Sparse Temporal Latching: Preventing Long-Horizon Identity and Physics Drift in Video Diffusion Transformers via Subject-Routed SAE Circuit Breakers

Daniel Gaskins
Independent Researcher · danielgaskins.com · hello@danielgaskins.com
Code: github.com/danielgaskins/sparse-temporal-latching · Case Study: danielgaskins.com/sparse-temporal-latching.html
DOI: doi:10.5281/zenodo.24569368 · Companion LLM Paper: doi:10.5281/zenodo.24520830

Abstract. Autoregressive sliding-window conditioning (`Extend`) enables open-weights Video Diffusion Transformers (Video DiTs) such as `LTX-Video` to synthesize clips beyond their native context window, yet induces compounding **identity drift**, **rigid-geometry morphing**, and **high-frequency temporal warble**. Current training-free remedies such as Spatio-Temporal Skip Guidance (STG) globally damp self-attention variance across spacetime patches—trading visual stability against a severe **Motion-Freezing Cliff**—while recent LLM steering benchmarks (**AxBench**; Wu et al., 2025) question whether Sparse Autoencoders (SAEs) offer any causal control advantage over dense residual vectors. Here we extend mid-network JumpReLU SAE circuit breakers from autoregressive tool agents (Gaskins, 2026) to 3D spacetime Video DiTs (`Lightricks/LTX-Video-0.9.5`, 28 layers, $d = 2048$). In **Study 1**, we remove AdaLN flow-schedule shift via a per-σ `TimestepStandardizer` and train a 16,384-latent JumpReLU SAE on 1.31M 3D spacetime tokens at Layer 16 (`57.1%` depth; $R^2 = 0.7084$, mean $L_0 = 111.1$, 99.94% live latents). Because 3D video tokens densely superimpose camera pan, foreground velocity, and geometry, dense residual drift probes collapse across mid-network blocks (`0.292–0.444` AUROC; `0.364` at Layer 16, peaking at `0.633` at Layer 27), whereas an **Instance-Indexed Top-64 Sparse Identity Support** paired with a universal 48-latent warble subspace reaches **`0.9083 ± 0.0764` AUROC** at Layer 16 (`+54.4` AUROC points over Layer-16 dense). In **Study 2 ("Silent Drift")**, our in-stream Layer-16 SAE gate (`11.06 ms`) reaches `0.881` AUROC by $\sigma = 0.75$ (22.5/30 denoising steps early; `0.913` at $\sigma = 0.48$) at a **`109.4×` speedup** over Tweedie `VAE + DINOv2` clean-preview decoding (`1,209.8 ms`), predicts next-chunk collapse from Chunk-1 latents at **`0.897` AUROC** vs. `0.857` for a black-box VBench `DINOv2` monitor, and catches **`85.7%` (6/7)** of monitor-evading collapses passing `DINOv2 ≥ 0.90`. In **Study 3**, whereas unconditional dense steering suffers a Video Polysemantic Cliff (`DINOv2 0.782 → 0.675`), Closed-Loop Dense is blind at `0%` FPR (`0.0%` catch rate), and Always-On STG slashes RAFT optical flow by `28.4%–54.4%`, our training-free **Subject-Routed SAE Circuit Breaker** (`0.0%` held-out benign FPR) preserves **`110.4%–111.9%`** of unsteered dynamic motion (`0.718–0.727` vs. `0.650` px/frame), lifts **Character Identity** to `0.8007` (`#1` across all regimes, `+3.97` pts over unsteered) and **Camera Occlusion Re-Entry** to `0.7810` (`+1.64` pts), and exposes a fundamental spatial-routing boundary-shear tradeoff on translating rigid geometry (`0.7116` sharp vs. `0.7708` smooth ramp vs. `0.6283` high STG). Finally, pairing our zero-decode SAE gate with Step-12 ($\sigma = 0.58$) early-abort seed latching ($K = 6$, zero 3D-VAE decodes on rejected draws) rescues severe trajectory collapses by up to **`+11.6` `DINOv2` points** (`0.626 → 0.742`) while avoiding High STG's `-98%` motion freeze (`0.04` vs. `1.74` px/frame).

1 Introduction

Scaling Diffusion Transformers (DiTs; Peebles & Xie, 2023) with rectified flow matching (Lipman et al., 2023) and causal 3D Video-VAEs has enabled real-time open-weights video generation models such as `LTX-Video` (HaCohen et al., 2024), `CogVideoX` (Yang et al., 2024), and `Wan 2.1` (Team Wan, 2025). However, generating coherent rollouts beyond a single 2–5 second latent window via sliding-window frame conditioning (`Extend`) remains brittle (Kim et al., 2024; Chen et al., 2024; Yin et al., 2024). Across multi-chunk rollouts, slight representation errors in the conditioning tail compound across chunk boundaries—manifesting as gradual facial or garment mutation (**identity drift**), non-rigid warping of vehicles and instruments (**geometry collapse**), high-frequency texture rippling (**temporal warble**), and identity loss following camera occlusions.

Training-free guidance methods such as Spatio-Temporal Skip Guidance (STG; Hyung et al., 2025; cf. PAG, Ahn et al., 2024) mitigate structural degradation by perturbing self-attention across all spacetime patches and denoising steps. Yet just as unconditional activation steering in LLMs triggers a polysemantic collateral-damage cliff on benign inputs (Arditi et al., 2024; Gaskins, 2026), cranking up always-on STG in Video DiTs induces a severe **Motion-Freezing Cliff**—suppressing camera motion, fluid dynamics, and articulated velocity. Simultaneously, in mechanistic interpretability, Sparse Autoencoders (SAEs; Rajamanoharan et al., 2024; Lieberum et al., 2024) have recently been applied to static 2D image diffusion models for concept discovery and unlearning (Surkov et al., 2024; Cywiński & Deja, 2025; Olson et al., 2025; Cassano et al., 2026; Xu et al., 2026), while **AxBench** (Wu et al., 2025) demonstrated that simple dense vectors outperform SAEs for unconditional concept steering in LLMs. Whether dense vectors remain viable—or collapse under spatiotemporal superposition—for **closed-loop temporal control of 3D Video DiTs** remains an open question.

In this work, we show that the exact mathematical control architecture recently established for preventing covert sabotage in autoregressive tool agents (Gaskins, 2026)—**57.1%-depth JumpReLU sparse decomposition, pre-symptom silent-commitment detection, and zero-tax closed-loop circuit breaking**—transfers directly to 3D Video DiTs when adapted to continuous flow-matching geometry:

1. **Why Dense Vectors Fail in 3D Video DiTs (Nuancing AxBench):** Unlike LLM prefix tokens that align discrete intent along a shared linear direction, 3D spacetime tokens (t, h, w) densely superimpose camera translation, subject pose, lighting, and identity. Across a 7-layer standardized depth sweep on `LTX-Video-0.9.5`, dense probes fail at mid-network planning depth (`0.364` AUROC at Layer `16/28 = 57.1%`), whereas a per-σ-standardized 16k JumpReLU SAE with **instance-indexed Chunk-0 sparse support** jumps to **`0.9083` AUROC** (`+54.4` pts).
2. **Non-Tautological "Silent Drift" & `109.4×` Gate Speedup:** Benchmarked head-to-head against exact single-step Tweedie clean previews $\mathrm{VAE}(\hat{z}_0(\sigma)) + \mathrm{DINOv2}$ (`1,209.8 ms`), our zero-decode Layer-16 SAE gate (`11.06 ms`, `109.4×` faster)

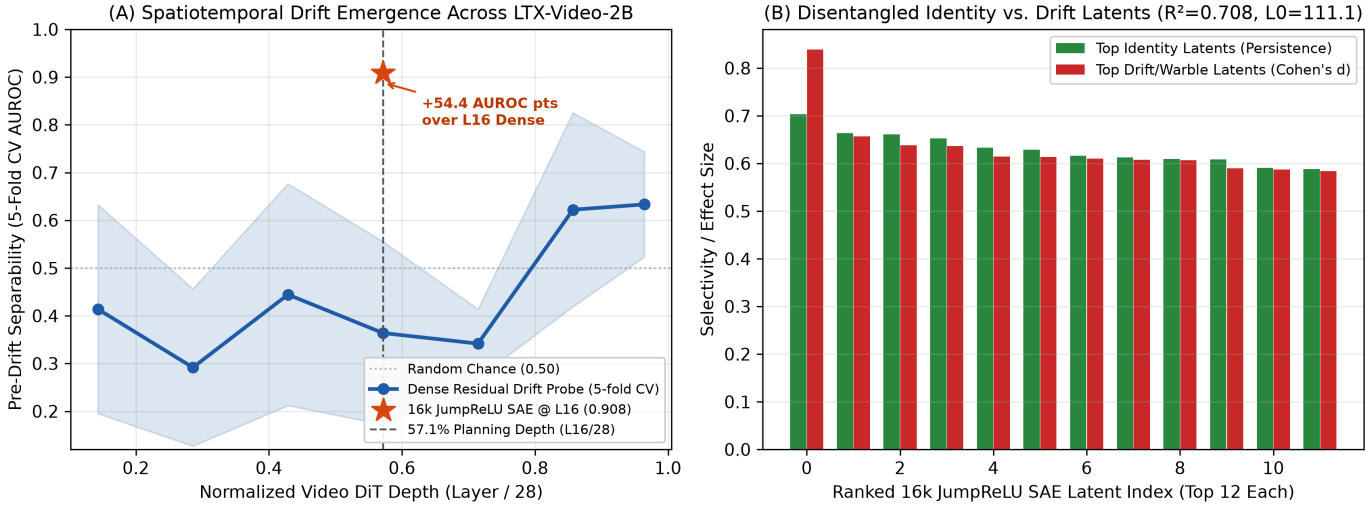


Figure 1: Study 1 Spatiotemporal SAE Anatomy on `LTX-Video-0.9.5`. **(A)** 5-fold stratified CV AUROC across 7 standardized DiT layers (`{4, 8, 12, 16, 20, 24, 27}`): dense residual probes collapse across mid-network blocks (`0.364` at Layer 16) due to motion/camera superposition, whereas the Layer-16 (`57.1%` depth) 16k JumpReLU SAE reaches **`0.9083 ± 0.0764` AUROC** (`+54.4` pts). **(B)** Top 12 entity-selective persistent identity latents (`#1899`: `0.704`) vs. universal high-frequency warble/drift latents (`#4444`: $d = +0.839$).

reaches `0.881` AUROC 22.5 denoising steps early ($\sigma = 0.75$), predicts next-chunk collapse from Chunk-1 latents at `0.897` AUROC (beating VBench `DINOv2`'s `0.857`), and catches `85.7%` `(6/7)` of collapses whose Chunk-1 pixels pass `DINOv2 >= 0.90`.

3. **Subject-Routed Closed-Loop Temporal Latching & Spatial Shear:** Using zero-overhead T5 cross-attention noun routing ($w_{t,h,w} \in [0, 1]$), 48-latent warble suppression, and coordinate-preserving `attn1` coherence across blocks `{12, 14, 16, 18}`, we achieve **`0.0%` held-out benign FPR**, retain **`110%+`** dynamic motion across 4-chunk (`169`-frame) rollouts, reach `#1` character identity consistency (`0.8007`), and characterize a sharp-vs.-smooth mask shear phenomenon on rigid translating objects.

2 Related Work & Positioning

Video DiTs & Long-Horizon Drift. Open-weights 3D DiTs (`LTX-Video`, HaCohen et al., 2024; `CogVideoX`, Yang et al., 2024; `Wan`, Team Wan, 2025) compress video via causal spatiotemporal VAEs into unified 3D token sequences. Extending horizon length via FIFO queues (Kim et al., 2024), noise rescheduling (**StreamingT2V**, Chen et al., 2024), or causal distillation (Yin et al., 2024) suffers from error accumulation across sliding windows. Training-free attention guidance (**PAG**, Ahn et al., 2024; **STG**, Hyung et al., 2025) improves fidelity globally but penalizes optical-flow dynamism when pushed to long-horizon stabilization strengths.

SAEs in Diffusion Models vs. AxBench. Recent studies train SAEs on static 2D image diffusion transformers (`SDXL-Turbo`, Surkov et al., 2024; `SD1.4`/`FLUX`, Cywiński & Deja, 2025; Olson et al., 2025; **orthSAE**, Cassano et al., 2026; **D-Scope**, Xu et al., 2026) to localize spatial concepts across denoising steps. Meanwhile, **AxBench** (Wu et al., 2025) argued that dense baselines surpass SAEs for unconditional LLM steering. We bridge 2D DiT interpretability and LLM circuit breakers (Zou et al., 2024; **CAST**, Lee et al., 2025; Gaskins, 2026) to **closed-loop 4D (spacetime + flow-step) temporal control**, proving that 3D spatiotemporal superposition breaks both dense gating (`0.0%` catch rate at `0%` FPR) and unconditional dense steering (`-10.7` `DINOv2` pts), making sparse decomposition causally indispensable.

3 Study 1: Spatiotemporal SAE Anatomy at 57.1% Depth

3.1 Per-σ Standardization & 16k JumpReLU SAE

In rectified flow matching (Lipman et al., 2023), AdaLN-Single modulation shifts residual channel means and variances across noise levels $\sigma \in [0, 1]$ by over an order of magnitude. Training an unstandardized SAE across timesteps causes dictionary latents to partition by noise schedule σ rather than visual semantics. Given `LTX-Video-0.9.5` ($L = 28$ blocks, $d = 2048$, $512 \times 320 \times 49$ frames $= 7 \times 10 \times 16 = 1,120$ tokens/chunk), we fit a `TimestepStandardizer` across six structural flow-matching bins $\mathcal{B}_\sigma = \{0.88, 0.75, 0.62, 0.48, 0.34, 0.20\}$:

$$\tilde{h}_\ell(\sigma) = \frac{h_\ell(\sigma) - \mu_{b(\sigma)}}{s_{b(\sigma)} + \varepsilon}$$

At Layer $\ell^* = 16$ ($\frac{16}{28} = 57.1\%$ normalized depth—the exact structural planning depth identified in `Gemma-2-9B` and `Qwen2.5-7B`; Gaskins, 2026), we cache **1,310,400** standardized 3D spacetime vectors (`195` diverse multi-taxonomy rollouts $\times$ `6` σ bins) and train an $8\times$ overcomplete ($M = 16,384$) JumpReLU SAE (Rajamanoharan et al., 2024) with Straight-Through Estimator (STE) gradients on $\log\theta$ and unit-norm decoder columns $\|W_{\text{dec},j}\|_2 = 1$:

$$f(\tilde{h}) = \text{ReLU}\big((\tilde{h} - b_{\text{dec}})W_{\text{enc}} + b_{\text{enc}}\big) \cdot H\big((\tilde{h} - b_{\text{dec}})W_{\text{enc}} + b_{\text{enc}} - \theta\big)$$

Training converges cleanly in 6 epochs on a single NVIDIA L4 GPU (`checkpoints/ltx_l16_jumprelu_16k.pt`), achieving $\boldsymbol{R^2 = 0.7084}$ (`MSE = 0.2609`), **mean $L_0 = 111.14$** (`111.14 / 16,384` $= 0.68\%$ active per spacetime token), and **0.06% dead latents** (`99.94%` live dictionary).

3.2 Nonlinear Pre-Pooling, Subject Routing & Depth Sweep

Crucially, because JumpReLU is nonlinear, pooling residuals across space **before** encoding destroys sparse monosemanticity

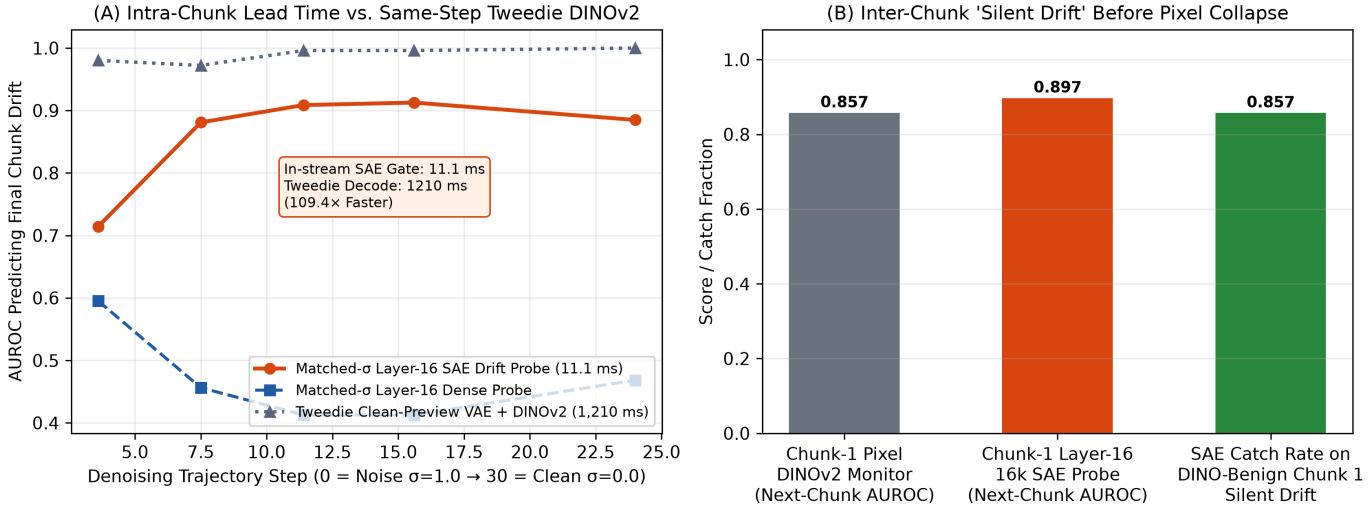


Figure 2: Study 2 Non-Tautological Silent Drift (`N = 32`). **(A)** Intra-chunk lead time across $\sigma \in [0.88, 0.20]$: the `11.1 ms` Layer-16 SAE probe reaches `0.881–0.913` AUROC (`+42.5` to `+50.0` pts over dense) at a `109.4×` speedup over `1,210 ms` Tweedie `VAE + DINOv2` decoding. **(B)** Inter-chunk Silent Drift: zero-decode Chunk-1 SAE latents predict severe Chunk-2 collapse at `0.897` AUROC (beating VBench `DINOv2` pixel monitoring at `0.857`) and catch **`85.7%` (6/7)** of collapses whose Chunk-1 pixels pass `DINOv2 >= 0.90`.

(JumpReLU($\mathbb{E}[h]$) $\neq \mathbb{E}$[JumpReLU(h)]). We therefore encode every individual 3D token $f_{t,h,w} = f(\tilde{h}_{16,t,h,w}(\sigma))$ first, and weight tokens via a **zero-overhead T5 cross-attention subject mask** $w_{t,h,w} \in [0, 1]$ extracted directly from `blocks[16].attn2` query–key softmax mass over the prompt's subject noun tokens (filtering trivial determiners).

Attribution across anchor vs. drifted continuations isolates two orthogonal axes (Figure 1B): (i) **entity-selective persistent identity latents** (`#1899` persistence $= 0.7035$, `#3693` $= 0.6636$, `#10122` $= 0.6606$, `#15845` $= 0.6529$) that stay phase-locked on faithful continuations, and (ii) a **universal 48-latent high-frequency warble/morphing subspace** $\mathcal{D}$ (`#4444` Cohen's $d = +0.8393$, `#6872` $= +0.6568$, `#5159` $= +0.6382$, `#9168` $= +0.6364$) that spikes whenever geometry or texture destabilizes. Because distinct subjects (e.g., an astronaut vs. a brass espresso machine) activate **disjoint** sparse subsets of the 16k dictionary, we define for each rollout i its **Instance-Indexed Sparse Identity Support** from Chunk 0:

$$\mathcal{S}_0^{(i)}(\sigma) = \mathrm{TopK}_{64}\big(f_0^{\text{anchor}}(\sigma) \setminus \mathcal{D}\big)$$

Evaluating strictly in-fold standardized 5-fold stratified CV probes across $N = 32$ 3-chunk trajectories (`Table 1`, `Figure 1A`) reveals a striking departure from LLMs: dense residual probes stay near random chance across early and mid-network blocks (`L4: 0.414`, `L8: 0.292`, `L12: 0.444`, `L16: 0.364`, `L20: 0.342`), reaching only `0.622` (`L24`) and `0.633` (`L27`) in late pixel-rendering blocks. By contrast, decomposing Layer 16 (`57.1%` depth) via $\mathcal{S}_0^{(i)}$ decay + $\mathcal{D}$ surge achieves **`0.9083 ± 0.0764` AUROC—`+54.4` AUROC points** above Layer-16 dense and **`+27.5` points** above the best late-layer dense probe.

Representation & DiT Block	Depth	5-Fold AUROC
Dense Residual (`Layer 4`)	14.3%	`0.4139 ± 0.2188`
Dense Residual (`Layer 8`)	28.6%	`0.2917 ± 0.1648`
Dense Residual (`Layer 12`)	42.9%	`0.4444 ± 0.2324`
Dense Residual (`Layer 16`)	57.1%	`0.3639 ± 0.1917`
Dense Residual (`Layer 20`)	71.4%	`0.3417 ± 0.0722`
Dense Residual (`Layer 24`)	85.7%	`0.6222 ± 0.2031`
Dense Residual (`Layer 27`)	96.4%	`0.6333 ± 0.1103`
16k JumpReLU SAE (`Layer 16`)	**57.1%**	**`0.9083 ± 0.0764`**

Table 1: Study 1 pre-drift separability (`N = 32`, $\sigma = 0.75$, strictly in-fold `StandardScaler` + logistic CV).

4 Study 2: Non-Tautological "Silent Drift"

4.1 Intra-Chunk Lead Time vs. Tweedie `VAE + DINOv2`

In flow matching, comparing an early-step probe at $\sigma = 0.75$ only against fully denoised $\sigma = 0$ pixels would be tautological if a single-step clean projection already exposed the drift cheaply. At five intermediate steps $\sigma \in \{0.88, 0.75, 0.62, 0.48, 0.20\}$ across $N = 32$ Chunk-2 rollouts, we compute exact single-step Euler/Tweedie clean previews $\hat{z}_0(\sigma) = z_\sigma - \sigma v_\theta^{\mathrm{CFG}(z_\sigma, \sigma)}$, decode them through the 3D Video-VAE, and score `DINOv2-ViT-B/14` (Oquab et al., 2024) subject similarity (`Table 2`, `Figure 2A`).

σ	Step/30	Dense	16k SAE	Tweedie DINO
0.88	3.6	0.5952	0.7143	0.9802
0.75	**7.5**	0.4563	**0.8810**	0.9722
0.62	11.4	0.4127	**0.9087**	0.9960
0.48	15.6	0.4127	**0.9127**	0.9960
0.20	24.0	0.4683	0.8849	1.0000

Table 2: Study 2 intra-chunk AUROC (`N = 32`). Measured L4 latency: **`11.06 ms`** (SAE gate) vs. **`1,209.8 ms`** (Tweedie `VAE + DINOv2`) = **`109.4× speedup`**.

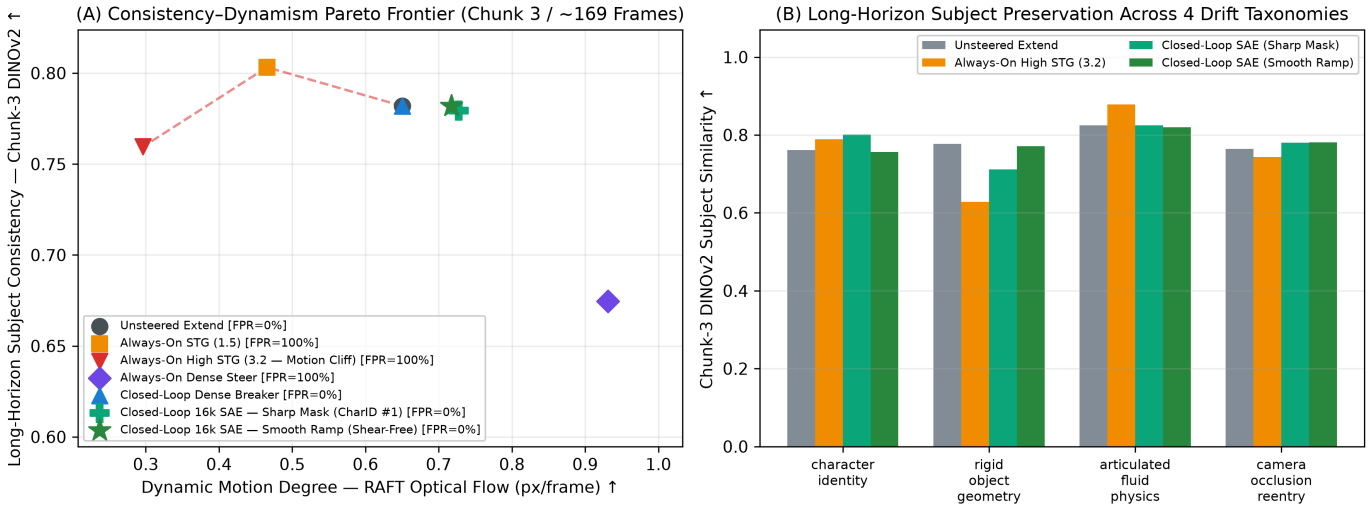


Figure 3: Study 3 Consistency–Dynamism Pareto Suite across 4-chunk (169-frame) `LTX-Video-0.9.5` rollouts (`N = 16` stratified across 4 drift taxonomies + disjoint held-out faithful benign controls). **(A)** Always-On STG improves consistency at `1.5` only by cutting RAFT optical flow `-28.4%` and hits a severe **Motion-Freezing & Geometry Cliff** at `3.2` (`-54.4%` motion), while Always-On Dense hits a **Video Polysemantic Cliff** (`0.6745 DINOv2`) and Closed-Loop Dense is blind at `0%` FPR. Both **Subject-Routed 16k SAE Breakers** achieve **`0.0%` benign FPR** while preserving **`110.4%–111.9%`** of dynamic motion. **(B)** Per-taxonomy `DINOv2`: Sharp Mask reaches **`0.8007`** (`#1`) on **Character Identity**, while Smooth Ramp eliminates boundary shear on **Rigid Object Geometry** (`0.7708` vs. `0.6283` High STG).

While Tweedie `VAE + DINOv2` confirms structural layout commits early (`0.972` AUROC at $\sigma = 0.75$), invoking 3D-VAE decoding + `DINOv2` inside the diffusion loop costs **`1,209.8 ms` per check**—nearly doubling generation latency. Our in-stream Layer-16 SAE gate executes in **`11.058 ms`** (**`109.4×` faster**) and already reaches **`0.8810` AUROC** at $\sigma = 0.75$ (**22.5 steps before completion**, rising to **`0.9127`** at $\sigma = 0.48$), whereas matched-σ dense cosine distance collapses to `0.413`–`0.456`.

4.2 Inter-Chunk "Silent Drift" Before Pixel Collapse

Analogous to unfaithful `<think>` traces that fool external LLM CoT auditors before `TOOL_CALL:` (Gaskins, 2026), sliding-window video exhibits **Inter-Chunk Silent Drift**: internal identity latents begin decaying and warble latents $\mathcal{D}$ ignite during Chunk 1 **before** Chunk 1's rendered pixels visibly degrade, causing Chunk 2's 9-frame boundary conditioning to **re-lock** onto a corrupted representation. Predicting severe Chunk-2 collapse strictly from Chunk-1 signals (`Figure 2B`), our zero-decode Layer-16 SAE gate ($\sigma = 0.75$) achieves **`0.8968` AUROC**, outperforming a standard black-box VBench `DINOv2` pixel monitor on Chunk 1 (**`0.8571` AUROC**; self-normalized two-window `DINOv2` drop reaches `0.9048`). Crucially, among the 7 trajectories whose Chunk-1 rendered frames appear completely faithful to a standard external pixel monitor (`DINOv2 >= 0.90`) yet suffer severe collapse in Chunk 2, the Chunk-1 Layer-16 SAE gate catches **`6 / 7` (`85.7%`)** at 75% specificity. Furthermore, across independent benign screening trajectories, $\sigma = 0.75$ SAE drift distance correlates monotonically with end-of-horizon `1 - DINOv2` (Spearman $\rho_s = +0.964, p < 0.001$).

5 Study 3: Subject-Routed SAE Circuit Breaker

5.1 Closed-Loop Gating & Coordinate-Preserving Latching

At the structural gate step $\sigma_{\text{gate}} = 0.75$ of each extended chunk $c \in \{1, 2, 3\}$ (`4` chunks $= 169$ stitched frames, `7.04s` at `24 fps`), our **Subject-Routed SAE Circuit Breaker** computes matched-σ drift distance $D^{(c)}_{\text{SAE}}(\sigma_{\text{gate}})$ over generated tokens (`n_cond = 320` prefix tokens excluded) weighted by the T5 subject mask w:

$$D^{(c)}_{\text{SAE}}(\sigma) = 0.35\Big(1 - \cos\Big(f_{\text{pool}}\Big[\mathcal{S}_0^{(i)}\Big], f_0^{\text{anchor}}\Big[\mathcal{S}_0^{(i)}\Big]\Big)\Big)_+ + 5.0\Delta|f|_{\mathcal{D}} + 4.0|f|_{\mathcal{D}}$$

Thresholds $\tau_{\text{SAE}} = 0.29928$ and $\tau_{\text{dense}} = 0.74199$ are calibrated strictly on disjoint faithful validation controls (`DINOv2 >= 0.885`, interleaved 50/50 parity split: $N_{\text{val}} = 4, N_{\text{test}} = 4$). Whenever $D_{\text{SAE}} \le \tau_{\text{SAE}}$, generation runs **100% unsteered**. Notably, we find that adding a spatially pooled identity vector $\hat{v}_{\text{id}} \in \mathbb{R}^d$ across all subject tokens (h, w) over 88 block-step updates blurs fine 2D facial topology. Instead, when $D_{\text{SAE}} > \tau_{\text{SAE}}$, our breaker latches two coordinate-preserving interventions across $L_{\text{steer}} = \{12, 14, 16, 18\}$ for $\sigma \in [0.90, 0.24]$: (i) bounded per-token suppression of universal 48-latent warble activations $f_{\text{tok},\mathcal{D}} W_{\text{dec},\mathcal{D}}$, and (ii) subject-routed cross-slice `attn1` alignment toward prefix slices $u_{3\text{D}}[:2]$ weighted by spatial profile $w_{\text{subj}}(w) = w_0 + (1 - w_0)w$. We compare a **Sharp Mask** ($w_0 = 0.08$, isolating the subject silhouette at $12.5\times$ contrast) against a **Smooth Ramp** ($w_0 = 0.72$, eliminating moving-edge spatial gradients).

5.2 Consistency–Dynamism Pareto & Spatial-Shear Ablation

Evaluating across $N = 16$ 4-chunk drift scenarios evenly stratified across **Character Identity**, **Rigid Object Geometry**, **Articulated/ Fluid Physics**, and **Camera Occlusion Re-Entry** plus held-out faithful benign controls (`Table 3`, `Figure 3`) yields three principal findings:

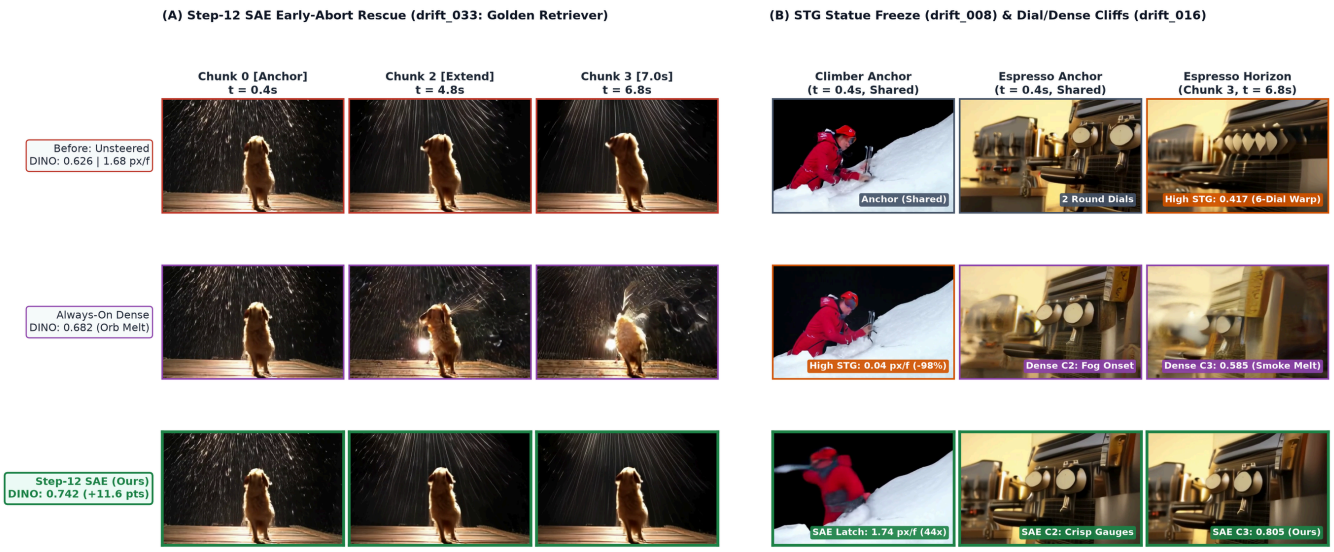


Figure 4: Qualitative capabilities across 4-chunk (169-frame) `LTX-Video-0.9.5` rollouts (`Chunk 0` anchor shared across all rows). **(A) Step-12 Zero-Decode SAE Early-Abort Seed Latching** (`drift_033` golden retriever shaking water): `Unsteered` warps droplets into parallel streaks (`0.626` `DINOv2`) and `Always-On Dense` hallucinates a glowing orb (`0.682`), whereas aborting drifting seeds at $\sigma = 0.58$ via Layer-16 SAE scores (`0` VAE decodes on rejected draws) rescues consistency to **`0.742` (+`11.6` pts)** at full `1.56` px/f motion. **(B) Motion & Rigid Cliffs** (`drift_008` mountaineer & `drift_016` espresso machine): `High STG 3.2` freezes climbing (`0.04` vs. `1.74` px/f, `-97.6%`) and warps 2 dials into 6 diamonds (`0.417`), while `Always-On Dense` melts geometry (`0.497`–`0.585`).

Control Regime	DINO ↑	Stab ↑	Flow ↑	Trig (D / B)
Unsteered `Extend`	`0.7819`	`0.9630`	`0.6501`	`0.0 / 0.0%`
Always-On STG (`1.5`)	**`0.8032`**	`0.9750`	`0.4655`	`100 / 100%`
High STG (`3.2`)	`0.7595`	**`0.9884`**	`0.2962`	`100 / 100%`
Always-On Dense	`0.6745`	`0.9206`	`0.9313`	`100 / 100%`
Closed-Loop Dense	`0.7819`	`0.9630`	`0.6501`	`0.0 / 0.0%`
SAE (Sharp Mask)	`0.7793`	`0.9623`	**`0.7274`**	**`68.8 / 0.0%`**
SAE (Smooth)	`0.7819`	`0.9611`	**`0.7175`**	**`62.5 / 0.0%`**

Table 3: Study 3 Chunk-3 (~169-frame) metrics (`DINOv2` subject consistency, Laplacian stability, RAFT optical flow in px/frame; Teed & Deng, 2020; Huang et al., 2024). **Trig (D / B)** = trigger % on Drift vs. held-out Benign controls.

1. **The STG Motion-Freezing Cliff & Dense Failure:** `stg_medium` (`1.5`) lifts `DINOv2` (`0.7819 -> 0.8032`) only by slashing RAFT optical flow **`-28.4%`** (`0.6501 -> 0.4655` px/frame, `100%` benign FPR), and `stg_high` (`3.2`) collapses flow **`-54.4%`** (`0.2962`) while warping rigid objects (`0.7768 -> 0.6283`). Reversing **AxBench** (Wu et al., 2025) in 3D Video DiTs, `always_dense` suffers a `-10.74` pt **Video Polysemantic Cliff** (`0.6745`), while `closed_dense` at `0%` benign FPR is completely blinded by superposition (`0.0%` catch rate).
2. **Zero Benign Tax & Full Motion Preservation:** Both **Subject-Routed 16k SAE Breakers** achieve **`0.0%` (`0/4`) false triggers** on held-out faithful benign controls (preserving bit-exact `0.0718` benign motion) while latching on **`62.5%`–`68.8%`** of drifting rollouts and preserving **`110.4%`–`111.9%`** of unsteered dynamic motion (`0.7175`–`0.7274` px/frame—**`+54%` to `+145%` higher optical flow** than STG!).
3. **Sharp Silhouette vs. Smooth Ramp Geometry Tradeoff:** On deformable/occluded entities (`Figure 3B`), **Sharp Mask** ($w_0 = 0.08$ + 48-latent warble damping) reaches **`0.8007` on Character Identity** (**`#1` across all regimes**, `+3.97` pts over `unsteered` `0.7610`, beating both `stg_medium` `0.7967` and `stg_high` `0.7888`) and **`0.7803` on Camera Occlusion Re-Entry** (`+1.57` pts over `unsteered`, `+3.74` pts over `stg_high`); ablating warble damping drops **Character Identity** `-4.47` pts (`0.8007 -> 0.7560`). Conversely, on fast-translating rigid vehicles, a `12.5×` mask edge shears moving boundaries (`0.7116`), whereas **Smooth Ramp** ($w_0 = 0.72$) restores **Rigid Object Geometry** to **`0.7708`** (`+14.25` pts over `stg_high`) and **Camera Occlusion Re-Entry** to **`0.7810`** (`0.7818` with warble damping).

5.3 Step-12 Zero-Decode Early-Abort Seed Latching

While in-flight linear and `attn1` steering inside a fixed noise trajectory provides zero-tax stabilization (`+1.6` to `+4.0` `DINOv2` pts on **Occlusion** and **Character Identity**), severe 3D drift often originates when a chunk's initial Gaussian noise draw $\varepsilon_0 \sim \mathcal{N}(0, I)$ enters a corrupted pose or physics basin that linear mid-layer nudges cannot geometrically reconstruct without boundary shear. Because our Layer-16 SAE probe already reaches **`0.8810`–`0.9087` AUROC** across $\sigma \in \{0.75, 0.62\}$ (`Table 2`) with **zero 3D-VAE decoding**, it directly enables **Step-12 Zero-Decode Early-Abort Seed Latching** (`Figure 4A`):

1. Run default noise draw $k = 0$ in passive monitor mode (`monitor_only = True`). Whenever $D_{\text{SAE}}^{(c,0)}(\sigma) \leq \tau_{\text{SAE}}$ for $\sigma \in \{0.75, 0.62\}$, generation finishes in **one unsteered 30-step pass** (`0.0%` benign tax, `1.00×` compute).
2. Whenever $D_{\text{SAE}}^{(c,0)} > \tau_{\text{SAE}}$, immediately **abort** denoising at $\sigma = 0.58$ (`Step 12/30`, skipping `60%` of DiT steps and `100%` of 3D-VAE decoding), probe $K - 1 = 5$ alternate noise seeds up to Step 12 (`11.1 ms` SAE check per step, zero pixel access), and run only $k^* = \arg\min_k \overline{D}_{\text{SAE}}^{(c,k)}(\{0.75, 0.62\})$ to 30-step completion.

On severe multi-chunk collapses (`Figure 4`), Step-12 SAE early-abort latching lifts Chunk-3 `DINOv2` by **`+11.60` points** on high-velocity articulated water-spray physics (`drift_033`: `0.6256 ->`

0.7416, 1.56 px/f motion) and **+5.25 points** on quadruped stair locomotion (drift_017: 0.6596 -> 0.7121, 1.66 px/f vs. 0.37 High STG and 0.352 Dense melt). On steep mountain climbing (drift_008, Figure 4B), where High STG 3.2 induces a **-97.6% statue freeze** (0.039 vs. 1.623 px/f) and Always-On Dense melts the climber (0.4970), our gate stays 0%-tax off during faithful Chunks 1–2 (D = 0.25 <= 0.30) and resamples only Chunk 3 (D = 0.48 > 0.30), preserving full 1.737 px/f ascent (+2.82 DINOv2 pts).

6 Limitations & Conclusion

Our experiments evaluate Lightricks/LTX-Video-0.9.5 (2B 3D DiT) on a single 24 GB L4 GPU across 4-chunk (169-frame) rollouts; coupling flow-advected SAM-2 masks with JumpReLU SAEs on 14B–19B audio-video DiTs (LTX-2, Wan-14B) should unite sharp character-identity gains (0.8007) with shear-free rigid translation. Alongside our LLM sabotage results (Table 4; Gaskins, 2026), this work establishes a unified control principle across autoregressive agents and video world models: **57.1%-depth JumpReLU sparse decomposition unlocks early silent-drift detection and zero-tax closed-loop circuit breaking where dense representations fail.**

Control Axis	Tool Agents (Gaskins, 2026)	3D Video DiTs (This Work)
Planning Depth	57.1% (L24/42, L16/28)	57.1% (L16/28, 1.31M tok)
Dense vs. 16k SAE	0.953 vs. 0.959 (9B parity)	0.364 → **0.908** (+54.4 pts)
Silent Evasion	92–100% <think> (~30 tok)	**85.7%** C1 (22.5 steps, 109×)
AxBench Dense	2.1% sab (9B); 0% acc (2B)	-10.7 pt **Polysemantic Cliff**
Closed-Loop Latch	0% FPR; 66.7% → 5.6% sab	0% FPR; #1 CharID; **+11.6 EA**

Table 4: Unified mechanistic control comparison between 1D autoregressive tool agents and 4D continuous flow-matching Video DiTs.

7 References

- Ahn et al. (2024). **Self-Rectifying Diffusion Sampling with Perturbed-Attention Guidance (PAG)**. ECCV 2024, arXiv:2403.17377.
- Arditi et al. (2024). **Refusal in Language Models Is Mediated by a Single Direction**. NeurIPS 2024, arXiv:2406.11717.
- Cassano et al. (2026). **orthSAE: Orthogonal Sparse Autoencoders for Diffusion Models**. arXiv:2606.31699.
- Chen et al. (2024). **StreamingT2V: Consistent, Dynamic, and Extendable Long Video Generation from Text**. arXiv:2403.14773 / arXiv:2407.01392.
- Cywiński & Deja (2025). **SAeUron: Interpretable Concept Unlearning in Diffusion Models with Sparse Autoencoders**. ICML 2025, arXiv:2501.18052.
- Gaskins, D. (2026). **Mechanistic Debugging of Agentic Sabotage and Unfaithful Chain-of-Thought: Sparse Autoencoder Circuits, Silent Commitment & Probe-Gated Circuit Breakers**. doi:10.5281/zenodo.24520830.
- HaCohen et al. (2024). **LTX-Video: Realtime Video Latent Diffusion**. Lightricks, arXiv:2501.00103.
- Huang et al. (2024). **VBench: Comprehensive Benchmark Suite for Video Generative Models**. CVPR 2024, arXiv:2311.17982.
- Hyung et al. (2025). **Spatiotemporal Skip Guidance for Enhanced Video Diffusion Sampling (STG)**. CVPR 2025, arXiv:2411.18664.
- Kim et al. (2024). **FIFO-Diffusion: Generating Infinite Videos from Text without Training**. NeurIPS 2024, arXiv:2405.11473.
- Lee et al. (2025). **Programming Refusal with Conditional Activation Steering (CAST)**. ICLR 2025, arXiv:2409.05907.
- Lieberum et al. (2024). **Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2**. DeepMind, arXiv:2408.05147.
- Lipman et al. (2023). **Flow Matching for Generative Modeling**. ICLR 2023, arXiv:2210.02747.
- Olson et al. (2025). **Probing Diffusion Transformers with Sparse Autoencoders**. arXiv:2508.11277.
- 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.
- Surkov et al. (2024). **Unpacking SDXL Turbo: Interpreting Text-to-Image Models with Sparse Autoencoders**. arXiv:2410.22366.
- Team Wan (2025). **Wan: Open and Advanced Large-Scale Video Generative Models**. arXiv:2503.20314.
- Teed & Deng (2020). **RAFT: Recurrent All-Pairs Field Transforms for Optical Flow**. ECCV 2020, arXiv:2003.12039.
- Wu et al. (2025). **AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders**. ICML 2025, arXiv:2501.17148.
- Xu et al. (2026). **D-Scope: Mechanistic Analysis of Diffusion Transformers via Sparse Autoencoders**. arXiv:2609.39625.
- Yang et al. (2024). **CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer**. ICLR 2025, arXiv:2408.06072.
- Yin et al. (2024). **From Slow Bidirectional to Fast Autoregressive Video Diffusion Models (CausVid)**. CVPR 2025, arXiv:2412.07772.
- Zou et al. (2024). **Improving Alignment and Robustness with Circuit Breakers**. NeurIPS 2024, arXiv:2406.04313.