GAUGE AND DRIFT: SEPARATING COMMANDED CHANGE FROM ERROR ACCUMULATION IN AUTOREGRESSIVE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive generators let a user steer a rollout in flight — move the camera, switch an action, take a branch — while the same autoregressive loop that grants this control also accumulates error. Drift scores measure how far the current frame has moved from a reference, and so charge a model for change the user *asked for*. Existing remedies work *around* the measure, gating scores on whether the action executed or building action-free probe protocols; the drift quantity itself still treats a rollout's change as one indivisible number. We propose an *internal* fix. Modelling admissible interventions as a group acting on the generator's representation places commanded change, to first order, in a low-dimensional *gauge subspace*, leaving drift as displacement in the *invariant complement*. The norm of that projection is *gauge-invariant drift* (GID): it depends only on the subspace and not on the control basis, vanishes to first order on purely commanded trajectories, and is monotone in the subspace dimension. On a controlled autoregressive testbed with exactly parameterized camera commands, the decomposition is real and estimable from short paired rollouts *without drift labels*: measured on paired base-versus-intervened differences, in which shared seeds cancel the ordinary generation step, a per-reference subspace captures commanded change while absorbing almost none of the drift (separation 0.74), against 0.00 both for the same subspace at a mismatched reference and for a random subspace of equal dimension. It repeats on a video-to-video chain (0.63), on a photometric control family, on discrete branch selection, and in a third encoder whose training exerts no invariance pressure. The subspace is a genuinely *local* object in the representation we measure in — the same command at two images moves features along nearly orthogonal directions — which turns estimator design into a concrete engineering question: a projector re-estimated at the current state retains separation ≈ 0.35 at every depth tested, $80\times$ chance. A factorial experiment on 500 rollouts, with depth and command count randomized independently, then quantifies what GID is built for: every displacement-based score loads substantially on command count, i.e. is partly reporting obedience rather than degradation. Carried through to a whole-rollout score a maintained projector removes about a further point of command inflation, and the appendices report what governs that amount, a negative result for the regularizer form, and the diagnostics that locate both.

1 INTRODUCTION

Video generation is becoming *interactive*: autoregressive and streaming models emit frames causally and expose the loop to a user who steers it in flight—pan the camera, change the action, take a branch (Chen et al., 2024; Huang et al., 2025; Yin et al., 2025; He et al., 2025b; Zhao et al., 2026). The same causal structure that grants this control makes the model its own worst enemy: each step conditions on its own previous output, so errors compound and subjects deform over a long rollout. This is *drift*, the dominant failure mode of long interactive generation (Liu et al., 2026; Zhang et al., 2026b).

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.

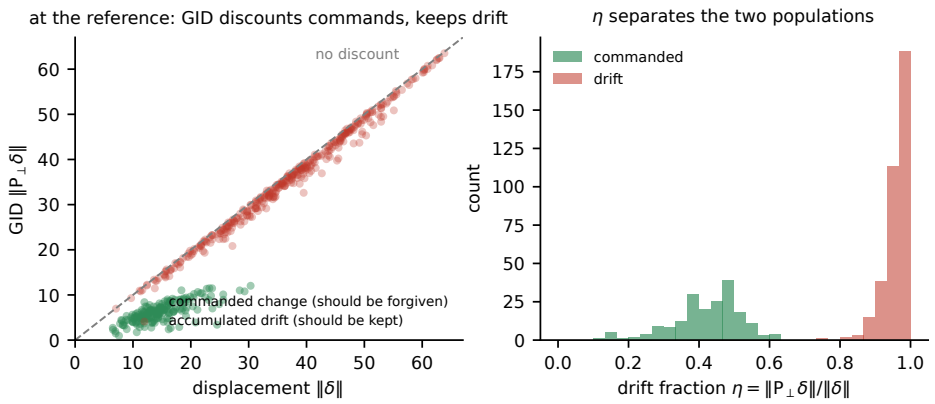


Figure 1: The failure, and the fix, measured at a reference. **Left:** each point is one change vector — a commanded camera move (green) or a step of accumulated drift (red) — plotted as displacement against gauge-invariant drift. Commanded change falls far below the diagonal: GID discounts it. Drift stays on it: GID keeps it. **Right:** the drift fraction η separates the two populations that a displacement score reports identically.

A concrete failure of the ruler. Consider the cleanest interactive command: the user pans the camera. Every pixel moves, displacement from the first frame is large, and subject/background consistency and the drift-error variants built on them (Huang et al., 2024; Zhang et al., 2026b) all report severe degradation — yet *nothing has gone wrong*. Now hold the control fixed and roll out long until the subject dissolves: displacement may be comparable, but this time the model failed. A ruler returning the same reading for both is measuring motion, not drift. The consequences are practical: such a leaderboard ranks an unresponsive model above a controllable one, and the same quantity used as a regularizer is literally an instruction not to move.

What the community already does, and what is missing. This tension is recognized, and answered with *external* devices: gating a score on whether the commanded action executed, or building action-free probe segments so residual change must be degradation (AMAP CV Lab, 2026); on the training side, constraining supervision whose semantic shift departs too far from the generator's own trajectory (Wu et al., 2026). Both are real progress and both work *around* the measure, needing a designed protocol or an intervention on supervision while the drift quantity still treats a rollout's change as one indivisible number. Subtracting the commanded component *from the measurement* is what would make an arbitrary mixed rollout measurable without annotation.

Our claim. The fix is structural. Change induced by a legitimate intervention is not arbitrary: it is constrained by what the interface can express, so panning a camera or switching an action moves the representation along a particular, low-dimensional set of directions, while drift is displacement no admissible intervention could have produced—movement *transverse* to those directions. This is the language of gauge freedom: admissible interventions span a gauge subspace, an equivalence class of "the same content, legitimately restyled," and drift is the component leaving the class. Reading admissible transformations as a group and quotienting it out is a developed program in learning (Cohen & Welling, 2016; Cohen et al., 2019; Bronstein et al., 2021); we bring it to interactive drift, where the group comes from a user-facing control interface rather than an architectural symmetry.

Contributions. In decreasing order of confidence. **H1 (decomposition).** We formalize representation change under interactive control as a group action and posit that it splits, to first order, into a low-dimensional gauge subspace plus an invariant complement (Section 3), with three propositions: GID depends on the interface only through the subspace, not the basis; it vanishes to first order on purely controlled trajectories; and it is monotone in the admissible subspace, making r an explicit false-alarm/miss knob. That control effects are legible as displacement in a frozen encoder's features is not our conjecture but the operating assumption of recent latent-action work (Jiang et al., 2026), which we adopt and cite as such. **H2 (a more honest metric).** We define gauge-invariant drift as

displacement projected onto the estimated complement and ask whether it responds to degradation rather than to obedience. We test this with a factorial design in which rollout depth (which generates the error) and command count (the confound) are randomized independently, so "what is this measure responding to?" has an answer rather than a correlation. **H3 (drift-only regularization).** The same projector becomes a penalty on complement-space displacement only; we give its form and test it as inference-time guidance, with a control arm that calibrates what the lever can move (Appendix G).

What we find. The decomposition is real and cheap to estimate — eight short paired rollouts per reference, no drift labels, no annotation — and it is a *local* object, attached to a state rather than to the control interface, so the design question becomes how often to refresh it. A projector re-estimated at the current state separates commanded change from drift at every depth we tested, and a factorial design quantifies how much existing displacement scores respond to obedience rather than degradation.

2 Related Work

Interactive video generation and drift. Long-horizon generation has moved from bidirectional diffusion of a fixed clip (Ho et al., 2022; Blattmann et al., 2023; Brooks et al., 2024) to causal formulations steerable mid-rollout: per-frame independent noise levels enable autoregressive rollout (Chen et al., 2024), self-forcing closes the train–test gap on the model's own outputs (Huang et al., 2025), and distilling a bidirectional teacher into a causal few-step student brings latency into the interactive regime (Yin et al., 2025); open world models expose actions, camera trajectories, and branches as first-class inputs (Bruce et al., 2024; Valevski et al., 2024; He et al., 2025b; Zhao et al., 2026; He et al., 2025a), realizing in a generator the learned-simulator programme of earlier world models (Ha & Schmidhuber, 2018; Hafner et al., 2023; LeCun, 2022). Error accumulation under self-conditioning is exposure bias (Bengio et al., 2015; Ranzato et al., 2016; Ross et al., 2011) transplanted into video; mitigations use rolling windows and attention sinks (Liu et al., 2026), chunk-wise noise scheduling (Zhang et al., 2026b), and cache-level re-anchoring. Closest to H3, *trust-region steering* gates supervision by the magnitude of its semantic shift, having identified that condition-aligned but trajectory-agnostic supervision induces persistent drift after a condition change (Wu et al., 2026). That work and ours share a target and differ in kind: it gates a scalar magnitude on the *supervision signal*; we decompose the *state change itself* into commanded and transverse parts, first as a measurement and only then as a penalty. Splitting a representation into factors one wants to control and factors one does not is an old ambition (Higgins et al., 2017); what is available now, and was not, is a control interface that names the factors directly.

Measuring drift, and existing attempts to discount control. Video generators are scored with dimension-wise benchmarks whose consistency axes are feature similarities across frames (Huang et al., 2024), built on perceptual and embedding distances (Zhang et al., 2018; Wang et al., 2004; Radford et al., 2021; Caron et al., 2021; Oquab et al., 2024) and distributional scores (Heusel et al., 2017; Unterthiner et al., 2018; Salimans et al., 2016); for long video these become drift errors over clarity, motion, aesthetics, background and subject (Zhang et al., 2026b). The interactive setting has already forced the community to confront commanded change contaminating such scores: open-world stability evaluation answers with per-frame action metrics, segment-based drift measures catching non-monotonic collapse, *controllability-gated* physics scoring crediting plausibility only under faithful execution, and *action-decoupled* memory protocols (AMAP CV Lab, 2026). We regard this as the right diagnosis and an incomplete cure: gating conditions *when* a score counts and a probe protocol constructs segments where commanded change is absent by design, so both need a curated harness, and on a mixed rollout the drift quantity is still a single norm over undecomposed change. Our departure is to put the discount inside the definition.

Control directions in feature space, and identifiability. Our decomposition assumes control effects are legible as displacement in a frozen encoder's feature space—not a conjecture we introduce, but the premise of latent-action learning, which aligns integrated latent actions to temporal feature differences of a frozen video encoder and shows the space transfers across contexts (Jiang et al., 2026). We take that as independent support for Assumption 1 and adopt the same finite-difference construction for our simplest estimator rather than claiming it as ours. Controlled latent models are moreover generically identifiable only up to an orthogonal transformation of coordinates (Locatello

et al., 2019; Khemakhem et al., 2020; Zhang et al., 2026a), which is why we build the metric on a projector rather than coordinates.

Symmetry, gauge, and quotient geometry. Treating admissible transformations as a group and building measures invariant to it runs from group-equivariant networks to geometric deep learning (Cohen & Welling, 2016; Cohen et al., 2019; Weiler & Cesa, 2019; Bronstein et al., 2021), with reparameterization symmetries studied as both nuisance and structure (Hu et al., 2022) and standard machinery for optimizing over subspaces and their quotients (Edelman et al., 1998; Absil et al., 2008; Boumal, 2023; Björck & Golub, 1973). Our novelty is not this machinery but the *source* of the group—a user-facing control interface estimated from paired rollouts—and the *target*: a drift measure for interactive generation.

3 Problem Formulation

We formalize interactive change as a group action on a generator's representation, isolate a *gauge subspace* of admissible interventions, and define drift as displacement in its complement. Throughout, "representation" means a fixed, identity-relevant feature map $\phi : \mathcal{X} \to \mathbb{R}^D$ applied to a generated frame or pooled window; the construction is agnostic to ϕ, whose choice is an experimental variable (Section D). An interactive generator produces frames $y_1, y_2, \dots$ autoregressively, each conditioned on previous frames and on a possibly time-varying control c_t from an interface $\mathcal{C}$: an action set, a camera pose, a prompt edit, a branch choice. Write $z_t = \phi(y_t) \in \mathbb{R}^D$ and z_0 for the reference; the observed change $\delta_t = z_t - z_0$ (or $z_t - z_{t-1}$) mixes *control-induced change*, caused by $c_{1:t}$ and legitimate, with *drift*, caused by error accumulation and a failure. A metric reporting $\|\delta_t\|$ scores both; we want a functional $\mathrm{Drift}(z_t \mid c_{1:t})$ responding to the second and approximately blind to the first.

3.1 The control group and the gauge subspace

Assumption 1 (Admissible interventions act linearly to first order)**.** There is a neighborhood of the reference z_0 and a smooth map $\Psi : \mathcal{C} \times \mathbb{R}^D \to \mathbb{R}^D$ such that applying control c moves the representation by $z \mapsto \Psi(c, z)$, with $\Psi(\mathbf{0}, z) = z$. Its differential at z_0 spans a linear subspace

$$\mathcal{G}_{z_0} \;=\; \mathrm{span}\big\{\, \partial_c \Psi(c, z_0)|_{c=\mathbf{0}}[e_k] : k = 1, \dots, K \,\big\} \subseteq \mathbb{R}^D,$$

the *gauge (control) subspace* of admissible first-order change, of dimension $r = \dim \mathcal{G}_{z_0} \ll D$. Write $\mathcal{G}_{z_0}^{\perp}$ for its orthogonal complement, the *invariant subspace*.

Assumption 1 is the load-bearing modelling choice, flagged as an assumption rather than a theorem: real control effects are nonlinear and state-dependent, so $\mathcal{G}$ is at best a local, first-order object, and estimating it is the main risk. It is not unsupported — recent latent-action work builds its training signal on the premise that a control's effect is recoverable from temporal feature differences of a frozen encoder (Jiang et al., 2026) — but Section 5.3 measures just how local "local" turns out to be. When admissible interventions form a Lie group G, $\mathcal{G}_{z_0}$ is the tangent space at z_0 to the orbit $G \cdot z_0$, and the drift we want is displacement *transverse* to it.

Definition 1 (Gauge and invariant projections)**.** Let $\mathsf{P}_{\mathcal{G}}$ and $\mathsf{P}_{\mathcal{G}^\perp} = I - \mathsf{P}_{\mathcal{G}}$ be the orthogonal projectors onto $\mathcal{G}_{z_0}$ and $\mathcal{G}_{z_0}^{\perp}$. Decompose any change $\delta = \mathsf{P}_{\mathcal{G}}\delta + \mathsf{P}_{\mathcal{G}^\perp}\delta$ into its *gauge part* (explained by admissible control) and *invariant part* (unexplained).

Definition 2 (Gauge-invariant drift)**.** The gauge-invariant drift at step t is the norm of the invariant part,

$$\mathrm{GID}(z_t \mid c_{1:t}) \;=\; \big\|\mathsf{P}_{\mathcal{G}_{z_0}^{\perp}}\,(z_t - z_0)\big\|, \tag{1}$$

optionally normalized by $\|z_t - z_0\|$ to give a unitless *drift fraction* $\eta_t = \mathrm{GID}(z_t)/\|z_t - z_0\| \in [0, 1]$.

The one-size-fits-all baseline is the special case $\mathcal{G} = \{0\}$, for which GID $= \|z_t - z_0\|$ is ordinary total drift. GID therefore never exceeds total drift and reduces to it when nothing is admissible: it only ever *forgives* change a legitimate intervention explains. This one-sidedness matters empirically—H2 cannot be satisfied by GID being merely different, it must be better precisely where it forgives.

3.2 WELL-POSEDNESS AND INVARIANCE

Proofs are sketched here and carried out in Appendix A, which also states the regularity each quantitative bound assumes.

Proposition 1 (Invariance to reparameterization of the control basis)**.** $\mathrm{GID}(z_t \mid c_{1:t})$ *depends on the control interface only through the subspace* $\mathcal{G}_{z_0}$*, not through the choice of generators* $\{\partial_c \Psi[e_k]\}$ *spanning it: any two full-rank bases of the same* $\mathcal{G}_{z_0}$ *give the same* $\mathsf{P}_{\mathcal{G}_{z_0}^{\perp}}$ *and hence the same GID.*

Sketch. The orthogonal projector onto a subspace is a function of the subspace alone: $UU^{+} = U'U'^{+}$ whenever U, U' share a column space (Golub & Van Loan, 2013), so $\mathsf{P}_{\mathcal{G}^{\perp}} = I - \mathsf{P}_{\mathcal{G}}$ and Eq. (1) are unchanged. □

Proposition 2 (Blindness to pure admissible interventions, to first order)**.** *If between the reference and step* t *the representation moves only by admissible controls and no error accumulates, i.e.* $z_t - z_0 \in \mathcal{G}_{z_0} + o(\|z_t - z_0\|)$*, then* $\mathrm{GID}(z_t \mid c_{1:t}) = o(\|z_t - z_0\|)$*. Conversely, if the movement is entirely transverse,* $z_t - z_0 \in \mathcal{G}_{z_0}^{\perp}$*, then* $\mathrm{GID} = \|z_t - z_0\|$ *and* $\eta_t = 1$*.*

Sketch. If $\delta \in \mathcal{G}$ then $\mathsf{P}_{\mathcal{G}^{\perp}}\delta = 0$; the $o(\cdot)$ collects the second-order remainder of Assumption 1, and the converse is the projector acting as the identity on $\mathcal{G}^{\perp}$. Appendix A makes this uniform, bounding $\widehat{\mathrm{GID}}$ by an orbit-curvature term in $\|\delta\|^2$ plus an estimation term $\sin\theta_{\max}\|\delta\|$ (Proposition 4); the second is the quantity our experiments measure, and rank-deficient estimates are shown to err only in the safe direction. □

Proposition 3 (Monotone refinement)**.** *If* $\mathcal{G} \subseteq \tilde{\mathcal{G}}$ *are nested admissible subspaces, then* $\mathrm{GID}_{\tilde{\mathcal{G}}}(z_t) \leq \mathrm{GID}_{\mathcal{G}}(z_t)$ *for every* z_t*.*

Sketch. $\mathcal{G} \subseteq \tilde{\mathcal{G}} \Rightarrow \tilde{\mathcal{G}^{\perp}} \subseteq \mathcal{G}^{\perp}$, and orthogonal projection is nonexpansive, so enlarging the admissible subspace can only lower GID. □

Proposition 3 makes the central tension quantitative: too small an r leaves commanded change in the drift term, too large an r absorbs genuine drift into the gauge subspace. Because GID is monotone in r, sweeping it traces one curve whose ends are the two failure modes, so r is an explicit reportable parameter, swept rather than tuned (Section D).

4 METHOD

Section 3 defines GID given $\mathcal{G}_{z_0}$; the method is how to *estimate* it from a real model, *compute* the drift, and *regularize* with it. Each step is deliberately simple, and we are explicit about each estimator's assumptions and about which parts are ours.

4.1 ESTIMATING THE GAUGE SUBSPACE

We lack $\partial_c \Psi$ in closed form, so we estimate its column space empirically with three estimators of increasing sophistication and decreasing independence; contrasting them is a primary ablation. A fourth question — at *which* state to estimate — turns out to matter more than the choice among them (Section 5.3).

(E-a) Intervention-difference basis. For each admissible primitive k ("pan camera right," "take branch B"), generate from the same reference z_0 a base and an intervened rollout differing only in c_k, both short enough that error accumulation is negligible. The difference $d_k = \phi(y^{c_k}) - \phi(y^{\text{base}})$ estimates the k-th control direction; the top-r left singular vectors of the $D \times K$ stack give an orthonormal U and $\mathsf{P}_{\mathcal{G}} = UU^{\top}$ (Golub & Van Loan, 2013; Halko et al., 2011). *We claim no novelty for this construction*—anchoring control effects to temporal feature differences of a frozen encoder is the operating principle of recent latent-action learning (Jiang et al., 2026), of which E-a is the finite-difference special case. New is the *use*: projecting displacement onto the complement of that span to measure drift.

Algorithm 1 Gauge-invariant drift of an interactive rollout

Input: rollout frames $y_{0:T}$, controls $c_{1:T}$, feature map ϕ, basis $U \in \mathbb{R}^{D\times r}$
$z_0 \leftarrow \phi(y_0)$; $\mathsf{P}_{\mathcal{G}^\perp} \leftarrow I - UU^\top$
for $t = 1, \dots, T$ **do**
  $z_t \leftarrow \phi(y_t)$; $\delta_t \leftarrow z_t - z_0$
  $\mathrm{GID}_t \leftarrow \|\mathsf{P}_{\mathcal{G}^\perp}\,\delta_t\|$ ▷ Eq. (1)
  $\eta_t \leftarrow \mathrm{GID}_t/(\|\delta_t\| + \epsilon)$ ▷ drift fraction
end for
return $\{\mathrm{GID}_t\}$, $\{\eta_t\}$, summary $\overline{\mathrm{GID}} = \frac{1}{T}\sum_t \mathrm{GID}_t$

(E-b) Pooled control PCA averages differences within each control type across references and takes the top-r components of the type-mean matrix (Jolliffe, 2002), trading locality for stability — appropriate only if $\mathcal{G}_{z_0}$ varies slowly with z_0, which Section 5.2 tests and rejects. **(E-c) Learned projector** optimizes U on the Stiefel manifold (Edelman et al., 1998; Absil et al., 2008) to maximize captured control variance minus captured drift variance; it is the only estimator needing drift labels, and it carries a circularity risk we address by held-out fitting. Implementation details, including a centering trap a synthetic self-check caught, are in Appendix B.

Choosing r. By Proposition 3, r trades false alarms against misses; we take the elbow of the singular spectrum, fix it on a held-out split, and report the full sweep (Section D) rather than the best value.

4.2 Computing gauge-invariant drift

Given U, GID is a projection and a norm (Eq. (1)); Algorithm 1 is $O(Dr)$ per frame, so it adds negligible cost to generation and can run online. Optionally ϕ-space may be whitened by the within-content covariance before projecting, making the distance Mahalanobis.

Note what Algorithm 1 does *not* need: it never reads $c_{1:T}$ except to report them — no per-frame annotation of which steps were intervened, no action-free probe construction, no gating rule. That is the operational content of putting the discount in the definition rather than the protocol.

4.3 Drift-only regularization

The same projector turns the metric into a training or sampling signal that penalizes drift *without* penalizing commanded change,

$$\mathcal{L}_{\mathrm{drift}} \;=\; \lambda \left\| \mathsf{P}_{\mathcal{G}_{z_0}^{\perp}} \big(\phi(y_t) - z_0\big) \right\|^2, \tag{2}$$

leaving gradients in $\mathcal{G}$ untouched, in contrast to $\mathcal{L}_{\mathrm{total}} = \lambda\|\phi(y_t) - z_0\|^2$, which resists all movement and so fights controllability. It differs in kind from trust-region steering (Wu et al., 2026), which gates supervision by the *magnitude* of its semantic shift (Schulman et al., 2015): Eq. (2) is indifferent to magnitude and asks only *in which directions* the state moved. Section 5.3 determines which projector it should carry, and Appendix G reports what the penalty does when applied at inference time, alongside the control arm that calibrates the comparison.

5 Experiments

5.1 A controlled autoregressive testbed

GID is defined for any autoregressive generative rollout with a control interface; interactive video is the application that motivates it. We validate on a small, fully controlled instance: a few-step latent diffusion model applied repeatedly to its own output — a self-conditioning loop with real error accumulation — steered by eight exact 2-D camera moves, and we repeat the central measurement on a video backbone (Section 5.2, Appendix I). Camera motion is the cleanest instance of what Assumption 1 posits: low-dimensional, continuously parameterized, exactly repeatable, enormous in pixel space and null in scene content. The testbed is fast enough to *randomize* the design rather than sample it, which is what makes the factorial test possible.

Table 1: Subspace estimation at the reference. A reference's own control differences separate cleanly; the same subspace at a *different* reference, and a random subspace of equal dimension, separate nothing — so the effect is structure rather than dimension counting.

Estimator	drift labels?	Ctrl var. ↑	Drift var. ↓	Separation ↑
E-a, per reference (ours)	no	0.808	0.072	0.736
same U, mismatched reference	no	—	—	-0.000
random subspace, same r	no	—	—	0.000
E-b, pooled across references	no	0.043	0.010	0.033
E-c, pooled + learned on drift labels	**yes**	0.007	0.211	-0.204

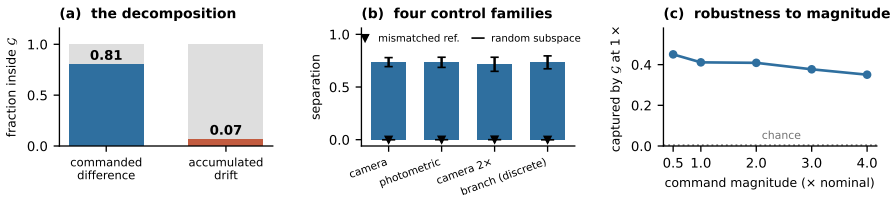


Figure 2: **(a)** a commanded difference lies almost entirely inside $\mathcal{G}$, accumulated drift almost entirely outside it. **(b)** this holds across four control families sharing little with one another, one of them discrete, while a mismatched-reference or random subspace separates nothing. **(c)** command magnitude costs it nothing: over an eightfold range capture matches the held-out-command rate at the nominal magnitude.

From 24 references we build $\mathcal{S}_{\text{ctrl}}$ (short paired rollouts differing only in one commanded move, with shared per-step seeds so the difference isolates the command and cancels common-mode drift), $\mathcal{S}_{\text{drift}}$ (long rollouts, control held fixed), and $\mathcal{S}_{\text{mix}}$ (500 rollouts in which **depth and number of commands are randomized independently**), so depth — which generates the error — serves as ground truth for accumulated degradation. The local control-difference spectrum has a clear elbow and we take $r = 4$ throughout.

Pre-registered decision rules. Fixed before running. **Gate 1:** short horizons must be approximately drift-free and long horizons must visibly drift. **Gate 2:** the estimated $\mathcal{G}$ must capture much control variance and little drift variance, beating a mismatched-reference and a random-subspace control. **Gate 3:** GID must respond to depth and not to command count. Gates 1 and 2 pass; Section 5.3 shows what Gate 3 additionally requires of the estimator, and Section 5.4 reports the measurement that establishes it.

5.2 The gauge subspace is real, and it is local (E1)

Measured on the paired differences of $\mathcal{S}_{\text{ctrl}}$ — where shared seeds cancel the generation step, so what is decomposed is the commanded effect itself rather than a rollout's accumulated change — a per-reference subspace captures most commanded change while absorbing almost no drift: separation 0.736 ± 0.043. Scoring reference i's data with reference j's subspace gives -0.000, a random subspace of the same r gives 0.000, and the paired difference against the mismatched control is significant at $p = 0.00005$. **Gate 2 passes**, and it passes using an estimator that needs no drift labels and no annotation — eight short paired rollouts per reference.

The same decomposition under four different control families. The construction refers only to the span of admissible change, never to what the controls mean, and the measurement bears that out (Figure 2b): separation is 0.734 for a photometric family carrying no geometric content, 0.716 at twice the magnitude, and 0.734 for *discrete branch selection*. That last one differs in kind from the others: it applies no transform to the frame, and instead changes the generator's own conditioning input, so the commanded effect is one the model produces natively rather than one we impose on its output. Mismatched and random controls sit at zero throughout. Command magnitude is also

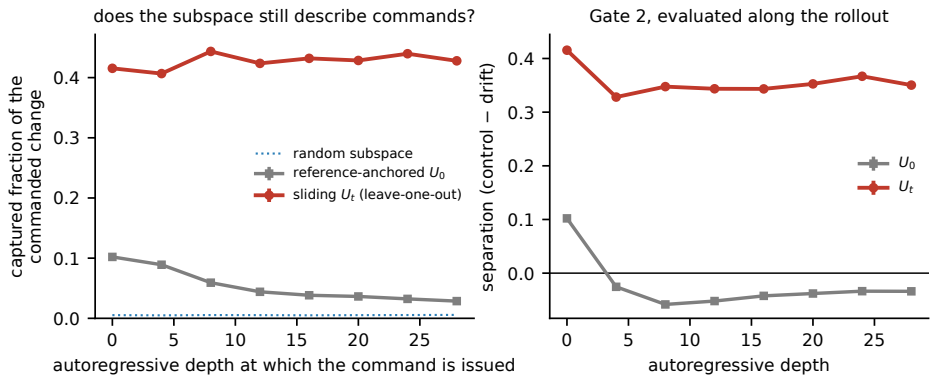


Figure 3: A projector re-estimated at the current state holds ≈ 0.35 separation at every depth, at $80\times$ chance and with no decay. One fixed at the reference describes commands only near the reference, which is what sets the maintenance schedule.

not an extra source of error: a subspace estimated at the nominal magnitude captures commanded differences from $0.5\times$ to $4\times$ that magnitude at 0.451 down to 0.351 (Figure 2c), which brackets the 0.411 it achieves on a held-out command at the nominal magnitude itself. It is the held-out-command rate that this matches, not a stronger invariance — Appendix F gives the control that sets that ceiling.

What the subspace is attached to. The structure is local, though how strongly depends on the encoder: in the self-supervised embedding the cosine between the *same* command at *different* references is $+0.003$ against $+0.432$ between *different* commands at the *same* reference, while in the language-aligned embedding the between-image figure is 0.216 (Table 4). The decomposition is robust to that choice — separation is 0.679, 0.702 and 0.736 in three independently trained encoders, highest in a masked autoencoder whose reconstruction objective exerts no invariance pressure at all, so the effect is not an artifact of crop augmentation. How *local* the subspace is does track the encoder: a pooled projector is useless under crop-augmented training (0.033, and -0.204 if allowed to *learn* on pooled data with drift labels, having fit the one component consistent across references, namely drift) and usable under reconstruction (0.307). The local estimator is the one that works in every encoder we tried (Table 4).

The same measurement on a video backbone. Because the construction is motivated by interactive video, we repeat E1 on a text-to-video diffusion model driven as an autoregressive video-to-video chain, so accumulation passes through a *spatiotemporal* denoiser. Gate 2 passes there too: separation 0.627 ± 0.111, mismatched and random controls again at zero. Captured control variance is unchanged (0.803 against 0.808), so the subspace describes commanded change just as well, and the whole gap is the drift term (0.175 against 0.072): more of a video rollout's accumulated error lies in directions the commands also excite. Locality survives ($+0.034$). Appendix I gives the setup and what a video-to-video chain does not reproduce of a streaming world model.

5.3 Maintaining the projector along a rollout (E1b)

Because the subspace belongs to a state, the practical question is how often it must be refreshed, and that is directly measurable. A subspace re-estimated at the current state captures 0.415 of a held-out commanded direction at depth 0 and 0.428 at depth 28 against a chance level of 0.005, while absorbing under 0.10 of accumulated drift; separation is 0.356 pooled over depths (95% bootstrap CI (Efron, 1979) $[0.351, 0.360]$, $n = 3840$), and every depth clears zero with a CI of width ≈ 0.025, so the 0.328–0.415 spread is between-depth structure rather than noise (Table 5). The property GID relies on is not fragile, simply attached to the current state: a projector fixed at the reference tracks commands near it (0.102 at depth 0) and then follows the state away (0.029 by depth 28). Refreshing costs nine one-step branches; Appendix H measures the refresh interval that buys and what the estimate is attached to.

Table 2: Factorial test on 500 mixed rollouts. Depth (which generates the error) and command count (the confound) are randomized independently; β are standardized regression coefficients. A measure that tracks drift should have large β_{depth} and $\beta_{\text{cmd}} \approx 0$.

Measure	$\beta_{\text{depth}} \uparrow$	$\beta_{\text{cmd}} \downarrow$	$\lvert\beta_{\text{cmd}}/\beta_{\text{depth}}\rvert \downarrow$
Total displacement	0.447	0.164	0.367
Frame-to-frame displacement	0.174	0.155	0.889
Cosine drop to reference	0.392	0.163	0.415
Subject consistency	0.186	0.150	0.808
CLIP distance to reference	0.276	0.211	0.763
CLIP displacement	0.288	0.216	0.749
GID, reference-anchored U_0	0.458	0.159	0.348
GID, pooled U	0.448	0.164	0.366
Oracle ($1-$SSIM, saturating)	0.025	-0.070	2.760

5.4 What each measure responds to (E2)

Randomizing the confound rather than observing it lets us ask directly what a score responds to, and the answer confirms this paper's premise: every displacement-derived measure loads substantially on command count (Table 2) — total displacement at $\beta_{\text{cmd}} = 0.164$ against $\beta_{\text{depth}} = 0.447$, CLIP displacement more confounded still (0.216 vs 0.288). These scores are partly reporting obedience; GID with a projector frozen at the reference sits with them (0.159), as Section 5.3 predicts of a stale projector.

Appendix E reports a second factorial run in which the projector is maintained along every rollout: the projection removes about a point of command inflation from a whole-rollout score (path form $+11\%$ to $+10\%$), which is what an r of 4 in $D = 768$ can reach, and Appendix H measures why.

6 Discussion

Locality is a design rule, not an obstacle. In the encoder we measure in, a gauge subspace is attached to an image rather than to a control interface: the same command at two references moves features along nearly orthogonal directions ($+0.003$), while different commands at one reference share a large component ($+0.432$). Assumption 1 was always stated locally, and here the locality is the substance — which converts an open modelling question into a costed one: a short paired rollout per primitive, no drift labels, no annotation. A third encoder sharpens it. The decomposition survives in a reconstruction-trained representation with no invariance pressure (0.736, highest of three), and there a pooled projector becomes usable (0.307; Table 4), so locality is a property of the representation and one the practitioner can trade.

Scope and limitations. The main substrate is an autoregressive image rollout with 2-D camera commands, chosen because the transforms are exactly parameterized and fast enough to randomize the design rather than sample it; Appendix I repeats the central measurement on a video-to-video chain and Appendix F on three further control families and an eightfold magnitude range. Of those four families only branch selection is consumed by the generator natively rather than applied to its output, so a model trained on an action interface is the natural next test, alongside streaming chunk-causal accumulation, 3-D commanded change, and interfaces whose commands overlap with degradation — a case Proposition 3 anticipates and Gate 2 tests for. The decomposition is established on *isolated* commanded change: Appendix E records what it contributes to a whole-rollout score, Appendix H locates why that contribution is the size it is, and Appendix G evaluates Eq. (2) as inference-time guidance, where the training-time form the equation is written for remains untested. We report agreement against randomized depth rather than human judgments of breakdown.

In sum. Drift scores charge a model for change its user asked for; where existing remedies work around the measure, we put the discount inside it. The result is basis-independent, first-order blind to commanded change, monotone in r, and needs no protocol.

Reproducibility Statement

The measure is deterministic and cheap: given a frozen feature map ϕ and a basis U, Algorithm 1 is a projection and a norm, so every number in this paper is reproducible from stored feature trajectories without re-running a generator. Concretely: Definition 2 and Eq. (1) fix the metric; Section 4.1 and Appendix B fix the three reference-anchored estimators and the sliding estimator, including the leave-one-out protocol under which every sliding number is reported; Appendix C states the testbed, the eight control primitives with their magnitudes, and the Gate 1 noise-floor measurement; Appendices D and E give the r-sweep, the feature-map ablation, and the maintained-projector factorial run with its two controls; Appendix F states the two additional control families and the magnitudes defining them, and Appendix G states the guidance procedure, the strength sweep and the displacement counter by which its control arm was calibrated; and Appendix I states the video backbone, the video-to-video chain, the clip resolution and length, and the pooling by which a clip is mapped to a feature. The pre-registered decision rules are stated in Section 5.1 *before* any result, so the difference between a confirmation and a negative result is fixed independently of the outcome; the negative outcomes they produced (the pooled and learned estimators, Table 1) are reported as such. All runs are single-GPU inference on publicly available generators and frozen encoders on one 24 GB device; every experiment in the paper together is under 3.5 GPU-hours, and the stored artifacts are a few megabytes because we store features rather than frames. We release the rollout construction, the estimators, the sliding-subspace probe, the diagnostics and the analysis, together with a synthetic self-check that verifies the implementation against Propositions 1–3 numerically on planted data before any measured trajectory is touched. Every table body and every numerical macro quoted in the text is emitted directly by the analysis code, so a number cannot drift from the run that produced it.

Ethics Statement

This work proposes an evaluation measure and an auxiliary training penalty for generative rollouts; it introduces no new generative capability, and its intended use — distinguishing degradation from commanded change — is diagnostic. Two considerations deserve statement. First, a metric that "forgives" change lying in an estimated control subspace can be gamed: a system could be tuned so that its failures align with the directions the metric excuses, which is why we report the drift fraction and the whole-rollout factorial results alongside GID rather than GID alone, and why we treat a metric-only improvement as insufficient evidence. Second, **this study involves no human subjects and no human ratings**: all reported quantities are computed from model-generated frames and frozen encoders, and the ground truth for accumulated error is the randomized rollout depth rather than a human judgment. Identity-style encoders are applied to synthetic frames for measurement only, never for recognition of real individuals. All backbones, encoders and evaluation assets are publicly available and used in accordance with their licenses.

Use of Large Language Models

Large language models were used as writing and engineering assistance: drafting and copy-editing prose, refactoring LaTeX and analysis scaffolding, and literature triage — surfacing candidate related work, which the authors then verified against primary sources. Entries that could not be confirmed against a paper page, project page or proceedings entry were removed rather than retained on the strength of a model's summary; where a release is attributable to a lab but an individual author list could not be confirmed, we cite the releasing lab rather than guess names. The research contribution — the gauge/complement decomposition, the definition of GID, the propositions, the experimental design, and the pre-registered decision rules — is the authors' own, and the authors take full responsibility for all content, including any errors. No results, data, or citations were generated by a language model.

REFERENCES

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

AMAP CV Lab. Worldroambench: An open-world benchmark for long-horizon stability of interactive world models. *arXiv preprint arXiv:2606.31672*, 2026.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

Åke Björck and Gene H. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, 1973.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI technical report, 2024.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Boyuan Chen, Diego Martí Monsó, 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 (NeurIPS)*, 2024.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Taco S. Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4th edition, 2013.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

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

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025a.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Cyrus Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025b.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

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

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 (ICLR)*, 2022.

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

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yuxin Jiang, Yuchao Gu, Ivor W. Tsang, and Mike Zheng Shou. Olaf-world: Orienting latent actions for video world modeling. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

Yann LeCun. A path towards autonomous machine intelligence. Open Review position paper, 2022.

Kunhao Liu, Wenbo Hu, Jiale Xu, and Ying Shan. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations (ICLR)*, 2026.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Maurice Weiler and Gabriele Cesa. General e(2)-equivariant steerable cnns. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yuheng Wu, Xiangbo Gao, Tianhao Chen, Xinghao Chen, Qing Yin, Zhengzhong Tu, and Dongman Lee. Delta forcing: Trust region steering for interactive autoregressive video generation. *arXiv preprint arXiv:2605.14382*, 2026.

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

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. *arXiv preprint arXiv:2607.22430*, 2026a.

Zeyu Zhang, Shuning Chang, Yuanyu He, Yizeng Han, Jiasheng Tang, Fan Wang, and Bohan Zhuang. Blockvid: Block diffusion for high-quality and consistent minute-long video generation. *arXiv preprint arXiv:2511.22973*, 2026b.

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

A Proof Details

Every step is carried out here. Where a bound needs regularity beyond Assumption 1, that assumption is stated at the point of use and the resulting constant is named, so it is visible what each bound is buying.

A.1 Proposition 1 (basis invariance)

Let $U, U' \in \mathbb{R}^{D\times r}$ have the same column space $\mathcal{G}_{z_0}$. For full-column-rank U the orthogonal projector onto $\mathrm{col}(U)$ is $UU^+ = U(U^\top U)^{-1}U^\top$, and the orthogonal projector onto a given subspace is unique: if $\mathsf{P}_1, \mathsf{P}_2$ are both symmetric, idempotent and have range $\mathcal{G}_{z_0}$, then for any x we have $\mathsf{P}_1 x - \mathsf{P}_2 x \in \mathcal{G}_{z_0}$ while $\langle \mathsf{P}_1 x - \mathsf{P}_2 x, v\rangle = 0$ for all $v \in \mathcal{G}_{z_0}$, whence $\mathsf{P}_1 = \mathsf{P}_2$ (Golub & Van Loan, 2013). So $\mathsf{P}_{\mathcal{G}}$, $\mathsf{P}_{\mathcal{G}^\perp} = I - \mathsf{P}_{\mathcal{G}}$ and Eq. (1) are unchanged under any full-rank change of spanning set, and in particular under the orthogonal indeterminacy inherent to identifiable latent models (Khemakhem et al., 2020; Zhang et al., 2026a). We can now state the estimation term the paper's measurements make load-bearing. Let $\mathcal{G}, \hat{\mathcal{G}}$ be r-dimensional subspaces with orthogonal projectors $\mathsf{P}, \hat{\mathsf{P}}$ and principal angles $\Theta(\mathcal{G}, \hat{\mathcal{G}})$.

Proposition 4 (Estimation error of GID)**.** *For every change δ,*

$$\big|\, \|\mathsf{P}_{\mathcal{G}^\perp}\delta\| \;-\; \|\hat{\mathsf{P}}_{\mathcal{G}^\perp}\delta\| \,\big| \;\le\; \|\mathsf{P}_{\mathcal{G}^\perp} - \hat{\mathsf{P}}_{\mathcal{G}^\perp}\|_2\, \|\delta\| \;=\; \sin\theta_{\max}(\mathcal{G}, \hat{\mathcal{G}})\, \|\delta\|,$$

where $\theta_{\max}$ is the largest principal angle. If $\hat{\mathcal{G}}$ is the top-r invariant subspace of a matrix perturbed by E from one whose r-th and $(r{+}1)$-th singular values are separated by a gap $\gamma > 0$, then $\sin\theta_{\max} \le \|E\|_2/\gamma$.

Proof. The reverse triangle inequality gives $\big|\|\mathsf{P}_{\mathcal{G}^\perp}\delta\| - \|\hat{\mathsf{P}}_{\mathcal{G}^\perp}\delta\|\big| \le \|(\mathsf{P}_{\mathcal{G}^\perp} - \hat{\mathsf{P}}_{\mathcal{G}^\perp})\delta\| \le \|\mathsf{P}_{\mathcal{G}^\perp} - \hat{\mathsf{P}}_{\mathcal{G}^\perp}\|_2\|\delta\|$. Since $\mathsf{P}_{\mathcal{G}^\perp} - \hat{\mathsf{P}}_{\mathcal{G}^\perp} = (I - \mathsf{P}) - (I - \hat{\mathsf{P}}) = \hat{\mathsf{P}} - \mathsf{P}$ and, for two subspaces of equal dimension, $\|\mathsf{P} - \hat{\mathsf{P}}\|_2 = \sin\theta_{\max}$ (Golub & Van Loan, 2013; Björck & Golub, 1973), the equality follows. The final bound is the Davis–Kahan $\sin\Theta$ theorem in its singular-subspace form. □

This is the quantity our experiments actually measure. Section 5.2 reports principal angles between subspaces estimated at different references, and Section 5.3 reports how the captured fraction decays with depth; by Proposition 4 those are exactly the terms that control how far a computed GID can sit from the GID of the true subspace. It also explains the locality result without appeal to anything beyond linear algebra: a reference-anchored $\hat{\mathcal{G}}$ has a large $\theta_{\max}$ against the subspace at the current state, so the bound is loose there and tight when the projector is maintained.

The rank-deficient case, and why it is the safe one. Proposition 4's equality $\|\mathsf{P} - \hat{\mathsf{P}}\|_2 = \sin\theta_{\max}$ requires the two subspaces to have equal dimension. When the sampled intervention differences do not excite all r directions — $\hat{U}$ has numerical rank $\hat{r} < r$ — we take the projector through a thin SVD thresholded at τ, keeping $\hat{U}_\tau$ of rank $\hat{r}$, and the estimated subspace is then a *proper* subspace of the true one, $\hat{\mathcal{G}} \subset \mathcal{G}$. Write $\mathcal{M} = \mathcal{G} \ominus \hat{\mathcal{G}}$ for the $(r - \hat{r})$-dimensional part that was missed. The two complements differ by exactly that piece, $\hat{\mathsf{P}}_{\mathcal{G}^\perp} = \mathsf{P}_{\mathcal{G}^\perp} + \mathsf{P}_{\mathcal{M}}$, and the summands project onto orthogonal subspaces, so for any δ

$$\|\hat{\mathsf{P}}_{\mathcal{G}^\perp}\delta\|^2 = \|\mathsf{P}_{\mathcal{G}^\perp}\delta\|^2 + \|\mathsf{P}_{\mathcal{M}}\delta\|^2, \qquad 0 \;\le\; \|\hat{\mathsf{P}}_{\mathcal{G}^\perp}\delta\| - \|\mathsf{P}_{\mathcal{G}^\perp}\delta\| \;\le\; \|\mathsf{P}_{\mathcal{M}}\delta\|.$$

Corollary 5 (Rank deficiency is one-sided)**.** *If the estimated gauge subspace is a proper subspace of the true one, GID is* over-*estimated, never under-estimated, and the error is at most the energy of δ along the unrecovered directions.*

This is Proposition 3 seen from the estimation side, and it is the direction one wants: a projector that misses part of the control subspace forgives less commanded change and so may raise a false alarm, but it cannot launder drift into the admissible part. Under-estimation requires $\hat{\mathcal{G}} \not\subseteq \mathcal{G}$ — an estimated direction that is not admissible at all — which is what the mismatched-reference and random-subspace controls of Table 1 are there to detect.

A.2 Proposition 2 (blindness to admissible change)

Write $\delta = z_t - z_0 = \delta_\parallel + \rho$ with $\delta_\parallel \in \mathcal{G}_{z_0}$ and $\|\rho\| = o(\|\delta\|)$ the second-order remainder of Assumption 1. Then $\mathsf{P}_{\mathcal{G}^\perp}\delta = \mathsf{P}_{\mathcal{G}^\perp}\rho$ and, by nonexpansiveness, $\|\mathsf{P}_{\mathcal{G}^\perp}\delta\| \le \|\rho\| = o(\|\delta\|)$; the converse is immediate.

Making the $o(\cdot)$ uniform. That bound is asymptotic; the two constants that make it quantitative are the curvature of the orbit and the estimation error, and both follow from smoothness of the same map Assumption 1 already posits. Suppose $\Psi(\cdot, z_0)$ is C^2 on a ball $\|c\| \le R$ with $\sup \|\partial_c^2 \Psi\| \le \kappa$, and that the differential $A = \partial_c \Psi|_{c=0}$ has smallest nonzero singular value $\sigma > 0$. Taylor expansion gives $\delta = Ac + \rho$ with $\|\rho\| \le \frac{\kappa}{2}\|c\|^2$, so for $\|c\| \le \min\{R, \sigma/\kappa\}$,

$$\|\delta\| \;\ge\; \sigma\|c\| - \tfrac{\kappa}{2}\|c\|^2 \;\ge\; \tfrac{\sigma}{2}\|c\| \;\Longrightarrow\; \|c\| \le \tfrac{2}{\sigma}\|\delta\| \;\Longrightarrow\; \|\rho\| \;\le\; \tfrac{2\kappa}{\sigma^2}\,\|\delta\|^2.$$

Adding the estimation term of Proposition 4 gives a uniform statement on that ball:

$$\widehat{\mathrm{GID}}(z_t \mid c_{1:t}) \;\le\; \underbrace{\tfrac{2\kappa}{\sigma^2}\,\|\delta\|^2}_{\text{orbit curvature}} \;+\; \underbrace{\sin\theta_{\max}(\mathcal{G}, \hat{\mathcal{G}})\,\|\delta\|}_{\text{subspace estimation}}. \tag{3}$$

The two terms fail in different regimes and the experiments separate them. The linear term dominates at small excursions and is what the locality result governs: $\sin\theta_{\max}$ is large for a reference-anchored projector and small for a maintained one (Section 5.3). The quadratic term is the one a larger command exercises, and Appendix F reports the doubled-magnitude run — separation holds, and the drift admitted into $\mathcal{G}$ rises by a fifth. Eq. (3) also says where the construction would stop working: an interface whose orbit is sharply curved relative to its own scale, meaning $\kappa\|\delta\|/\sigma^2$ of order one, has no useful first-order gauge subspace at any estimation quality.

A.3 Proposition 3 (monotone refinement)

For nested $\mathcal{G} \subseteq \tilde{\mathcal{G}}$ we have $\tilde{\mathcal{G}^\perp} \subseteq \mathcal{G}^\perp$, hence $\mathsf{P}_{\tilde{\mathcal{G}^\perp}} = \mathsf{P}_{\tilde{\mathcal{G}^\perp}}\mathsf{P}_{\mathcal{G}^\perp}$ and $\|\mathsf{P}_{\tilde{\mathcal{G}^\perp}}\delta\| \le \|\mathsf{P}_{\mathcal{G}^\perp}\delta\|$ for every δ. Monotonicity holds only along a *nested* family; an r-sweep by SVD truncation is nested by construction, a sweep across different estimators is not.

B Estimator Details

E-a (intervention-difference basis), per reference. Differences are ℓ_2-normalized, one per (reference, primitive) pair, and the $D \times K$ matrix is factorized by a randomized thin SVD (Halko et al., 2011) truncated at r. This is the finite-difference case of anchoring control effects to temporal feature differences of a frozen encoder (Jiang et al., 2026); the contribution is the use, not the construction. Base and intervened rollouts share per-step seeds, which is what cancels common-mode drift from the difference.

E-b (pooled), E-c (pooled and learned). E-b averages within control type across references and takes components of the type-mean matrix (Jolliffe, 2002). One trap our synthetic validation caught: mean-*centering* within type rather than averaging removes exactly the signal, since under Assumption 1 the type mean *is* the control direction; the centered variant recovers nothing on planted data and makes a label permutation score *higher* than the truth. E-c optimizes U on the Stiefel manifold (Edelman et al., 1998; Absil et al., 2008) to maximize captured control variance minus captured drift variance, fitted on references disjoint from evaluation. Both are reported as negative results in Table 1.

The sliding estimator. At depth d we branch nine one-step continuations off the current frame — one base and eight commanded, all sharing a seed — and estimate U_d from the eight paired differences. Every number reported for U_d is leave-one-out: the tested direction is excluded from the fit, so the ≈ 0.43 capture is out-of-sample.

C TESTBED DETAILS

Substrate. A distilled few-step latent text-to-image diffusion model (Rombach et al., 2022; Song et al., 2021) in image-to-image mode at 512^2, applied to its own output; two denoise steps per autoregressive step. The feature maps are a frozen self-supervised encoder (Caron et al., 2021; Oquab et al., 2024) and a frozen language-aligned encoder (Radford et al., 2021). All runs are single-GPU inference on one 24 GB card; every experiment in the paper together is under 3.5 GPU-hours and the stored artifacts are a few megabytes because we store features, not frames.

Baseline measures in Table 2. Every measure is computed from the same stored feature trajectories, so the factorial comparison is between definitions rather than between implementations. Write z_t for the self-supervised feature of frame t, $\hat{z}_t$ for its unit normalization, and c_t for the language-aligned feature.

Measure	Definition
Total displacement	$\|z_T - z_0\|$
Frame-to-frame displacement	$\frac{1}{T}\sum_t \|z_t - z_{t-1}\|$
Cosine drop to reference	$1 - \hat{z}_T^\top \hat{z}_0$
Subject consistency (running)	$\frac{1}{T}\sum_t \big(1 - \hat{z}_t^\top \hat{z}_{t-1}\big)$
CLIP distance to reference	$1 - \hat{c}_T^\top \hat{c}_0$
CLIP displacement	$\|c_T - c_0\|$
Oracle	$1 - \mathrm{SSIM}\big(y_T, A_T(y_0)\big)$

Two of these are *in the spirit of* the consistency family used by video benchmarks (Huang et al., 2024) rather than those benchmarks' implementations: we compute them on whole-frame features of our own rollouts, with no subject/background segmentation, so they should be read as representative displacement-derived scores and not as reported VBench numbers. We name them for what they compute for the same reason — an earlier draft called the third row "identity drop," which would have suggested a face-recognition embedding it does not use.

Control primitives. Eight exact 2-D similarity transforms: pan by $\pm 4.5\%$ of the frame in each of four directions, zoom by $\pm 5\%$, and roll by $\pm 3^\circ$. Magnitudes are small so the valid region stays inside the evaluation crop under composition, which also means this study does not probe the large-excursion regime where the linearization would be expected to fail.

Gate 1 measurement. Fixed-control rollouts at increasing depth across four seeds give displacement 15.1 at depth 1 against an across-seed spread of 2.4 at depth 16, rising to 32.4 by depth 16. Drift is therefore detectable almost immediately, so $\mathcal{S}_{\mathrm{ctrl}}$ at horizon 2 is not perfectly drift-free and the paired-seed construction is doing real work rather than being a convenience.

D ADDITIONAL RESULTS

Ablations. The result does not hinge on one encoder: local separation is essentially the same in two independently trained embeddings (Table 4). The r-sweep required by Proposition 3 shows the predicted trade-off directly, with drift absorbed by $\mathcal{G}$ climbing from 0.016 at $r = 1$ to 0.108 at $r = 8$, so r buys forgiveness at a measurable price. A geometric oracle we built as a second referent — $1 - \mathrm{SSIM}$ against the exactly accumulated warp — saturates after a single autoregressive step, reaching 0.463 at depth 1 and only 0.853 by the last depth we measured, with a standardized loading on depth of just 0.025. It is reported here rather than relied on, and randomized depth serves that role in the body.

An encoder with no invariance pressure. The first two encoders are both trained with objectives that reward invariance to crops and rescalings — which is exactly what a camera command performs — so a reader is entitled to ask whether the decomposition, and the locality in particular, are facts about the world or about augmentation. The third row is the control for that question: a masked autoencoder reconstructs pixels rather than matching augmented views, so nothing in its training tells it to ignore a pan or a zoom. All three encoders are read from the *same* frames in the same run, so the

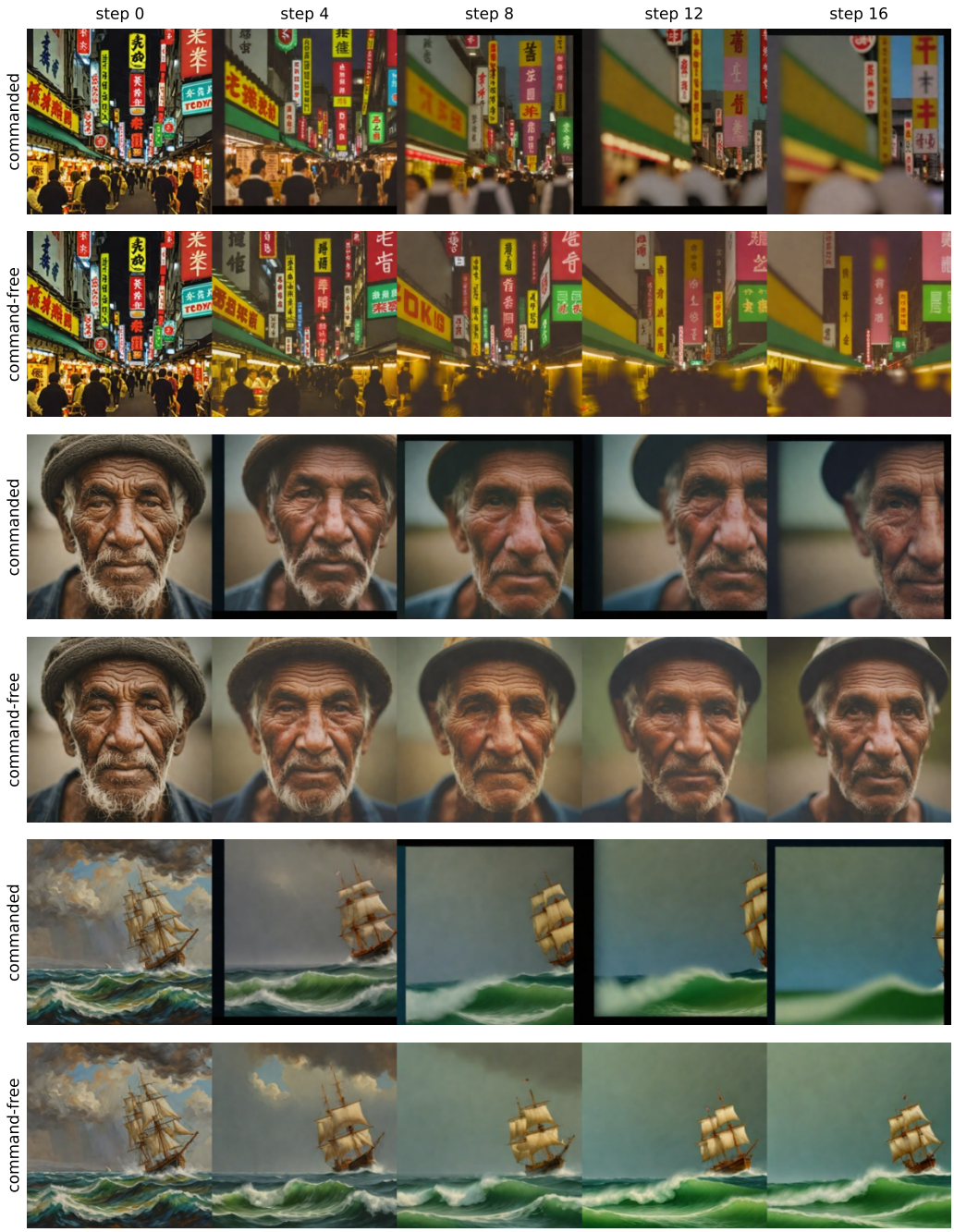


Figure 4: Three references, each rolled out twice from the same frame with the same seeds: once under a commanded camera path and once with no command at all. The commanded rows move because the user asked and because the loop accumulates error; the command-free rows isolate the second. These are the trajectories the paired construction of $\mathcal{S}_{\text{ctrl}}$ differences away.

comparison is between representations rather than between rollouts; absolute values differ slightly from Table 1 because this run uses a shorter drift chain.

Two things follow, and they point in different directions. First, the decomposition is *not* an augmentation artifact: separation is 0.736 in the reconstruction-trained encoder, the highest of the three,

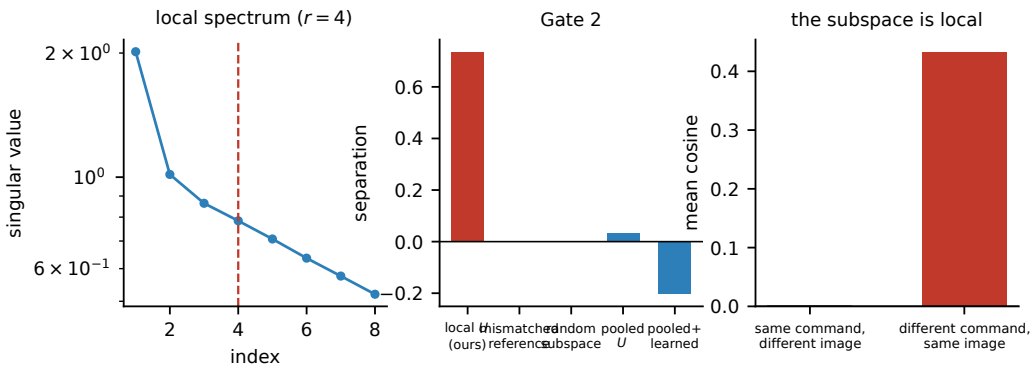


Figure 5: **Left:** the per-reference control-difference spectrum has an elbow, so commanded change is low-dimensional locally. **Middle:** separation for the local subspace against the mismatched-reference, random and pooled controls. **Right:** the mechanism — the same command at different images points in unrelated directions, while different commands at one image share a large component.

Table 3: Subspace dimension sweep at the reference (E4), averaged over references. Separation rises monotonically because $K = 8$ primitives cap the attainable capture; the trade-off Proposition 3 predicts is visible in the middle row, where the drift absorbed by $\mathcal{G}$ grows with r.

r	1	2	3	4	5	6	7	8
control captured	0.515	0.644	0.736	0.808	0.869	0.921	0.965	1.000
drift absorbed	0.016	0.039	0.058	0.072	0.084	0.092	0.099	0.108
separation	0.499	0.605	0.678	0.736	0.785	0.829	0.866	0.892

Table 4: The same analysis in two embeddings. Both support a local subspace; they differ sharply in how much commanded structure transfers between images, which is consistent with DINOv2's crop-augmented training making it partly invariant to the pan and zoom we command.

Feature map ϕ	local separation ↑	pooled separation	cos(same cmd, diff. image)
DINOv2 (crop-augmented)	0.679 ± 0.062	0.032	$+0.000$
CLIP ViT-B/32 (language-aligned)	0.702 ± 0.073	0.225	$+0.204$
MAE ViT-B/16 (reconstruction)	0.736 ± 0.062	0.307	$+0.257$

with the random-subspace control at zero. Second, *how local* the subspace is does depend on the encoder, and strongly: the cosine between the same command at different references is $+0.000$ under crop-augmented training, $+0.204$ under language alignment and $+0.257$ under reconstruction, and a pooled projector goes from useless (0.032) to usable (0.307) along the same ordering. Locality is therefore a property of the representation one measures in, not of the control interface, and it is a property one can trade: an encoder that keeps the information a crop-invariant one discards admits a subspace that transfers across references, at the price of a representation that is itself less invariant to things one may not want to forgive. We report the local estimator throughout the body because it is the one that works in every encoder we tried.

Reproducibility. Given stored feature trajectories, every number above is a projection and a norm: no generator is re-run. We release the rollout construction, the four estimators, the sliding-subspace probe, the diagnostics and the analysis, together with a synthetic self-check that verifies the implementation against Propositions 1–3 numerically before any real data is touched.

Table 5: Validity along the rollout. At each depth d we branch nine one-step continuations off the current frame (one base, eight commanded, shared seeds) to obtain *paired* control differences at that state, and ask how much of each lies in a subspace fixed at the reference (U_0) versus one re-estimated at the current state (U_d, leave-one-out, so the tested direction is never in the fit). 95% bootstrap CIs over the 480 held-out directions at each depth.

Depth d	0	4	8	12	16	20	24	28
separation, sliding U_d	0.415	0.328	0.348	0.344	0.343	0.353	0.367	0.350
95% bootstrap CI	[0.40,0.43]	[0.32,0.34]	[0.33,0.36]	[0.33,0.36]	[0.33,0.36]	[0.34,0.37]	[0.35,0.38]	[0.34,0.36]
separation, reference U_0	0.102	-0.025	-0.058	-0.052	-0.042	-0.038	-0.034	-0.034

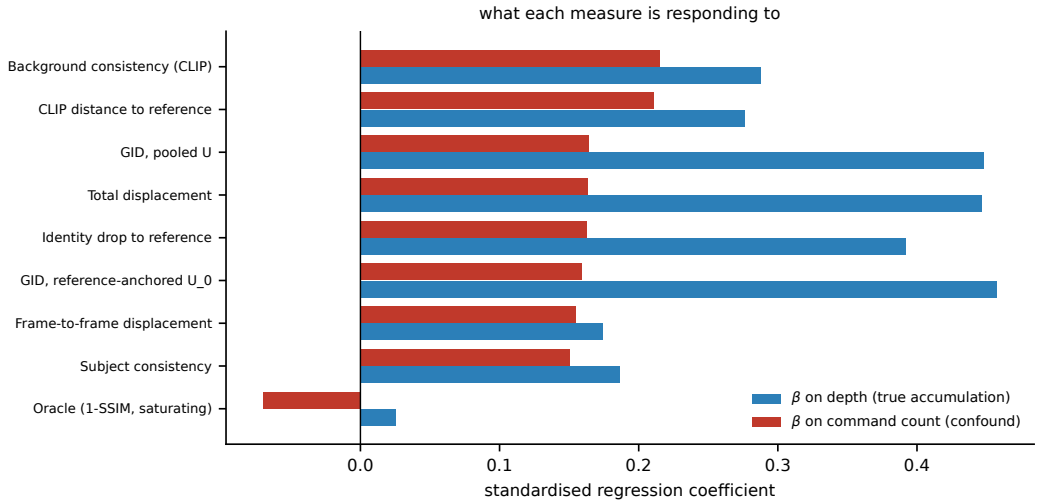


Figure 6: Factorial loadings for every measure (the numbers are Table 2). Depth and command count were randomized independently; a measure that tracks drift should load on the former only.

E Carrying the Decomposition to a Whole-Rollout Score

Section 5.2 establishes the decomposition on *isolated* commanded change: paired differences in which the ordinary generation step is cancelled by shared seeds. Section 5.3 shows a projector re-estimated at the current state retains that property at every depth. This appendix reports the composition of the two — the factorial test itself, run with the projector maintained throughout — so that the size of the resulting correction to a whole-rollout score is on record rather than inferred.

Table 6: Carrying the decomposition to a whole-rollout score (120 rollouts, depth and command count randomized independently, projector maintained throughout). "Command inflation" is how much a measure grows, per step, between zero-command and ≥ 5-command rollouts, so step count is divided out.

Measure	$\beta_{\text{depth}} \uparrow$	$\beta_{\text{cmd}} \downarrow$	command inflation $\downarrow$
Total displacement $\|z_T - z_0\|$	0.336	0.176	+33.1%
+ gauge projection, maintained U_T	0.359	0.183	+32.7%
Path length $\sum_t \|z_t - z_{t-1}\|$	0.863	0.111	+11.2%
+ gauge projection, maintained U_t	0.865	0.104	+10.2%

Result. Commands inflate a displacement score by +33% per step and a projected one by +33%. Accounting per step rather than by chord takes the inflation to +11%, and adding the projection on top gives +10%. The correction the projection contributes is therefore real but modest at this level,

and its size has a direct explanation: the projection removes 7% of the measured magnitude, since $r = 4$ of D directions carry the commanded component, and that component — on which Table 1 shows the decomposition to be sharp — is a small share of what a long rollout accumulates.

Two controls that the reading depends on. First, we also ran the path form with the *stale* projector. It gives $\beta_{\text{cmd}} = 0.111$, identical to path length with no projection at all, which is what separates the contribution of the path form from the contribution of maintaining the projector; without this control the two are confounded and the maintained projector would appear to be doing work that the path accounting is doing. Second, path length is 74% explained by rollout depth alone, i.e. it is close to a step counter, so a low β_{cmd} is easier to achieve in the path family than in the chord family. We therefore report the per-step inflation column beside the standardized coefficients rather than relying on either alone.

What this locates. The decomposition is established, cheap and label-free at the level of commanded change; turning it into a correspondingly large correction of an accumulated score is the open problem, and these measurements say where to push — a larger or better-conditioned $\mathcal{G}$, since $r = 4$ touches only 7% of the magnitude, or a score built on the per-step decomposition rather than on the chord.

F Coverage Across Control Families

The main experiments use one family of admissible interventions: eight small 2-D camera moves. Two questions follow. Does the decomposition depend on that family being geometric, and does it survive a larger excursion, where Assumption 1's first-order picture is under more strain? We rebuilt $\mathcal{S}_{\text{ctrl}}$ for two further families and reused the existing drift corpus, which is control-family independent by construction — it is a fixed-control chain, so nothing in it depends on which interventions the other corpus exercises.

Photometric replaces every geometric move with an appearance move: brightness, contrast, saturation and gamma, each in both directions, at magnitudes typical of a photo-editing control. It shares no geometric content with the camera family, and it is also *smaller*: mean commanded displacement in ϕ-space is 7.7 against 14.8 for the camera family. **Camera-large** repeats the original eight moves at twice the parameter magnitude, which raises mean displacement to 17.0 — the encoder is compressive, so doubling the transform does not double the feature excursion. **Branch selection** is discrete and applies no transform to the frame at all: at the branch point the rollout is continued under a different scene condition, expressed as a discrete edit to the conditioning text, with the starting frame and every seed held fixed so that the only difference is which branch was taken. It is the family the introduction names alongside camera motion, and it is the one that differs from the others in *kind* rather than in content: the camera and photometric families are transforms we apply to the generator's output, whereas a branch is a change to the generator's own conditioning input, executed by the model. That the same estimator recovers a gauge subspace in both regimes is the closest evidence we have that the construction does not depend on the intervention being externally applied, which is the form it would take in a system whose control interface the model consumes natively.

Table 7: The decomposition across control families, $r = 4$, $n = 24$ references each. Mismatched and random controls are at zero throughout, so each row is a self-contained repetition of Gate 2 rather than a variation on a shared fit.

Control family	Separation ↑	Mismatched	Random	Ctrl var. ↑	Drift var. ↓
Camera, small (main)	0.736 ± 0.043	-0.000	-0.001	0.808	0.072
Photometric	0.734 ± 0.048	$+0.000$	-0.000	0.813	0.079
Camera, $2\times$ magnitude	0.716 ± 0.067	$+0.001$	-0.000	0.804	0.088
Branch selection (discrete)	0.734 ± 0.061	-0.001	-0.001	0.808	0.074

Result. Separation is essentially unchanged on the photometric family (0.734 ± 0.048 against 0.736 ± 0.043), and the estimator is not told which family it is looking at. It reaches that separation at roughly half the commanded displacement, so separation is not a function of how large the excursion

is. The reading is that the construction responds to *a family of admissible interventions* rather than to camera geometry in particular, which is the property the formulation claims: Definition 1 refers only to the span of $\partial_c \Psi$, never to what the controls mean.

Command magnitude does not degrade the estimate. A gauge subspace should describe *which way* a command moves the state, not how far, and that is directly testable: estimate the subspace at the nominal magnitude and ask how much of a commanded difference at a different magnitude it captures. Over an eightfold range the answer is essentially flat.

Table 8: Robustness to command magnitude — not invariance; see the control below. The subspace is estimated at $1\times$ (leave-one-out) and scores commanded differences generated at other magnitudes; 16 references, $r = 4$.

Command magnitude	$0.5\times$	$1\times$	$2\times$	$3\times$	$4\times$
Mean displacement $\|\delta\|$	14.7	15.5	17.3	18.6	20.5
Captured by $\mathcal{G}$ at $1\times$	0.451	0.411	0.409	0.377	0.351
Chance	0.004	0.004	0.004	0.004	0.005

Capture stays between 0.351 and 0.451 across the range, bracketing the 0.411 the same estimator reaches on a held-out command at the nominal magnitude, so changing the magnitude by a factor of eight costs no more than asking about a different command at the same magnitude. That is the right reading, and it is weaker than it may look, so we give the control that bounds it.

The control, and what it rules out. One could read Table 8 as saying the subspace is *invariant* to magnitude. It does not say that. The displacement row already shows the response saturating — an eightfold command produces only a 40% larger excursion — and the direction cosines make the limit explicit: for one reference and one primitive, $\cos(\delta_{0.5\times}, \delta_{4\times}) = 0.382$, while two *different* primitives at the same magnitude and reference sit at 0.398. A command at $0.5\times$ and the same command at $4\times$ are therefore no better aligned than two different commands are, and the flat capture curve reflects the family-common component that an r-dimensional subspace picks up rather than a magnitude-independent direction per command. What Table 8 does establish is the operational point — magnitude does not degrade the estimate relative to its own held-out baseline — and it is consistent with Proposition 2 needing a statement about direction rather than amplitude, with the saturating magnitude channel keeping the second-order term of Eq. (3) from growing as an unbounded excursion would let it.

The large-excursion row carries a predicted signal. Doubling the command magnitude leaves separation high (0.716) but raises the drift absorbed into $\mathcal{G}$ by a fifth, from 0.072 to 0.088. This is the behaviour Proposition 2 anticipates rather than an anomaly: blindness holds up to the second-order remainder of Assumption 1, which grows with $\|\delta\|$, so a larger excursion should let slightly more genuine drift into the admissible subspace. The measured cost is mild at $2\times$, and the trend gives the direction in which the first-order model would eventually be the binding constraint.

G Using the Projector as a Training-Time Signal

Eq. (2) turns the same projector into a penalty that resists drift while leaving commanded directions free. We evaluate it in the cheapest available form — inference-time guidance on the rendered frame, descending $\nabla_y \|\mathsf{P}_{\mathcal{G}^\perp}(\phi(y) - z_0)\|^2$ for a fixed number of steps between rollout iterations — over four arms on 10 references at depth 20 with 5 commanded moves each.

None is the unguided rollout. $\mathcal{L}_{\text{total}}$ penalizes all displacement and is the control that shows what the lever can do at all. **Drift, stale** projects with a fixed reference-anchored basis and **drift, sliding** re-estimates the basis at the current state; the pair separates "a projected penalty helps" from "maintaining the projector helps," which Section 5.3 shows are different claims.

Calibrating the lever first. A first sweep returned four identical arms *including* $\mathcal{L}_{\text{total}}$ — the arm that resists all movement and should therefore dominate. A pixel-displacement counter located the cause: the guidance gradient had been normalized by the ℓ_2 norm of the whole $3 \times 512 \times 512$ tensor,

which divides the step by $\sqrt{3HW}$ and moves each pixel by about 2×10^{-5} per iteration. With max-norm normalization the control arm responds. Across a three-point strength sweep its mean per-pixel movement rises from $0.01/255$ to $0.24/255$ and its mean displacement falls from 37.0 to 27.8; at the setting we report, movement is $1.44/255$ and displacement 22.6 against 36.5 for the unguided rollout on the same references. We report at that setting, where the control demonstrably acts; below it no comparison between arms carries information, and reporting one would have turned a dead lever into a finding about GID.

Table 9: Eq. (2) as inference-time guidance, at the strength where the $\mathcal{L}_{\text{total}}$ control demonstrably acts. Oracle drift is $1 - \text{SSIM}$ against the accumulated warp; controllability is SSIM to that commanded target minus SSIM to the reference, so it is positive exactly when the commands took effect; displacement is measured in ϕ-space. Ten references, depth 20, five commanded moves each.

Arm	Oracle drift ↓	Controllability ↑	Displacement
None (unguided)	0.645 ± 0.165	0.0133 ± 0.017	36.5
$\mathcal{L}_{\text{total}}$ (control)	0.705 ± 0.145	0.0084 ± 0.026	22.6
$\mathcal{L}_{\text{drift}}$, stale projector	0.700 ± 0.147	0.0074 ± 0.021	22.5
$\mathcal{L}_{\text{drift}}$, sliding projector	0.706 ± 0.141	0.0075 ± 0.017	22.2

Result, and the arithmetic behind it. The three guided arms coincide on every column: the projection changes nothing relative to the plain penalty, and maintaining it changes nothing further. All three also sit above the unguided rollout on oracle drift and below it on controllability, so at this strength the guidance costs on both axes.

The mechanism is one the rest of the paper already measures. With $r = 4$ of $D = 768$, the projector removes about 7% of a typical displacement's magnitude (Appendix E); a penalty gradient altered by that much is not a different penalty. The decomposition is sharp on *isolated* commanded change (0.736), but that component is a small share of an accumulated rollout's displacement, so a downstream use that consumes the whole displacement inherits the small effect size. The same arithmetic sets the size of the whole-rollout correction in Appendix E, and stating it once explains both.

What this settles. It settles that post-hoc pixel-space guidance through a frozen encoder is not a useful implementation of Eq. (2) at this r/D ratio, and the $\mathcal{L}_{\text{total}}$ control makes that a measurement rather than a null result of unknown origin. It does not settle Eq. (2) as a training objective, which is the form the equation is written for: a fine-tuned model carries the penalty inside its own optimization, where the arms would differ in what the model learns rather than in an after-the-fact nudge to a rendered frame. We did not run that experiment and make no claim about its outcome. In the form we tested, H3 is unsupported, and we report it as such.

H What the Maintained Estimator Depends On

Section 5.3 shows that a projector re-estimated at the current state keeps its separation at every depth, and Appendix E reports that carrying it through to a whole-rollout score buys about a point. Those two facts sit together more comfortably once one knows what the estimate is actually attached to, and this appendix measures that. Both experiments below are diagnostics rather than results we build on; we report them because they explain the size of the correction in Appendix E rather than leaving it as a bare number.

How stale may the projector be? Walking a rollout and collecting the paired difference matrix at every step lets the projector be aged by a controlled number of steps.

There is no intermediate schedule: the maintained estimator works and anything older does not, so the refresh interval is one step and the amortized cost is the nine branches quoted in Section 5.3 at every step rather than spread across several. The $\sin\theta_{\max}$ row instantiates the estimation term of Eq. (3) on this testbed: it is zero for a maintained projector and essentially one for anything staler, so the bound is tight exactly where the estimator works and vacuous exactly where it does not. Being

Table 10: Separation against projector lag, 12 rollouts of depth 16, $r = 4$, leave-one-out throughout. Chance is 0.005.

Projector lag	0 (maintained)	1 step	2 steps	4 steps	anchored at reference
Separation	$+0.325$	$+0.004$	-0.009	-0.053	-0.002
Captured control	0.418	0.110	0.106	0.075	0.079
$\sin\theta_{\max}$ vs. current state	0.000	0.996	0.997	0.998	0.936

a worst-case angle over r directions it saturates readily, which is why we report captured fraction alongside it rather than in its place.

Is that the state moving, or the draw changing? A projector aged by one step differs from the current one in two ways at once — the state advanced, and the branches were taken with a different seed — so the table above does not by itself say which. The discriminating control is to build two difference matrices at *one* state with two different branch seeds.

	Captured control
Same state, same seed (leave-one-out)	0.413
Same state, *different* seed	0.118
Next state, different seed	0.107

The second row sits with the third, not the first. The commanded direction a single set of branches recovers is therefore partly specific to that draw: eight differences that share one base frame and one seed also share a draw-specific component, and leave-one-out does not remove it because it is present in all eight. Averaging over seeds narrows the gap only slightly (capture 0.133 at one seed to 0.162 at eight), so this is a property of the estimator's construction rather than of sampling noise.

What this explains. This is the mechanism behind a number the paper already reports. The seed-transferable part of the estimate is the part that survives into a real rollout, whose commanded change was produced by its own draws, and it is a modest fraction of the whole — which is why a maintained projector improves a whole-rollout score by about a point (Appendix E) even though the decomposition is sharp on isolated commanded differences. The two numbers were always consistent; this locates why.

The obvious cancellation does not work, and we checked. The paired construction already cancels the generation step by sharing seeds between base and intervened, so the natural next move is to cancel the draw the same way. We collected several branch sets per state and searched estimator variants offline. Removing the component common to all primitives within one draw is the obvious candidate, and it is *not* the centering trap of Appendix B — that one averaged within a control type across references; this removes an offset shared across control types within one draw.

Estimator fitted on held-out seeds	Ctrl captured	Drift captured	Separation
Single draw, as in the body	0.147	0.106	$+0.042$
Single draw, across-primitive mean removed	0.142	0.099	$+0.043$
Four draws averaged	0.170	0.140	$+0.030$
Four draws averaged, mean removed	0.162	0.144	$+0.019$
Four draws stacked, each mean removed	0.207	0.192	$+0.014$
Ceiling: same draw, leave-one-out	0.426	0.147	$+0.279$

None of them helps. Capture can be raised — stacking four centered draws reaches 0.207 — but drift capture rises with it, so separation does not improve; the draw-specific structure is not an additive offset that centering removes. We report this because it forecloses the fix a reader would try first, and because a diagnostic that only says "something is lost" is less useful than one that also says which repairs do not recover it.

Seed-matched branching, measured. The branch seed is not an adversarial unknown: it is an input the surrounding system chooses, so a deployed scorer may branch with the same draw the next step is

about to use. That puts it in the same-draw regime of the last row rather than the cross-draw regime above, and it is cheap to test: we maintain two projectors along the *same* 120 rollouts, differing only in whether their branches use an unrelated seed or the seed the upcoming step will use, and read off the same command inflation the body reports.

Command inflation per step ↓	no projection	sliding, cross-draw	sliding, seed-matched
Chord, $\|z_T - z_0\|$	$+33.1\%$	$+32.7\%$	$+30.9\%$
Path, $\sum_t \|z_t - z_{t-1}\|$	$+11.2\%$	$+10.2\%$	$+13.5\%$

On the chord form matching the draw helps, and by more than the cross-draw projector does: the correction over an unprojected score grows from 0.4 to 2.2 percentage points, a fivefold larger effect, though a small one in absolute terms. On the path form it does not help — it is worse than the cross-draw projector and worse than no projection at all. The mechanism is the drift column of the table above: a same-draw projector captures more drift (0.147 against 0.106), and a per-step difference is dominated by the generation step of exactly that draw, so the projector removes part of the signal the score is supposed to keep. Matching the draw buys back some of what cross-draw estimation loses on the chord form and overshoots on the path form, and since the path form is where most of the confound reduction lives, the net effect on the score we would recommend is negative. We report this because the argument that seed control is available to a deployed system is correct and the conclusion drawn from it is not: availability is not benefit.

I The Same Measurement on a Video Backbone

The construction is motivated by interactive video, and the main testbed is an autoregressive image rollout. This appendix repeats E1 on a video diffusion backbone so that the substrate and the motivation meet at least once, and states precisely what the repetition does and does not establish.

Setup. A text-to-video latent diffusion model with a transformer backbone (Peebles & Xie, 2023; Esser et al., 2024) generates eight reference clips (33 frames, 320×192), and the rollout is a video-to-video chain: each iteration re-denoises the model's own previous clip at fixed strength, so accumulation runs through a *spatiotemporal* denoiser rather than a per-frame one. The control family is unchanged — the same eight 2-D camera moves, now applied to every frame of the clip — so the comparison against the image testbed is like-for-like on the control side and differs only in the substrate. The representation is $\phi(\text{clip}) =$ the mean-pooled self-supervised embedding over the clip's frames. $\mathcal{S}_{\text{ctrl}}$ is built exactly as before, from paired base-versus-intervened continuations sharing a seed; $\mathcal{S}_{\text{drift}}$ is a fixed-control chain of depth 8.

Table 11: E1 on two substrates, $r = 4$. The control term is unchanged across substrates; the whole difference in separation is the drift term.

Substrate	Separation ↑	Mismatched	Random	Ctrl var. ↑	Drift var. ↓
Image rollout (main)	0.736 ± 0.043	-0.000	-0.001	0.808	0.072
Video-to-video chain	0.627 ± 0.111	-0.001	-0.000	0.803	0.175

Result. Gate 2 passes on the video substrate: separation 0.627 ± 0.111, with the mismatched-reference and random-subspace controls at zero, so the effect is again structure rather than dimension counting. The informative part is *where* the difference from the image testbed sits. Captured control variance is essentially identical (0.803 against 0.808), so the estimated subspace describes commanded change just as well on a spatiotemporal denoiser. The entire gap comes from the drift term, which rises from 0.072 to 0.175: more of the accumulated error lies in directions the commands also excite. That is the behaviour one would predict from a richer accumulation mechanism, and it is the quantity Proposition 3 identifies as the price of admitting an r-dimensional subspace. Locality survives as well: the cosine between the same command at different references is $+0.034$, still near-orthogonal, so the estimator must again be local. The control-difference spectrum keeps the same shape, with a leading singular-value ratio of 1.95 against 1.98 on images, so the commanded effect is low-dimensional to the same degree once it is expressed through a temporal model.

What this does not establish. A video-to-video chain is not a streaming interactive world model. It re-denoises a whole clip per iteration rather than extending a sequence chunk by chunk, it carries no KV cache across iterations, and its control interface is still an applied 2-D warp rather than an action the model was trained to consume. Those are the three properties under which a deployed system accumulates error, and this experiment reproduces only the first-order one — that accumulation passes through a denoiser with spatiotemporal structure. We read Table 11 as evidence that the decomposition is not an artifact of per-frame image rollouts, and not as evidence about any particular deployed world model.

J NOTATION

ϕ: feature map to $\mathbb{R}^D$; $z_t = \phi(y_t)$; z_0: reference; $\delta_t = z_t - z_0$; $\mathcal{C}$: control interface; c_t: control at step t; $\mathcal{G}_{z_0}$: gauge subspace, $\dim = r$; $\mathcal{G}_{z_0}^{\perp}$: invariant complement; $\mathsf{P}_{\mathcal{G}}, \mathsf{P}_{\mathcal{G}^{\perp}}$: orthogonal projectors; U_0: basis estimated at the reference; U_d: basis re-estimated at depth d; GID: Eq. (1); η_t: drift fraction; $\mathcal{S}_{\mathrm{ctrl}}, \mathcal{S}_{\mathrm{drift}}, \mathcal{S}_{\mathrm{mix}}$: the three corpora; K: number of control primitives.