selection.

**Evidence inputs.** The gate consumes static-camera identity, capture poses, and residuals on clean context pairs from the noised model input. Section 3.3 defines how these inputs form the influence evidence.

**Zero influence at initialization.** The added path reproduces the base model before any gradient step, so later comparisons are against the frozen network itself. The warmup and the zeroed LoRA output projections implement that condition.

**Separated signals.** Dynamics evidence, telemetry labels, and evaluation metrics are computed by separate implementations for their respective roles.

## 3.2 SETTING

The primary backbone is a rectified-flow video world model (Liu et al., 2023; Zhao et al., 2026) with per-frame diffusion-forcing timesteps (Chen et al., 2024). Given clean latents $x_0$ and noise $\epsilon$, the input at noise level $t$ is $z = (1-t)\,x_0 + t\,\epsilon$, the network predicts a velocity $v \approx x_0 - \epsilon$, the clean-state estimate is $\hat{x}_0 = z + t\,v$, and the sampler update is $z \leftarrow z - \Delta t\,v$. Frames are encoded by a causal video VAE (Wan Team, 2025) with 16× spatial and 4× temporal compression into $N$ tokens per latent frame; a window of $T$ latent frames yields $M = TN$ tokens processed by a block-causal DiT with fused-QKV attention. Conditioning (robot actions on DROID (Khazatsky et al., 2024), camera poses on RealEstate10K (Zhou et al., 2018)) enters through modulation. The transferability backbone (Section 4.2) is DFoT, a pixel-space continuous-diffusion U-ViT with per-pixel ray conditioning and no VAE (Song et al., 2025; Bao et al., 2023). The mechanism attaches to whichever mid-level token grid the backbone exposes; nothing in the gate or hold assumes a latent codec.

**One interface, two parameterizations.** The retrofit needs only the map from a network output to a clean-state estimate and its inverse. For the rectified-flow backbone these are $\hat{x}_0 = z + t\,v$ and $v = (\hat{x}_0 - z)/t$. For the variance-preserving $v$-parameterization of DFoT, with $z = \alpha_t x_0 + \sigma_t \epsilon$ and $v = \alpha_t \epsilon - \sigma_t x_0$, they are $\hat{x}_0 = \alpha_t z - \sigma_t v$ and $v = (\alpha_t z - \hat{x}_0)/\sigma_t$. The hold edits $\hat{x}_0$ and returns through the inverse map, so each backbone keeps its own sampler, and a zero weight leaves its velocity unchanged up to floating-point rounding.

### 3.3 INFLUENCE EVIDENCE

Each token $i$ carries a small meta-state: a patch-grid anchor $p_i$, a frame index, a visibility flag, and a dynamics probability $d_i$ from a frozen residual detector that compares the token to its correspondence-aligned previous frame. Training follows a *model-input evidence protocol*: the detector consumes exactly the noised input the network sees, and its residual is trusted only where both frames of a pair are effectively clean context. Everywhere else the dynamics evidence is zeroed. At inference the detector reads the evolving generated state, so the evidence available at training time and at rollout time is the same object.

Three channels combine into pairwise support $s_{ij}$ for query $i$ attending to key $j$. Closed forms live in Appendix A.3; the semantics are what matter here.

**Correspondence support.** A correspondence field maps each token to its previous-frame source. A static exterior camera provides an exact identity correspondence. Under ego-motion we use the rotation-only reprojection of patch rays (Section 4.1), the dominant term for the angular-departure windows we evaluate. The same field aligns the dynamics detector and the hold prior.

**Action support.** Given an effector calibration, actions are projected into patch coordinates and support keys within a learnable reach radius. On camera-pose captures, correspondence and propagation carry the evidence directly.

**Propagation support.** Keys with high dynamics probability $d_j$ extend support to spatial neighbors within a learnable radius, so a licensed change can spread at a controlled rate instead of being treated as an isolated pixel event.

The union is a soft OR, $G_{ij} = 1 - \prod_k (1 - g^k_{ij})$, taking values in $(0, 1]$. An unknown pathway opens when geometric confidence is low, and a same-chunk term treats rendering the current frame as a read. The gate has two outputs. The per-token change evidence $e_i$ enters the residual hold. The other output is the additive attention-logit bias $b_{ij} = \alpha\, m_h \log G(s_{ij})$, with per-head magnitudes $m_h$ and a global warmup $\alpha$, injected into a candidate subset of layers on the backbone's native SDPA kernel.

### 3.4 CONFIDENCE-WEIGHTED RESIDUAL HOLD

The residual hold is the authorization step. It is applied to the clean-state estimate $\hat{x}_0$, and a weight of zero leaves that estimate unchanged:

$$w_i = \alpha\, \sigma(\theta_s)\, (1 - e_i)\, c_i\, \nu_i, \qquad \hat{x}_0 \leftarrow \hat{x}_0 + w_i \big(\mathcal{W}\hat{x}_0^{\text{prev}} - \hat{x}_0\big). \tag{1}$$

Here $\mathcal{W}\hat{x}_0^{\text{prev}}$ is the previous-frame clean estimate warped by the correspondence field and detached from the gradient graph; $e_i$ is the gate's evidence that token $i$ may change; $c_i$ and $\nu_i$ are anchor and visibility confidences; $\theta_s$ is a learnable scale initialized at zero, so $\sigma(\theta_s) = 0.5$; and the warmup satisfies $\alpha(0) = 0$, so $w_i = 0$ at the start. In the reported runs $c_i$ is a dataset-level correspondence confidence and $\nu_i$ is warp validity. The anchored estimate is converted back to a velocity, so the hold acts inside every denoising step. At inference the trained strength can be re-dosed by a logit shift, which gives a single stability–plasticity knob.

The hold combines gated interpolation (Hochreiter & Schmidhuber, 1997) with a zero-initialized control path (Zhang et al., 2023). Physical evidence determines the per-token anchoring weight $w_i$.

### 3.5 ANCHORING AS BOUNDED DRIFT

The weight in Eq. 1 has a simple reading once drift is split into its two sources. Consider tokens that stay visible and that no influence reaches, so their true state only undergoes transport. Let $\mathcal{T}_t$ carry a state from step $t-1$ into the geometry of step $t$, and measure error with the token-wise norm $\|x\|_\infty = \max_i \|x_i\|_2$. One step has two error terms: the backbone's departure from transport, $M_{t,i} = \|\hat{x}^{\text{base}}_{0,t,i} - (\mathcal{T}_t x_{t-1})_i\|_2$, and the error of the warped incumbent, $\eta_{t,i} = \|(\mathcal{W}x_{t-1})_i - (\mathcal{T}_t x_{t-1})_i\|_2$.

**Proposition 1 (drift decomposition).** *If every $\mathcal{T}_t$ is non-expansive in $\|\cdot\|_\infty$, which holds for any token reindexing such as a nearest-patch warp, the anchored rollout $x_t = (1-w_t)\,\hat{x}^{\text{base}}_{0,t} + w_t\,\mathcal{W}x_{t-1}$*

Table 2: Retrofit footprint of one 4000-step run, from the run records.

| Backbone | LoRA (+ bridge) | gate | hold | share | time, twin / full | memory, twin / full |
|---|---|---|---|---|---|---|
| MiniWorld-0.5B | 774,508 | 85 | 1 | 0.14% | 32 / 44 min | 9.2 / 16.7 GB |
| DFoT 0.46B | 147,456 | 50 | 1 | 0.03% | 28 / 38 min | 7.7 / 10.3 GB |

*started from the observed context $x_0$ satisfies*

$$\left\| x_T - \mathcal{T}_T \cdots \mathcal{T}_1 \, x_0 \right\|_\infty \;\le\; \sum_{t=1}^{T} \max_i \Big[ (1 - w_{t,i})\, M_{t,i} + w_{t,i}\, \eta_{t,i} \Big]. \tag{2}$$

The proof is a triangle inequality per step followed by induction (Appendix B). Without the hold ($w = 0$) the bound is the accumulated backbone drift $\sum_t \max_i M_{t,i}$. With it, each token pays the smaller of its two errors when $w_{t,i}$ follows their order: anchoring helps exactly where the incumbent is more reliable than the proposal. Every factor of $w_i$ serves that condition. The evidence term $(1 - e_i)$ removes tokens whose state is licensed to change, where transport is the wrong reference, and $c_i$ and $\nu_i$ remove tokens whose warp error is large or undefined. The bound also predicts where the gain should concentrate: where the rotation warp is most accurate, that is, at small angular departures (Figure 2b).

### 3.6 Retrofit training

Trainable parameters are rank-8 LoRA on the fused Q and K projections of every block (V receives an exact zero delta), gate parameters, hold parameters, and, for pose-conditioned data, a zero-initialized linear bridge from raw relative poses into the frozen conditioning space. The total is under 0.4 percent of the backbone; Table 2 gives the exact counts, 0.14 percent of MiniWorld and 0.03 percent of DFoT, together with the measured cost. The loss is the backbone's masked flow loss on $v$, plus a radius-prior regularizer and a sparse head-magnitude regularizer. The gate and hold receive gradient from the prediction loss through their effect on $v$.

With $\alpha(0) = 0$ and LoRA-B zeroed, gate and hold contributions vanish even though $m_h = \sigma(-6) \approx 0.0025$ and $\sigma(\theta_s) = 0.5$. Forward-pass comparisons verify equivalence with the frozen base at initialization, accounting for bf16 kernel tolerance (Appendix A).

## 4 Experiments

**Backbones and data.** The primary backbone is the frozen MiniWorld-0.5B checkpoint (Zhao et al., 2026), an action-conditioned latent flow-matching world model trained on DROID (Khazatsky et al., 2024), with its Wan 2.2 VAE (Wan Team, 2025) and block-causal rollout. DROID experiments use 96 train and 16 validation episodes. RealEstate10K (Zhou et al., 2018) experiments adapt the same backbone to camera-pose conditioning on 360 training scenes, with evaluation on strictly held-out revisit windows (Section 4.1). The transferability backbone is the official 0.46B DFoT RealEstate10K release (Song et al., 2025), retrofitted and evaluated independently with the same recipe (Section 4.2).

**Variants.** *lora_only*: fused-QK LoRA, matched in rank, data, and seeds. *full_iga*: LoRA + gate + hold. *gated_hold_no_attn*: identical to full_iga except the attention-logit bias is disabled; this is the identifying ablation for the routing pathway. *hold_only*: hold without evidence (blind warp-and-blend with a learned strength). All variants share initialization, data order, and generation seeds within a seed cohort; we report three cohorts (s0, s1, s2) on MiniWorld and two (s0, s1) on DFoT. The sampler is identical for all entries; only the mechanism toggles. All numbers use the evidence protocol of Section 3.3 with captured RGB targets, except where a result is marked *extended sweep*: that configuration trains the detector on clean context, scores against VAE reconstructions, and covers the wide dose and baseline sweeps in Appendix E.

**Metrics.** Pixel MSE and PSNR to the captured target, split by revisit phase, always reported together with the view-change tracking ratio. Statistics are scene-level paired sign-flip permutation tests; the scene is the sampling unit. The persistence numbers below are ego-motion revisits on RealEstate10K, where capture poses give a correspondence field at every frame.

Table 3: MiniWorld RealEstate10K revisit: paired change in pixel MSE of full_iga against its LoRA twin (fraction of scenes improved, sign-flip $p$), rollout noise seed 0.

| Pool | Seed | Outbound | Away | Return |
|---|---|---|---|---|
| primary (56) | s0 | $-4.5\%$ (89%, $p<10^{-4}$) | $-6.3\%$ (82%, $p<10^{-4}$) | $-7.5\%$ (79%, $p=.005$) |
| | s1 | $-3.2\%$ (75%, $p<10^{-4}$) | $-2.4\%$ (62%, $p=.08$) | $-13.3\%$ (80%, $p<10^{-4}$) |
| | s2 | $-2.6\%$ (79%, $p<10^{-3}$) | $-3.3\%$ (68%, $p=.015$) | $-7.4\%$ (73%, $p=.004$) |
| stronger (128) | s0 | $-4.0\%$ (83%, $p<10^{-4}$) | $-6.3\%$ (80%, $p<10^{-4}$) | $-5.3\%$ (69%, $p=.075$) |
| | s1 | $-2.7\%$ (77%, $p<10^{-4}$) | $-2.6\%$ (65%, $p=.019$) | $-12.4\%$ (74%, $p<10^{-4}$) |
| | s2 | $-2.2\%$ (66%, $p<10^{-3}$) | $-3.9\%$ (66%, $p=.002$) | $-10.9\%$ (73%, $p<10^{-4}$) |

### 4.1 REALESTATE10K: PALINDROMIC REVISIT

RealEstate10K paths at this horizon are monotone sweeps, so we build revisit windows from the captures themselves: an outbound run of 31 frames (stride 2) followed by the same frames reversed. Scenes are static, so time reversal is physically valid, every frame-pose pair is a real capture, and the return phase has exact ground truth (Figure 1a). Evaluation is strictly held out: revisit windows are scanned only over the 80 scenes never touched by adaptation or validation, yielding 56 windows with angular departures of at least 15 degrees. The backbone is adapted to pose conditioning with a zero-initialized 12-dimensional relative-pose bridge; the correspondence field and hold prior use the rotation warp of Section 3.3.

#### 4.1.1 A REPRESENTATION FINDING

A temporally compressing VAE encodes each latent frame from a group of RGB frames, so a latent token represents a short spatiotemporal chunk including its motion direction. On palindromes this is directly measurable. Pairs of latent frames with the same pose, the same content, and the same encoder state (both group-encoded) still differ by a raw latent residual of about 1.0, comparable to different content under this metric, because outbound and return motion directions are mirrored. We therefore score revisits in decoded pixels against captured frames. The frozen detector remains a mechanism input. Its output is not used as an evaluation label.

#### 4.1.2 PIXEL-SPACE REVISIT RESULTS

Table 3 reports flow-matching results under the model-input evidence protocol. On both held-out pools, all three matched twin pairs improve in all three phases at rollout noise seed 0. On the stronger pool, 128 scenes with departures of at least 40 degrees and fully disjoint from adaptation, validation, and the primary pool, return error drops by 5.3 to 12.4 percent. Away and outbound error drop by 2.2 to 6.3 percent, with the same sign in every pool-seed cell.

For every RealEstate10K variant and seed, the validation flow loss falls by 14 to 30 percent over the 4000 adaptation steps (from 0.28 to 0.33 at step 0 to 0.22 to 0.26), so the zero-initialized pose bridge learns real conditioning.

Figure 2c resolves Table 3 by phase, pool, and training seed with scene-bootstrap 95 percent intervals. All 18 point estimates are below zero, and 16 of the 18 intervals lie entirely below zero. The return phase carries the largest effect in five of the six pool-seed pairs, which is where a correct incumbent matters most: the return frames revisit content the model has already generated once.

### 4.2 SECOND BACKBONE: PIXEL-SPACE DFOT

We then retrofit the official DFoT RealEstate10K release (Song et al., 2025): pixel-space continuous diffusion (pred-$v$, cosine log-SNR), a U-ViT with per-pixel ray conditioning, per-frame independent noise levels, and no VAE. The representation, the diffusion parameterization, the conditioning pathway, and the training regime all differ from MiniWorld. The mechanism attaches to the $16\times16$ mid-level token grid. The evidence protocol, hold, LoRA, zero-initialization, and matched-twin design are unchanged. Revisit windows are the same 61-frame palindromes subsampled to DFoT's

Table 4: DFoT revisit: paired change in pixel MSE (fraction of scenes improved, sign-flip $p$). The dose-2 row compares with the frozen base.

| Comparison | Away | Return | Tracking ratio |
|---|---|---|---|
| full vs lora (s0, noise 0) | $-67.5\%$ (100%, $p<10^{-4}$) | $-14.3\%$ (93%, $p<10^{-4}$) | 1.92 vs 2.76 |
| full vs lora (s0, noise 1) | $-67.5\%$ (100%, $p<10^{-4}$) | $-14.3\%$ (93%, $p<10^{-4}$) | 1.92 vs 2.77 |
| full vs lora (s1, noise 0) | $-67.0\%$ (100%, $p<10^{-4}$) | $-12.3\%$ (86%, $p<10^{-4}$) | 1.95 vs 2.75 |
| full vs base (dose 2) | $-79.9\%$ (100%, $p<10^{-4}$) | $-19.2\%$ (84%, $p<10^{-4}$) | 1.52 vs 2.47 |

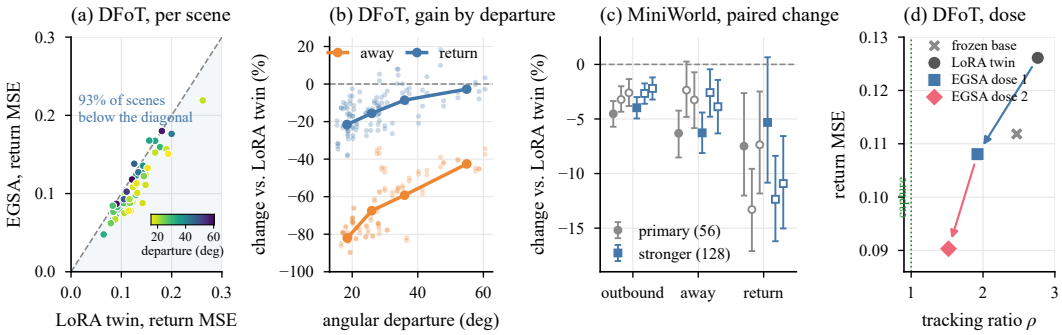



Figure 2: **Revisit results from the per-scene records.** (a) DFoT return MSE per scene. (b) DFoT change against angular departure. (c) MiniWorld paired change with 95 percent scene-bootstrap intervals (filled: s0). (d) DFoT return MSE against tracking ratio across dose.

native 8-frame window (4 context frames end just past the away peak; the 4 generated frames span the away tail and the near-origin return). Scoring is in pixels from the start, since DFoT has no VAE.

The gain transfers (Table 4). Look-away error drops by 67 percent against the matched twin, with every held-out scene improving, and return error drops by 12 to 14 percent. The result holds across training seeds and rollout noise seeds. At dose 2 the retrofit beats the frozen base on return by 19 percent, and the tracking ratio moves from 2.47 toward 1: the hold removes unlicensed change while the commanded motion remains. On the flow-matching backbone the same evidence channels lower error in every pool, seed, and phase of Table 3.

Figure 2a shows the per-scene picture behind Table 4: 52 of the 56 held-out scenes lie below the diagonal. Figure 2b sorts the same scenes by angular departure. Averaged over the departure bins of 15 to 20, 20 to 30, 30 to 45, and above 45 degrees, look-away error falls by 82, 68, 59, and 43 percent and return error by 22, 16, 9, and 3 percent. Every bin improves, and the ordering is the one Proposition 1 predicts: the gain is largest where the rotation warp is the most accurate incumbent. Figure 2d places each DFoT model by fidelity and by executed view change. The LoRA twin moves the view 2.76 times as much as the capture; EGSA at the trained dose and at dose 2 moves toward the captured motion while return error falls monotonically.

### 4.3 Ablations and controls

**Fidelity with tracking.** Indoor palindromes return to a nearby view, so a low pixel error has to be read next to how much view change the model actually executed. The tracking ratio is pixel change relative to the context view during look-away, divided by the same quantity in the capture. On MiniWorld, EGSA stays in its twin's range (0.38 to 0.49 versus 0.30 to 0.54 per cohort and pool) while fidelity improves. The frozen base has no pose bridge and has a tracking ratio of $\rho = 0.007$ to 0.009. Revisit comparisons therefore use the pose-conditioned LoRA twin, which executes the commanded motion. On DFoT the ratio moves from 2.76 toward 1.92 at the trained dose, and to 1.52 at dose 2.

**Anchoring, with the bias removed.** The gated_hold_no_attn variant computes the same evidence and the same hold, with the attention-logit bias set to zero. Across three training seeds on each MiniWorld pool, and across training seeds and rollout noise seeds on DFoT, its largest deviation

Table 5: Synthetic over-change palindrome, 64 scenes (Appendix J).

| Variant | Static MSE ↓ | Cause MSE ↓ | Tracking |
|---|---|---|---|
| unanchored predictor | .725 | .755 | 1.43 |
| evidence-free hold | .298 | 1.580 | 1.17 |
| evidence-gated hold | .309 | .755 | 1.18 |

from full_iga in any phase is 0.26 percent on MiniWorld and 0.07 percent on DFoT. The revisit gain is therefore carried by the anchoring pathway, and the bias is an optional routing term.

**Evidence versus an evidence-free hold.** hold_only applies the same warp-and-blend with a learned strength and with the evidence term removed. On static-scene revisits, where nothing in view is caused to change, both forms anchor the view. The evidence term matters where a token is licensed to change. On the synthetic palindrome of Table 5, where the predictor rewrites static tokens and one token is a known cause, the gated hold cuts static return error by 57 percent on all 64 scenes and keeps the caused token at the predictor's error. The evidence-free hold reaches the same static reduction and raises cause-token error by 109 percent.

**Dose.** On DFoT, doubling the dose moves the away and return gains from 67 and 12 percent against the twin to 80 and 19 percent against the frozen base, and pulls the tracking ratio from 2.47 toward 1. At the trained dose the learned hold weight averages 0.29 to 0.30 per token across the three DFoT training seeds, so the operating point is a partial blend, and the dose moves it toward the incumbent. The extended sweep shows the same single-knob behavior up to $3\times$ trained strength (Figure 3b, c). Shifting the hold initialization by $\pm 1.5$ logits changes the trained operating point and leaves the dose direction intact. The strength is a learned scalar that can be re-dosed at inference, in the same practical sense as a guidance scale (Ho & Salimans, 2022).

## 5 Discussion

A frozen video world model can take a write-authorization layer with zero initial influence, LoRA-scale trainable state, and a detector that sees only the noised input. Evidence-gated anchoring then improves real-capture revisit fidelity against matched twins on two disjoint backbones. On DFoT the gain is large and repeats across training seeds, noise seeds, and dose. On MiniWorld it holds on a 128-scene disjoint pool, and removing the attention bias leaves both results in place. The retrofit adds that persistence to a frozen generator at well under one percent of its parameters.

## 6 Limitations

Reported correspondence is a rotation-only nearest-patch field on angular-departure windows. Translation-dominant motion lies outside that field, and Appendix J records where the field matches a full pose. The tables anchor on the previous finalized estimate along a continuous reverse path. A directional last-valid memory for a hard cut is implemented (Appendix A.5) and is not used in those tables. The palindromes are static, so a known cause separates evidence-gated anchoring from an evidence-free blend only in Table 5. Table 3 uses rollout noise seed 0. Both backbones have about half a billion parameters.

## 7 Conclusion

Evidence-gated state anchoring treats persistence as write authorization: where no verifiable evidence licenses a change, the write proposed by a frozen world model is pulled back toward its own warped previous estimate. The retrofit trains under 0.4 percent of the parameters, starts at zero influence, reads only the noised input, and improves real-capture revisit fidelity on two disjoint backbones, with one dose knob and a gain that the no-bias ablation places on the anchoring pathway.

### Reproducibility statement

The supplementary materials provide the implementation, configurations, data preparation and evaluation scripts, and per-scene records for the main experiments. Separate seed streams control initialization, data order, and rollout noise. Appendices A, D, and K specify the backbone settings, revisit construction, and hyperparameters. Canonical RealEstate10K checkpoints are retained, and seeded retraining reproduces their losses to three decimals.

### Ethics statement

This work evaluates existing world-model checkpoints on the public DROID and RealEstate10K datasets. No new human subjects data were collected. The datasets contain residential imagery and robot demonstrations, so their existing privacy considerations apply to this work. The retrofit controls temporal state persistence; content moderation is a separate system function. Persistent video simulators can support both benign and dual-use downstream applications.

### The use of large language models

Large language models were used only for language editing and polishing of the manuscript. The authors checked the revised wording before inclusion.

## References

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025.

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

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. $\pi_0$: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. OpenAI Technical Report, 2024. URL `https://openai.com/index/video-generation-models-as-world-simulators/`.

Jake Bruce, Michael 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 *International Conference on Machine Learning*, 2024.

Boyuan Chen, Diego Martí Monsé, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, 2024.

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

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. FLATTEN: Optical flow-guided attention for consistent text-to-video editing. In *International Conference on Learning Representations*, 2024.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International Conference on Machine Learning*, 2017.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Samuel Garcin, Thomas Walker, Steven McDonagh, Tim Pearce, Hakan Bilen, Tianyu He, Kaixin Wang, and Jiang Bian. Beyond pixel histories: World models with persistent 3D state. *arXiv preprint arXiv:2603.03482*, 2026.

Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, et al. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538 (7626):471–476, 2016.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Bohai Gu, Yueyang Yuan, Taiyi Wu, Dazhao Du, Jian Liu, Xiaoyi Pang, Jie Zhang, Xiaocheng Lu, Haobin Zhong, Xiaotong Zhao, Alan Zhao, and Song Guo. WorldCycle: Self-verifiable reinforcement learning for long-horizon video world models. *arXiv preprint arXiv:2608.04964*, 2026.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

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

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Robotics: Science and Systems*, 2024.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Wenchao Ma, Changran Liu, Sharon X. Huang, and Haomiao Jiang. Closing the loop: Training-free revisit consistency for autoregressive generative rendering. *arXiv preprint arXiv:2607.21848*, 2026.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *International Conference on Learning Representations*, 2023.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, et al. Genie 2: A large-scale foundation world model. Google DeepMind Technical Report, 2024. URL https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.

Zihan Qiu, Zekun Wang, Bo Zheng, Zeyu Huang, Kaiyue Wen, Songlin Yang, Rui Men, Le Yu, Fei Huang, Suozhi Huang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Gated attention for large language models: Non-linearity, sparsity, and attention-sink-free. In *Advances in Neural Information Processing Systems*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 2020.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-informed world-consistent video generation with precise camera control. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *International Conference on Machine Learning*, 2025.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y. Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026.

Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3D representation for consistent world modeling. *arXiv preprint arXiv:2507.07982*, 2025a.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *Advances in Neural Information Processing Systems*, 2025b.

Xindi Wu, Sven Elflein, James Lucas, Olga Russakovsky, Laura Leal-Taixé, Despoina Paschalidou, Jonathan Lorraine, and Aljoša Ošep. Addressable memory for video world models. *arXiv preprint arXiv:2608.07408*, 2026.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. WorldMem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. CogVideoX: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. MIND: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.

Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. WorldKV: Efficient world memory with world retrieval and compression. *arXiv preprint arXiv:2605.22718*, 2026.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. *arXiv preprint arXiv:2506.03141*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 2023.

Yian Zhao, Ruochong Zheng, Hongcan Guo, Yu Yan, Jian Zhang, and Jie Chen. MiniWorld: Democratizing the training of video world models from scratch. *arXiv preprint arXiv:2608.01127*, 2026.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4), 2018.

# A Implementation details

## A.1 Backbone and VAE

MiniWorld-0.5B (Zhao et al., 2026) is a 0.55B-parameter block-causal DiT with 28 blocks, 16 heads, fused QKV, per-frame diffusion-forcing noise levels, and chunk size 2 latent frames, trained on

DROID at 8 latent frames (29 RGB frames at $240\times320$) with a rectified-flow objective (Ho et al., 2020; Liu et al., 2023). The VAE is Wan 2.2 (Wan Team, 2025): $16\times$ spatial, $4\times$ temporal, 48 latent channels, 300 tokens per latent frame. DFoT (Song et al., 2025) is a 0.46B pixel-space U-ViT (Bao et al., 2023) with per-pixel Plücker-ray conditioning, pred-$v$ parameterization, and a cosine log-SNR schedule; there is no VAE.

### A.2 Retrofit parameterization and optimization

LoRA rank 8, $\alpha = 16$, on fused Q and K slices of all 28 MiniWorld blocks (V delta exactly zero); the DFoT attachment uses the same rank on the mid-level token grid. Gate: per-head magnitudes on 8 candidate layers (indices 2, 5, 8, 11, 14, 17, 20, 23 on MiniWorld), warmup 600 steps. Hold: scale logit zero-initialized. Optimizer AdamW (Loshchilov & Hutter, 2019), learning rates $5\times10^{-5}$ (LoRA) and $1\times10^{-4}$ (gate, hold), 4000 steps, batch 1, gradient clipping 1.0, with bf16 autocast on A800 and fp16 with dynamic loss scaling on V100. Wall-clock and memory per variant are in Table 2 and Appendix G.

### A.3 Gate: evidence terms, union, and logit bias

Each token $i$ carries a patch-grid anchor $p_i$, latent-frame index $t_i$, visibility flag $\nu_i$, anchor confidence, and dynamics probability $d_i$ from the frozen detector. All reported runs use the `noisy_context` mode, which feeds the detector the actual noised model input and accepts its evidence only where adjacent frames are clean context; the extended sweep of Appendix E uses the clean-context mode. Pose-conditioned runs store world-space Plücker rays and use ray distance for $g^{\text{self}}$, while the extended sweep uses patch-grid distance. For a query $i$ and key $j$ define $\Delta_{ij} = t_i - t_j$, the past indicator $\pi_{ij} = \mathbf{1}[\Delta_{ij} > 0]$, and the edge confidence $c_{ij}$ (product of the endpoint anchor confidences). The support terms, all in $[0, 1]$, are

$$g_{ij}^{\text{self},h} = c_{ij}\,\exp\!\big(-D_{ij}^2/2\sigma_h^2\big)\,\pi_{ij}, \qquad \sigma_h \text{ learnable per head}, \tag{3}$$

$$g_i^{\text{act}} = a^{\text{conf}}\,\sigma\big((r_a - \|p_i - p^{\text{agent}}\|)/\tau_a\big)\,\mathbf{1}[\text{action present}], \qquad r_a \text{ per action type}, \tag{4}$$

$$g_{ij}^{\text{prop}} = d_j\,c_{ij}\,\sigma\big((r_0 + \kappa\,\Delta_{ij} - \|p_i - (p_j + v_j\Delta_{ij})\|)/\tau_p\big)\,\pi_{ij}, \tag{5}$$

where $D_{ij}$ is the anchor distance and $r_a$, $\tau_a$, $r_0$, $\kappa$, $\tau_p$ are softplus-parameterized learnable scalars. Three further terms complete the set: an unknown pathway $g_{ij}^{\text{unk}} = \text{clamp}(1 - c_{ij} + p_i^{\text{glob}}, 0, 1)$, so low geometric confidence leaves the gate open; a whitelist term $g_j^{\text{glob}}$ for global columns; and a same-chunk transport term $g_{ij}^{\text{same}} = \mathbf{1}[\Delta_{ij} = 0]$, because rendering the current frame is a read, not a state write. The union is a soft OR with no double penalty,

$$G_{ij} = 1 - \prod_k \big(1 - g_{ij}^k\big), \qquad k \in \{\text{self}, \text{act}, \text{prop}, \text{unk}, \text{glob}, \text{same}\}, \tag{6}$$

and the additive logit bias on scaled dot-product attention (Vaswani et al., 2017) is

$$b_{ij}^h = \beta_h\,g_{ij}^{\text{self},h} + \xi_h\,g_{ij}^{\text{prop}} + \alpha(s)\,m_h\log\big(\epsilon + G_{ij}\big), \qquad m_h = \sigma(\tilde{m}_h),\ \ \epsilon = 10^{-3}, \tag{7}$$

which the main text abbreviates by its dominant penalty term. At initialization $\beta_h = \xi_h = 0$ and $\tilde{m}_h = -6$ (so $m_h \approx 0.0025$), and the warmup satisfies $\alpha(0) = 0$, so the bias tensor is exactly zero at step 0. The per-query cause evidence consumed by the hold is

$$e_i = 1 - \big(1 - g_i^{\text{act}}\big)\big(1 - \max_j g_{ij}^{\text{prop}}\big)\big(1 - p_i^{\text{glob}}\big). \tag{8}$$

The bias is added to the backbone's existing additive block-causal mask on the candidate layers, so the pretrained kernels run unchanged.

### A.4 Residual hold

With $e_i$ from the gate, anchor confidence $c_i$, visibility $\nu_i$, and occlusion probability $o_i$, the hold weight and update are

$$w_i = \alpha\,\sigma(\theta_s)\,(1 - e_i)\,c_i\,\nu_i\,(1 - o_i), \qquad \hat{x}_0 \leftarrow \hat{x}_0 + w_i\big(\mathcal{W}\hat{x}_0^{\text{prev}} - \hat{x}_0\big), \tag{9}$$

written in this residual form so that $w_i = 0$ reproduces the base estimate exactly, which the zero-initialization check requires bitwise on this path. $\mathcal{W}$ is the correspondence warp of Appendix D, and the prior is detached from the gradient graph. At inference a dose $\delta$ rescales the trained strength in probability space, $w_i \propto \text{clamp}(\delta\,\sigma(\theta_s))$, with $\delta = 1$ the trained operating point.

### A.5 Rolling sampler with mechanism injection

The MiniWorld window holds $T$ latent frames with chunk size 2 and 13 flow steps per chunk on the backbone's shifted timestep schedule; the context frame is pinned at noise level 0 and frames beyond the active chunk stay at pure noise. In `chunk` mode the sampler builds evidence from finalized generated frames once per chunk and holds the resulting bias fixed across its denoising steps:

1. build meta-state from finalized generated chunks (future and active dynamics probabilities are zero);
2. compute the gate bias and add it to the block-causal mask on the candidate layers; the forward pass yields the velocity $v$;
3. form the clean estimate $\hat{x}_0 = z + t\,v$ and apply the residual hold;
4. convert back, $v \leftarrow (\hat{x}_0 - z)/t$, and take the Euler step $z \leftarrow z - \Delta t\, v$.

Finished chunks join the context and the window rolls. The sampler is identical for all entries; only the mechanism toggles. The bias is present in every forward but fixed within a chunk, while the hold updates every denoising step. The `step` mode refreshes detector evidence and bias from each previous clean-state estimate. The MiniWorld results in Table 3 use step mode, and the extended sweep uses chunk mode. The released sampler additionally offers a `last_valid` hold anchor: finalized frames write their latent tokens into a world-direction memory (an azimuth/elevation grid keyed by each patch's camera ray), and revisited directions are re-anchored to their most recent finalized observation instead of the neighbor frame. This is the inference-time control that a hard-jump revisit requires; all reported tables use the `prev` anchor.

### A.6 Zero-initialization equivalence

Before optimization, a retrofit forward is compared with the frozen base on the same batch. The float32 CPU outputs are bitwise identical. On CUDA, a bf16 tolerance of $2\times10^{-2}$ accounts for the different reduction orders of the SDPA math and memory-efficient kernels. Exact zero influence follows from $\alpha(0) = 0$ and zeroed LoRA-B. The same forward-pass comparison is used for the extended sweep.

### A.7 Evaluation setup

DROID: 32 validation clips (8-frame) or 16 clips (16-frame). Static/dynamic masks derived from ground-truth latent residuals at threshold 0.35; SCR excludes the context frame. RealEstate10K: 56 primary held-out revisit windows (scanned only over the 80 scenes never used for adaptation or validation) plus 128 stronger-look-away windows from the 6728 full-split test scenes (protocol in Appendix D, full tables in Appendix E); pixel metrics on decoded $240\times320$ frames in $[-1, 1]$. Scene-level statistics: per-scene paired differences averaged over the canonical cohorts, sign-flip permutation test (100000 permutations), bootstrap CIs over scenes (20000 resamples).

## B Proof of Proposition 1

Write $y_t = \mathcal{T}_t \cdots \mathcal{T}_1 x_0$ for the transported context and $d_t = \|x_t - y_t\|_\infty$, so $d_0 = 0$. For every token $i$,

$$x_{t,i} - y_{t,i} = (1-w_{t,i})\big(\hat{x}^{\text{base}}_{0,t,i} - (\mathcal{T}_t x_{t-1})_i\big) + w_{t,i}\big((\mathcal{W}x_{t-1})_i - (\mathcal{T}_t x_{t-1})_i\big) + \big((\mathcal{T}_t x_{t-1})_i - (\mathcal{T}_t y_{t-1})_i\big),$$

because $(1 - w_{t,i}) + w_{t,i} = 1$ and $y_t = \mathcal{T}_t y_{t-1}$. With $w_{t,i} \in [0, 1]$, the triangle inequality bounds the norm of the first two terms by $(1 - w_{t,i})M_{t,i} + w_{t,i}\eta_{t,i}$. The third term is at most $\|\mathcal{T}_t x_{t-1} - \mathcal{T}_t y_{t-1}\|_\infty \le d_{t-1}$ by non-expansiveness. Taking the maximum over tokens gives $d_t \le d_{t-1} + \max_i[(1 - w_{t,i})M_{t,i} + w_{t,i}\eta_{t,i}]$, and induction from $d_0 = 0$ gives Eq. 2. A token reindexing $(\mathcal{T}x)_i = x_{\pi(i)}$ is non-expansive because $\max_i \|x_{\pi(i)} - y_{\pi(i)}\|_2 \le \max_j \|x_j - y_j\|_2$. For a fixed step the per-token term is linear in $w_{t,i}$, so it is minimized at $w_{t,i} = 1$ when $\eta_{t,i} < M_{t,i}$ and at $w_{t,i} = 0$ otherwise.

# C SIGNAL ISOLATION

The dynamics probability that feeds the gate comes from a frozen, version-pinned residual detector. Evaluation masks and metrics are computed by separate code from decoded pixels or ground-truth latents, and detector outputs are not reused as those labels. The step budget is 4000 for every variant and for the matched twin.

# D REVISIT PROTOCOL DETAILS

## D.1 CONSTRUCTED REVISITS

RealEstate10K camera paths at this horizon are monotone sweeps, so each window is a palindrome: 31 outbound frames at stride 2 followed by the same frames in reverse order, 61 RGB frames in total, which encode to 16 latent frames under the VAE's $4\times$ temporal compression. Every frame-pose pair is a capture from the same static scene, and the return frame is that capture. Reported revisit scores use the captured RGB. Each window stores its departure peak index $a$ and return-onset index $r$; the outbound phase covers latent frames 1 to $a$, the away phase $a$ to $r$, and the return phase $r$ onward, with latent indices mapped to RGB indices under the $4\times$ temporal factor for pixel metrics.

## D.2 DIRECTIONAL LAST-VALID ANCHOR

The continuous reverse path measures short-horizon persistence. The released builder also provides a `jump` protocol: one real start view, a distant real view held across a gap, then a cut back to the start view. The sampler pairs this protocol with a directional last-valid-write anchor. Finalized frames store their tokens on an azimuth/elevation grid keyed by camera ray, and a revisited direction reads the latest finalized observation along that ray. The palindrome tables in the paper use the previous-frame anchor, which is the one exercised by the continuous reverse path.

## D.3 POOLS AND EXCLUSION DISCIPLINE

The primary pool scans only the 80 scenes never used for adaptation or validation and keeps 56 windows with angular departure of at least 15 degrees. The stronger pool used in the extended sweep excludes the 420 adaptation/validation scenes but not the remaining 80 mini-split identifiers, then keeps the 256 largest departures (all at least 40 degrees). Four identifiers overlap the primary pool among the first 128 windows (eight among all 256); merged scene statistics average repeated identifiers once. The stronger-pool builder for Table 3 excludes all 500 mini-split scenes, making it fully disjoint from the primary pool, adaptation, and validation. The 5000-scene adaptation pool excludes the union of all 500 mini scenes and all 256 stronger-pool scenes (748 unique identifiers), so no evaluated scene enters scale-up adaptation.

## D.4 POSE CONDITIONING

Relative world-to-camera poses (row-flattened rotation and translation, 12 dims) per latent-frame anchor, frame 0 zeroed to match the backbone's null-conditioning convention, mapped by a zero-initialized linear bridge into the frozen conditioning space. Zero initialization makes the bridge a null condition at the start of training, preserving zero-initialization equivalence with the base model.

## D.5 ROTATION-ONLY CORRESPONDENCE

Patch-center rays are unprojected with dataset-median normalized intrinsics $(f_x, f_y, c_x, c_y)$ of the aspect-preserving crop,

$$d_i = \left(\tfrac{u_i - c_x}{f_x},\ \tfrac{v_i - c_y}{f_y},\ 1\right), \qquad d'_i = R_{t\to t-1}\, d_i = R_{t-1} R_t^{\top}\, d_i, \tag{10}$$

reprojected to $u' = f_x\, d'_x/d'_z + c_x$ and $v' = f_y\, d'_y/d'_z + c_y$, and assigned to the nearest patch center. Tokens with $d'_z \le 10^{-4}$ or a reprojection outside the patch grid are invalid for the residual detector and hold. This warp does not define the reported attention anchors, which remain patch-grid coordinates. Translation is ignored; palindrome windows are selected by angular departure, so

rotation dominates. On a $320\times240$ pinhole grid with 16-pixel patches the rotation-only assignment matches a full pose warp on every patch when the baseline is zero, and still matches on every patch at a 4 m depth with a 5 cm lateral baseline (Appendix J).

### D.6 View-change tracking ratio

For each window let $x_1$ denote the context frame. The ratio is

$$\rho = \frac{\text{mean}_{t\in\text{away}} \|\hat{x}_t - \hat{x}_1\|^2}{\text{mean}_{t\in\text{away}} \|x_t^{\text{gt}} - x_1^{\text{gt}}\|^2} \tag{11}$$

in decoded pixel space over the away-phase frames. A rollout that executes the commanded trajectory has $\rho$ near 1, and we report $\rho$ next to pixel error.

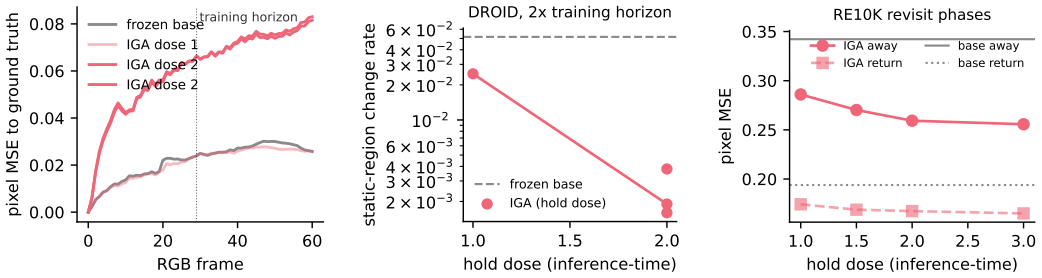
### D.7 Training-free heuristic baselines

All three are derived from the frozen base rollout with identical decoding and metrics. *Freeze* repeats the context latent for the whole window. *EMA* smooths, $y_t = \beta y_{t-1} + (1-\beta) b_t$ with $\beta \in \{0.5, 0.7\}$, where $b_t$ is the base rollout. *Warp-and-blend* applies the same correspondence warp as the hold, $y_t = \beta\, \mathcal{W} y_{t-1} + (1-\beta)\, b_t$. Their tracking ratios fall between 0.01 and 0.17. The retrofitted model stays in its LoRA twin's range while lowering pixel error.

### D.8 The latent-residual measurement

Palindromes provide pairs of latent frames with identical pose, identical content, and identical encoder state (both group-encoded away from the window boundary). Their latent residual is about 1.0 in the backbone's latent metric, comparable to entirely different content, because the temporally compressing VAE encodes the motion direction of each latent frame's RGB group, and outbound versus return motion is mirrored. Revisit scores in the paper are therefore computed on decoded pixels.

## E Extended sweep: dose, scale-up, and per-cohort tables




(a) Persistence over 61 frames. (b) DROID dose-response. (c) RealEstate10K dose-response.

Figure 3: **Persistence and dose control, extended sweep.** (a) At twice the training horizon the frozen base accumulates error fastest; EGSA tracks ground truth more closely across the rollout. (b) Static-region change rate versus inference dose on DROID at the 16-frame horizon (log scale; points are independent noise draws). (c) On the held-out revisit protocol both phases improve monotonically with dose up to $3\times$ trained strength. All panels come from the extended sweep configuration defined in Section 4.

This appendix collects the dose sweep, the fourteen-fold adaptation scale-up, and per-cohort absolute values. Figure 3 summarizes the persistence and dose sweeps: rollout error over 61 RGB frames (Figure 3a), and dose responses on DROID (Figure 3b) and RealEstate10K (Figure 3c). The configuration trains the detector on clean context and scores VAE reconstructions. The captured-RGB results using model-input evidence are Tables 3 and 4. Under the extended configuration the

same twin design improves outbound, away, and return error by 9.0, 10.8, and 13.3 percent over 180 scenes, each $p < 10^{-4}$, and the gain remains after the scale-up. Tables 6, 7, and 8 give phase-split pixel MSE, return-phase PSNR, and the view-change tracking ratio for every checkpoint on both pools, all at rollout noise seed 0 under the identical deterministic sampler.

Table 6: Primary pool (56 held-out windows), per cohort. MSE lower is better, PSNR higher is better; tracking near 1 is better.

| Cohort | Model | Outbound ↓ | Away ↓ | Return ↓ | Return PSNR ↑ | Tracking |
|---|---|---|---|---|---|---|
| – | frozen base | .2550 | .3422 | .1939 | 13.63 | .007 |
| s0 | lora_only | .2526 | .3284 | .2023 | 13.40 | .317 |
| | full_iga | **.2284** | **.2857** | **.1732** | **14.05** | .391 |
| s1 | lora_only | .2462 | .3077 | .2270 | 12.98 | .511 |
| | full_iga | **.2232** | **.2769** | **.2154** | **13.20** | .500 |
| s2 | lora_only | .2488 | .3210 | .2135 | 13.16 | .457 |
| | full_iga | **.2252** | **.2830** | **.1828** | **13.88** | .461 |

Table 7: Stronger-look-away pool (first 128 of the 256 windows, all departures at least 40 degrees), per cohort.

| Cohort | Model | Outbound ↓ | Away ↓ | Return ↓ | Return PSNR ↑ | Tracking |
|---|---|---|---|---|---|---|
| – | frozen base | .3075 | .3686 | .1726 | 14.16 | .007 |
| s0 | lora_only | .3041 | .3493 | .1883 | 13.73 | .364 |
| | full_iga | **.2767** | **.3098** | **.1569** | **14.52** | .399 |
| s1 | lora_only | .2969 | .3243 | .2201 | 13.06 | .523 |
| | full_iga | **.2719** | **.2953** | **.1996** | **13.50** | .490 |
| s2 | lora_only | .2989 | .3376 | .2050 | 13.31 | .487 |
| | full_iga | **.2720** | **.3009** | **.1709** | **14.16** | .460 |

Table 8: Scale-up pair (5000 adaptation scenes, 16000 steps) on both pools. The stronger pool here is the full 256 windows, so the base row differs slightly from Table 7.

| Pool | Model | Outbound ↓ | Away ↓ | Return ↓ | Return PSNR ↑ | Tracking |
|---|---|---|---|---|---|---|
| primary (56) | frozen base | .2550 | .3422 | .1939 | 13.63 | .007 |
| | lora_only | .2426 | .3057 | .2172 | 13.17 | .485 |
| | full_iga | **.2269** | **.2817** | **.1987** | **13.59** | .484 |
| stronger (256) | frozen base | .3040 | .3661 | .1805 | 13.92 | .009 |
| | lora_only | .2960 | .3286 | .2270 | 12.97 | .563 |
| | full_iga | **.2780** | **.3012** | **.2042** | **13.44** | .549 |

**Per-cohort paired improvements.** Relative to the matched LoRA twin (cohorts s0/s1/s2): on the primary pool, outbound 9.6/9.3/9.5 percent, away 13.0/10.0/11.8 percent, return 14.4/5.1/14.4 percent; on the stronger pool, outbound 9.0/8.4/9.0 percent, away 11.3/8.9/10.9 percent, return 16.7/9.3/16.6 percent. Every cell improves. For the scale-up pair, 88 to 95 percent of primary-pool scenes and 79 to 98 percent of stronger-pool scenes improve per phase. Comparisons use the matched LoRA twin, which executes the commanded view change (tracking 0.32 to 0.56). EGSA stays in that tracking range while lowering return error.

**Noise robustness.** DROID dose 2 at the 16-frame horizon stays between $\mathrm{SCR} = .0010$ and $.0038$ across the rollout noise draws measured, 93 to 98 percent below the base value of $.0516$. The RealEstate10K scene-level effects in Table 3 use rollout noise seed 0. On DFoT the same look-away and return gains repeat at a second rollout noise seed (Table 4).

**Cross-device determinism.** The rolling sampler is bitwise deterministic across the two GPUs used, verified by identical per-scene metric arrays for the same checkpoint and seed.

# F External baseline protocol

DFoT (Song et al., 2025) publishes checkpoints trained natively on the full RealEstate10K training split; BD is its published baseline without history guidance. We evaluate both on our palindrome windows as follows. The windows (scene identifiers, frame indices, phase marks) are exported from our encoded shards; the official DFoT codebase is driven with its own sampler and configuration, consuming each 61-frame window through its single-window keyframe scheme followed by its interpolation stage; wandb runs offline; generated videos are dumped by output hooks. Scoring uses the same aspect-preserving crop, the same $240\times320$ resize, the same phase split, and the same view-change ratio. Section 4.2 is the comparison used in the paper: the retrofit and the matched LoRA twin, both run on the DFoT checkpoint with the model-input evidence protocol.

# G Compute budget

The extended sweep, the DFoT leg, and the evaluations ran on A800 40GB GPUs; the MiniWorld retrofits in Table 3 ran on a V100 32GB with fp16 loss scaling and take 32 to 44 minutes per variant (Table 2). On the A800, a 4000-step MiniWorld retrofit takes 15 minutes (lora_only) to 24 minutes (full_iga) per variant; the 16000-step scale-up pair takes 57 and 94 minutes; encoding the 5000-scene adaptation pool takes about two hours; a 56-window evaluation of seven entries takes about 2.4 hours, a 128-window evaluation about 5.3 hours, and a 256-window evaluation about 90 minutes per entry. The canonical release set (six twin checkpoints, two attribution checkpoints, and both pool evaluations) reproduces from scratch in under one GPU-day; the complete program including the extended sweep, the dose and noise probes, the DROID leg, and the external baselines fits in about two GPU-weeks.

# H Extended related work

## H.1 Memory-augmented world models

The dominant answer to look-away forgetting is to give the generator access to old state. Video World Models with Long-term Spatial Memory maintain a geometry-grounded point map plus sparse episodic keyframes (Wu et al., 2025b); WorldMem keys a bank of past frames by state and pose (Xiao et al., 2025); Context as Memory stores raw context frames and retrieves them by field-of-view overlap between camera poses (Yu et al., 2025); WorldKV retains evicted key/value chunks and reinstates scene-relevant ones (Yi et al., 2026); WorldTrace keeps a compressed cache addressable by assigning summary slots in-distribution virtual positions, and introduces LoopBench for episodic recall (Wu et al., 2026); Mirage lifts latent tokens into a persistent 3D cache and queries it by latent-space warping (Wang et al., 2026); PERSIST goes further and simulates an evolving latent 3D scene with an explicit camera and renderer (Garcin et al., 2026). All of these mechanisms govern what the model can read. EGSA governs whether a local write is accepted: when the denoiser proposes changing a token whose evidence indicates no influence reached it, the proposal is pulled back toward the aligned incumbent estimate. The two directions compose naturally: a retrieval or 3D memory could supply a better incumbent prior than our one-frame warp, and our authorization coefficient could protect retrieved content from being overwritten.

## H.2 Geometry-grounded and camera-controlled generation

GEN3C renders an explicit 3D point cache along the commanded trajectory and conditions generation on those renderings (Ren et al., 2025); Geometry Forcing aligns intermediate diffusion features with a geometric foundation model (Wu et al., 2025a); history-guided diffusion strengthens conditioning on past frames (Song et al., 2025). Additive attention biases have long encoded static structure such as relative position (Raffel et al., 2020; Press et al., 2022) and window geometry (Liu et al., 2021); FLATTEN restricts attention to optical-flow trajectories (Cong et al., 2024); Closing the Loop biases attention toward depth-reprojected correspondences in retrieved chunks (Ma et al., 2026). Relative to this line, EGSA needs only a lightweight rotation-only nearest-patch field, derives its support from learned terms that include detected change rather than geometry alone, and reuses the same signal to modulate the state update itself.

### H.3 GATING AS A COMPUTATIONAL PRIMITIVE

The forget/update gate is one of the few ideas in sequence modeling that has been rediscovered, profitably, in every architectural generation: LSTMs and GRUs (Hochreiter & Schmidhuber, 1997; Cho et al., 2014), highway networks (Srivastava et al., 2015), gated convolutions (Dauphin et al., 2017), external differentiable memory (Graves et al., 2016), and, most recently, head-specific sigmoid gates after scaled dot-product attention in large language models (Qiu et al., 2025). Selective state spaces are a continuous relative (Gu & Dao, 2023). In this paper the source is cross-frame physical evidence, and the site is the clean-state write of a frozen video denoiser. The same evidence feeds a routing bias and a write weight, so the ablation in Section 4.3 can turn the bias off and leave the hold on.

### H.4 WHERE THE AUTHORIZATION ACTS

Correspondence can also be used to reshape the attention graph, as in FLATTEN and Closing the Loop. Softmax attention still reconstructs a token from the remaining keys, the residual stream, or the velocity head. The clean-state mix in $\hat{x}_0$ is the step that commits the write. Section 4.3 measures that split directly: with the logit bias removed and the hold left on, revisit error stays within 0.5 percent of the full model on both backbones.

### H.5 REVISIT PROTOCOLS AND BENCHMARKS

MIND builds symmetric forward/reverse trajectories and scores memory consistency at matched path positions (Ye et al., 2026); WorldCycle turns reversible action cycles into self-verifying rewards and releases CycleBench (Gu et al., 2026); LoopBench probes recall of a scene after a long detour (Wu et al., 2026); WorldModelBench scores instruction following and physics adherence at scale (Li et al., 2025). Our palindromic instantiation differs in one property we need for measurement: every target frame is a real capture with a real pose, so the return phase has a captured reference rather than a model-derived one. The accompanying view-change magnitude check and the hard-jump diagnostic address the two known degenerate strategies, stalling and adjacent-frame smoothing.

### H.6 WORLD MODELS AS DOWNSTREAM SIMULATORS

Diffusion and flow policies (Chi et al., 2023; Black et al., 2024) trained on robot demonstrations (Khazatsky et al., 2024) treat a simulated observation as a function of state and action. Classical latent world models wrote that contract into the dynamics (Ha & Schmidhuber, 2018; Hafner et al., 2020; Micheli et al., 2023; Alonso et al., 2024). EGSA adds a checked write to a frozen video generator, using a one-frame correspondence warp as the incumbent prior.

## I PRETRAINED HEADS ALREADY ALIGN WITH EVIDENCE

On the frozen MiniWorld backbone, attention heads with incoming alignment to causal evidence are present (paired permutation $p \approx 10^{-4}$). Ablating the top-aligned heads changes the static-region change rate by 10 percent. The retrofit acts on structure the pretrained model already carries. In the extended sweep, dose 2 holds that change rate between .0010 and .0038 across noise draws, 93 to 98 percent below the base value of .0516.

## J CONTROLLED CPU PROBES

The two measurements below isolate the residual hold and the rotation warp in controlled settings with exact ground truth.

### J.1 RESIDUAL HOLD ON SYNTHETIC PALINDROMES

The scene is a $12\times12$ grid with a fixed texture and one caused token. The camera steps out by three cells and back, so the last frame revisits the first. The predictor in the over-change regime emits the true frame plus isotropic noise ($\sigma = 0.85$); the hold prior is the previous prediction warped by the

known camera shift, not the current target. Evidence is an exact mask on the caused token, which is the action channel under perfect calibration. The hold is `ResidualHold` with $\sigma(\theta_s) = 0.5$ and dose $\alpha = 1$. Sixty-four scenes, paired sign-flip test, 20,000 permutations.

Static return MSE in Table 5 excludes the caused token, and tracking is away-phase change relative to ground truth. Gated anchoring reduces static return error by 57.4 percent against the unanchored predictor (64 of 64 scenes, $p < 10^{-4}$) and leaves the caused token exactly at the predictor's error, because the mask sets the hold weight to zero there. The evidence-free hold reaches a similar static error (.298 versus .309) and raises cause-token error from .755 to 1.580, a 109.3 percent increase ($p < 10^{-4}$). Tracking moves from 1.43 toward 1.18, so the hold is removing unlicensed change rather than motion. Static error falls with dose over the trained range: .725, .464, .309 at $\alpha \in \{0, 0.5, 1\}$.

### J.2 Rotation-only patch assignment

Patch centers on a $320\times240$ image (16-pixel patches, $f_x = f_y = 280$) are reprojected by a yaw of 15 or 40 degrees, once with a lateral baseline and once without. Agreement is the fraction of patches whose nearest source patch is the same under the rotation-only warp and the full pose warp. Table 9 summarizes this agreement for the evaluated camera rotations, baselines, and depths.

Table 9: Rotation-only versus full pose, nearest-patch agreement on the departures used in the paper, at zero baseline and at a 5 cm lateral baseline.

| Yaw | Baseline (m) | Depth (m) | Agreement |
|---|---|---|---|
| $15^\circ$ | 0 | 2 | 1.00 |
| $15^\circ$ | 0 | 4 | 1.00 |
| $40^\circ$ | 0 | 2 | 1.00 |
| $40^\circ$ | 0 | 4 | 1.00 |
| $15^\circ$ | 0.05 | 4 | 1.00 |
| $40^\circ$ | 0.05 | 2 | 1.00 |
| $40^\circ$ | 0.05 | 4 | 1.00 |
| $15^\circ$ | 0.05 | 2 | 0.79 |

At zero baseline the two warps assign every patch identically, at both departures used in the paper. At a 4 m depth, a 5 cm baseline keeps agreement at 1.00, with median reprojection error 3.6 px at $15^\circ$ and 4.6 px at $40^\circ$, under half a patch. At 2 m and the same 5 cm baseline, the $15^\circ$ yaw still agrees on 79 percent of patches (median 7.3 px).

## K Hyperparameters and seeds

Unless noted, MiniWorld retrofits use: LoRA rank 8, LoRA $\alpha = 16$, AdamW, LoRA learning rate $5\times10^{-5}$, gate and hold learning rate $1\times10^{-4}$, weight decay $0.01$, warmup 600 steps (linear on $\alpha$), 4000 steps, gradient clip 1.0, bf16 on A800 or fp16 with loss scaling on V100, batch size 1, fused-QK LoRA on all 28 blocks, V LoRA disabled, candidate gate layers $\{2, 5, 8, 11, 14, 17, 20, 23\}$, $m_h$ initialized at $\sigma(-6)$, $\theta_s$ initialized at 0, radius priors as in the released config. DFoT uses the same LoRA rank and the same gate/hold initialization, attached to the $16\times16$ mid-level grid. Training seeds $\{0, 1, 2\}$ control weight initialization, data shuffling, and dropout (none, in the reported runs). Rollout noise seeds are independent. Each evaluation report records the training cohort and rollout noise seed.