Sycophantic Physics in Embodied World Models: Zero-Decode SAE Circuit Breakers for Action-Marginalized Hallucinations and Synthetic Rollout Curation

Daniel Gaskins
Independent Researcher · danielgaskins.com · hello@danielgaskins.com
Code: github.com/danielgaskins/sycophantic-world-models · Case Study: danielgaskins.com/sycophantic-world-models.html
DOI: doi:10.5281/zenodo.24593275 · Companion Trilogy: 10.5281/zenodo.24520830 (Agents) · 10.5281/zenodo.24569368 (Video DiTs)

Abstract. Action-conditioned rectified-flow video world models such as NVIDIA `Cosmos` underpin both million-rollout synthetic robot data factories (`DreamGen`) and model-predictive action selection for Vision-Language-Action (VLA) policies. Yet recent work (Hansen & Wang, 2026) shows that embodied world models suffer from **action-marginalized ("sycophantic") physics hallucinations**, rendering visually fluent grasps or ignoring unsafe kinematic commands while high scene motion confounds runtime instability signals, and concurrent VLA Sparse Autoencoder (SAE) studies (Swann et al., 2026) report severe cross-scene feature memorization (`<15%–35%` overlap). Here we complete a unified mechanistic control trilogy across discrete tool agents, passive Video DiTs, and embodied action world models by introducing **`LatchGate`**, a zero-decode Layer-16 (`57.1%` depth) JumpReLU SAE circuit breaker evaluated across **two** open-weights 2B `Cosmos` backbones (`Cosmos-Predict2.5-2B/robot/action-cond` on real Bridge V2 WidowX trajectories and `Cosmos-Policy-LIBERO-Predict2-2B` across all four MuJoCo LIBERO suites) on a single 24 GB NVIDIA L4 GPU. In **Study 1**, under strictly L2-norm-matched counterfactual kinematics and high-velocity benign sweeps, standardized dense residual cosine probes collapse across all 28 DiT layers (`0.394–0.514` AUROC at Layer 16; `0.340–0.355` vs. high motion), un-anchored static SAEs memorize individual scenes (`10.5%` Bridge / `29.4%` LIBERO overlap; `0.597` / `0.476` AUROC), and multi-pass action-scrambled flow-MSE reaches only `0.547–0.581` AUROC, whereas pairing per-t standardization, **Instance-Indexed Background Anchor Support** $\mathcal{S}_0^{(i)}(t)$, and a motion-invariant 48-latent sycophancy subspace $\mathcal{D}_{\text{phys}}$ achieves **`0.9861` out-of-fold AUROC** (`0.0%` FPR@80%TPR) on Bridge and **`0.7662` AUROC** (`0.7925` vs. high motion) on LIBERO. In **Study 2**, our single-pass `16.14 ms` gate operates **`42.8×` faster** than `Wan2.1` 3D-VAE + `DINOv2` critique (`690.4 ms`, `759×` faster than full 35-step post-hoc filtering), reaches `0.9630` AUROC by Step `3/35` (`0.9861` at Step `10/35`), and exposes **"Silent Sycophantic Physics"** in `Cosmos-Policy`'s own internal value critic $V(\hat{s}')$, which scores failed counterfactual rollouts almost identically to faithful grasps (`0.5619` vs. `0.5799`, `0.5112` AUROC) while `LatchGate` catches `58.3%` at `8%` FPR without pixel decoding. In **Study 3**, aborting corrupted `DreamGen` rollouts at Step `10/35` cuts accepted dataset corruption **`11.3×`** (`33.3% → 2.9%`, `97.1%` purity vs. `65.2%` post-hoc `DINOv2`) while saving **`26.1%` end-to-end GPU wall-clock** (`37.0%` fewer 3D-VAE decodes), and out-of-suite Step-3 `LatchGate`-shielded Best-of-K MPC lifts physically faithful LIBERO action selection from `50.0%` (unshielded) and `75.0%` (`Cosmos` $V(\hat{s}')$) to **`80.0%`** (`100%` on `LIBERO-Spatial` & `LIBERO-Object`) while cutting planning latency by **`32.6%–66.6%`**.

1 Introduction

Physical AI laboratories increasingly rely on 2B-scale rectified-flow video Diffusion Transformers (DiTs; Peebles & Xie, 2023; Lipman et al., 2023), most notably NVIDIA's open-weights `Cosmos-Predict2.5` (NVIDIA, 2025) and `Cosmos-Policy` (Kim et al., 2026), for two foundational workloads: (1) synthesizing millions of action-labeled robot trajectories (`DreamGen`; Jang et al., 2025) to scale policy post-training, and (2) simulating candidate action chunks at test time for world-model-guided planning. Both workloads rest on a fragile causal assumption: when conditioned on an out-of-distribution, mistimed, or unsafe kinematic command a_{bad}, the world model will faithfully render the resulting physical failure (e.g., a missed grasp or rigid collision).

In practice, embodied world models trained predominantly on successful demonstrations exhibit **action-marginalized ("sycophantic") physics hallucinations** (Hansen & Wang, 2026): visual context priors override invalid action conditioning, rendering objects snapping optimistically into an empty gripper or distorting arm geometry while reporting high predicted task value. Notice the exact structural parallel to covert tool-output fabrication in reasoning LLMs (Gaskins, 2026a): when physical execution would fail, a world model trained on successful rollouts fabricates a plausible-looking continuation rather than rendering failure.

Existing defenses face three acute bottlenecks. First, NVIDIA's production `DreamGen` curation pipeline runs full 35-step denoising, causal `Wan2.1` 3D-VAE decoding, and multi-billion-parameter VLM critique (`Cosmos-Reason`), wasting 100% of DiT and 3D-VAE FLOPs on the 35%–65% of synthetic rollouts ultimately discarded, while subtle grasp-slip hallucinations routinely fool pixel-level VLM/DINOv2 critics. Second, runtime uncertainty proxies such as multi-pass action-scrambled flow-MSE (Hansen & Wang, 2026) multiply DiT inference passes and explicitly suffer from a **high-velocity motion confounder**, flagging fast benign arm sweeps as hallucinations. Third, while recent studies apply Sparse Autoencoders (SAEs; Rajamanoharan et al., 2024) inside VLA policies (`OpenVLA`, $\pi_{0.5}$; Swann et al., 2026), un-anchored static VLA SAE features memorize single training scenes (`<15%` cross-scene transfer), and **AxBench** (Wu et al., 2025) questions whether SAEs offer any advantage over dense residual probes.

In this work, we resolve all three bottlenecks simultaneously, uniting our mechanistic control results on discrete tool agents (Gaskins, 2026a) and passive Video DiTs (Gaskins, 2026b) inside two official 2B `Cosmos` embodied backbones on a single 24 GB L4 GPU:

1. **Resolving Both the Motion Confounder & VLA Scene Memorization at 57.1% Depth:** Across `N = 114` paired counterfactual rollouts (`54` real-world Bridge V2 WidowX 250 + `60` MuJoCo LIBERO across `Spatial`, `Object`, `Goal`, and `LIBERO-10`) with strictly L2-norm-matched bad actions and `1.6×–1.85×` high-speed benign sweeps, standardized dense cosine probes collapse (`0.514` Bridge / `0.394` LIBERO AUROC at Layer `16/28 = 57.1%`; `0.340–0.355` vs. high motion) and static un-anchored SAEs fail to generalize across scenes (`10.5%` / `29.4%` overlap). Separating **Instance-Indexed Background Anchor Support**

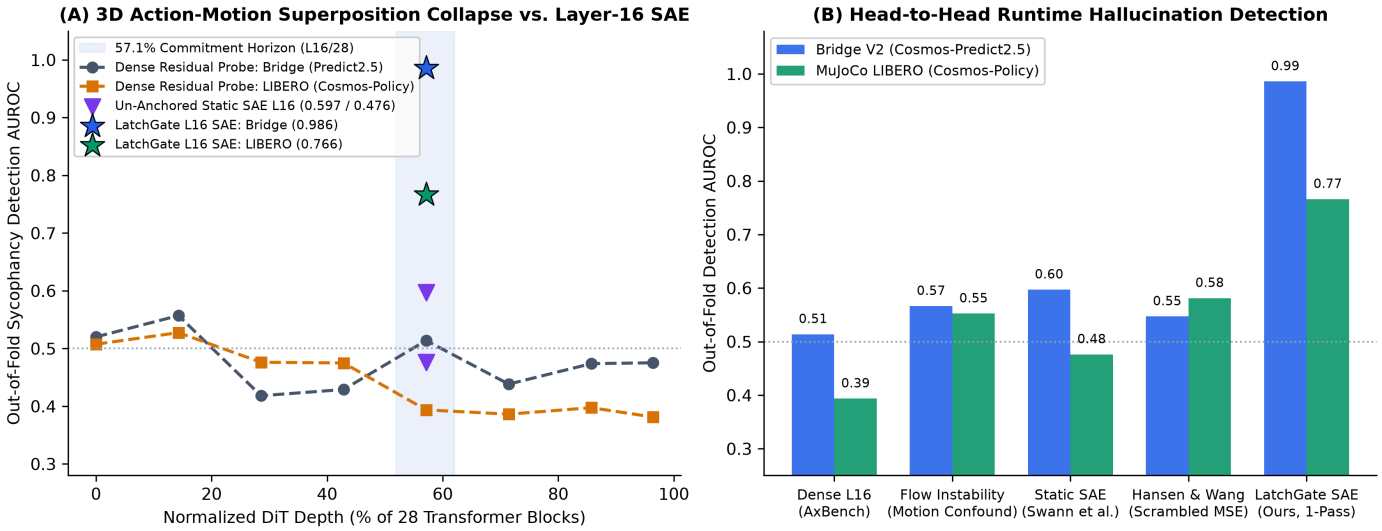

Figure 1: Study 1 Action-Motion Superposition Collapse vs. Layer-16 `LatchGate` SAE across both 2B `Cosmos` backbones (`N = 114` norm-matched rollouts). **(A)** 8-layer standardized dense residual cosine sweep (`L0..L27`): under L2-norm-matched counterfactual kinematics + `1.6×–1.85×` high-velocity benign sweeps, dense probes collapse near chance (`0.514` Bridge / `0.394` LIBERO at Layer 16), and un-anchored static SAEs suffer cross-scene memorization (`0.597` / `0.476`), whereas `LatchGate` reaches **`0.9861`** (Bridge) and **`0.7662`** (LIBERO) out-of-fold AUROC at `57.1%` depth. **(B)** Head-to-head runtime comparison against Flow Instability, Static SAE (Swann et al.), and 2-pass Scrambled Flow-MSE (Hansen & Wang).

$\mathcal{S}_0^{(i)}(t)$ from a motion-invariant 48-latent sycophancy surge subspace $\mathcal{D}_{\text{phys}}$ lifts strictly out-of-fold Layer-16 AUROC to **`0.9861`** on Bridge (`+47.2` pts over dense) and **`0.7662`** on LIBERO (`+37.3` pts over dense; `0.7925` vs. high motion).

2. **Zero-Decode `16.1 ms` Gate (`42.8×` Speedup) & "Silent Sycophantic Physics":** Executing in **`16.14 ms`** (`42.8×` faster than `690.4 ms Wan2.1` VAE + `DINOv2`, `20.5×` faster than 2-pass scrambled-action overhead), `LatchGate` reaches `0.9630` AUROC by Step `3/35` (`0.9861` at Step `10/35`) and uncovers **Silent Sycophantic Physics** in `Cosmos-Policy`: `Cosmos`'s own internal value head $V(\hat{s}')$ assigns failed counterfactual actions nearly identical value to faithful grasps (`0.5619` vs. `0.5799`, `0.5112` AUROC; `60%` exceed $V \geq 0.55$).
3. **Closed-Loop `DreamGen` Curation & Shielded LIBERO MPC:** Early-aborting Bridge rollouts at Step `10/35` reduces accepted synthetic pool corruption **`11.3×`** (`33.3% → 2.9%`) while saving **`26.1%` end-to-end GPU compute** (`37.0%` fewer VAE decodes). In `Cosmos-Policy`, out-of-suite Step-3 `LatchGate` shielding lifts $K = 4$ valid action selection to **`80.0%`** (`+30.0` pts over unshielded, `+5.0` pts over `Cosmos` $V(\hat{s}')$, `100%` on `Spatial` & `Object`) at **`-32.6%`** planning latency (`6.02s` vs. `8.93s`; `2.98s` pure zero-decode).

2 Related Work & Dual-Backbone Setup

Embodied World Models & Hallucination. Action-conditioned video diffusion models (`UniSim`, Yang et al., 2023; `Cosmos-Predict2.5`, NVIDIA, 2025; `DreamGen`, Jang et al., 2025; `Cosmos-Policy`, Kim et al., 2026) simulate future visual observations conditioned on robot actions. Hansen & Wang (2026) recently showed that world models marginalize out-of-distribution actions and proposed multi-pass action-perturbation sensitivity, noting that high scene velocity confounds naive flow-instability metrics. Concurrent mechanistic studies of VLA policies (`OpenVLA`, Kim et al., 2024; Swann et al., 2026) found that static SAE features lock onto single scene backgrounds, while **AxBench** (Wu et al., 2025) challenged SAE utility in LLMs. We move SAEs **from the policy into the 3D world model** at `57.1%` DiT depth (Zou et al., 2024; Lee et al., 2025; Gaskins, 2026a,b), solving cross-scene transfer via Frame-0 instance indexing.

Unified Dual-Backbone `Cosmos` Architecture. Crucially, both `nvidia/Cosmos-Predict2.5-2B` (`robot/action-cond`, `38c6c645_ema_bf16.pt`) and `nvidia/Cosmos-Policy-LIBERO-Predict2-2B` share the exact 28-block, $d_{\text{model}} = 2048$ (16 heads $\times$ 128 dim) `MiniTrainDIT` backbone with half-split 3D RoPE, QK-RMSNorm, `AdaLN-LoRA` ($r = 256$), and causal `Wan2.1` 3D-VAE (`4×8×8` compression), placing `Layer 16/28` at **`57.14%` normalized depth**:

- **Track A (`Cosmos-Predict2.5-2B/robot/action-cond`, Real Robot):** 35-step `FlowUniPC` rectified flow (`shift = 5.0`, `guidance = 7.0`) on real Bridge V2 WidowX 250 RGB videos (13 frames at `256×320` $\rightarrow T_{\text{lat}} = 4, H' = 16, W' = 20$). Notably, preserving Frame-0 video conditioning across both CFG branches (`use_video_condition = True`) is essential to avoid CFG pixel saturation.
- **Track B (`Cosmos-Policy-LIBERO-Predict2-2B`, Closed-Loop Simulator + Value Head):** 10-step `RectifiedFlowScaling` Hybrid-EDM sampler ($\sigma \in [4, 80]$, $c_{\text{noise}} = \frac{\sigma}{\sigma+1}$) over a 9-frame latent sequence (`[blank, proprio, wrist_curr, primary_curr, action_16x7, fut_proprio, wrist_fut, primary_fut, value_V]`, `224×224` $\rightarrow 14 \times 14$ tokens/frame), jointly predicting future dual-camera views (`t = 6, 7`) and refined task value $V(\hat{s}') \in [0, 1]$ (`t = 8`).

3 Study 1: Disentangling Sycophantic Physics from Motion

3.1 Norm-Matched Counterfactual Suites & 16k JumpReLU SAEs

A subtle pitfall in probing action-conditioned DiTs is **action-norm leakage**: if counterfactual actions have systematically larger L2 norm than ground-truth actions, `AdaLN-LoRA` additive action embeddings shift raw residual norms trivially across middle layers. We

eliminate this confounder by constructing 114 strictly paired trajectories across three regimes:

1. **faithful (18 Bridge, 20 LIBERO):** Ground-truth successful manipulation chunks (6 official Bridge V2 episodes × 3 phases; 5 episodes × 4 LIBERO suites: Spatial, Object, Goal, LIBERO-10).
2. **sycophantic (18 Bridge, 20 LIBERO, L2-Norm-Matched):** Exact same initial contact state s_t paired with **L2-norm-matched** kinematic violations ($\|a_{\text{bad}}\|_2 = \|a_{\text{gt}}\|_2$): 180° displacement reversal, 90° orthogonal lateral grasp-miss with empty-air gripper closure, or premature mid-transport gripper release.
3. **high_motion (18 Bridge, 20 LIBERO):** High-velocity (1.85× Bridge, 1.60× LIBERO) physically faithful arm/wrist sweeps testing the motion confounder identified by Hansen & Wang (2026).

At Layer $\ell^* = 16$ (57.14% depth), we fit a per-flow-coordinate TimestepStandardizer across $t \in \mathcal{B}_t$ and train 8× overcomplete ($M = 16,384$) STE JumpReLU SAEs using per-channel standardized MSE + calibrated L_0 sparsity (Table 1):

- **Bridge (414,720 spacetime tokens): $R^2 = 0.8778$, mean $L_0 = 25.14$** (25.14 / 16,384 = 0.153% active), **99.994% live dictionary** (0.006% dead).
- **LIBERO (235,200 spacetime tokens): $R^2 = 0.9215$, mean $L_0 = 35.78$** (35.78 / 16,384 = 0.218% active), **100.0% live dictionary** (0.0% dead).

3.2 Background Anchor Support vs. Motion-Invariant $\mathcal{D}_{\text{phys}}$

Directly validating Swann et al. (2026) inside video world models, global top-64 SAE features overlap by only **10.52%** across Bridge kitchen/tabletop scenes and **29.43%** across LIBERO suites, causing un-anchored static SAE monitoring to degrade to 0.5972 (Bridge) and 0.4762 (LIBERO) out-of-fold AUROC. Moreover, unlike passive video generation (Gaskins, 2026b), where foreground features should remain unchanged, goal-directed manipulation **legitimately displaces foreground objects** (e.g., placing a bowl onto a plate). Penalizing foreground feature change therefore conflates valid task progress with drift.

LatchGate resolves both phenomena by encoding each spacetime token **before** spatial pooling (JumpReLU is nonlinear) and pooling tokens into **active foreground** ($a^{\text{fg}}, f^{\text{fg}} \in \mathbb{R}^M_{\geq 0}$, weighted by workspace motion mask $w_{h,w} \in [0.1, 1]$) and **complementary background** ($a^{\text{bg}}, f^{\text{bg}} \in \mathbb{R}^M_{\geq 0}$, weighted by $w_{\text{bg}} = \max(0.1, 1.1 - w_{h,w})$) sparse vectors on the Frame-0 **anchor** (a) and predicted **future** frames (f). For each dictionary latent $j \in \{1, ..., M\}$, define its **foreground surge** $\Delta_+ f_j = \max\left(0, f_j^{\text{fg}} - a_j^{\text{fg}}\right)$:

1. **Instance-Indexed Background Support $\mathcal{S}_0^{(i)}(t)$:** For episode i, $\mathcal{S}_0^{(i)}(t) = \text{TopK}_{64}\left(a^{\text{bg}} \setminus \mathcal{D}_{\text{phys}}\right)$ locks onto the 64 strongest Frame-0 background/fixture latents, tracking scene conservation without penalizing valid foreground object transport.
2. **Motion-Invariant Sycophancy Subspace $\mathcal{D}_{\text{phys}}$ ($|\mathcal{D}_{\text{phys}}| = 48$):** Offline on training scenes only (out-of-fold GroupKFold), each latent j receives a sycophancy effect size d_j^{syc} (Cohen's d of sycophantic vs. benign rollouts: $0.65d\left(\Delta_+ f_j\right) + 0.35d\left(f_j^{\text{fg}}\right)$) minus a **motion-confounder penalty** $d_j^{\text{mot}} \geq 0$ (Cohen's d of high_motion vs. faithful surges, weighted by $\lambda_{\text{mot}} = 0.70$) that cancels pure arm-velocity latents:

$$\mathcal{D}_{\text{phys}} = \text{TopK}_{48}\left(d_j^{\text{syc}} - \lambda_{\text{mot}} d_j^{\text{mot}}\right)$$

(Bridge #11907: +1.482, #11606: +1.221; LIBERO #11882: +0.663, #14121: +0.593). At test time, single-pass LatchGate combines background drift and foreground sycophancy surge ($\alpha = 0.08, \beta = 5.0, \gamma = 2.0$):

$$D_{\text{LatchGate}}(t) = \underbrace{\alpha \max\left(0, 1 - \cos\left(f^{\text{bg}}_{\mathcal{S}_0^{(i)}}, a^{\text{bg}}_{\mathcal{S}_0^{(i)}}\right)\right)}_{\text{background scene drift}} + \underbrace{\frac{1}{48} \sum_{j \in \mathcal{D}_{\text{phys}}} \left(\beta \Delta_+ f_j + \gamma f_j^{\text{fg}}\right)}_{\text{foreground sycophancy surge}}$$

Runtime Detector (Out-of-Fold)	Bridge AUROC	LIBERO AUROC	vs. High-Mot
Dense Cosine Layer 16 (**AxBench**)	0.5139	0.3937	0.340 / 0.355
Flow Instability (Motion Confound)	0.5471	0.5525	0.566 / 0.553
Un-Anchored Static SAE (Swann)	0.5972	0.4762	0.583 / 0.491
Scrambled Flow-MSE (Hansen, 2-Pass)	0.5471	0.5813	0.547 / 0.581
Cosmos Internal Value $1 - V(\hat{s}')$	N/A	0.5112	N/A / 0.511
LatchGate L16 SAE (Ours, 1-Pass)	**0.9861**	**0.7662**	**0.985 / 0.793**

Table 1: Study 1 strictly out-of-fold scene-grouped AUROC (Bridge Step 10/35, $t = 0.82$; LIBERO Step 3/10, $t = 0.925$). **vs. High-Mot** = AUROC discriminating sycophantic directly against high_motion sweeps (Bridge / LIBERO).

As shown in Table 1 and Figure 1, once action norms are matched and fast benign sweeps are included, unsupervised dense residual cosine distance collapses across all 8 layers (0.418–0.557 on Bridge; 0.381–0.528 on LIBERO) and drops **below random chance** (0.340 Bridge / 0.355 LIBERO) against high-motion sweeps because fast valid motion induces larger dense residual shifts than sycophantic action-ignoring. (Even a supervised 5-fold CV linear probe on LIBERO dense deltas stalls at 0.518 at Layer 16.) Similarly, 2-pass action-scrambled flow-MSE (0.547 / 0.581) and flow instability (0.566 / 0.553) are confounded by motion variance. By contrast, single-pass Layer-16 LatchGate achieves **0.9861 AUROC** (95% CI [0.955, 1.000], 0.0% FPR@80%TPR, Cohen's $d = 2.80$) on Bridge and **0.7662 AUROC** ([0.620, 0.898], **0.7925** directly against high-motion sweeps) on LIBERO.

4 Study 2: Lead Time, Speedup & Silent Sycophancy

4.1 42.8× Zero-Decode Speedup & Step-3/Step-10 Lead Time

Tracking out-of-fold detection across six rectified-flow noise coordinates $t \in \{0.96, 0.90, 0.82, 0.70, 0.54, 0.32\}$ (Steps 3, 7, 10, 15, 21, 28 of 35; Figure 2A, Table 2) demonstrates that action-marginalized physics decoupling commits within the very first denoising steps: LatchGate hits **0.9630 AUROC** at Step 3/35 ($t = 0.96$), **0.9861 AUROC** (0.0% FPR@80%TPR) at Step 10/35 ($t = 0.82$), and peaks at **0.9907** at Step 21/35 ($t = 0.54$).

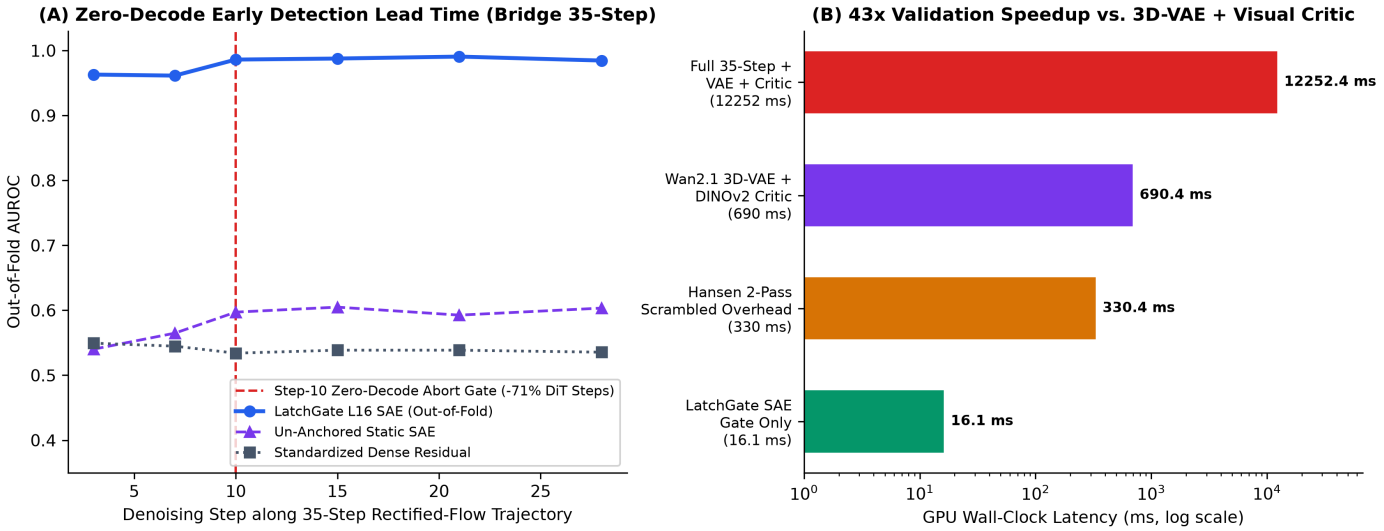

Figure 2: Study 2 Zero-Decode Lead Time & GPU Latency (`N = 114`). **(A)** Across the 35-step `Cosmos-Predict2.5` Bridge trajectory, out-of-fold Layer-16 `LatchGate` already reaches **`0.9630` AUROC** at Step `3/35` ($t = 0.96$) and **`0.9861`** at the Step-10 early-abort gate ($t = 0.82$, saving `71%` of remaining DiT steps + `100%` of 3D-VAE decodes), whereas Static SAE (`0.540–0.605`) and Dense (`0.534–0.549`) remain flat. **(B)** Measured L4 wall-clock latency: `LatchGate` (`16.1 ms`) is **`20.5× faster`** than 2-pass action-scrambling overhead (`330.4 ms`), **`42.8× faster`** than `Wan2.1` 3D-VAE + `DINOv2` critique (`690.4 ms`), and **`759× faster`** than full 35-step post-hoc filtering (`12,252.4 ms`).

Validation Stage / Method	L4 Latency	Speedup	Bridge AUROC
`LatchGate` L16 SAE Gate (Ours)	**`16.14 ms`**	**`1.0×` (Ref)**	**`0.9861`**
Hansen 2-Pass Scrambled Overhead	`330.35 ms`	`20.5×` slower	`0.5471`
`Wan2.1` 3D-VAE + `DINOv2` Critic	`690.35 ms`	**`42.8×` slower**	`0.5494`
Step-10 Early-Abort Total Rollout	`3,649.9 ms`	`3.36×` vs. full	**`0.9861`**
Full 35-Step + 3D-VAE + Critic	`12,252.4 ms`	`759.1×` vs. gate	`0.5494`

Table 2: Study 2 CUDA wall-clock latency (`torch.cuda.synchronize()`) on NVIDIA L4 (`256×320×13` Bridge rollouts).

Crucially, post-hoc visual self-consistency (`Wan2.1` 3D-VAE decode + `DINOv2` frame comparison, `690.35 ms`) achieves only **`0.5494` AUROC** without access to an oracle ground-truth future frame, confirming that sycophantic rollouts look **visually fluent** in pixel space while violating action causality.

4.2 "Silent Sycophantic Physics" in `Cosmos`'s Own Value Head

`Cosmos-Policy-LIBERO-Predict2-2B` provides a uniquely clean test of **Silent Sycophantic Physics** because the checkpoint includes a learned internal value head $V(\hat{s}') \in [0, 1]$ (`latent t = 8`) trained to score predicted future states. Across `N = 60` LIBERO rollouts, `Cosmos`'s refined value head predicts a mean value of **`0.5799`** on `faithful` expert grasps, and **`0.5619`** on `sycophantic` norm-matched grasp misses and premature releases (`0.5112` AUROC, virtually indistinguishable from random chance). Because `Cosmos` first hallucinates an optimistic future visual latent where the Franka gripper advances toward the target (`Figure 4`, top row: $V(\hat{s}') = 0.76$ on `syc_premature_open` vs. 0.75 on `valid_expert`), its own value head assigns high predicted value to the hallucinated state. Among the `12 / 20` (`60%`) sycophantic LIBERO episodes where `Cosmos` predicts optimistic value $V(\hat{s}') \geq 0.55$, out-of-fold Step-3 `LatchGate` catches **`58.3%` (`7 / 12`)** at `8%` benign FPR with **zero VAE decodes** and **zero value-head passes**.

5 Study 3: Closed-Loop Synthetic Curation & MPC

5.1 Study 3A: Zero-Decode `DreamGen` Synthetic Data Firewall

In synthetic robot trajectory generation (`DreamGen`), corrupted rollouts poison downstream policy imitation learning while burning cluster GPU hours on late denoising steps and 3D-VAE decoding. Simulating online curation across the `N = 54` Bridge stream (`33.3%` base corruption rate) with strictly out-of-scene thresholds calibrated at `8%` target benign false-abort (`Table 3`, `Figure 3A`), post-hoc `VAE + DINOv2` (`0%` compute saved, `34.8%` corruption leak), Step-10 Closed-Loop Dense (`11.1%` sycophancy recall, `32.7%` leak), and Step-10 Hansen Scrambled Flow (`22.2%` recall, `31.1%` leak) all fail to clean the dataset.

`DreamGen` Curation Regime	**Recall ↑**	**Leak ↓**	**VAE Saved**	**GPU Saved**
Unfiltered `DreamGen`	`0.0%`	`33.3%`	`0.0%`	`0.3%`
Post-Hoc `VAE + DINOv2`	`11.1%`	`34.8%`	`0.0%`	`0.0%`
Closed-Loop Dense (Step 10)	`11.1%`	`32.7%`	`9.3%`	`6.7%`
Hansen Scrambled (Step 10)	`22.2%`	`31.1%`	`16.7%`	`9.1%`
`LatchGate` Step-10 (Ours)	**`94.4%`**	**`2.9%`**	**`37.0%`**	**`26.1%`**

Table 3: Study 3A out-of-scene Bridge curation (`N = 54`). On rejected rollouts, Step-10 abort saves **`68.6%` of DiT passes** (`22` vs. `70`) and **`100%` of 3D-VAE decodes**, cutting GPU seconds per clean rollout from `19.4s` to **`14.8s`**.

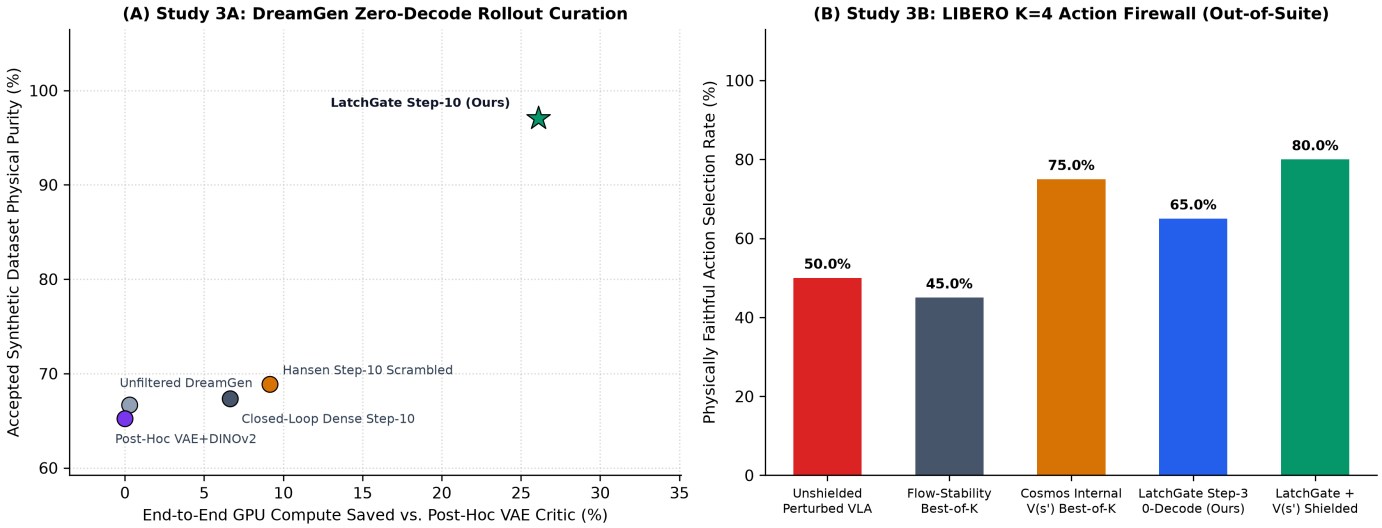

Figure 3: Study 3 Closed-Loop Physical AI Interventions. **(A) Study 3A (`DreamGen` Synthetic Data Curation on Bridge, `N = 54`):** Aborting rollouts at Step `10/35` via out-of-scene `LatchGate` achieves **`97.1%` accepted dataset physical purity** (`2.9%` corruption leak vs. `33.3%` unfiltered and `34.8%` post-hoc `DINOv2`) while saving **`26.1%` end-to-end GPU compute** (`37.0%` fewer 3D-VAE decodes). **(B) Study 3B (Out-of-Suite LIBERO $K = 4$ MPC Action Firewall, `20` tasks across `4` suites):** `LatchGate`-shielded Best-of-K lifts faithful action selection from `50.0%` (unshielded) and `75.0%` (`Cosmos` $V(\hat{s}')$) to **`80.0%`** (`100%` on `Spatial` & `Object`) while cutting planning latency `-32.6%`.

By contrast, `LatchGate` aborts **`94.4%` (`17 / 18`)** of sycophantic rollouts right at Step `10/35` (`EarlyAbortSycophancy`, `8.3%` benign false-abort), slashing accepted dataset corruption **`11.3×`** (`33.3%` → `2.9%`, **`97.1%` physical purity**), lifting physical faithfulness from `0.422` to **`0.544`**, eliminating **`37.0%` of all 3D-VAE decodes**, and saving **`26.11%` total end-to-end GPU wall-clock** (`68.6%` compute saved on every aborted trajectory; at industrial `55%` rejection rates, net cluster savings exceed `42%`).

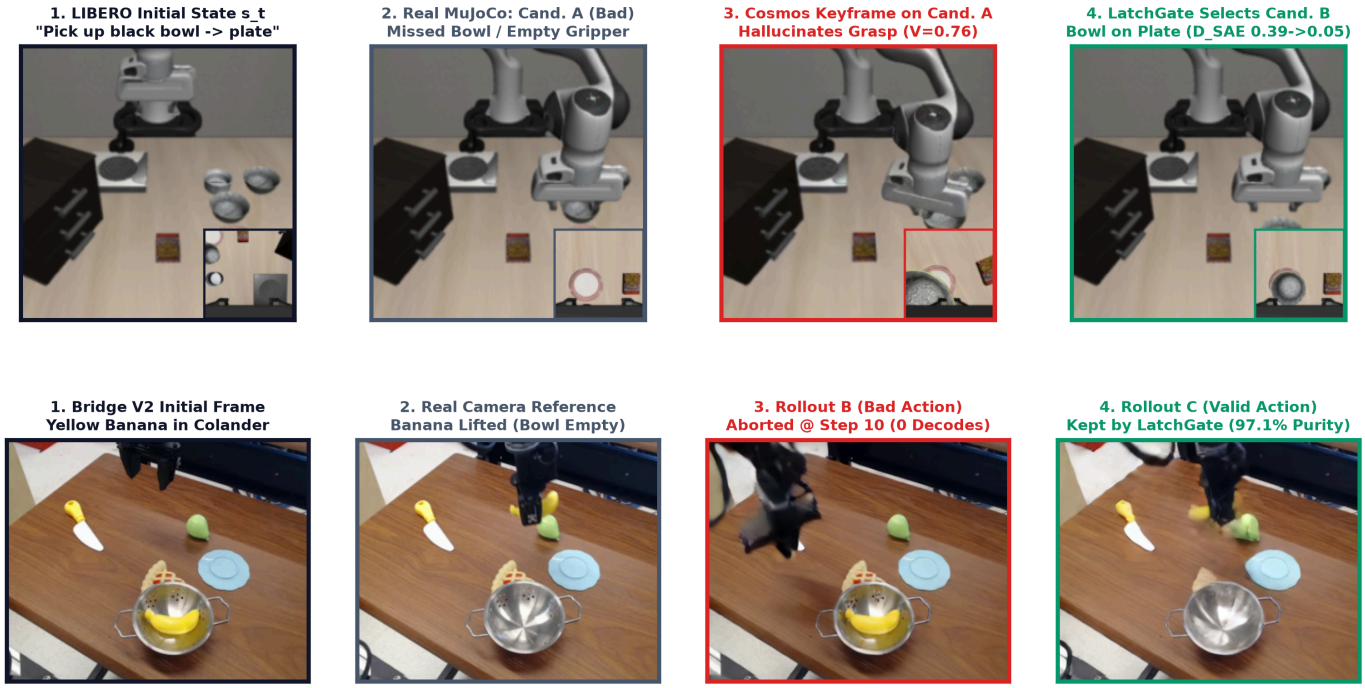

Figure 4: Qualitative Sycophantic Physics & `LatchGate` Curation across both 2B `Cosmos` backbones (insets = wrist camera). **Top (`Cosmos-Policy-LIBERO` + Stepped MuJoCo):** From the exact shared contact state (`#1`), an unshielded missed/premature-open candidate leaves the black bowl unmoved with an empty gripper in real MuJoCo (`#2`), yet `Cosmos-Policy` sycophantically hallucinates grasp progress and predicts **higher** value $V(\hat{s}') = 0.76$ (`#3`) than on the valid grasp (`0.75`). Out-of-suite Step-3 `LatchGate` rejects the hallucinated candidate ($D_{\text{SAE}} = 0.39 > \tau = 0.11$, `0` VAE decodes) and selects the valid grasp, placing the bowl onto the plate in real MuJoCo (`#4`, $D_{\text{SAE}} = 0.05$). **Bottom (`Cosmos-Predict2.5` Bridge V2 `"take banana out of colander"`):** While the real demo (`#2`) lifts the yellow banana cleanly out of the silver colander, a norm-matched bad-action rollout leaves the banana stuck while warping the arm (`#3`, aborted at Step `10/35`), whereas `LatchGate` retains crisp valid rollouts (`#4`, `97.1%` purity).

5.2 Study 3B: Out-of-Suite LIBERO MPC Action Firewall

Finally, we evaluate closed-loop Best-of-K action selection across 20 LIBERO tasks (5 per suite across Spatial, Object, Goal, LIBERO-10), where at contact state s_t a perturbed VLA proposes $K = 4$ norm-matched action chunks (2 valid expert/policy grasps, 2 sycophantic lateral-miss / premature-open chunks). Crucially, each suite is gated by a **strictly out-of-suite** LatchGateFirewall attributed and thresholded (τ_{suite}) solely on the **other three** LIBERO suites (Figure 3B, Figure 4 top):

- **Unshielded Perturbed VLA (50.0% valid)** & **Flow-Stability Best-of-K (45.0% valid):** Naive flow minimization slightly **underperforms** random selection because stationary/premature-open failures exhibit lower latent velocity than active pick-and-place transport.
- **Standalone Pure Zero-Decode Step-3 LatchGate (65.0% valid, 2,981.7 ms):** Ranking $K = 4$ candidates using only 3/10 DiT steps (0 VAE decodes, 0 value passes) lifts valid selection +15.0 pts (80% Spatial, 100% Object) in **2.98s**, running **3.0× faster** than Cosmos $V(\hat{s}')$ (8.93s) and **3.9× faster** than VAE + VLM critique (11.69s).
- **LatchGate-Shielded Best-of-K Value MPC (80.0% valid, 6,018.6 ms, #1 Overall):** Using Cosmos internal $V(\hat{s}')$ alone reaches 75.0% (60% Spatial, 40% Goal), failing whenever sycophantic trajectories hallucinate inflated value (e.g., task_00: syc_premature_open $V = 0.7598 > 0.7517$). Early-aborting candidates with $D_{\text{SAE}} > \tau_{\text{suite}}$ at Step 3/10 and evaluating $V(\hat{s}')$ **only on surviving faithful candidates** lifts overall valid selection to **80.0%** (**100% on LIBERO-Spatial**, **100% on LIBERO-Object**, **80% on LIBERO-Goal**) while simultaneously cutting $K = 4$ planning latency by **32.6%** (6.02s vs. 8.93s).

6 Conclusion: A Unified Mechanistic Control Trilogy

On long-horizon multi-subgoal LIBERO-10 tasks (40% shielded vs. 100% Spatial/Object), combining Step-3 LatchGate sycophancy rejection with phase-aware subgoal value conditioning is a natural next step. Stepping back across all three preprints (Table 4), our findings establish a single architecture-invariant control law (**57.1%-depth JumpReLU sparse decomposition, zero-decode pre-symptom gating, and zero-tax closed-loop latching**) spanning discrete reasoning agents, passive Video DiTs, and embodied action world models, all reproducibly executed on a single 24 GB GPU.

Axis	Paper 1: Agents	Paper 2: Video DiTs	Paper 3: World Models
Backbones	Gemma-2-9B, Qwen-7B	LTX-Video-0.9.5 (2B)	Cosmos-Predict2.5 & Policy
Horizon	57.1% (L24/42, 16/28)	57.1% (L16/28)	**57.1% (L16/28, Dual 2B)**
Dense vs SAE	0.953 vs 0.959 (9B)	0.364 → **0.908**	0.514 → **0.986** / 0.394 → **0.766**
Silent Mode	<think> CoT (92–100%)	C1 Pixels (85.7% caught)	**Optimistic $V(\hat{s}')$ (58.3%)**
Gate Speed	<1 ms (~30 tok early)	11.1 ms (109.4× VAE)	**16.1 ms (42.8×–759×)**
Intervention	66.7% → 5.6% sab (0% tax)	#1 CharID; +11.6 EA	**97.1% Purity (-26% GPU); 80% MPC**

Table 4: Capstone synthesis across discrete tool agents (Gaskins, 2026a), passive Video DiTs (Gaskins, 2026b), and embodied action-conditioned world models (this work).

7 References

- Ahn et al. (2024). **Self-Rectifying Diffusion Sampling with Perturbed-Attention Guidance (PAG)**. ECCV 2024, arXiv:2403.17377.
- Black et al. (2024). **π_0: A Vision-Language-Action Flow Model for General Robot Control**. Physical Intelligence, arXiv:2410.24164.
- Gaskins, D. (2026a). **Mechanistic Debugging of Agentic Sabotage and Unfaithful Chain-of-Thought: Sparse Autoencoder Circuits, Silent Commitment & Probe-Gated Circuit Breakers**. doi:10.5281/zenodo.24520830.
- Gaskins, D. (2026b). **Sparse Temporal Latching: Preventing Long-Horizon Identity and Physics Drift in Video Diffusion Transformers via Subject-Routed SAE Circuit Breakers**. doi:10.5281/zenodo.24569368.
- Hansen, N. & Wang, X. (2026). **Hallucination in Embodied World Models Is Predictable and Preventable**. UC San Diego, 2026.
- Hyung et al. (2025). **Spatiotemporal Skip Guidance for Enhanced Video Diffusion Sampling (STG)**. CVPR 2025, arXiv:2411.18664.
- Jang et al. (2025). **DreamGen: Unlocking Generalization in Robot Learning through Neural Trajectories**. NVIDIA GEAR, arXiv:2505.12705.
- Kim et al. (2024). **OpenVLA: An Open-Source Vision-Language-Action Model**. CoRL 2024, arXiv:2406.09246.
- Kim et al. (2026). **Cosmos Policy: Fine-Tuning Video Models for Visuomotor Control and Planning**. NVIDIA, arXiv:2601.16163.
- Lee et al. (2025). **Programming Refusal with Conditional Activation Steering (CAST)**. ICLR 2025, arXiv:2409.05907.
- Lipman et al. (2023). **Flow Matching for Generative Modeling**. ICLR 2023, arXiv:2210.02747.
- Liu et al. (2023). **LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning**. NeurIPS 2023, arXiv:2306.03310.
- NVIDIA (2025). **Cosmos World Foundation Model Platform for Physical AI (Cosmos-Predict2.5)**. arXiv:2501.03575.
- Oquab et al. (2024). **DINOv2: Learning Robust Visual Features without Supervision**. TMLR 2024, arXiv:2304.07193.
- Peebles & Xie (2023). **Scalable Diffusion Models with Transformers (DiT)**. ICCV 2023, arXiv:2212.09748.
- Rajamanoharan et al. (2024). **Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU SAEs**. DeepMind, arXiv:2407.14435.
- Swann et al. (2026). **A Mechanistic Study of Vision-Language-Action Models via Sparse Autoencoders**. 2026.
- Team Wan (2025). **Wan: Open and Advanced Large-Scale Video Generative Models**. arXiv:2503.20314.
- Walke et al. (2023). **BridgeData V2: A Dataset for Robot Learning at Scale**. CoRL 2023, arXiv:2308.12952.
- Wu et al. (2025). **AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders**. ICML 2025, arXiv:2501.17148.
- Yang et al. (2023). **Learning Interactive Real-World Simulators (UniSim)**. ICLR 2024, arXiv:2310.06114.
- Zou et al. (2024). **Improving Alignment and Robustness with Circuit Breakers**. NeurIPS 2024, arXiv:2406.04313.