# CACHE HOMEOSTASIS: DIAGNOSING AND CONTROLLING FEATURE DRIFT IN AUTOREGRESSIVE VIDEO DIFFUSION

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

## ABSTRACT

Autoregressive video diffusion supports streaming generation by producing video segments sequentially and reusing historical features through a key–value cache (KV Cache). However, cached feature statistics can drift during long rollouts, potentially affecting video quality. To understand this problem, we study how cached value features evolve during long video generation. We find that their energy becomes increasingly concentrated in a small subset of feature directions. Crucially, this change is mainly driven by the growth of a component shared across tokens, rather than by the disappearance of differences between them. To distinguish these effects, we derive an exact decomposition that separates the contributions of shared energy, alignment, and centered variation. Experiments show that the shared component explains most of the observed concentration increase and that its growth is associated with lower aesthetic quality. Guided by this finding, we propose Cache Homeostasis, a training-free method that regulates newly written cache features relative to an early reference. Its lightweight version controls the shared component, while its full version also regulates excess energy in selected centered directions. Experiments demonstrate that reference-directed intervention suppresses abnormal cache growth, while the full version improves text consistency over the lightweight version, which introduces little computational cost.

## 1 INTRODUCTION

Video diffusion models generate videos through progressive denoising. Models with bidirectional attention process an entire clip jointly, incurring substantial computation and memory costs. In contrast, autoregressive video diffusion generates frames or chunks sequentially, enabling streaming and long-video generation (Chen et al., 2024; Yin et al., 2025; Huang et al., 2025). Within this framework, the key–value cache (KV cache) acts as a working memory, storing historical internal representations for reuse. These cached representations provide the historical context needed to maintain consistency in subjects, scenes, and motion throughout autoregressive generation.

However, this reliance on cached history also allows generation errors to accumulate over long rollouts, as errors from each chunk are stored in the KV cache and carried into subsequent generation. Existing approaches mainly mitigate this degradation through self-generated training (Huang et al., 2025) or inference-time cache manipulation, but rarely examine how the cached representations themselves evolve. This limited understanding makes it difficult to intervene precisely and prevent long-horizon degradation. This motivates a central question: *what changes in the KV cache accompany progressive quality degradation during long-horizon generation?*

To understand how the cache evolves as a video grows longer, we first apply principal component analysis (PCA) to cache representations over long autoregressive rollouts. We observe that the energy of cache features grows progressively along a few dominant directions. To identify the source of this growth, we decompose each token representation in the value cache into two parts: the mean across tokens, which captures the component shared by all tokens, and a centered residual, which captures token-specific variation. This decomposition reveals that, in the rollouts we examine, the shared component steadily grows in magnitude and accounts for an increasing fraction of the total feature energy. We refer to this phenomenon as **shared-component amplification**.



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.

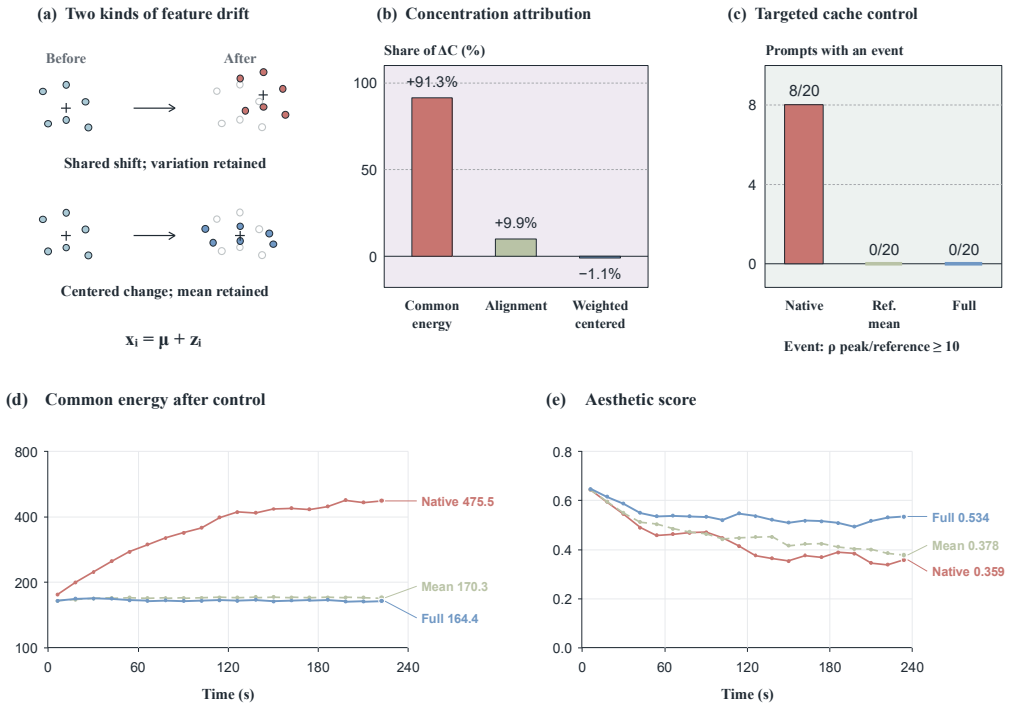



Figure 1: From temporal feature drift to targeted cache control. (a) Schematic decomposition $x_i = \mu + z_i$: a shared shift changes the mean while retaining centered variation, whereas a centered change preserves the mean. (b) In four 120 s audits, common energy accounts for $91.3\%$ of the increase in concentration $\Delta C$, with alignment and weighted centered terms contributing $+9.9\%$ and $-1.1\%$, respectively; these percentages attribute $\Delta C$ rather than total feature energy. (c) In a separate 20-prompt, 240 s cache-probe study, extreme-growth events ($\rho_{\text{peak/reference}} \geq 10$; exploratory retrospective threshold) occur in $8/20$ Native runs and $0/20$ runs under either Ref. mean or Full. (d,e) Equal-weight averages over six completed pilot prompts (p20–p25, seed 0), comparing Native, Mean and Full. (d) Post-control layer-median common energy, averaged on the original scale and shown on a logarithmic axis at the mean consumer times of the preceding cache writes. (e) Aesthetic scores at the centers of the corresponding video windows. Curves are unsmoothed; each contains 19 valid energy points or 20 aesthetic-score points. Ref. mean in (c) denotes the historical reference-directed intervention and differs from the Mean trust-region controller in (d,e). These descriptive analyses do not establish that lower cache growth causes improved visual quality or preservation of intended content.

This amplification matters because it skews the distribution of energy across feature directions. As the shared component grows, it increasingly dominates the representation and concentrates energy in a few directions, even when substantial token-specific variation is preserved. We quantify this effect with **spectral concentration**, which measures how unevenly energy is distributed across feature directions, and derive an exact decomposition that attributes changes in spectral concentration to individual representational components. This analysis identifies shared-component amplification as the dominant contributor to the observed rise in spectral concentration. Furthermore, a joint analysis of cache statistics and the quality of the corresponding videos reveals a negative association between the growth of shared-component energy and aesthetic quality. Together, these findings establish shared-component amplification as a concrete intervention target and provide a principled basis for targeted cache correction.

Based on these findings, we propose **Cache Homeostasis**. The method uses an early feature state as a reference and corrects newly written cache values during subsequent generation. Its lightweight **Mean** controller limits how far the shared component can move from the reference while preserving

differences between tokens. The **Full** extension additionally suppresses energy growth beyond the reference level along selected directions in the centered residuals. Both variants operate during inference, require no additional training, and maintain the same cache capacity.

We evaluate the method through changes in cache features, generated-video quality, and computational cost. Targeted correction suppresses abnormal growth of the shared component and improves average aesthetic quality and consistency with text prompts in the evaluated long videos. Mean introduces little additional computation. Full further improves text consistency, at a higher computational cost and with trade-offs in subject consistency. These results show how a mathematical analysis of cache evolution can guide targeted interventions and translate corrections to internal features into improvements in generated-video quality.

Our contributions are threefold:

- **Characterizing cache evolution during long video generation.** Starting from observations obtained through principal component analysis, we mathematically model the value cache, identify the growing dominance of its shared component, and characterize the association with aesthetic degradation.
- **Explaining the source of spectral concentration.** We derive an exact decomposition that separates the contributions to spectral concentration and identifies shared-component growth as its main source in the studied rollouts, providing a mathematical basis for selecting an intervention target.
- **Developing cache correction guided by the diagnosis.** We introduce Mean and its Full extension to suppress abnormal cache-feature growth during generation, improve selected dimensions of video quality, and assess the resulting quality and computational trade-offs.

## 2 Related work

**Autoregressive generation and cache management.** Diffusion Forcing (Chen et al., 2024) supports flexible causal conditioning, CausVid (Yin et al., 2025) distills a few-step causal generator, and Self-Forcing (Huang et al., 2025) trains on generated history. Rolling Forcing (Liu et al., 2026) and LongLive (Yang et al., 2026) combine rollout training with streaming context management. At inference, StreamingLLM (Xiao et al., 2024) retains initial sinks and recent tokens. Video methods add positional handling and pruning (Yi et al., 2025), head-specific memory (Tian et al., 2026), learned token salience (Chen et al., 2026b), or training-free future-aware merging (Luo et al., 2026). These methods manage the context available to the model. We keep the model frozen and study the statistics of values written into a fixed-capacity cache; Sink+Full tests composition with retention.

**Reference-based feature correction.** AdaIN (Huang & Belongie, 2017) and whitening/coloring transforms (Li et al., 2017) provide precedents for aligning feature statistics. TetherCache (Meng et al., 2026) reports reference-relative mean/std drift during rollout; its TAME partially aligns recalled keys and values to per-head, per-channel statistics of a trusted sink-and-memory pool. Our distinction is the diagnosis and correction geometry: Mean constrains the token mean relative to a fixed early reference at the new-value write site; Full adds selected centered-covariance corrections. A same-position mean/std baseline tests simpler alignment at this site (Appendix E.1), without reproducing TetherCache's recall system.

**Spectra and stability.** FreeLong (Lu et al., 2024) blends global and local attention in spatiotemporal Fourier bands. FreqForcing (Li et al., 2026) diagnoses Fourier-energy drift during rollout and guides local attention with early anchors' low frequencies. Our spectra describe energy across feature directions. FreeSpec (Chen et al., 2026a) studies attention-output concentration under enlarged windows; we study value-cache evolution over time at fixed token count. Depth-wise rank collapse in pure attention (Dong et al., 2021) and training-time attention-entropy instability (Zhai et al., 2023) concern different observables. Participation ratio also differs from singular-value-entropy effective rank (Roy & Vetterli, 2007). The following decomposition makes the specific ambiguity in value-feature concentration explicit before any intervention is chosen.

# 3 MEASURING SHARED AND CENTERED FEATURE CONCENTRATION

This section establishes how to distinguish shared amplification from a change in token-specific variation. We compare the uncentered second-moment spectrum with the centered covariance spectrum; the distinction is essential because centering before PCA removes the shared component we seek to measure.

## 3.1 MEASUREMENT OBJECT AND ENERGY SPLIT

At one layer and one autoregressive chunk, let $X \in \mathbb{R}^{n \times D}$ contain the newly written value vectors, with $n$ tokens and $D = Hd_h$ channels obtained by concatenating $H$ heads. We hold the measured token count fixed instead of pooling an expanding cache. Values form attention's output and, in the inspected Self-Forcing implementation, do not undergo the normalization and rotary encoding applied to keys. Appendix C.5 reports the site measurements; Appendix C.6 gives the token-count control.

Writing $\mathbf{1}$ for the $n$-vector of ones, define

$$\mu = \frac{1}{n} X^\top \mathbf{1}, \qquad X = \mathbf{1}\mu^\top + X_c, \qquad \mathbf{1}^\top X_c = 0. \tag{1}$$

The common component repeats the token mean; the centered component retains all pairwise token differences. Their cross term vanishes, so

$$M = \frac{X^\top X}{n} = \mu\mu^\top + \Sigma, \quad \Sigma = \frac{X_c^\top X_c}{n}, \quad \frac{\|X\|_F^2}{n} = \|\mu\|^2 + \operatorname{tr}\Sigma. \tag{2}$$

Thus common energy is $\|\mu\|^2$ and centered energy is $s = \operatorname{tr}\Sigma$. This split also has an attention interpretation. For one head's weights $a_h$ on this token block, $a_h^\top X_h = m_h \mu_h^\top + a_h^\top X_{c,h}$, where $m_h = \mathbf{1}^\top a_h$ is the block's attention mass. A shared value shift therefore changes this contribution in proportion to $m_h$, even though all within-block token differences are preserved. This local statement holds at fixed queries and keys; $m_h$ varies across queries and heads.

For nonzero $M$, the participation ratio and its inverse are

$$P(M) = \frac{(\operatorname{tr} M)^2}{\operatorname{tr}(M^2)}, \qquad C(M) = \frac{1}{P(M)}. \tag{3}$$

$P$ ranges from 1 for a single energetic direction to $D$ for equal energy in all directions. Lower $P$ need not imply lower algebraic rank: increasing a shared offset can concentrate $M$ while leaving $\Sigma$ unchanged.

## 3.2 EXACT ATTRIBUTION OF CONCENTRATION

For $s > 0$ and a nonzero mean, put

$$\rho = \frac{\|\mu\|^2}{s}, \quad f = \frac{\rho}{1+\rho}, \quad \beta = \frac{\hat{\mu}^\top \Sigma \hat{\mu}}{s}, \quad P(\Sigma) = \frac{s^2}{\operatorname{tr}(\Sigma^2)}, \qquad \hat{\mu} = \frac{\mu}{\|\mu\|}. \tag{4}$$

Here $f$ is the fraction of total energy shared across tokens, $\beta$ measures centered energy aligned with the mean, and $P(\Sigma)$ describes the centered spectrum. Expanding $\operatorname{tr}(M^2)$ yields

$$P(M) = \frac{(1+\rho)^2}{\rho^2 + 2\rho\beta + P(\Sigma)^{-1}}, \tag{5}$$

whose inverse separates three additive terms:

$$C(M) = \underbrace{f^2}_{\text{common energy}} + \underbrace{2f(1-f)\beta}_{\text{alignment}} + \underbrace{\frac{(1-f)^2}{P(\Sigma)}}_{\text{weighted centered concentration}}. \tag{6}$$

This is an identity, without distributional or dynamical assumptions (Appendix B.1). Its temporal differences sum exactly to $\Delta C$. A share of $\Delta C$ is distinct from the energy fraction $f$; signed shares may exceed one when another term decreases. In particular, centered concentration can increase while its weighted contribution falls. The identity supplies an accounting of observed change, rather than a causal explanation. We now use it to locate the change in recorded rollouts.

# 4 FROM SHARED-COMPONENT GROWTH TO AN INTERVENTION TARGET

We ask whether the shared term explains the measured concentration increase, whether it accompanies poorer video quality, and whether directing a correction toward an early reference controls its growth. These questions use separate cohorts; matching a prompt and seed alone does not make their trajectories interchangeable.

## 4.1 SHARED GROWTH EXPLAINS THE AUDITED CONCENTRATION INCREASE

At the value-projection site of four 120 s Self-Forcing/Wan2.1 audits, $C(M)$ increases by a factor of 3.3 between the first and last $15\%$ of post-startup chunks. The common term explains $91.3\%$ of the increase, alignment $9.9\%$, and the weighted centered term $-1.1\%$. The centered spectrum becomes more concentrated in all four runs, but its energy weight $(1-f)^2$ falls faster. The dominant observed change is therefore shared amplification, not disappearance of centered variation.

Two selected 240 s extreme-growth cases show a $99.6\%$ common-term contribution, late $P(M) = 1.325$, and common energy fraction $f = 0.872$. They illustrate the extreme regime rather than its prevalence. Table 4 and Figure 2 give the full attribution and depth profiles; Appendix C.4 reports the algebraic consistency checks across 711,030 measurement rows.

## 4.2 SHARED ENERGY AND QUALITY ON THE SAME VIDEOS

The complete main cohort links cache measurements with quality scores from 500 generated videos. For Native trajectories, the mean within-prompt correlation between preceding log common energy and aesthetic quality is $-0.697$. This association weakens to $-0.314$ after linear detrending and to $-0.140$ after first differencing. Reverse-time analysis yields similarly negative correlations of $-0.673$, $-0.307$, and $-0.096$, indicating that a substantial portion of the raw association is explained by a shared temporal trend rather than by a clear predictive relationship. The corresponding raw associations are much weaker for Mean and Full, with correlations of $-0.118$ and $0.071$, respectively. All forward and reverse comparisons use the same 17 eligible quality windows for each trajectory, with seeds averaged within each prompt and prompts weighted equally. These analyses therefore connect the representation-level diagnosis to output quality, while the strong attenuation after temporal adjustment and the symmetry of the reverse-time results do not support interpreting common-energy growth as an early warning signal.

The same cohort also allows direct characterization of how the controllers regulate cache statistics. We define a sustained common-energy event as post-control common energy remaining at least twice its reference level from the first 36 seconds for three consecutive 12-second supports. Under this criterion, sustained events occur in 86 of 100 Native trajectories, compared with none of the 100 Mean trajectories, none of the 100 Full trajectories, 2 of 100 Sink trajectories, and none of the 100 Sink+Full trajectories. Three Native trajectories and one Sink trajectory cross the threshold near termination but do not contain sufficient subsequent observations for confirmation. This analysis uses the median common energy across layers and is distinct from the historical event definition based on $\rho$. The resulting counts show that Mean and Full effectively suppress sustained post-control common-energy growth, while the corresponding effects on generated-video quality are evaluated separately in Section 6.

Beyond threshold-based event counts, we further compare the late-to-early growth of post-write common energy. The geometric mean ratios across prompts are 3.811 for Native, 1.006 for Mean, and 1.192 for Sink. These results indicate that Mean constrains the growth of stored common energy more tightly than either Native or Sink. However, before the write-time projection is applied, the corresponding ratio for Mean is 1.258, slightly higher than the value of 1.192 observed for Sink. This distinction suggests that the improved post-write stability of Mean is partly attributable to its direct projection mechanism. Therefore, reduced post-write common-energy growth should be interpreted as evidence of effective cache regulation rather than as sufficient evidence of improved video generation quality.

### 4.3 REFERENCE-DIRECTED CORRECTION CONTROLS THE DIAGNOSED GROWTH

In a four-prompt direction study, mean-based and singular-subspace interventions are matched to the same pre-cast relative Frobenius correction norm of $0.1550$. Moving the token mean toward an early reference reduces the latent cosine distance to the initial frame by $0.182$, with $t = -8.93$. Moving the mean toward zero also reduces this drift proxy, but increases the normalized latent standard deviation by $0.524$, with $t = 9.00$. This latent statistic is distinct from the cache feature energy analyzed in our representation-level diagnosis, as detailed in Appendix C.6. By contrast, matched interventions along singular directions do not yield a significant drift reduction in this small study, as reported in Table 9. These results motivate preserving a reference shared component rather than merely suppressing its magnitude.

A separate cache-probe study evaluates whether reference-directed correction suppresses extreme common-component growth over 240 seconds. Across 20 prompts, we define an extreme-growth event using a retrospectively selected threshold at ten times the reference level of the deep-layer median $\rho$. Such events occur in 8 of 20 Native trajectories, but in none of the 20 trajectories using reference-directed mean correction and none of the 20 trajectories using Full control, as shown in Table 5 and Figure 3. The historical reference-directed mean intervention applies a matched-magnitude correction toward the early reference with $\alpha = 0.5$ and differs from the Mean trust-region controller introduced below. Because the randomized direction probe alters the generation random-number state, its event labels cannot be directly paired with quality measurements from separately generated videos. We therefore interpret this experiment as evidence that reference-directed correction can suppress the diagnosed shared-component growth, while evaluating its effects on video quality and computational cost separately in the following sections.

## 5 CACHE HOMEOSTASIS: REFERENCE-BASED FEATURE CONTROL

We specify two updates that preserve the distinction in equation 1. Mean bounds deviation of the shared component from an early reference; Full adds correction of selected centered directions. Both leave model weights, keys and cache capacity unchanged, and edit only newly written values once, after the clean-context pass. Previously corrected rows are not rewritten.

### 5.1 REFERENCE CONSTRUCTION AND MEAN PROJECTION

Each of the 30 layers has its own reference within the current trajectory. Zero-based unmodified chunks $\mathcal{R} = \{2, \dots, 6\}$ define its mean, variation and centered covariance:

$$\mu_{\mathrm{ref}} = \frac{1}{|\mathcal{R}|} \sum_{c \in \mathcal{R}} \mu_c, \quad \tau^2 = \frac{1}{|\mathcal{R}|} \sum_{c \in \mathcal{R}} \|\mu_c - \mu_{\mathrm{ref}}\|^2, \quad G_{\mathrm{ref}} = \frac{1}{|\mathcal{R}|} \sum_{c \in \mathcal{R}} \Sigma_c. \tag{7}$$

Control begins after this window. Let $b = k\tau$ be the trust-region radius, where $k$ is a dimensionless multiplier ($k = 1$ by default). With $d = \mu - \mu_{\mathrm{ref}}$, project the mean as

$$\mu' = \begin{cases} \mu, & \|d\| \le b, \\ \mu_{\mathrm{ref}} + \dfrac{b}{\|d\|} d, & \|d\| > b, \end{cases} \qquad \Delta_\mu = \mathbf{1}(\mu - \mu')^\top. \tag{8}$$

Mean uses $X' = X - \alpha \Delta_\mu$, with $\alpha = 1$ by default. At this strength the new mean lies in the radius-$b$ ball; $0 < \alpha < 1$ relaxes the projection and need not reach the ball in one update. At $\alpha = 1$, this is also the unique minimum-Frobenius-change update among all matrices whose token mean lies in the reference ball: it spends no correction on centered differences (Appendix B.2). The reference assumes that the early state remains useful; an intentional scene transition can violate that assumption.

### 5.2 FULL: SELECTED CENTERED-COVARIANCE EXCESS

The shared-growth diagnosis directly motivates Mean, which leaves $\Sigma$ unchanged. Full tests whether additionally limiting centered covariance excess improves output quality. It selects orthonormal eigenvectors $U$ associated with the $r$ largest algebraic eigenvalues of $\Sigma - G_{\mathrm{ref}}$, ranking covariance

excess rather than the energy of $\Sigma$ alone. These directions can mix channels across heads. For each selected direction, set

$$e_i^{\text{now}} = u_i^\top \Sigma u_i, \quad e_i^{\text{ref}} = u_i^\top G_{\text{ref}} u_i, \quad s_i = \min\left(1, \sqrt{\frac{\lambda e_i^{\text{ref}}}{e_i^{\text{now}}}}\right), \tag{9}$$

with $s_i = 1$ when current energy is zero. The dimensionless tolerance $\lambda$ is an allowed energy multiple (1 by default), distinct from reference spread $\tau$ and radius multiplier $k$. The update is

$$\Delta_{\text{lr}} = X_c U \operatorname{diag}(1 - s) U^\top, \qquad X' = X - \alpha(\Delta_\mu + \Delta_{\text{lr}}). \tag{10}$$

Defaults are $r = 16$ and $\alpha = 1$. At full strength in exact arithmetic, $u_i^\top \Sigma' u_i = \min(e_i^{\text{now}}, \lambda e_i^{\text{ref}})$ for each selected direction. This does not bound covariance in every direction.

**What the local guarantees establish.** The mean correction has identical rows and the centered correction has zero row mean, so they act on separate components of $X$. Their derived concentration terms remain coupled through $f$ and $\beta$. Mean projection is idempotent at $\alpha = 1$ with a fixed reference; recomputing $U$ makes Full non-idempotent in general (Appendices B.4 and B.5). These are local, exact-arithmetic properties, not guarantees of rollout stability or perceptual quality. The next section tests those practical effects using float32 corrections written back to bfloat16 caches; numerical safeguards and the once-per-chunk schedule are specified in Appendix B.3.

## 6 Long-video quality, cost, and transfer

The evaluation tests three consequences of the proposed correction: its output tradeoffs at fixed capacity, the value of the centered extension, and its computational cost. Additional generators test the scope of the diagnosis.

### 6.1 Paired four-minute evaluation

The main evaluation is conducted on Self-Forcing with Wan2.1-T2V-1.3B (Huang et al., 2025; Wan Team, 2025). We use 50 fixed prompts and two random seeds for each of five methods: Native, Mean, Full, Sink, and Sink+Full. This results in 500 completed trajectories, with no terminal failures. The prompt set contains 13 subject-motion scenes, 13 camera-motion scenes, 12 scene-change cases, and 12 stable scenes. All methods are evaluated under matched prompts, seeds, checkpoints, initial noise, and generation settings. Each trajectory contains 960 latent frames and is continuously decoded into 3,837 video frames at a resolution of $480 \times 832$ and a frame rate of 16 frames per second, corresponding to approximately 240 seconds of video. All methods use the same cache capacity of 21 latent frames. The Sink baseline retains four early frames together with the 17 most recent frames, while Mean and Full use the default configurations defined in Section 5.

We evaluate generated videos using the fixed VBench protocol (Huang et al., 2024). The reported dimensions include subject consistency, background consistency, motion smoothness, aesthetic quality, imaging quality, and text consistency, while dynamic degree is used only to characterize motion activity. Each video is divided into 20 non-overlapping temporal windows, consisting of 19 windows of 192 frames and one final window of 189 frames. Scores are first aggregated within each trajectory using frame-count weighting and are then averaged equally across the two seeds associated with each prompt. Imaging-quality scores are normalized by a factor of 100. The prompt serves as the independent statistical unit, yielding $n = 50$ paired observations for each method comparison. The six pre-specified Full–Mean comparisons across quality and text dimensions are evaluated using two-sided paired $t$-tests with Holm correction for multiple testing, while the reported confidence intervals remain unadjusted. Other method contrasts and dynamic-degree comparisons are treated as descriptive analyses. Appendix D.1 provides the complete set of 70,000 metric-window records together with all primary endpoints.

Table 1: Long-video quality at 239.8 s. Each method uses 50 prompts and two seeds; $n$ counts complete two-seed prompts. Scores use all 3,837 frames with equal seed weights. Imaging scores are divided by 100. Dynamic degree measures activity.

| Method | $n$ | Failed | Subj. | Backg. | Smooth. | Dyn. | Aes. | Img. | Text |
|---|---|---|---|---|---|---|---|---|---|
| Native | 50/50 | 0/100 | 0.951 | 0.966 | 0.985 | 0.297 | 0.412 | 0.621 | 0.122 |
| Mean | 50/50 | 0/100 | 0.950 | 0.958 | 0.986 | 0.380 | 0.436 | 0.614 | 0.162 |
| Full | 50/50 | 0/100 | 0.940 | 0.954 | 0.987 | 0.403 | 0.444 | 0.586 | 0.186 |
| Sink | 50/50 | 0/100 | 0.917 | 0.938 | 0.983 | 0.475 | 0.553 | 0.688 | 0.238 |
| Sink+Full | 50/50 | 0/100 | 0.922 | 0.942 | 0.984 | 0.395 | 0.513 | 0.604 | 0.233 |

### 6.2 Quality differences and their tradeoffs

The low-cost Mean controller has measurable benefits and losses (Table 2). Relative to Native, aesthetic quality increases by $0.0238$ (95% interval $[0.0047, 0.0429]$) and text by $0.0401$ ($[0.0280, 0.0522]$), while background consistency falls by $0.0080$. Relative to Sink, Mean retains higher subject/background consistency and smoothness, but lower aesthetic, imaging and text scores. These added post-hoc intervals retain all dimensions without changing the primary family.

The original four scene strata show the same subject/background advantage and aesthetic/text cost relative to Sink. For subject-motion prompts, Mean's activity is lower by $0.2289$ ($[-0.3962, -0.0616]$). All 56 stratum intervals are in Appendix D.2; they are descriptions, not interaction tests or motion-fidelity evidence. Mean offers a particular quality tradeoff at similar measured cost to Sink (Section 6.4), not a universal win.

The pre-specified Full–Mean comparison tests the centered extension (Table 13): text increases by $0.0241$ (95% CI $[0.0115, 0.0367]$, Holm $p = 0.0022$), while subject consistency falls by $0.0105$ ($[-0.0172, -0.0039]$, Holm $p = 0.0129$). The other four primary endpoints do not pass the six-test correction. Dynamic degree changes by $0.0234$ ($[-0.0658, 0.1127]$), not a motion non-inferiority result. Full–Native secondary estimates are $+0.0318$ in aesthetic ($[0.0068, 0.0569]$) and $+0.0642$ in text ($[0.0488, 0.0795]$). All 21 original secondary contrasts remain in Appendix D.1.

Sink has the highest aesthetic, imaging and text means; Native the highest subject and background means. Adding Full to Sink raises subject/background means but lowers aesthetic and imaging by $0.0401$ and $0.0833$. Full time profiles and fixed examples remain in Appendices D.4 and D.5.

### 6.3 Does correction geometry matter?

A separate six-prompt development study compares Mean, Full, matched random centered correction, uniform centered contraction and same-position mean/std alignment: all 30 videos and 4,200 scores are retained. Mean/std alignment improves aesthetic, imaging and text means over Mean by $0.0262$, $0.0239$ and $0.0402$, recovering part of Full's gains. Full exceeds the random control by $0.0679$, $0.0351$ and $0.0337$ on these dimensions, with lower subject consistency and smoothness. All 56,340 random post-bfloat16 centered-dose matches pass the fixed tolerance; 28 infeasible uniform targets remain included. These descriptive results compare correction geometries at the same write site. Diverging later inputs and small post-cast common increments prevent isolating direction as the sole causal difference (Appendix E.1).

### 6.4 Generation cost

Eight exclusive-A100 runs compare two fixed prompts across Native, Mean, Full and Sink at the main study's duration. Mean generation times are $470.27$, $478.33$, $736.43$ and $469.98$ s; averaging within-prompt ratios gives Mean $+1.7\%$ and Full $+56.6\%$ relative to Native. Throughputs are $8.16$, $8.02$, $5.21$ and $8.16$ nominal output frames/s, excluding VAE decoding. Peak reserved memory is $9.54$–$9.86$ GiB (Table 29). Mean stays close to Native runtime, whereas Full's per-layer, per-chunk eigendecomposition adds substantial cost. The estimate covers this implementation and two prompts, not end-to-end video throughput or a hardware-independent overhead.

Table 2: Mean's paired tradeoffs in the complete main cohort ($n = 50$ prompts). These added post-hoc 95% intervals are unadjusted, not a new confirmatory family. Dynamic degree measures activity, with no uniformly preferred direction.

| | Mean–Native | | Mean–Sink | |
|---|---|---|---|---|
| Metric | $\Delta$ | 95% CI | $\Delta$ | 95% CI |
| Subject | $-0.0012$ | $[-0.0061, +0.0036]$ | $+0.0333$ | $[+0.0241, +0.0424]$ |
| Background | $-0.0080$ | $[-0.0114, -0.0046]$ | $+0.0195$ | $[+0.0145, +0.0245]$ |
| Smoothness | $+0.0010$ | $[-0.0007, +0.0027]$ | $+0.0037$ | $[+0.0022, +0.0053]$ |
| Aesthetic | $+0.0238$ | $[+0.0047, +0.0429]$ | $-0.1173$ | $[-0.1378, -0.0969]$ |
| Imaging | $-0.0069$ | $[-0.0451, +0.0314]$ | $-0.0733$ | $[-0.1069, -0.0397]$ |
| Text | $+0.0401$ | $[+0.0280, +0.0522]$ | $-0.0760$ | $[-0.0894, -0.0626]$ |
| Dynamic degree | $+0.0822$ | $[+0.0053, +0.1590]$ | $-0.0951$ | $[-0.1616, -0.0286]$ |

### 6.5 Transfer across trained generators

MAGI-1 4.5B adds a distinct architecture, with two known development prompts and 18 video assignments across Native, four Mean and four Full settings at 24 s (Appendix F.4). All settings improve six quality/text scores on the camera-motion example, but none passes the fixed cross-prompt joint targets. Mean $(\alpha, k) = (0.5, 2)$ raises pooled aesthetic/imaging scores by $0.00158/0.00175$, with text lower by $0.00038$. Full's largest pooled text increment over that anchor is $0.000205$, with lower aesthetic and imaging scores.

Causal Forcing (Zhu et al., 2026) shares the Wan1.3B family but uses different training. All 48 videos (eight prompts, two seeds, three arms) are scored. Native common fraction stays near $0.203$; both controllers reduce mean-reference drift, but none of four fixed quality tests is significant (all Holm $p = 1$). Full–Mean text is $+0.00330$ ($[-0.00593, 0.01252]$). A separate 16-trajectory Native study finds a startup increase without confirmed later energy dominance (Appendices F.6 and F.7). Reference drift is controllable here without a significant quality gain. Together with the mixed MAGI results, this limits the transfer claim; it does not establish which native feature changes predict a benefit from control.

## 7 Discussion

The evidence supports a component-specific design principle. Shared growth explains the audited Self-Forcing concentration change; a reference-directed intervention suppresses extreme cache events; and current Mean and Full show selected quality gains with distinct costs. These are linked findings, not a demonstrated causal chain from shared growth to perceptual failure. The historical event arm differs from current Mean; the main cohort now supplies a separate descriptive event count for the implemented controller. Temporal correlations weaken after detrending, and the main confirmatory result is Full–Mean's text–subject tradeoff. The main protocol was fixed before its outcomes, but complete disjointness from all historical prompt use is not certified.

A fixed early reference may suppress legitimate transitions, and activity scores do not establish motion fidelity. Transfer and controller-aware adaptation remain mixed (Appendix G.1). Self-Forcing quality evidence ends at $239.8$ s; $480$ s runs exceeded the implementation's temporal-position limit. These results support Mean as a low-cost controller and Full as a costlier extension, while motivating motion-aware evaluation and reference adaptation.

## 8 Conclusion

Shared-component amplification accounts for most of the concentration increase in the audited long Self-Forcing rollouts. An exact decomposition distinguishes that growth from centered variation, and same-video measurements connect it to aesthetic decline as an association. Cache Homeostasis uses this diagnosis to regulate newly written values relative to an early reference. Complete quality and runtime comparisons show what each extension buys: inexpensive mean control, and an additional text-consistency gain from Full with subject consistency and computational costs. The useful intervention target is the feature component that changes, with output benefits assessed separately.

## AI USE STATEMENT

Generative AI tools assisted with developing the conceptual framework and mathematical arguments, refining hypotheses and experimental protocols, implementing methods and analysis code, generating evaluation prompts, organizing and reformatting experimental records, interpreting quantitative results and sampled visual outputs, literature search, figure and table preparation, and drafting, translating, and editing the manuscript. Quantitative results are computed from recorded model outputs by executable analysis; AI-assisted visual descriptions are not human evaluations. The authors are responsible for the final content, claims, and artifacts.

## ETHICS STATEMENT

The main-text controllers modify inference-time caches with frozen model weights; no human participants are involved. They share synthetic video's dual-use risks. Misuse prevention, fairness and semantic controllability remain unevaluated, and downstream use must respect the underlying model conditions.

## REPRODUCIBILITY STATEMENT

Appendices B.1 and B.5 give the derivations and counterexample. Executable analysis recomputes recorded attribution, event counts, proxy and VBench results, retaining input hashes, coverage, intervals and multiplicity checks. The source includes all $70{,}000$ main metric-window records, final paired statistics, temporal profiles and figure reproduction scripts. The MAGI scans additionally retain all $252$ distinct metric-window records, and the Causal Forcing comparison retains all $6{,}720$. The Native follow-up releases complete prompt-level diagnostic summaries, all sampled layer/head descriptions and fixed analysis protocols. Pilot, transfer and adaptation records remain separate. Missing records and unsupported cross-harness joins are explicit. Generation requires the corresponding weights and GPU resources.

## REFERENCES


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.

Fangda Chen, Shanshan Zhao, Longrong Yang, Chuanfu Xu, Zhigang Luo, and Long Lan. FreeSpec: Training-free long video generation via singular-spectrum reconstruction. *arXiv preprint arXiv:2605.06509*, 2026a. URL https://arxiv.org/abs/2605.06509.

Hanmo Chen, Chenghao Xu, Xu Yang, Xuan Chen, and Cheng Deng. Past- and future-informed KV cache policy with salience estimation in autoregressive video diffusion. *arXiv preprint arXiv:2601.21896*, 2026b.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning (ICML)*, 2021.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1501–1510, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Huang_Arbitrary_Style_Transfer_ICCV_2017_paper.html.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jiatong Li, Leo Liang, Linghe Kong, and Yulun Zhang. Freqforcing: Autoregressive long video generation via spectral self-anchoring. *arXiv preprint arXiv:2607.27110*, 2026. URL https://arxiv.org/abs/2607.27110.

Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/49182f81e6a13cf5eaa496d51fea6406-Abstract.html.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling Forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.25161.

Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. FreeLong: Training-free long video generation with SpectralBlend temporal attention. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/ed67dff7cb96e7e86c4d91c0d5db49bb-Abstract-Conference.html.

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

Yu Meng, Xiangyang Luo, Letian Li, Wenyuan Jiang, Chen Gao, Xinlei Chen, Yong Li, and Xiao-Ping Zhang. TetherCache: Stabilizing autoregressive long-form video generation with gated recall and trusted alignment. *arXiv preprint arXiv:2606.13035*, 2026. doi: 10.48550/arXiv.2606.13035. URL https://arxiv.org/abs/2606.13035.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *European Signal Processing Conference (EUSIPCO)*, 2007.

Jiahao Tian, Yiwei Wang, Gang Yu, and Chi Zhang. Head forcing: Long autoregressive video generation via head heterogeneity. *arXiv preprint arXiv:2605.14487*, 2026.

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

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. LongLive: Real-time interactive long video generation. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2509.22622.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

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

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning (ICML)*, 2023.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

# A EVIDENCE MAP AND INTERPRETATION

This appendix identifies which study supports each step of the argument. Cohorts retain their original statistical units and are never pooled merely because they share prompts, seeds or a controller name.

## A.1 STUDY RELATIONSHIPS AND SCOPE OF INTERPRETATION

**Conventions and units.** Common and shared denote the same token-mean component. Native denotes uncontrolled generation; historical tables retain the alias open loop. The radius is $b = k\tau$, with reference spread $\tau$ and dimensionless multiplier $k$; $\lambda$ is the centered-energy tolerance and $r$ the rank. Implementation identifiers are mapped to descriptive names in the reproduction documentation. Unless a study states otherwise, Mean/Full use $\alpha = k = \lambda = 1$ and Full uses $r = 16$. The historical reference-directed mean intervention is denoted Ref. mean (implementation identifier: `constant_ref`) and is distinct from Mean. A nominal 240 s Self-Forcing/CF video contains 3,837 frames at 16 fps (239.8125 s); a MAGI 24 s video contains 576 frames at 24 fps. Quality differences use evaluator score units, with imaging divided by 100 once, except historical Table 22, which retains raw MUSIQ units. They are not percentages or a pooled leaderboard score.

Table 3: Evidence map. Rows are separate studies, with overlap retained rather than summed into a total sample size. Windows, layers and intervention events are repeated measurements. All scored cohorts retain every assigned outcome.

| Question | Cohort and independent unit | Evidence and scope |
|---|---|---|
| Concentration attribution | Four 120 s audits; two selected 240 s cases | Table 4; descriptive attribution, not prevalence |
| Cache-event control | 20 paired prompts, 240 s | Table 5; historical reference mean and Full |
| Main quality | 50 prompts, two seeds, five arms; 500 videos | Table 13; six Full–Mean tests; $n = 50$ |
| Developmental quality | Six prompts, one seed, three arms; 18 videos | Table 23; no Holm-significant primary endpoint |
| Correction geometry | Same six development prompts, five arms; 30 videos | Tables 27–28; descriptive, actual-dose controls |
| Generation cost | Two fixed prompts, four methods; eight runs | Table 29; exclusive-device generation only |
| MAGI transfer | Two development prompts; 18 assignments at 24 s | Appendix F.4; complete scans, no joint-target winner |
| MAGI confirmation | 16 new prompts, one seed, two arms | Incomplete comparison; no quality result is reported here. |
| Causal Forcing | Eight prompts, two seeds, three arms; 48 videos | Appendix F.6; four primary tests, no significant gain |
| CF Native follow-up | Eight different prompts, two seeds; 16 trajectories | Appendix F.7; diagnostic only, one decoded video |
| Training adaptation | Separate update batches; final $4 \times 2 \times 6 = 48$ videos | Appendix G.1; developmental, mixed quality |

**Evidence for diagnosis, intervention, and output quality.** The four 120 s audits quantify the sources of concentration growth. The two 240 s attribution cases were selected for extreme cache growth and do not estimate its prevalence. The 20-prompt cache study compares event incidence across interventions. Its Ref. mean arm applies a correction toward the early reference with matched magnitude, $\alpha = 0.5$, and new-row writeback. It differs from the Mean trust-region projection in Section 5.1. The $0/20$ event count for this historical arm therefore supports the intervention target rather than supplying an event count for the current Mean controller. Full's cache-event variants are listed in Table 5.

The historical probe uses randomized SVD, which consumes the generation RNG. Matching prompt and seed is consequently insufficient to join its events to videos generated under a separate output-measurement harness. Cache-event labels are not used to group those videos. The main 500-video cohort instead records cache statistics and output quality on the same trajectories with non-perturbing probes; its complete paired quality and temporal analyses appear in Appendix D.1.

**Development and main evaluation.** The four-prompt direction study compares corrections at matched pre-cast Frobenius magnitude. Rank-one top-direction suppression changes latent drift by $+0.102$ with two-sided $p = 0.078$; it establishes neither significant harm nor a general result about spectral suppression. Alignment rotations belong to a separate, unmatched-budget comparison (Appendix C.7). The six-prompt video pilot and the 30-video centered-control study use development prompts. The main 50-prompt list and analysis protocol were frozen before its outcomes, but disjointness from every historical prompt use is not certified. All main assignments and windows are retained; no missing scores are imputed and no assignments are replaced.

**Temporal interpretation and generalization.** The main model is Self-Forcing/Wan2.1-T2V-1.3B. The same-video measurements establish temporal association, with substantial attenuation after detrending and similar reverse-time coefficients. They do not establish a perceptual failure mechanism or a validated early-warning rule. The secondary analysis rules were fixed after 92 scores existed. The historical horizon, MAGI and controller-aware adaptation studies retain their separate protocols and exploratory status. The two-prompt MAGI scans add complete paired quality measurements on development examples, but do not complete the separate four-scene controller cohort. A separate 16-prompt Full–Mean confirmation remains incomplete in the evidence reported here and contributes no quality-effect estimate or significance claim. The Causal Forcing long-model comparison adds 48 scored videos in the same backbone family and finds no significant primary quality gains. Its separate 16-trajectory Native follow-up measures cache and attention statistics, not controller efficacy (Appendices F.4–F.7).

**Reference, precision, and implementation.** The controller assumes that an early reference remains useful; intentional scene transitions may require a changing reference. Motion non-inferiority has not been established by the dynamic-degree comparison. The mathematical properties describe local updates in exact arithmetic, with the conditions stated in Section 5 and Appendix B.5; finite precision and recurrent generation require empirical assessment. The controlled runtime study covers two fixed prompts, and Full includes a per-layer, per-chunk eigendecomposition. At 240 s, the 960 latent frames use 94% of the implementation's 1024-entry temporal RoPE table; all 12 attempted 480 s runs failed at that limit. The released exploration ledger (Appendix C.10) preserves unsuccessful mechanism and control hypotheses alongside the completed results.

# B Mathematical details of measurement and control

This appendix derives the diagnostic identity and states the exact local properties of the implemented controller. It separates those properties from the stronger, unimplemented covariance envelope.

## B.1 Deriving the concentration identity

Let $X \in \mathbb{R}^{n\times D}$, with token mean $\mu$ and centered matrix $X_c = X - \mathbf{1}\mu^\top$. The zero-sum condition $\mathbf{1}^\top X_c = 0$ eliminates the cross term, giving $M = X^\top X/n = \mu\mu^\top + \Sigma$ with $\Sigma = X_c^\top X_c/n$. Put $a = \|\mu\|^2$ and $s = \operatorname{tr}\Sigma$. Then

$$\operatorname{tr} M = a + s, \tag{11}$$

$$\operatorname{tr}(M^2) = \operatorname{tr}((\mu\mu^\top)^2) + 2\operatorname{tr}(\mu\mu^\top\Sigma) + \operatorname{tr}(\Sigma^2) \tag{12}$$

$$= a^2 + 2\mu^\top\Sigma\mu + \operatorname{tr}(\Sigma^2). \tag{13}$$

The equality uses trace cyclicity and holds without independence, Gaussianity or a model-dynamics assumption. For $a, s > 0$, substitute $\rho = a/s$, $\beta = \mu^\top\Sigma\mu/(as)$ and $P(\Sigma) = s^2/\operatorname{tr}(\Sigma^2)$ into $(\operatorname{tr} M)^2/\operatorname{tr}(M^2)$ to obtain equation 5. Dividing the denominator by $(a+s)^2$ and writing $f = a/(a+s)$ gives equation 6.

If $a = 0$ and $s > 0$, then $M = \Sigma$ and $C(M) = 1/P(\Sigma)$; one may set $\beta = 0$ by convention because its coefficient vanishes. If $s = 0$ and $a > 0$, positive semidefiniteness implies $\Sigma = 0$, so $M$ has rank one and $C(M) = 1$; expressions requiring division by $s$ should not be evaluated. If $a = s = 0$, $M = 0$ and participation ratio is undefined. Such degenerate inputs must be marked rather than interpreted as evidence of concentration.

The three summands are nonnegative at a fixed state, but their temporal changes can have either sign. A contribution divided by total $\Delta C$ can therefore exceed one when another contribution decreases. This is an additive accounting identity, not a unique causal or counterfactual attribution: both $f$ and $\beta$ can change when a component of $X$ is modified.

### B.2 Minimum-change interpretation of Mean

For a fixed reference, radius $b \geq 0$ and $\alpha = 1$, the ideal Mean update is

$$\operatorname*{arg\,min}_{Y:\, \|\mu_Y - \mu_{\mathrm{ref}}\| \leq b} \|Y - X\|_F^2.$$

To see this, write $Y = \mathbf{1}\mu_Y^\top + Y_c$ and use orthogonality: $\|Y - X\|_F^2 = n\|\mu_Y - \mu\|^2 + \|Y_c - X_c\|_F^2$. The constraint concerns only $\mu_Y$, so the unique minimizer has $Y_c = X_c$ and $\mu_Y$ equal to the Euclidean projection of $\mu$ onto the closed ball. The property justifies the correction geometry at a single write; it does not make an early reference optimal for a future scene or certify video quality.

### B.3 Implementation and numerical safeguards

The implemented matrix concatenates head channels, $D = Hd_h$, and uses every selected layer's own reference; the main configuration selects all 30 layers. It gathers chunks 2–6 without editing them, then controls new rows from chunk 7 after the clean-context pass. Old cache rows are not corrected again. Means, covariances and corrections use float32. The covariance difference is cast to float64 for eigendecomposition and its selected eigenvectors are cast back to float32; the final corrected values are cast to the cache's bfloat16. The reference variance is computed as $\max(0, \mathrm{mean}_c \|\mu_c\|^2 - \|\mu_{\mathrm{ref}}\|^2)$ before its square root. Mean deviations below $10^{-12}$ produce no mean correction. Full clamps computed directional energies to be nonnegative and uses a $10^{-12}$ lower bound in its energy denominator. The implementation skips a non-finite correction. These finite-precision rules can leave small common or centered residuals; the minimum-change and decoupling statements describe the ideal operator, not exact identities after rounding.

### B.4 Exact decoupling of the two terms

$\Delta_\mu$ has identical rows, so it leaves $X_c$ untouched, and $\mathbf{1}^\top X_c = 0$ gives $\mathbf{1}^\top \Delta_{\mathrm{lr}} = 0$, so the low-rank term leaves $\mu$ untouched. Thus each update acts on one part of the orthogonal decomposition. This does *not* assign each update to one channel of equation 6: $\rho$, $\beta$ and $P(\Sigma)$ are derived quantities, and the low-rank term changes $\mathrm{tr}\,\Sigma$ and hence $\rho$, while the mean term changes $\hat{\mu}$ and hence $\beta$. The orthogonality concerns $(\mu, X_c)$; the derived statistics $(\rho, \beta, P(\Sigma))$ remain coupled.

### B.5 Conditional idempotence: statement and proof

This result concerns repeated application to the *same* matrix with a fixed reference, $\alpha = 1$ and $\lambda > 0$. It is not the default rollout schedule, which corrects only newly written rows. The mean projection is idempotent because its output lies in the closed reference ball.

**Theorem 1** (sufficient conditions for idempotence)**.** *Let $\Sigma, G_{\mathrm{ref}} \succeq 0$, let $U$ contain the selected orthonormal eigenvectors of $\Sigma - G_{\mathrm{ref}}$, and put $Q = I - UU^\top$. Suppose* (i) *$U^\top \Sigma U$ is diagonal,* (ii) *$U^\top \Sigma Q = 0$, and* (iii) *$Q(\Sigma - \lambda G_{\mathrm{ref}})Q \preceq 0$. With zero-energy directions left unchanged, the full operator satisfies $H(H(X)) = H(X)$ in exact arithmetic.*

*Proof.* Since $U$ is an eigensubspace of $\Sigma - G_{\mathrm{ref}}$, $U^\top(\Sigma - G_{\mathrm{ref}})Q = 0$. Condition (ii) therefore also gives $U^\top G_{\mathrm{ref}} Q = 0$. Condition (i) and the eigenvector choice imply that $U^\top G_{\mathrm{ref}} U$ is diagonal. Both covariances consequently separate into the selected coordinates and their complement. Shrinkage changes selected variances to $\min(e_i^{\mathrm{now}}, \lambda e_i^{\mathrm{ref}})$ and leaves the complement unchanged. By (iii), $\Sigma' \preceq \lambda G_{\mathrm{ref}}$ globally. Every direction selected on the second pass therefore has shrink factor one (zero-energy directions need no change). The mean is already in its trust region, and centered shrinkage preserves that mean, proving the result. □

**Role of cross-subspace covariance.** The other two conditions alone are insufficient. For

$$\Sigma = \begin{pmatrix} 2 & 0.8 \\ 0.8 & 1 \end{pmatrix}, \qquad G_{\mathrm{ref}} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1.5 \end{pmatrix}, \qquad r = 1, \quad \lambda = 1,$$

both matrices are positive definite and $U = (1, 0)^\top$. Within-subspace diagonality is automatic, and the complement is below tolerance. The first shrink factor is $1/\sqrt{2}$, but $\lambda_{\max}(\Sigma' - G_{\mathrm{ref}}) = 0.09264 > 0$; the second shrink factor is $0.92136$. Cross-subspace covariance creates a new excess direction. This counterexample is included as a CPU regression test. The theorem supplies sufficient conditions; their validity for a particular cache requires a separate check.

### B.6 Idempotence measurements on real caches

At chunk 20, we applied the operator twice and recomputed $U$ on the second pass. The relative residual $\|H^2 - H^1\|/\|H^1\|$ is $1.6 \times 10^{-11}$ at layer 5, $1.3 \times 10^{-3}$ at layer 14, and exactly 0 at layer 25. The residual is unchanged over $r = 4, 8, \ldots, 128$, and at layer 14 the mean term contributes less than the reported precision to the second-pass correction. This establishes non-idempotence for that cache, but does not identify which sufficient condition fails. Comparing separately sorted eigenvalues of two projected covariances cannot certify a Loewner inequality. The required checks are $\lambda_{\max}(Q(\Sigma - \lambda G_{\mathrm{ref}})Q)$ and cross-subspace covariance; these quantities remain unmeasured for the historical GPU runs. Synthetic mean-only tests and the two-dimensional counterexample can be reproduced on CPU. Repeated fixed-step subtraction continues moving toward its reference; its deviation is not unbounded merely because it is non-idempotent.

### B.7 Full-rank whitened clipping (not implemented)

A full covariance envelope, $\Sigma \preceq \kappa G_{\mathrm{ref}}$, provides a stronger constraint than the selected-direction checks of the implemented operator. Assume $G_{\mathrm{ref}} \succ 0$ and $\kappa > 0$. Whitening gives $A = G_{\mathrm{ref}}^{-1/2} \Sigma G_{\mathrm{ref}}^{-1/2}$, whose eigenvalues $q_i$ measure variance relative to the reference. Let $Q$ contain a complete orthonormal eigenbasis of $A$. Clipping $q_i \leftarrow \min(q_i, \kappa)$ uses $s_i = \min(1, \sqrt{\kappa/q_i})$, with $s_i = 1$ for $q_i = 0$. Mapping back with

$$W = G_{\mathrm{ref}}^{-1/2}\, Q\, S\, Q^\top\, G_{\mathrm{ref}}^{1/2}, \qquad S = \mathrm{diag}(s_i), \tag{14}$$

gives $X_c' = X_c W$ with $W^\top \Sigma W = \Sigma' \preceq \kappa G_{\mathrm{ref}}$ in exact arithmetic. The bound holds in all directions. For a fixed reference, this centered map is idempotent: after one pass $q_i' \leq \kappa$, so every second-pass factor satisfies $s_i' = 1$. This statement uses the full eigenbasis, not a rank-$r$ truncation.

If the reference is singular, replacing it throughout by $G_{\mathrm{ref}} + \varepsilon I$, with $\varepsilon > 0$, instead certifies the regularized envelope $\Sigma' \preceq \kappa(G_{\mathrm{ref}} + \varepsilon I)$. It does not certify the unregularized envelope. This construction is an unimplemented mathematical alternative to the operator in Section 5.2; its numerical stability, computational cost and effects on generated video remain unevaluated.

The derivations justify the measurement and update rules in Sections 3 and 5; none assumes or proves perceptual improvement along a generated trajectory.

## C Diagnostic audits and historical interventions

These studies establish what changes in the measured representation and which interventions control the diagnosed cache event. We retain the historical instrumentation and distinguish accounting identities, local perturbations and rollout observations.

### C.1 Concentration attribution and large energy excursions

This section retains the numerical attribution behind Figure 1 and the separate event-count intervention study. The latter tests cache regulation; it does not label failures in the main quality cohort.

**Trajectory alignment.** The direction probe's randomized SVD advances the generation RNG state. Consequently, matching prompt indices and seeds does not align its cache events with trajectories generated under the separate output-measurement protocol. Event-to-output correlations and comparisons grouped by those event labels are therefore excluded from the paired evidence.

Table 4: Attribution of the *change* in concentration at `v_pre`. Early/late are the first and last 15% of chunks with chunk $\geq 8$. Shares of concentration terms are distinct from the energy fraction $f$. The 240 s column is $n = 2$ prompts, both of which blow up.

| | 120 s (4 prompts) | 240 s (2 blow-up prompts) |
|---|---|---|
| $P(M)$ early $\rightarrow$ late | 26.18 $\rightarrow$ 7.91 | 26.13 $\rightarrow$ 1.325 |
| common term / $C$, early $\rightarrow$ late | 0.541 $\rightarrow$ 0.835 | 0.612 $\rightarrow$ 0.972 |
| centered term / $C$, early $\rightarrow$ late | 0.386 $\rightarrow$ 0.072 | 0.312 $\rightarrow$ 0.007 |
| common energy fraction $f$, late | 0.415 | 0.872 |
| share of $\Delta C$: common | $+\mathbf{0.913}$ | $+\mathbf{0.996}$ |
| share of $\Delta C$: alignment | $+0.099$ | — |
| share of $\Delta C$: centered | $\mathbf{-0.011}$ | — |

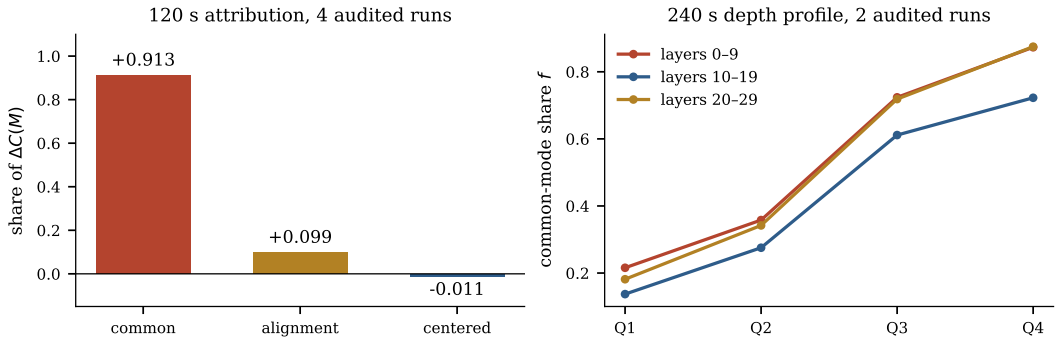



Figure 2: Left: attribution of the *change* in concentration at `v_pre` over 120 s. The common-mode channel accounts for 91.3% of the rise and the centered channel for $-1.1\%$. Right: the common-mode share by depth band and rollout quartile at 240 s; every band rises monotonically and the middle band remains lowest. Full numbers in Table 4 and Table 6.

Table 5: Large excursions of $\rho$ in deep layers (20–29) at 240 s. fold = peak/base, base = median of the first 10% of chunks, applied to the per-chunk median across layers 20–29. The $10\times$ threshold was chosen after inspecting these data, so counts and paired exact McNemar tests against open loop are exploratory. The controller modifies V only.

| Site | Arm | $\geq 10\times \downarrow$ | median fold $\downarrow$ | max fold $\downarrow$ | Paired $p$ |
|---|---|---|---|---|---|
| V | open loop | $\mathbf{8/20}$ | 3.87 | $\mathbf{466.00}$ | — |
| V | closed loop, $\lambda = 1$ | $\mathbf{0/20}$ | 1.16 | 4.49 | $\mathbf{0.0078}$ |
| V | closed loop, $\lambda = 4$ | $\mathbf{0/20}$ | 1.15 | 1.59 | $\mathbf{0.0078}$ |
| V | `constant_ref` (rank one) | $\mathbf{0/20}$ | 1.13 | 1.33 | $\mathbf{0.0078}$ |
| K | open loop | 0/20 | 1.64 | 5.60 | — |

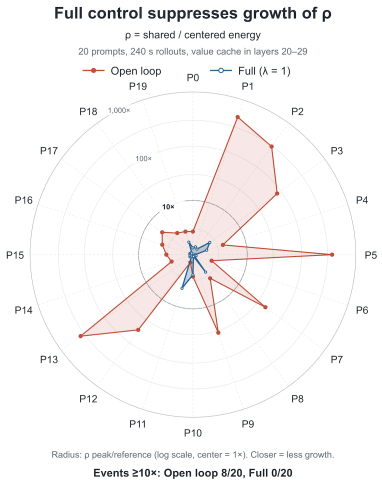



Figure 3: Prompt-level cache growth at 240 s. Each point is a prompt; $\rho$ fold is peak/reference in layers 20–29. Eight open-loop rollouts exceed $10\times$, compared with zero under Full control (paired exact McNemar $p = 0.0078$, exploratory retrospective threshold). Table 5 also reports the historical reference-directed mean arm.

### C.2 Concentration in the 240-second diagnostic runs

The two selected 240 s event prompts reach $P(M) = 1.325$, a strongly concentrated uncentered spectrum. The common term accounts for $99.6\%$ of the concentration increase, and the late common energy fraction is $87.2\%$. Neither statistic establishes a loss of algebraic rank. Every depth band rises monotonically across quartile averages, without preserving its exact shape (Table 6). Both 240 s audits concern event prompts; matched stable-prompt audits are unavailable. The comparison of attribution between stable and event trajectories remains an open experiment.

### C.3 Depth and time profile of the common-energy share

These profiles test whether the aggregate increase is confined to one depth band. The temporal averages rise in each band, but band ordering changes; this supports distributed growth rather than an invariant depth profile.

Table 6: Common-mode share $f$ at `v_pre` by depth band and rollout quartile, excluding chunks 0–7 and averaging runs equally. Left: 120 s, four runs. Right: 240 s, two selected blow-up runs. Every band rises and layers 10–19 remain lowest, but the outer bands exchange order. This is not an invariant-shape result.

| | 120 s | | | | 240 s (blow-up) | | | |
|---|---|---|---|---|---|---|---|---|
| Band | Q1 | Q2 | Q3 | Q4 | Q1 | Q2 | Q3 | Q4 |
| L0–9 | 0.1999 | 0.2284 | 0.2829 | 0.4226 | 0.2157 | 0.3580 | 0.7239 | 0.8731 |
| L10–19 | 0.1245 | 0.1502 | 0.2018 | 0.3252 | 0.1372 | 0.2755 | 0.6114 | 0.7225 |
| L20–29 | 0.1600 | 0.1855 | 0.2436 | 0.3935 | 0.1815 | 0.3419 | 0.7187 | 0.8749 |

### C.4 Algebraic consistency of the probe

The identity provides an algebraic consistency check. Across 711,030 released measurement rows from 13 runs, the stored pre-downcast relative residual has maximum $1.62 \times 10^{-15}$. Recomputing it from the released float32 scalar columns gives maximum $1.27 \times 10^{-7}$, reflecting release quantization (Table 7). Residuals beyond expected numerical error would indicate inconsistent measurements or implementation, rather than a departure from the identity. Two synthetic checks test the attribution

implementation: injecting a pure common mode yields $\mathrm{share}_\rho = 1.000$, and applying a pure channel gate yields $\mathrm{share}_\Sigma = 1.000$. These checks establish the expected separation for those constructed inputs.

### C.5 Identity residuals and static decomposition

Table 7 distinguishes the recorded and recomputed residuals; Table 8 gives the released-cohort static split of $C(M)$ into its three channels at five of the 16 instrumented probe sites, ordered by $\rho$.

Table 7: Maximum relative residual of the identity equation 5.

| Check | Volume | Max. relative residual |
|---|---|---|
| Logged identity check before release downcast | 711,030 rows | $1.62 \times 10^{-15}$ |
| Recomputed from released float32 scalars | 711,030 rows | $1.27 \times 10^{-7}$ |

Table 8: Static summary over 403,830 released rows with horizons at most 120 s (16 sites; 5 shown). Each column is a separate row-level median; median channel shares need not sum to one. The V-cache has a non-negligible centered channel. These are descriptive pooled-row statistics, not independent replicates or estimates of change.

| Site | $\rho$ | $\beta$ | $P(\Sigma)$ | $f$ | common/$C$ | align/$C$ | centered/$C$ |
|---|---|---|---|---|---|---|---|
| `v_pre` | 0.230 | 0.014 | 40.2 | 18.7% | 62.9% | 7.7% | 29.0% |
| `av` | 0.503 | 0.019 | 19.4 | 33.5% | 79.9% | 5.7% | 13.2% |
| `h_mlp` | 0.761 | 0.025 | 28.9 | 43.2% | 89.1% | 5.5% | 4.7% |
| `h_out` | 2.127 | 0.017 | 18.9 | 68.0% | 97.1% | 1.6% | 1.1% |
| `h_ln2` | 5.370 | 0.004 | 21.7 | 84.3% | 99.7% | 0.2% | 0.2% |

### C.6 Perturbation, replay and token-count controls

**Equal correction magnitude.** The fixed-budget mean and subspace arms target $\|\alpha\Delta\|_F/\|X\|_F = 0.1550$ before cache dtype conversion. This controls immediate relative correction magnitude, not the downstream trajectory. Alignment rotation has a different invariant and is excluded from this budget comparison. **Bitwise-exact replay.** Snapshots store CUDA and CPU RNG state alongside tensors, so a `null` arm reads exactly 0 everywhere. **Fixed token count.** In the i.i.d. chunk control, growing the cache from 1 to 8 chunks increases $P$ by $+8.4\%$ without changing the data-generating distribution. Token count must therefore be held fixed when interpreting participation-ratio changes as a property of the generated video.

**Definitions of historical latent proxies.** For flattened latent frame $z_t$, the released output analysis defines $D_t = 1 - \cos(z_t, z_0)$, $E_t = \mathrm{std}(z_t)/\mathrm{std}(z_0)$, and motion $m_t = \mathrm{mean}|z_t - z_{t-1}|/\mathrm{mean}_{u,j}|z_{u,j}|$. Reported tail summaries average the last $25\%$ of latent frames. Let $E$ denote the tail mean of $E_t$; the energy-deviation endpoint is $|E-1|$, with the absolute value applied after tail averaging. Historical tables retain the shorthand "energy" for $E$: it is a standard-deviation ratio, not squared feature energy or the common-to-centered energy ratio $\rho$. Lower cosine drift can also arise from a static video, so these proxies are read jointly and do not replace VBench or human evaluation.

**Direction controls at matched pre-cast magnitude.** The following table retains all recorded mean and subspace arms; its small paired sample motivates reference-directed control without establishing necessity of a particular direction.

Table 9: Historical fixed-relative-budget interventions, $n = 4$ paired prompts. The relative Frobenius budget is $0.1550$. Unadjusted two-sided tests are shown; non-significance does not establish no effect. Alignment rotation has a different invariant and is reported separately in Table 10.

| Channel | Arm | $d_{\text{drift}} \downarrow$ | $t$ | Verdict |
|---|---|---|---|---|
| $f$ | `constant_ref` (reference) | $-0.182$ | $-8.93$ | lower drift; motion $+0.110$ |
| | `constant` (zero) | $-0.130$ | $-3.55$ | lower drift; energy $+0.524$ |
| | `rand_rank1` (equal energy) | — | $-1.45$ | n.s. |
| | `shuffled` (equal energy) | — | $-1.66$ | n.s. |
| $P(\Sigma)$ | `top_svd`, $r = 1$ | $+0.102$ | $+2.63$ | n.s.; $p = 0.078$ |
| | `top_svd`, $r = 4$ / $r = 16$ | $+0.044$ / $-0.026$ | — | no demonstrated benefit |
| | `bottom_svd` / `rand_subspace` | $[-0.024, +0.102]$ | $\lvert t\rvert < 1.5$ | n.s. |

### C.7 Alignment rotation and exploratory temporal associations

`beta_rot` rotates $\mu$ relative to $\Sigma$. The CPU operator check holds $\|\mu\|$ and $f$ fixed to six digits, changes $\Sigma$ by at most $1.2 \times 10^{-7}$, spans a factor of $10.8$ in $\beta$, and matches equation 5 to $1.9 \times 10^{-7}$. Across rollout doses, within-prompt Spearman correlations average $+0.660$ for $|E - 1|$ ($20/20$ positive) and $-0.470$ for drift ($18/20$ negative). These averages describe direction, not a calibrated population test; one prompt lacks the zero-dose run. All treatment-minus-zero comparisons therefore use 19 matched prompts. The positive endpoint increases energy deviation, but decreases drift and increases motion: no scalar endpoint is a complete quality assessment. The logged endpoint angles differ ($\cos = 0.0017$ and $0.3246$); neither angular displacement nor cumulative Frobenius perturbation is matched. Preserving $\rho$ and $P(\Sigma)$ at one update does not preserve them along the feedback-coupled trajectory. A dose response alone is not causal mediation.

Table 10 reports the alignment-rotation contrasts; the separate early-to-late analysis is in Appendix C.8.

Table 10: Exploratory alignment-rotation contrasts at $240\,\text{s}$, $n = 19$ matched prompts per nonzero dose. The zero-dose arm has one missing record. Unadjusted $t$ statistics are shown; the release includes $95\%$ intervals and Holm correction across all 12 dose–metric tests. Endpoint cosines are historical operator diagnostics, not matched trajectory-level budgets.

| level | $\cos(\mu, \mu')$ | $\Delta\lvert E - 1\rvert \downarrow$ | $t$ | $\Delta\text{drift} \downarrow$ | $t$ | $\Delta\text{motion}$ | $t$ |
|---|---|---|---|---|---|---|---|
| $-1.0$ | $0.0017$ | $-0.011$ | $-0.15$ | $-0.003$ | $-0.13$ | $+0.081$ | $2.86$ |
| $-0.5$ | — | $+0.130$ | $1.92$ | $-0.016$ | $-0.70$ | $-0.024$ | $-1.14$ |
| $+0.5$ | — | $+0.506$ | $5.63$ | $-0.073$ | $-2.48$ | $+0.059$ | $1.98$ |
| $+1.0$ | $0.3246$ | $\mathbf{+0.575}$ | $4.91$ | $\mathbf{-0.111}$ | $-3.64$ | $\mathbf{+0.419}$ | $11.73$ |

### C.8 Exploratory early-to-late associations

The open-loop diagnostic logs $\|\mu_c - \mu_{\text{ref}}\|$ without the randomized direction probe. Across 19 corresponding output runs, its first-quarter growth rate correlates with late energy deviation ($\rho_s = 0.593$, raw $p = 0.00745$). A retrospective family of four available early features against two late targets gives Holm $p = 0.0596$ for this association; none of the eight tests survives that correction. We report these results as exploratory, not as a validated predictor or evidence that one mechanism uniquely explains drift. There is no held-out prediction experiment, and the cache-probe blow-up labels are not transferred across harnesses (Table 11).

Table 11: Exploratory early-to-late associations, 19 corresponding open-loop output runs at 240 s. Four early features and two late targets form one retrospective family of eight tests. No association survives Holm correction.

| Early feature | Late target | $\rho_s$ | Raw $p$ | Holm $p$ |
|---|---|---|---|---|
| Median cache deviation | $\lvert E-1\rvert$ | $+0.511$ | 0.0255 | 0.179 |
| Median cache deviation | late drift | $+0.053$ | 0.831 | 1 |
| Deviation / reference band | $\lvert E-1\rvert$ | $-0.077$ | 0.753 | 1 |
| Deviation / reference band | late drift | $+0.402$ | 0.0882 | 0.529 |
| Cache-deviation slope | $\lvert E-1\rvert$ | $+0.593$ | 0.00745 | 0.0596 |
| Cache-deviation slope | late drift | $-0.021$ | 0.932 | 1 |
| Early latent drift | $\lvert E-1\rvert$ | $+0.132$ | 0.591 | 1 |
| Early latent drift | late drift | $-0.247$ | 0.307 | 1 |

### C.9 Structural sources and transmission of a shared mean

In the inspected backbone, AdaLN's additive shifts are broadcast over spatial tokens within a frame, providing a structural source of common-component updates. A shift can reinforce or oppose the incoming mean, so its presence alone does not imply energy growth. For a fixed attention matrix $A$ and a shift shared by all attended values, softmax row sums give $A\mathbf{1} = \mathbf{1}$:

$$A(\mathbf{1}c^\top + \bar{V}) = \mathbf{1}c^\top + A\bar{V}. \tag{15}$$

Here $c$ is a shared feature vector and $\bar{V}$ denotes the remaining values. The shift passes through this mixing step unchanged. Distinct chunk means are instead weighted by the attention assigned to their chunks. This local algebra neither gives every cached chunk unit gain nor proves instability of a full residual block or an autoregressive rollout.

### C.10 Exploratory checks of candidate mechanisms

Table 12: Historical exploratory checks that informed the method. These summaries are not new experiments, universal refutations, or independent replications. In particular, null tests in F5 do not establish equivalence, and fixed gates in F3 can still affect time-varying inputs.

| # | Hypothesis | Measurement |
|---|---|---|
| F1 | Common mode accumulates across layers, $\rho_L \sim O(L)$ | $\rho$ *falls* with depth, $4.12 \to 1.30$ |
| F2 | The AdaLN additive shift raises $\rho$ | $\mathrm{d}\log\rho = -0.116$; it lowers it by $11\%$ |
| F3 | Channel gates compound multiplicatively across layers | Gates are bitwise identical across the whole rollout ($\max\lvert\Delta e\rvert = 0$), so gate values do not themselves vary with AR time |
| F4 | The cache-error recursion is linear | Gain $\times$ perturbation is constant over a $10\times$ range: the recursion saturates |
| F5 | The controller works by lowering propagation gain | 4 of 103 matched pairs cross the threshold; $t = -0.136$ |
| F6 | There exist high-gain "dangerous directions" | Gain is nearly isotropic (max/median 1.05–1.32) and uncorrelated with energy share ($\lvert r\rvert \le 0.13$) |
| F7 | $\rho$ is stationary at its working point | Observed mean at 240 s goes $0.208 \to 12.24$ (restriction, not refutation) |

### C.11 Data coverage and reproducibility

The CPU release contains 711,030 operator-audit rows from 13 runs, 1,612,800 direction-probe scalar rows from 90 runs, and per-prompt output and VBench tables. Coverage is reported separately for each analysis. The operator audits include four 120 s and two selected 240 s runs. The historical statistical reanalysis reuses recorded runs; the separately frozen pilot below is analyzed as a distinct

cohort. Seed reproducibility requires the same instrumentation and RNG protocol; a seed alone cannot align trajectories produced by different probes.

# D COMPLETE LONG-VIDEO EVALUATION AND DEVELOPMENTAL EVIDENCE

The complete main evaluation answers the output-quality question separately from the cache-event study. Primary tests, secondary time profiles and fixed examples are followed by the earlier developmental results; their sample sizes and inferential status remain distinct.

## D.1 COMPLETE MAIN EVALUATION AND ANALYSIS SCOPE

The frozen resource fallback uses the first $50$ prompts of a pre-sealed $100$-prompt list, two seeds and Native/Mean/Full/Sink/Sink+Full at nominally $240$ s. All $500$ assigned trajectories were scored. There are no terminal failures or missing seed pairs: each method contributes $100$ trajectories, and every primary comparison uses all $50$ paired prompts. Subject motion, camera motion, scene change and stable scenes contribute 13, 13, 12 and 12 prompts, respectively.

**Fixed paired analysis.** Every trajectory has $3{,}837$ frames at $16$ fps, partitioned into $20$ non-overlapping windows ($19 \times 192 + 189$). All windows are weighted by their frame counts within a trajectory, and the two seeds are averaged equally within a prompt. The prompt is the statistical unit. The six Full–Mean quality/text tests form one two-sided paired-$t$ Holm family; their $95\%$ intervals are unadjusted. Dynamic degree and the other method contrasts are descriptive. No assignment, window or seed was replaced or removed to obtain the reported coverage. Table 13 retains every primary endpoint, including unfavorable or non-significant differences. The released evidence also retains all $70{,}000$ metric-window records, method and scene-stratum summaries, and the complete assignment ledger.

**Secondary uncertainty beyond the primary Full–Mean tests.** Table 14 reports all 21 secondary paired estimates from the main quality analysis: Full–Native, Sink+Full–Sink, and Sink+Full–Full across seven dimensions. Its unadjusted intervals quantify paired variability without changing the six-test primary family. Additional post-hoc Mean–Native and Mean–Sink paired intervals are reported in Table 2 and Appendix D.2; they do not alter the primary testing family.

**Current-controller sustained energy descriptions.** The fixed secondary rule uses post-control values on the same trajectories as the main quality evaluation. For each clean-write chunk, it takes the median across the 30 recorded layers of $\|\mu\|^2$ and, separately, $\operatorname{tr}\Sigma$. For a video window starting at frame $s$, it then averages preceding writes satisfying $s-192 < 12(c+1)-3 \le s$. The first window has no preceding cache support. The analysis reference is the median of the first three available support values, covering $0$–$36$ s; it differs from controller reference chunks 2–6. An event requires at least twice this energy reference for three consecutive post-reference $12$ s supports. Confirmation occurs at the end of the third support, not the start of the first.

Under this rule, sustained common-energy events occur in $86$ of $100$ Native trajectories, $0$ of $100$ Mean trajectories, $0$ of $100$ Full trajectories, $2$ of $100$ Sink trajectories, and $0$ of $100$ Sink+Full trajectories. No sustained centered-energy event occurs in any of the five arms. These are trajectory counts, with two dependent seeds per prompt. The rules were fixed after $92$ trajectories' scores existed; these descriptions have no significance test or confirmatory event-rate claim. They also differ from the retrospective deep-layer $\rho$ threshold in Table 5. Cache non-events are right-censored at $228$ s and do not certify normal videos; quality declines can occur without a cache event. The supplementary analysis reports all $18$ sensitivity settings and both event orders.

**Same-trajectory descriptions.** The cache analysis joins logs and scores from these same videos, without additional generation. Table 15 summarizes common-energy/aesthetic associations using preceding cache and the fixed reverse-time comparison. Each trajectory uses the logarithm of post-control common energy and the same 17 quality windows (zero-based indices $w = 1, \ldots, 17$). Forward pairs use cache feature $x_w$; reverse pairs use $x_{w+2}$, whose cache support follows the quality window. Both seeds are averaged within a prompt before prompts are averaged. Every aesthetic coefficient in the table has $50$ complete two-seed prompts and no undefined pairs; the $10{,}000$ total

Table 13: Frozen primary Full-minus-Mean comparison: six two-sided paired $t$ tests with one six-test Holm correction. The paired unit is a prompt after equal averaging of its two fixed seeds. Intervals are unadjusted; $d_z$ is descriptive. All 50 prompts have both seeds; no pair is missing.

| Endpoint | $n$ | $\Delta$ | 95% CI | $p_{\mathrm{Holm}}$ | $d_z$ |
|---|---|---|---|---|---|
| Subject | 50 | -0.011 | [-0.017, -0.004] | 0.0129 | -0.449 |
| Background | 50 | -0.004 | [-0.007, -0.001] | 0.0754 | -0.344 |
| Smoothness | 50 | 0.001 | [-0.001, 0.002] | 0.4665 | 0.171 |
| Aesthetic | 50 | 0.008 | [-0.011, 0.027] | 0.4665 | 0.119 |
| Imaging | 50 | -0.028 | [-0.063, 0.007] | 0.3281 | -0.231 |
| Text | 50 | 0.024 | [0.011, 0.037] | 0.0022 | 0.542 |

Table 14: Secondary paired contrasts in the main quality receipt. Differences use the same score units as Table 1; seeds are averaged before pairing 50 prompts. Intervals are unadjusted 95% paired-$t$ intervals, without a secondary significance-testing family. Dynamic degree measures activity. These estimates are separate from the six primary Full–Mean tests and do not add confirmatory endpoints.

| Endpoint | $n$ | $\Delta$ | 95% CI |
|---|---|---|---|
| Full minus Native | | | |
| Subject | 50 | $-0.0117$ | $[-0.0178, -0.0057]$ |
| Background | 50 | $-0.0119$ | $[-0.0152, -0.0085]$ |
| Smoothness | 50 | $+0.0019$ | $[0.0005, 0.0034]$ |
| Dynamic degree | 50 | $+0.1056$ | $[0.0358, 0.1753]$ |
| Aesthetic | 50 | $+0.0318$ | $[0.0068, 0.0569]$ |
| Imaging | 50 | $-0.0349$ | $[-0.0793, 0.0094]$ |
| Text | 50 | $+0.0642$ | $[0.0488, 0.0795]$ |
| Sink+Full minus Sink | | | |
| Subject | 50 | $+0.0047$ | $[-0.0029, 0.0124]$ |
| Background | 50 | $+0.0040$ | $[0.0007, 0.0072]$ |
| Smoothness | 50 | $+0.0018$ | $[0.0004, 0.0033]$ |
| Dynamic degree | 50 | $-0.0795$ | $[-0.1436, -0.0155]$ |
| Aesthetic | 50 | $-0.0401$ | $[-0.0592, -0.0210]$ |
| Imaging | 50 | $-0.0833$ | $[-0.1243, -0.0423]$ |
| Text | 50 | $-0.0048$ | $[-0.0123, 0.0027]$ |
| Sink+Full minus Full | | | |
| Subject | 50 | $-0.0180$ | $[-0.0249, -0.0111]$ |
| Background | 50 | $-0.0117$ | $[-0.0155, -0.0078]$ |
| Smoothness | 50 | $-0.0028$ | $[-0.0037, -0.0019]$ |
| Dynamic degree | 50 | $-0.0079$ | $[-0.0869, 0.0712]$ |
| Aesthetic | 50 | $+0.0692$ | $[0.0503, 0.0881]$ |
| Imaging | 50 | $+0.0180$ | $[0.0034, 0.0326]$ |
| Text | 50 | $+0.0472$ | $[0.0358, 0.0585]$ |

windows are not independent statistical replicates. The evidence includes all seven score dimensions, raw, linearly detrended and first-difference coefficients, undefined cases, default events and all 18 fixed sensitivity settings. These secondary rules were fixed after 92 scores existed. They provide descriptions of the completed cohort, not confirmatory causal tests or a validated warning rule. Legitimate changes in scene or camera motion can affect both cache statistics and scores. The separate six-prompt pilot and its fixed-frame scene review remain in Appendix D.11.

**Relationship to the remaining evidence.** The development controls (Appendix E.1), published systems under their original protocols (Appendix E.5), transfer studies (Appendix F), and adaptation experiments (Appendix G.1) remain separate from this main comparison. Their completed results and remaining limitations are retained there, without extending the main cohort's confirmatory scope.

Table 15: Descriptive common-energy–aesthetic association on the same final videos. F uses preceding cache; R uses following cache under the fixed reverse comparison. Coefficients are averaged within complete two-seed prompts and then across eligible prompts. All seven dimensions, undefined-pair counts, default events and 18 fixed sensitivity settings are retained in the companion data. Rules were fixed after 92 scores existed; these associations are neither causal nor a validated early warning.

| Method | F raw | F detr. | F diff. | R raw | R detr. | R diff. |
|---|---|---|---|---|---|---|
| Native | -0.697 | -0.314 | -0.140 | -0.673 | -0.307 | -0.096 |
| Mean | -0.118 | -0.082 | 0.018 | -0.041 | -0.017 | 0.008 |
| Full | 0.071 | 0.030 | 0.052 | 0.111 | 0.078 | 0.066 |
| Sink | -0.239 | -0.044 | 0.100 | -0.216 | -0.048 | 0.113 |
| Sink+Full | 0.198 | 0.157 | 0.053 | 0.211 | 0.194 | 0.129 |

### D.2 MEAN'S PAIRED VALUE AND PREDEFINED SCENE STRATA

This additional analysis reuses all $500$ completed videos after the main results were known. Its plan was saved before calculating these new contrasts; it is not pre-data preregistration. Mean–Native and Mean–Sink are computed for all seven dimensions, overall and within the original subject-motion, camera-motion, scene-change and stable-scene labels ($13/13/12/12$ prompts). Table 2 gives the overall estimates; the next two tables retain every stratum. All $70$ intervals are marginal, unadjusted paired-$t$ intervals, without simultaneous coverage or a new significance claim. The original six Full–Mean tests and their Holm family remain unchanged.

Every score uses all $20$ frame-weighted windows, then equally averages the two seeds within each prompt. For a comparison with $n$ prompts, let $d_i$ be that prompt's two-seed mean difference. We report $\bar{d} \pm t_{0.975,n-1} s_d/\sqrt{n}$; windows and seeds are not independent replicates. The source retains all prompt-level differences and hashes. Reconstruction from the $70{,}000$ raw scores matches all $1{,}750$ released prompt–arm–metric means; an independent beta-CDF calculation verifies the Student-$t$ quantiles and intervals.

The comparisons specify what inexpensive shared-component control buys. Relative to Native, Mean has higher aesthetic and text means, with lower background consistency. Relative to Sink, Mean retains higher subject and background consistency but has lower aesthetic and text scores in every stratum. For subject-motion prompts, its dynamic-degree difference from Sink is $-0.2289$ ($95\%$ interval $[-0.3962, -0.0616]$). That reduction is activity, not a certified improvement in motion fidelity. Camera-motion aesthetic and stable-scene aesthetic point estimates differ relative to Native, while camera-motion and scene-change background consistency both decrease. Neither this variation nor overlapping intervals constitutes an interaction test. No stratum establishes a general preference for Mean over Sink.

Table 16: Mean–Native within every pre-existing scene stratum. Each cell gives $\Delta$ and its marginal, unadjusted $95\%$ paired-$t$ interval. All seven dimensions and all four strata are retained; these post-hoc estimates neither test interactions nor establish scene-specific superiority.

| | Subject motion $n = 13$ | Camera motion $n = 13$ | Scene change $n = 12$ | Stable scene $n = 12$ |
|---|---|---|---|---|
| Subject | $+0.0032$ $[-0.0108, +0.0171]$ | $-0.0082$ $[-0.0168, +0.0004]$ | $-0.0053$ $[-0.0139, +0.0033]$ | $+0.0056$ $[-0.0021, +0.0134]$ |
| Background | $-0.0053$ $[-0.0127, +0.0020]$ | $-0.0163$ $[-0.0242, -0.0083]$ | $-0.0103$ $[-0.0165, -0.0042]$ | $+0.0002$ $[-0.0031, +0.0036]$ |
| Smoothness | $-0.0006$ $[-0.0046, +0.0033]$ | $-0.0000$ $[-0.0050, +0.0050]$ | $+0.0020$ $[-0.0010, +0.0050]$ | $+0.0029$ $[+0.0006, +0.0052]$ |
| Aesthetic | $+0.0125$ $[-0.0150, +0.0400]$ | $+0.0714$ $[+0.0245, +0.1184]$ | $+0.0216$ $[-0.0058, +0.0490]$ | $-0.0134$ $[-0.0590, +0.0323]$ |
| Imaging | $+0.0264$ $[-0.0031, +0.0560]$ | $+0.0400$ $[-0.0130, +0.0930]$ | $-0.0194$ $[-0.1030, +0.0643]$ | $-0.0811$ $[-0.2115, +0.0492]$ |
| Text | $+0.0300$ $[+0.0057, +0.0544]$ | $+0.0561$ $[+0.0318, +0.0804]$ | $+0.0564$ $[+0.0385, +0.0742]$ | $+0.0174$ $[-0.0143, +0.0491]$ |
| Dynamic degree | $+0.0866$ $[-0.0444, +0.2176]$ | $+0.2694$ $[+0.0642, +0.4745]$ | $+0.0419$ $[-0.1082, +0.1921]$ | $-0.0852$ $[-0.1722, +0.0017]$ |

Table 17: Mean–Sink within every pre-existing scene stratum. Each cell gives $\Delta$ and its marginal, unadjusted $95\%$ paired-$t$ interval. All seven dimensions and all four strata are retained; these post-hoc estimates neither test interactions nor establish scene-specific superiority.

| | Subject motion $n = 13$ | Camera motion $n = 13$ | Scene change $n = 12$ | Stable scene $n = 12$ |
|---|---|---|---|---|
| Subject | $+0.0486$ $[+0.0281, +0.0691]$ | $+0.0379$ $[+0.0151, +0.0606]$ | $+0.0339$ $[+0.0168, +0.0510]$ | $+0.0109$ $[+0.0018, +0.0200]$ |
| Background | $+0.0262$ $[+0.0139, +0.0385]$ | $+0.0184$ $[+0.0064, +0.0304]$ | $+0.0222$ $[+0.0109, +0.0334]$ | $+0.0107$ $[+0.0067, +0.0147]$ |
| Smoothness | $+0.0073$ $[+0.0037, +0.0110]$ | $+0.0050$ $[+0.0022, +0.0077]$ | $+0.0017$ $[-0.0013, +0.0047]$ | $+0.0005$ $[-0.0019, +0.0028]$ |
| Aesthetic | $-0.0679$ $[-0.1064, -0.0294]$ | $-0.0956$ $[-0.1391, -0.0521]$ | $-0.1318$ $[-0.1633, -0.1003]$ | $-0.1799$ $[-0.2124, -0.1475]$ |
| Imaging | $-0.0283$ $[-0.0647, +0.0081]$ | $-0.0245$ $[-0.0570, +0.0080]$ | $-0.1072$ $[-0.1791, -0.0353]$ | $-0.1409$ $[-0.2526, -0.0291]$ |
| Text | $-0.0646$ $[-0.0853, -0.0438]$ | $-0.0541$ $[-0.0861, -0.0221]$ | $-0.0791$ $[-0.1019, -0.0563]$ | $-0.1091$ $[-0.1401, -0.0782]$ |
| Dynamic degree | $-0.2289$ $[-0.3962, -0.0616]$ | $-0.0618$ $[-0.2262, +0.1026]$ | $-0.0730$ $[-0.2104, +0.0643]$ | $-0.0083$ $[-0.0340, +0.0174]$ |

### D.3 CONTINUOUS CACHE GROWTH, EVENT TIMING AND QUALITY

This second added post-hoc plan compares Native, Mean and Sink on their same 300 videos. It fixes each energy baseline to the median of the first three available 12 s cache supports (0–36 s); late energy is the arithmetic mean over the last five of 19 eligible supports (168–228 s). Maximum energy uses all 19. Each trajectory's ratio uses its own baseline. We average log ratios across two seeds within prompt and then across 50 prompts; exponentiation gives the geometric summaries in Table 18. The baseline is arm-specific and already includes control, not a verified healthy or untreated reference. Common and centered energies are separate layer-median scalars rather than spectra of a pooled covariance matrix.

Table 18: Continuous growth beyond event counts. Ratios are geometric means across 50 prompt units, each averaging two seed log ratios. Post means after the cache-write intervention; pre means before it.

| Arm | Post common late | Post common max | Pre common late | Post centered late |
|---|---|---|---|---|
| Native | 3.8109 | 4.5144 | 3.8109 | 0.4970 |
| Mean | 1.0065 | 1.0422 | 1.2582 | 1.1230 |
| Sink | 1.1922 | 1.3773 | 1.1922 | 1.1097 |

The two planned descriptive log-ratio comparisons give Sink/Mean late growth of 1.1845 (95% interval $[1.1355, 1.2356]$) for common energy and 0.9881 ($[0.9613, 1.0157]$) for centered energy. These intervals are unadjusted, with no new $p$ values. All 600 prompt-level growth rows have both seeds and finite measurements. Mean's post-write stability includes its direct projection. Before that projection, Mean's late common-energy ratio exceeds Sink's; their absolute late pre-write energies are 260.73 and 222.35. Consequently, the post-write result does not establish uniformly lower drift throughout the generator or a sufficient condition for higher video quality.

**Event timing with right-censoring.** We retain the prior twice-baseline, three-support rule and its post-92-score registration status. Confirmation is the end of the third consecutive support. Native has 86 confirmed events, Mean none and Sink two. Their event-only median confirmation times are 120 s, undefined and 192 s, respectively; conditioning on event occurrence makes these unsuitable as paired effect estimates. Restricted confirmation time assigns the 228 s endpoint to unconfirmed and no-event traces and otherwise uses observed confirmation. Its equal-prompt means are 138.96, 228.00 and 227.28 s. This describes observed follow-up without a confirmed operational event, not time to perceptual failure. Three Native and one Sink terminal crossings remain unconfirmed; no-event trajectories are not certified healthy.

**Quality and associations on the same trajectories.** The late quality block uses windows 15–19 (180–239.8125 s), whose preceding cache supports cover 168–228 s. Late aesthetic means are $0.3487/0.3786/0.5301$ for Native/Mean/Sink, and late text means are $0.0788/0.1267/0.2240$. Thus Sink retains higher late aesthetic/text quality than Mean despite greater post-write common growth. The release reports all seven full and late quality dimensions, not just these favorable examples. For Sink, common-energy/aesthetic correlations are $-0.2389$ raw, $-0.0443$ after linear detrending and $+0.1002$ after differencing; reverse-time values are $-0.2160$, $-0.0482$ and $+0.1126$. These use the same 17 eligible windows and 50 complete two-seed prompt means. Attenuation and reverse-time symmetry again limit a warning or mediation interpretation. Every metric's finite and undefined coefficients is retained, without a lag or threshold search.

### D.4 INDEPENDENT LONG-VIDEO METRIC TRAJECTORIES

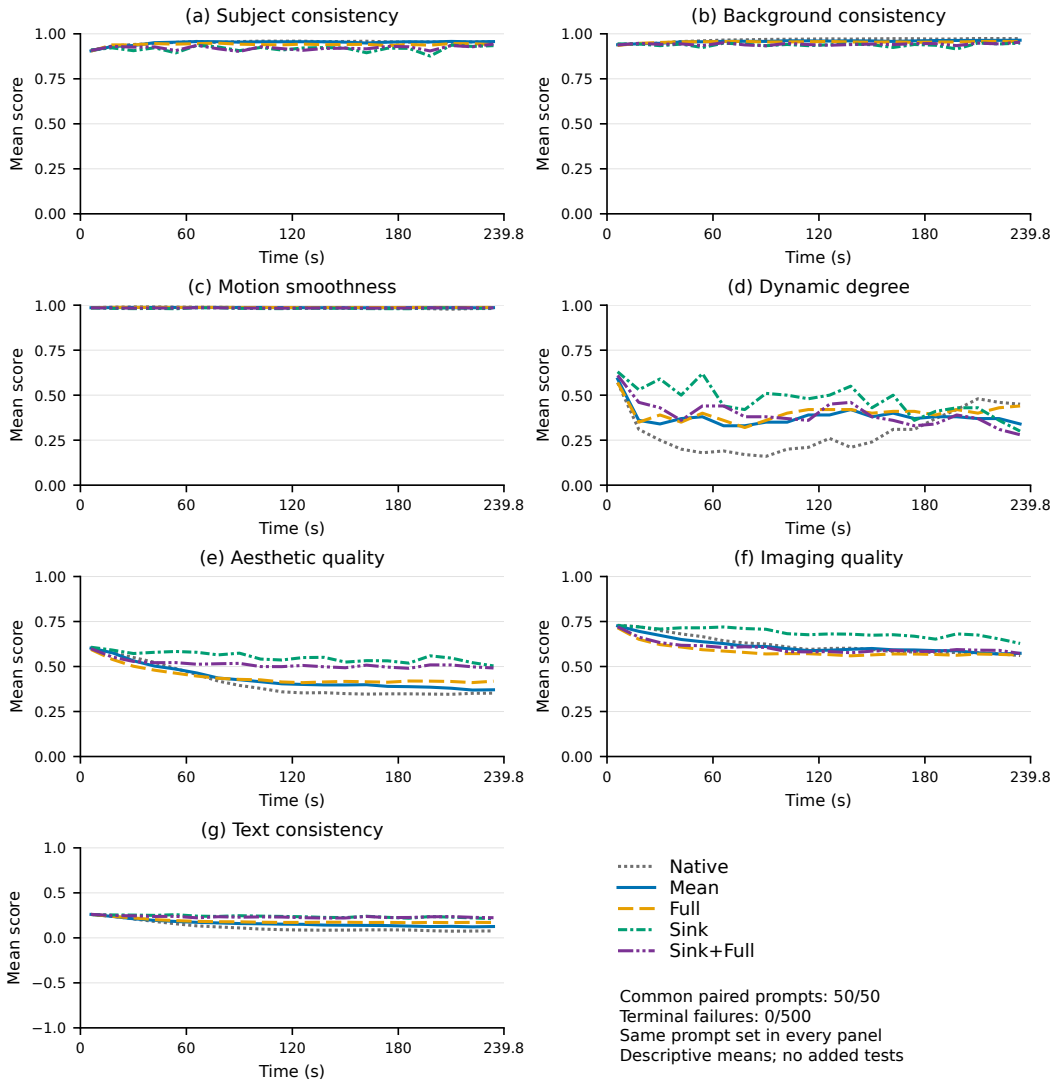



Figure 4: Long-video quality and motion for the common set of 50/50 prompts with both seeds completed for all five methods. At each of 20 windows, we average the two seeds within a prompt, then average prompts equally. Points mark window centers at 16 fps; lines connect the means without smoothing. Scores retain the normalization used in the main evaluation. The text-cosine scale is [-1,1]; other scores use [0,1]. Terminal failures: 0/500 trajectories (Native 0, Mean 0, Full 0, Sink 0, Sink+Full 0; 100 assigned per method). All assigned outcomes remain in the coverage report. These complete-case curves are descriptive: missing prompts are not imputed, windows are repeated observations, and greater dynamic degree does not imply better motion.

### D.5 FIXED EXAMPLES FROM THE MAIN EVALUATION

Figures 5 and 6 show the first two prompts in the frozen cohort, using seed 20270117 and all five methods. The prompt identities, seed and three timestamps were fixed before inspecting these frames or their scores. These examples do not estimate cohort performance or failure rates. Each panel retains the full frame and aspect ratio; images are uniformly downsampled for display, without cropping or enhancement. The source archive includes the original full-resolution stills.

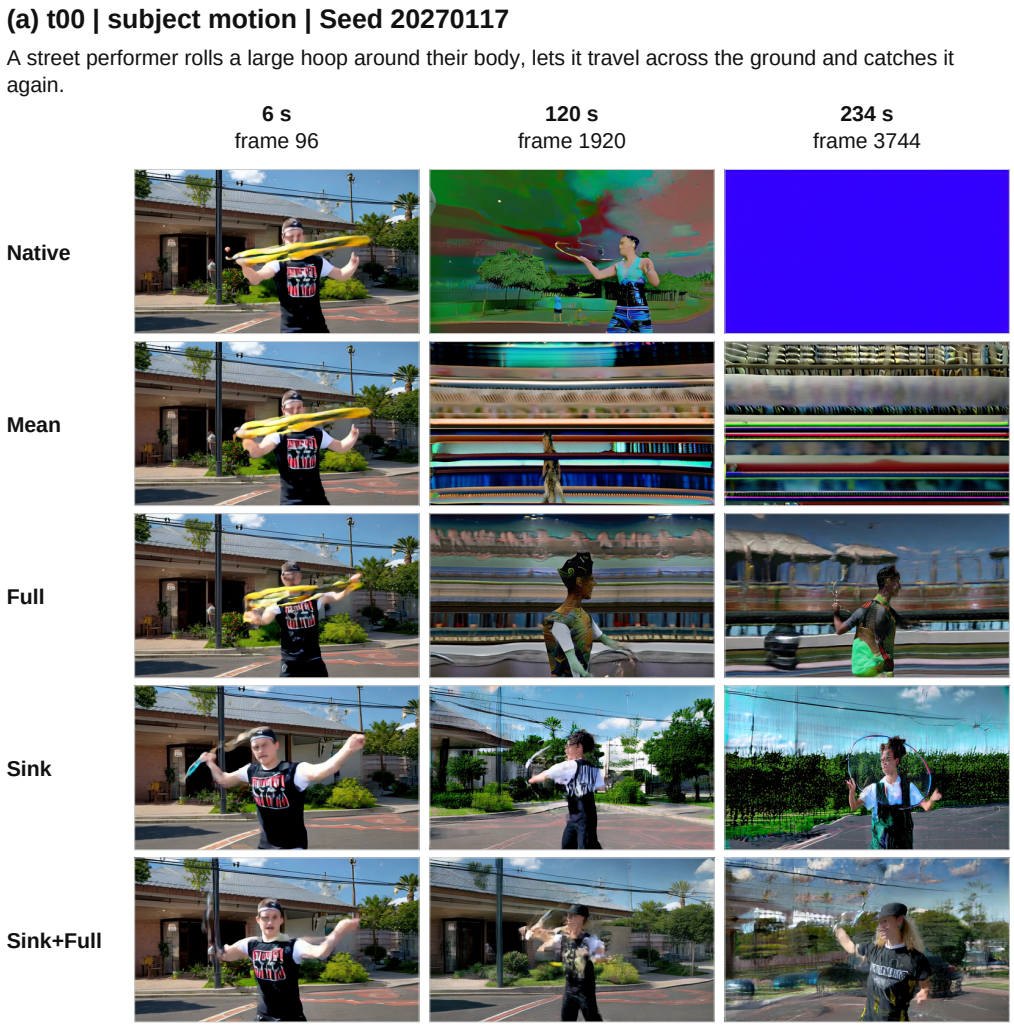



Figure 5: Fixed subject-motion example, cohort prompt t00. Rows show Native, Mean, Full, Sink and Sink+Full; columns show 6, 120 and 234 seconds, corresponding to zero-based frames 96, 1920 and 3744 at 16 fps. All methods use the same prompt and seed.

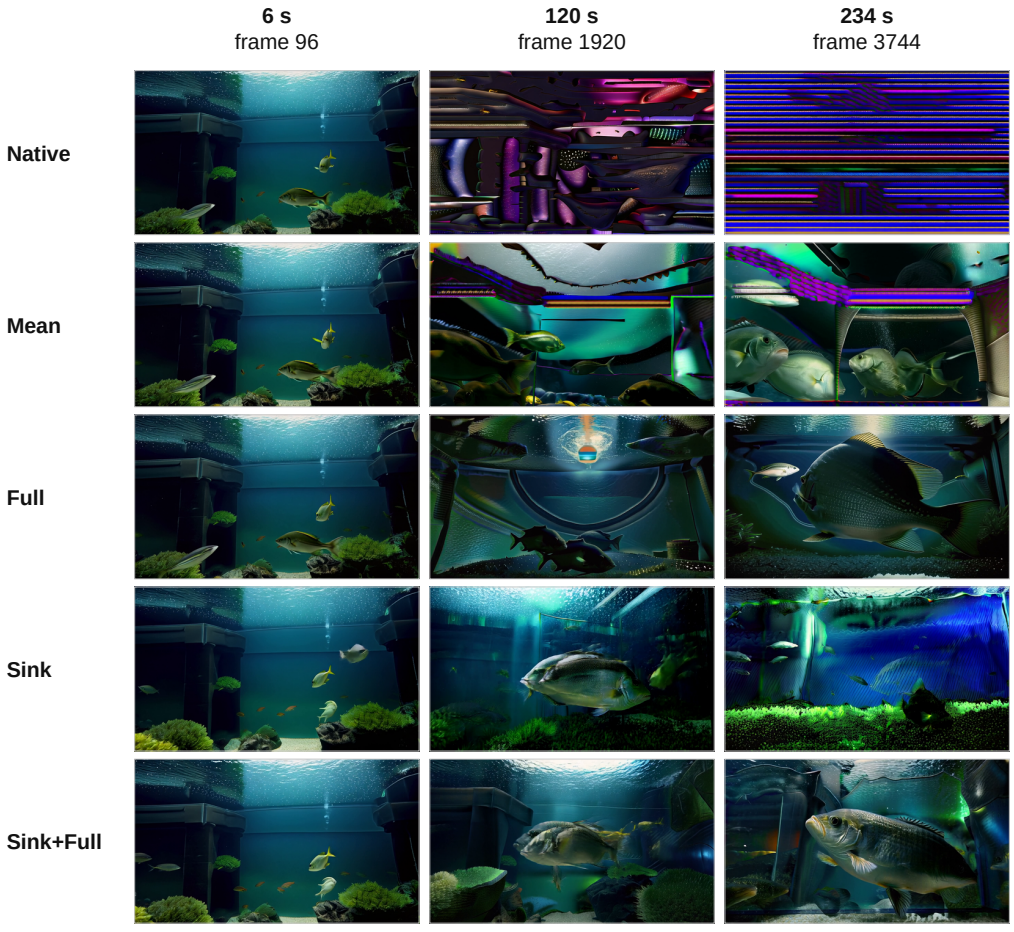



Figure 6: Fixed camera-motion example, cohort prompt t01. Methods, seed and timestamps follow Figure 5. Prompted camera and scene changes require interpretation alongside visible artifacts; isolated stills do not measure motion quality.

### D.6 Paired uncertainty, multiplicity and missing records

Reported paired-$t$ intervals are two-sided $95\%$ intervals for treatment minus control. Pairing retains prompt, seed and horizon within a fixed protocol; duplicate units are rejected. Seed 0 has 19 open-loop records and 20 full-controller records. Seeds 1 and 2 have 20 per arm. The three-seed summary averages effects within the 19 prompts with all six arm–seed records. It does not treat 59 prompt–seed pairs as independent prompts. Missing outcomes are not imputed as successes.

The nine seed–proxy tests form one Holm family. The three full-versus-mean tests form another; the three prompt-averaged tests form a sensitivity family. Each VBench comparison uses six dimensions as one family. The eight early-warning tests and twelve alignment-dose tests have their own families. These families were defined during retrospective reanalysis, not preregistered; adjustment does not erase configuration selection, metric selection or reuse of prompts. The release also gives Wilcoxon sensitivity tests, unmatched counts and input hashes. No global confirmatory claim across all exploratory experiments is made.

Table 19: Paired output-proxy estimates at 240 s. Seed rows compare full control to open loop. The full-versus-mean ablation uses 20 prompts. Prompt-averaged rows use 19 complete prompts, averaging three seeds first.

| Comparison, metric | $n$ | $\Delta$ | 95% CI | Raw $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Seed 0, $\lvert E-1\rvert$ | 19 | $-0.896$ | $[-1.124, -0.669]$ | 1.5e-07 | 1.0e-06 |
| Seed 0, drift | 19 | $-0.197$ | $[-0.290, -0.103]$ | 3.2e-04 | 0.0016 |
| Seed 0, motion | 19 | $+0.100$ | $[+0.021, +0.179]$ | 0.016 | 0.032 |
| Seed 1, $\lvert E-1\rvert$ | 20 | $-0.755$ | $[-0.932, -0.579]$ | 3.1e-08 | 2.7e-07 |
| Seed 1, drift | 20 | $-0.173$ | $[-0.262, -0.083]$ | 7.1e-04 | 0.00284 |
| Seed 1, motion | 20 | $+0.160$ | $[+0.075, +0.246]$ | 8.9e-04 | 0.00284 |
| Seed 2, $\lvert E-1\rvert$ | 20 | $-0.936$ | $[-1.164, -0.709]$ | 5.5e-08 | 4.4e-07 |
| Seed 2, drift | 20 | $-0.195$ | $[-0.271, -0.120]$ | 3.1e-05 | 1.9e-04 |
| Seed 2, motion | 20 | $+0.129$ | $[+0.024, +0.233]$ | 0.0182 | 0.032 |
| Full vs mean, $\lvert E-1\rvert$ | 20 | $-0.242$ | $[-0.347, -0.137]$ | 1.2e-04 | 3.5e-04 |
| Full vs mean, drift | 20 | $-0.114$ | $[-0.177, -0.051]$ | 0.00123 | 0.00245 |
| Full vs mean, motion | 20 | $+0.050$ | $[+0.013, +0.086]$ | 0.00986 | 0.00986 |
| Prompt-averaged, $\lvert E-1\rvert$ | 19 | $-0.856$ | $[-1.047, -0.665]$ | 2.2e-08 | 6.6e-08 |
| Prompt-averaged, drift | 19 | $-0.188$ | $[-0.274, -0.103]$ | 2.1e-04 | 4.3e-04 |
| Prompt-averaged, motion | 19 | $+0.127$ | $[+0.039, +0.215]$ | 0.00725 | 0.00725 |

### D.7 Development ablation: mean-only versus full control

The reference-directed mean intervention gives $0/20$ observed cache events (Table 5). A separate output-proxy ablation compares the current Mean trust-region controller with Full to measure the incremental effect of covariance control. At 240 s, keeping the mean trust region and $\alpha$ fixed, the direct full-minus-mean comparison on 20 prompts gives energy-deviation difference $-0.242$ ($p = 0.0001$), drift difference $-0.114$ ($p = 0.0012$), and motion difference $+0.050$ ($p = 0.0099$). All three survive Holm correction within this ablation family (Figure 7). Full control therefore reduces energy deviation and drift while increasing motion in this ablation. These latent diagnostics motivate the separate decoded-video comparison in Appendix D.10.

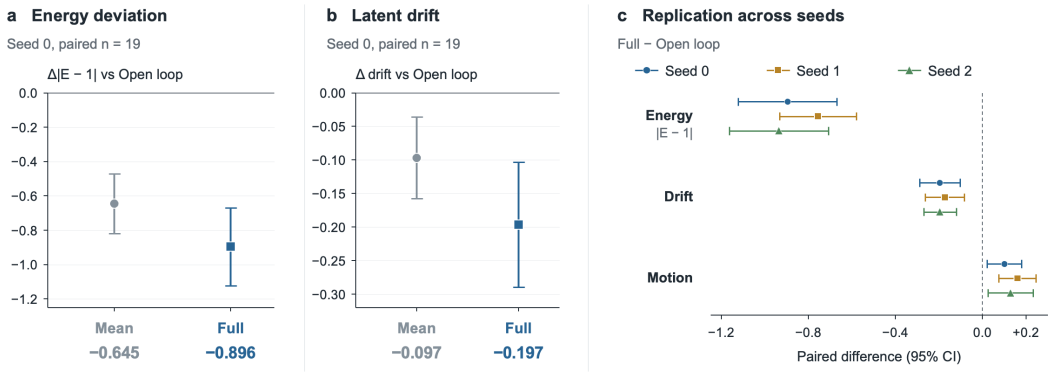



Figure 7: Mean-only and full control at 240 s. Left/center show paired output-proxy differences from open loop with 95% intervals ($n = 19$); the direct full–mean comparison uses $n = 20$. Right shows replication on two additional seeds with the same prompt set. "Term 2" is covariance correction; its proxy benefit is not a proof that it is necessary for visual quality.

Full-versus-native effects also repeat on seeds 1 and 2. Across the three seeds, all nine energy/drift/motion comparisons survive one Holm family ($p_{\text{Holm}} \leq 0.032$). Averaging seed effects within 19 complete prompts gives differences of $-0.856$, $-0.188$ and $+0.127$, respectively, with all three intervals excluding zero (Appendix D.6). These are seed replications on previously explored prompts, not independent prompt generalization. The separate six-prompt pilot has no Holm-significant result among its nine primary proxy tests (Appendix D.9).

### D.8 Horizon scaling and composition with token selection

Figure 8 uses the available per-prompt horizon records. Table 20 retains selected historical sink-sweep summaries; their underlying per-prompt inputs are not included in this release.

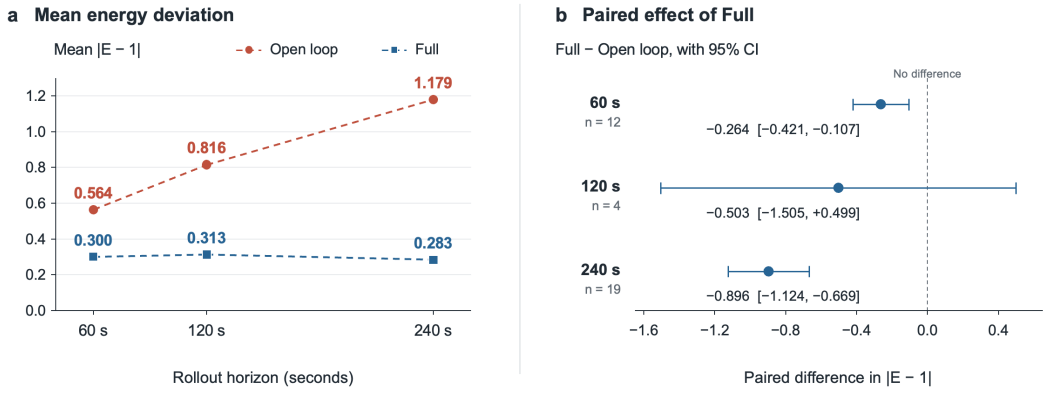



Figure 8: Available horizon cohorts, using only released per-prompt data. Left: mean latent energy deviation in each matched cohort. Right: paired treatment-minus-control differences with 95% confidence intervals. Sample sizes are 12, 4 and 19 at 60, 120 and 240 s. The cohorts differ; the middle estimate is uncertain and no scaling law is established.

Table 20: Historical sweep stacked on attention sink $= 4$ at 240 s, $n = 17$; $\Delta$ is relative to the sink-only arm. No stacked configuration has a statistically significant motion contrast.

| Config | $n$ | $\Delta\lvert E-1\rvert \downarrow$ | $\Delta$drift $\downarrow$ | $t$ |
|---|---|---|---|---|
| t025_b1 | 17 | −0.346 | −0.180 | −5.50 |
| t1_b025 | 17 | −0.337 | −0.142 | −4.45 |
| t05_b05 | 17 | −0.328 | −0.198 | −6.02 |
| t025_b025 | 17 | −0.317 | **−0.217** | **−8.50** |

### D.9 Pilot with a separately frozen protocol

After the historical analysis, we froze six additional benchmark prompts (indices 20–25), seed 0, three arms and 240 s horizons before observing their outcomes. All 18 rollouts completed, with no outcome-based replacement. The inspected historical records contained no matching prompt IDs or text; incomplete historical coverage prevents certification as a globally held-out set. The model, $r = 16$, reference chunks 2–6, mean-band and LR tolerances, new-V scope and attention window were fixed. This pilot uses a separately recorded generation RNG protocol and is not pooled with historical trajectories.

For each prompt, conditioning, initial noise, initial/final RNG state, checkpoint, shared generation settings and physical GPU match across arms. All latent outputs passed independent finite-value and hash checks, and all 19,200 rows of the joint scalar-probe grid passed the consistency checks. Controller events within the reference bands may correctly produce no change. A separate probe-off/on test established equivalence for one prompt and seed at 7.5 s; equivalence over 240 s remains untested.

The pre-specified readouts are final-quarter mean drift, normalized adjacent-frame motion and absolute deviation of the mean energy ratio from one. All three contrasts and all three metrics form one nine-test Holm family, with the prompt as the independent unit ($n = 6$). Table 21 reports every primary test. Full control has lower mean energy deviation and drift than open loop, but **none of the nine tests reaches Holm-adjusted** $p < .05$. Full–mean energy deviation has the smallest adjusted $p$ (.0736). Motion must be read alongside drift; neither lower drift nor higher motion alone

proves better generation. Unadjusted intervals and signed-rank sensitivity do not override the primary multiplicity correction. This is a small, single-seed pilot, not confirmatory generalization or causal-mediation evidence. Decoded-video evaluation is reported separately in Appendix D.10. Its total allocation, including failed prior attempts, was $4.627$ GPU-hours; shared-card timings are not controlled method-overhead measurements.

Table 21: New-cohort $240\,\mathrm{s}$ pilot, six exact prompt pairs per contrast. Differences are treatment minus control. Intervals are unadjusted two-sided $95\%$ paired-$t$ intervals; Holm adjustment covers all nine primary paired-$t$ tests. Exact two-sided signed-rank $p$ values are sensitivity readouts, not an alternative selection rule. This table contains latent proxies only.

| Comparison, proxy | $\Delta$ | $95\%$ CI | Raw $p$ | Holm $p$ | Signed-rank $p$ |
|---|---|---|---|---|---|
| Full–open, energy deviation | $-0.513$ | $[-0.908, -0.119]$ | 0.0204 | 0.1629 | 0.0625 |
| Full–open, drift | $-0.183$ | $[-0.327, -0.038]$ | 0.0229 | 0.1629 | 0.0312 |
| Full–open, motion | $+0.105$ | $[-0.074, +0.283]$ | 0.1914 | 0.6929 | 0.1562 |
| Mean–open, energy deviation | $-0.206$ | $[-0.520, +0.108]$ | 0.1523 | 0.6929 | 0.1562 |
| Mean–open, drift | $-0.103$ | $[-0.254, +0.047]$ | 0.1386 | 0.6929 | 0.0938 |
| Mean–open, motion | $+0.039$ | $[-0.107, +0.185]$ | 0.5203 | 0.8125 | 0.6875 |
| Full–mean, energy deviation | $-0.307$ | $[-0.494, -0.121]$ | 0.0082 | 0.0736 | 0.0312 |
| Full–mean, drift | $-0.079$ | $[-0.179, +0.020]$ | 0.0962 | 0.5770 | 0.0312 |
| Full–mean, motion | $+0.066$ | $[-0.121, +0.252]$ | 0.4063 | 0.8125 | 0.5625 |

Table 22: Decoded-video VBench comparisons at $240\,\mathrm{s}$, in the evaluator's native score units (imaging uses raw MUSIQ units; divide by $100$ to compare with normalized tables). Differences are treatment minus control. Higher scores are preferred for quality; dynamic degree describes activity with no uniformly preferred direction. Sample sizes reflect exact prompt overlap, including $16$, not $17$, for the stacked comparison.

| Dimension | $n$ | $\Delta$ | $95\%$ CI | Raw $p$ | Holm $p$ |
|---|---|---|---|---|---|
| Full controller vs open loop | | | | | |
| aesthetic | 19 | $+0.047$ | $[+0.012, +0.083]$ | 0.0107 | 0.0533 |
| dynamic | 19 | $+0.000$ | $[+0.000, +0.000]$ | 1 | 1 |
| imaging | 19 | $+3.131$ | $[-4.498, +10.760]$ | 0.4 | 1 |
| smoothness | 19 | $+0.00006$ | $[-0.00311, +0.00324]$ | 0.966 | 1 |
| subject | 19 | $+0.058$ | $[+0.025, +0.091]$ | 0.00164 | 0.00986 |
| flickering | 19 | $-0.00261$ | $[-0.00723, +0.00201]$ | 0.251 | 1 |
| Sink $= 4$ vs open loop | | | | | |
| aesthetic | 17 | $+0.181$ | $[+0.150, +0.211]$ | 9.8e-10 | 4.9e-09 |
| dynamic | 17 | $+0.059$ | $[-0.066, +0.184]$ | 0.332 | 0.332 |
| imaging | 17 | $+5.611$ | $[+2.387, +8.836]$ | 0.00199 | 0.00795 |
| smoothness | 17 | $-0.00632$ | $[-0.012, -0.00052]$ | 0.0345 | 0.0691 |
| subject | 17 | $+0.171$ | $[+0.146, +0.197]$ | 1.7e-10 | 1.0e-09 |
| flickering | 17 | $-0.014$ | $[-0.024, -0.00457]$ | 0.00651 | 0.0195 |
| Full controller + sink $= 4$ vs sink $= 4$ | | | | | |
| aesthetic | 16 | $-0.010$ | $[-0.040, +0.020]$ | 0.479 | 1 |
| dynamic | 16 | $-0.062$ | $[-0.196, +0.071]$ | 0.333 | 1 |
| imaging | 16 | $-1.451$ | $[-7.250, +4.348]$ | 0.602 | 1 |
| smoothness | 16 | $+0.00308$ | $[+0.00023, +0.00593]$ | 0.0362 | 0.181 |
| subject | 16 | $+0.00600$ | $[-0.019, +0.031]$ | 0.61 | 1 |
| flickering | 16 | $+0.00532$ | $[+0.00146, +0.00918]$ | 0.0102 | 0.0609 |

### D.10 Complete decoded-video development pilot

**Decoded-video evaluation of the complete pilot.** The evaluation includes all six prompts and three arms, with no outcome-based replacement. Each continuous-cache VAE decode contains $3{,}837$ frames at $16$ fps ($239.8125$ s). Using the original text prompt and a pinned VBench implementation,

we score seven dimensions over 19 windows of 192 frames and a final window of 189 frames. Frame-count-weighted means give one observation per prompt and arm. The 360 windows and 2,520 metric-window records therefore yield $n = 6$ paired prompts; $n = 360$ would incorrectly treat windows as independent samples. Imaging quality is rescaled by dividing by 100 once. These custom-prompt, window-aggregated scores assess individual dimensions; they are neither an official whole-video leaderboard total nor a measure of human preference.

The analysis plan was fixed before observing the pilot quality scores. Six full–mean paired $t$ tests form one Holm family (Table 23), separate from the nine-test latent-proxy family in Table 21. Unadjusted 95% intervals and tests over all 64 exact sign assignments provide sensitivity analyses. Text consistency has raw $p = 0.0123$ and Holm $p = 0.0739$; the aesthetic and text sign-flip tests have $p = 0.03125$. None of these sensitivity results overrides the primary Holm-adjusted decision rule. Full–native comparisons are descriptive.

Table 23: Complete six-prompt pilot: full minus mean-only VBench scores. Intervals are unadjusted paired-$t$ intervals; Holm adjusts $p$ across all six primary tests. None reaches adjusted $p < 0.05$. Imaging quality is divided by 100 once; details are in Appendix D.10.

| Dimension | $\Delta$ | 95% CI | Holm $p$ |
|---|---|---|---|
| Subject consistency | $-0.0084$ | $[-0.0366, +0.0197]$ | 1.0000 |
| Background consistency | $-0.0015$ | $[-0.0140, +0.0110]$ | 1.0000 |
| Motion smoothness | $-0.0011$ | $[-0.0102, +0.0080]$ | 1.0000 |
| Aesthetic quality | $+0.0769$ | $[-0.0052, +0.1590]$ | 0.3056 |
| Imaging quality | $+0.0339$ | $[-0.0115, +0.0793]$ | 0.4524 |
| Text consistency | $+0.0563$ | $[+0.0185, +0.0942]$ | 0.0739 |

Table 24: Complete pilot: equally weighted prompt means of frame-weighted scores (top) and descriptive full–native differences (bottom), all $n = 6$. Intervals are unadjusted 95% paired-$t$ intervals. Dynamic degree is reported separately because a higher value need not indicate correct motion.

| Dimension | Native | Mean | Full |
|---|---|---|---|
| Subject consistency | 0.94204 | 0.94816 | 0.93973 |
| Background consistency | 0.95784 | 0.95407 | 0.95258 |
| Motion smoothness | 0.98256 | 0.98005 | 0.97898 |
| Aesthetic quality | 0.43280 | 0.46343 | 0.54031 |
| Imaging quality | 0.61523 | 0.69746 | 0.73137 |
| Text consistency | 0.14844 | 0.15342 | 0.20974 |
| Dynamic degree | 0.26661 | 0.30831 | 0.49166 |

| Dimension | $\Delta$ | 95% CI |
|---|---|---|
| Subject consistency | $-0.0023$ | $[-0.0383, +0.0337]$ |
| Background consistency | $-0.0053$ | $[-0.0235, +0.0130]$ |
| Motion smoothness | $-0.0036$ | $[-0.0118, +0.0046]$ |
| Aesthetic quality | $+0.1075$ | $[-0.0528, +0.2678]$ |
| Imaging quality | $+0.1161$ | $[+0.0197, +0.2125]$ |
| Text consistency | $+0.0613$ | $[+0.0226, +0.1000]$ |

Table 25: Dynamic degree for every pilot prompt. Entries are frame-weighted window scores; the metric does not assess whether motion matches the prompt.

| Prompt | Native | Mean | Full |
|---|---|---|---|
| p20 | 0.00000 | 0.00000 | 0.00000 |
| p21 | 0.44957 | 0.10008 | 0.89992 |
| p22 | 0.10008 | 0.05004 | 0.79984 |
| p23 | 0.00000 | 0.00000 | 0.00000 |
| p24 | 1.00000 | 1.00000 | 1.00000 |
| p25 | 0.05004 | 0.69977 | 0.25020 |

**Same-trajectory mechanism measurements.** The scalar probe records common and centered energy without advancing the generation RNG state. Artifact hashes link each video window to preceding cache writes on the same trajectory. Opening windows without preceding cache measurements have missing cache features. This construction aligns summaries of earlier cache writes with quality scores from a later video interval, enabling temporal analysis. Establishing that cache growth precedes quality failure additionally requires a validated onset definition. A validated onset threshold, a normal/failure partition and a causal-mediation analysis remain unavailable.

### D.11 Same-trajectory cache and quality descriptions

**Cache growth and later video scores on the same trajectories.** We reuse all $18$ completed development trajectories and their $360$ video windows. Earlier common energy is often negatively associated with later quality, but the relationship is neither specific nor a validated warning (Figure 9). Across the six Native trajectories, the median Pearson correlation between log common energy and subsequent aesthetic score is $-0.762$; after removing linear time trends it is $-0.584$. Correlations with cache measurements that follow the same quality windows are also negative ($-0.543$ and $-0.378$). These forward and reverse descriptions retain identical quality windows and do not treat windows as independent samples.

An exploratory sustained-event definition gives common-energy events in $3/6$ Native, $0/6$ Mean and $0/6$ Full trajectories. It uses a twofold rise relative to the first $36$ seconds and three consecutive supports; a metric decline requires a decrease of $0.05$ for three consecutive windows, with imaging quality divided by $100$. No Native cache event is confirmed before the first aesthetic, imaging or text-decline support. Metric decreases also occur in trajectories without a detected energy event. All $18$ combinations of the existing sensitivity settings are retained; none is selected as a preferred warning rule. This outcome-aware development analysis supports an association in some trajectories, but not a reliable precursor or causal mediation claim.

**Content change is an alternative explanation.** The fixed-frame comparisons include all six prompts and all three arms. The dragon-dance prompt requests camera and subject motion; the gallery prompt requests multiple angles and sweeping pans. Such changes can alter cache statistics without constituting failure. Conversely, visible artifacts appear in the cat and orangutan Native samples without the default sustained common-energy event. Sparse frame inspection cannot certify continuous action completion or a normal/failure partition. These observations motivate reporting quality, text alignment and motion separately, and limit any claim that matching an early cache reference preserves all legitimate changes.

Table 26: Outcome-aware temporal descriptions for the six Native development trajectories. Each cell gives the median correlation with preceding / following log common energy, using the same $17$ quality windows per trajectory. "Detrended" removes a fitted linear time trend; "differenced" correlates adjacent changes. All dimensions have six defined correlations except dynamic degree, which has two; the other four constant series remain undefined. These are descriptions, not independent-window tests or causal estimates.

| Dimension | Raw | Detrended | Differenced |
|---|---|---|---|
| Subject | -0.247 / +0.368 | -0.285 / +0.074 | -0.196 / +0.201 |
| Background | +0.166 / +0.449 | -0.080 / +0.292 | -0.082 / +0.275 |
| Smoothness | -0.597 / -0.306 | -0.335 / +0.046 | -0.198 / +0.069 |
| Dynamic degree | -0.119 / -0.433 | -0.273 / -0.487 | -0.139 / -0.256 |
| Aesthetic | -0.762 / -0.543 | -0.584 / -0.378 | -0.388 / -0.246 |
| Imaging | -0.449 / -0.333 | -0.392 / -0.111 | -0.103 / +0.039 |
| Text | -0.496 / -0.288 | -0.292 / -0.303 | -0.347 / -0.309 |

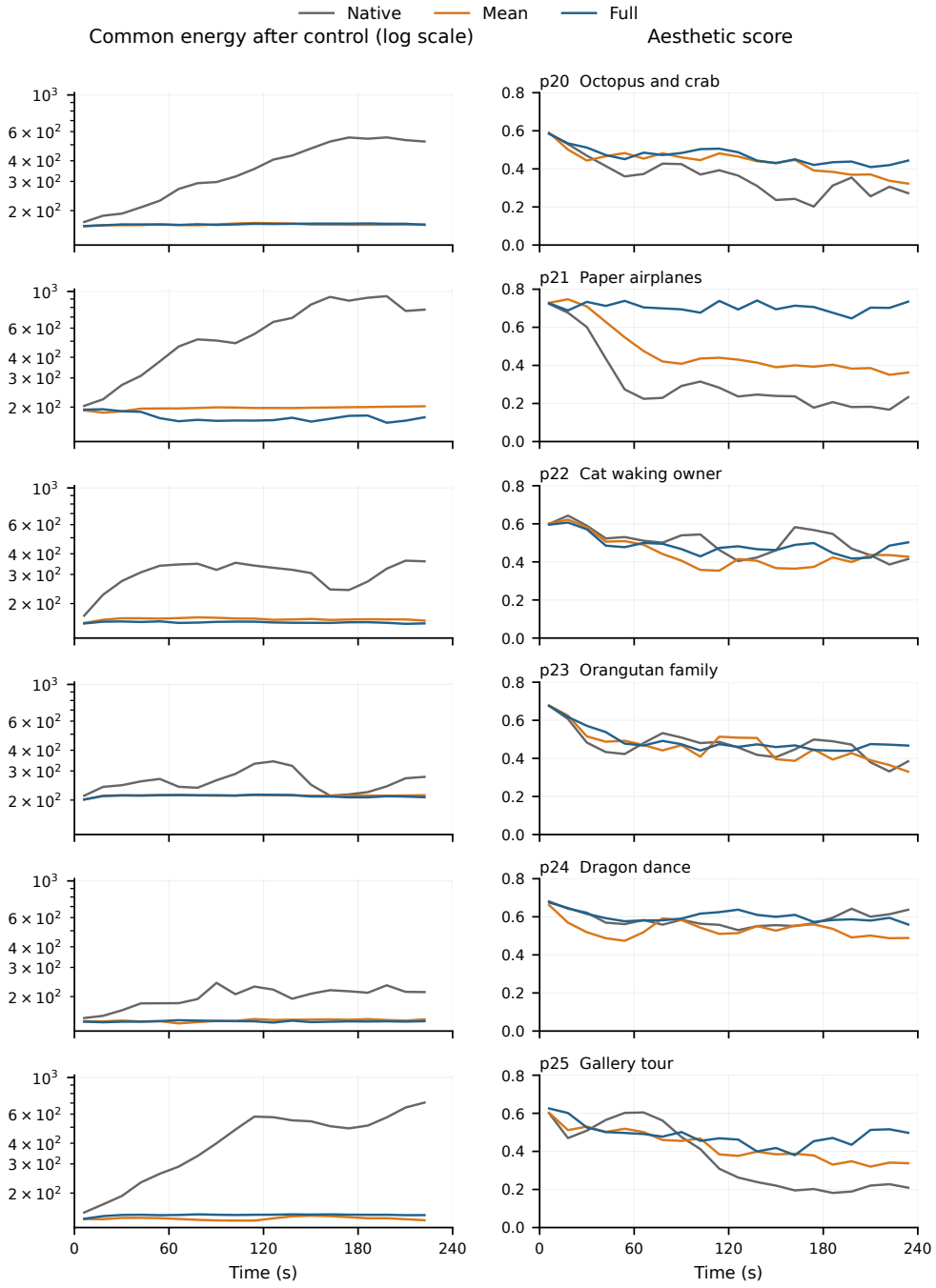
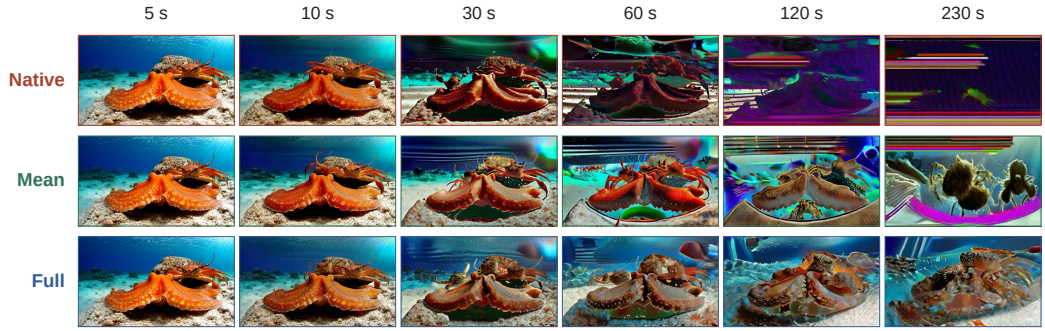



Figure 9: All six completed pilot trajectories, with Native, Mean and Full at seed 0. Left: layer-median common energy after control, on a shared log scale; points are located at the mean consumer time of the preceding cache writes. Right: aesthetic scores at the centers of the corresponding video windows. There is no smoothing or failure-based selection. The panels illustrate the heterogeneous association: large Native growth accompanies aesthetic decline in several scenes, while the dragon scene does not show that pattern. Lower cache growth alone does not establish quality or preservation of intended content. Table 26 retains all seven metric associations; the anonymous traces retain every window and metric.

### D.12 FIXED DEVELOPMENT EXAMPLES AND MOTION SCENES

These fixed-time panels retain every pilot prompt and three additional motion examples. They illustrate temporal content without treating frames as independent samples or certifying continuous action completion.

**(a) p20 | Octopus and crab**




**(b) p21 | Paper airplanes in a jungle**

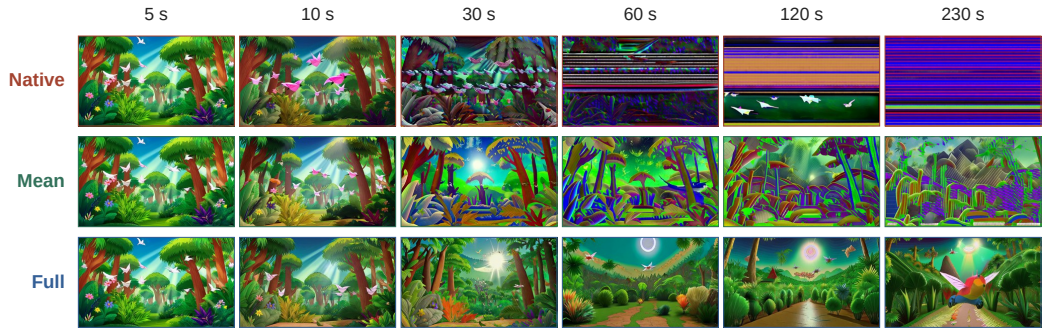



Figure 10: Fixed-time comparisons for the complete six-prompt development pilot (p20–p25; seed 0), part 1 of 3. Panels (a)–(b) show p20–p21; the next two pages show the remaining four prompts. Rows are Native, Mean and Full (default $\lambda = 1$), paired on prompt, initial noise, checkpoint and generation RNG protocol. Columns show 5, 10, 30, 60, 120 and 230 s from continuous VAE decodes (3,837 frames at 16 fps). All 18 pilot videos are included in prompt order at the same fixed times, with no outcome-based replacement. These development videos are separate from the independent cohort. The 108 pictured frames do not increase the six-prompt sample size or establish population quality or temporal motion fidelity.

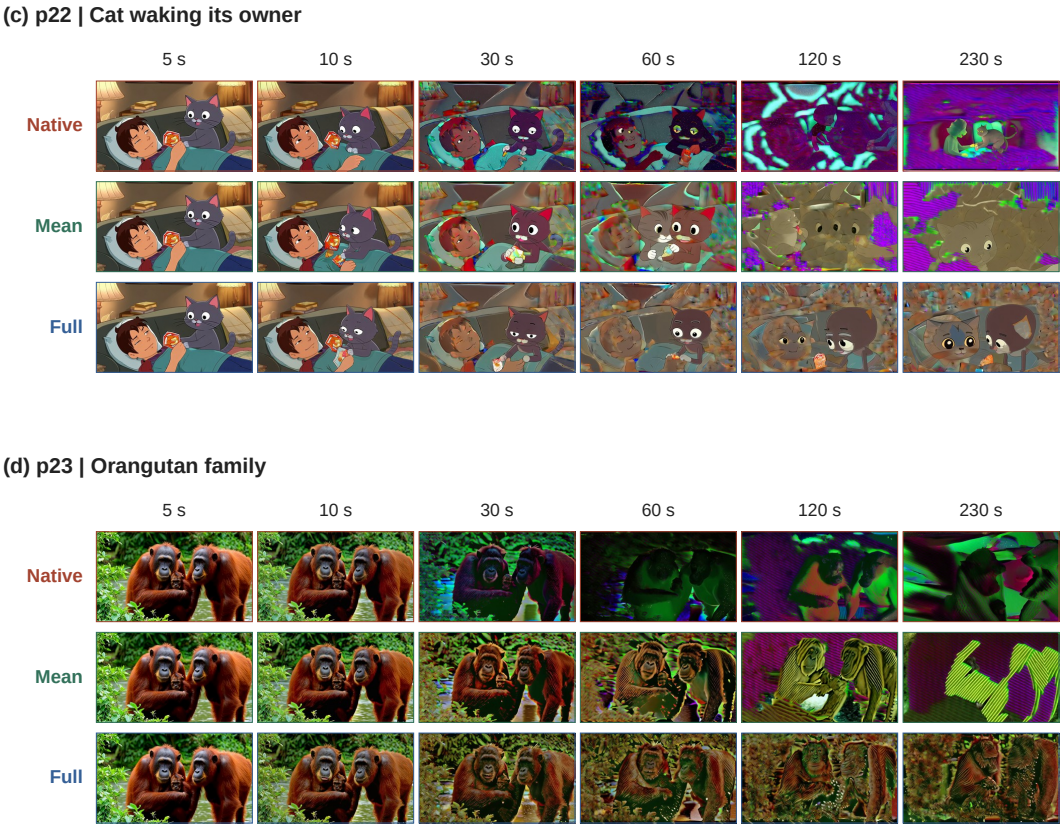



Figure 10: Complete development-pilot comparison (continued, part 2 of 3). Panels (c)–(d): p22–p23, seed 0. Rows are Native, Mean and Full ($\lambda = 1$); columns use the same fixed times as part 1. All visible artifacts are retained. See the first caption for pairing and interpretive limits.

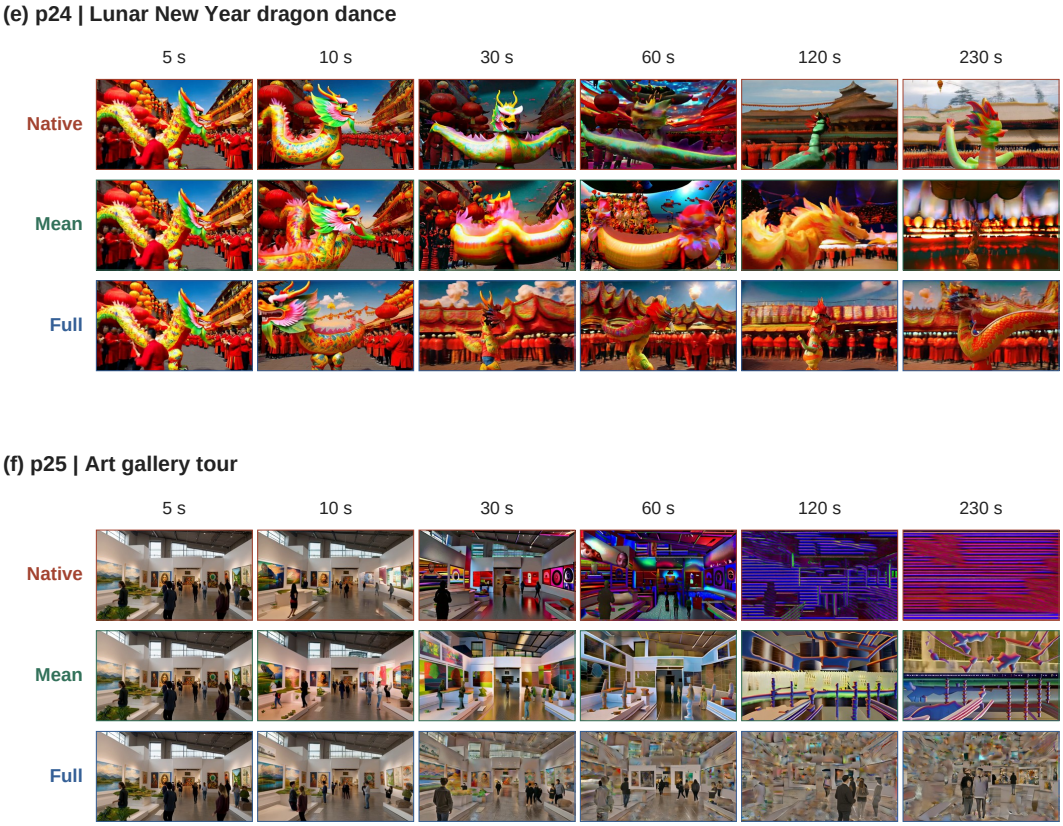



Figure 10: Complete development-pilot comparison (continued, part 3 of 3). Panels (e)–(f): p24–p25, seed 0. Rows are Native, Mean and Full ($\lambda = 1$); columns use the same fixed times as part 1. These frames complement the completed pilot scores without adding an independent experiment or changing the six Holm-adjusted primary decisions.

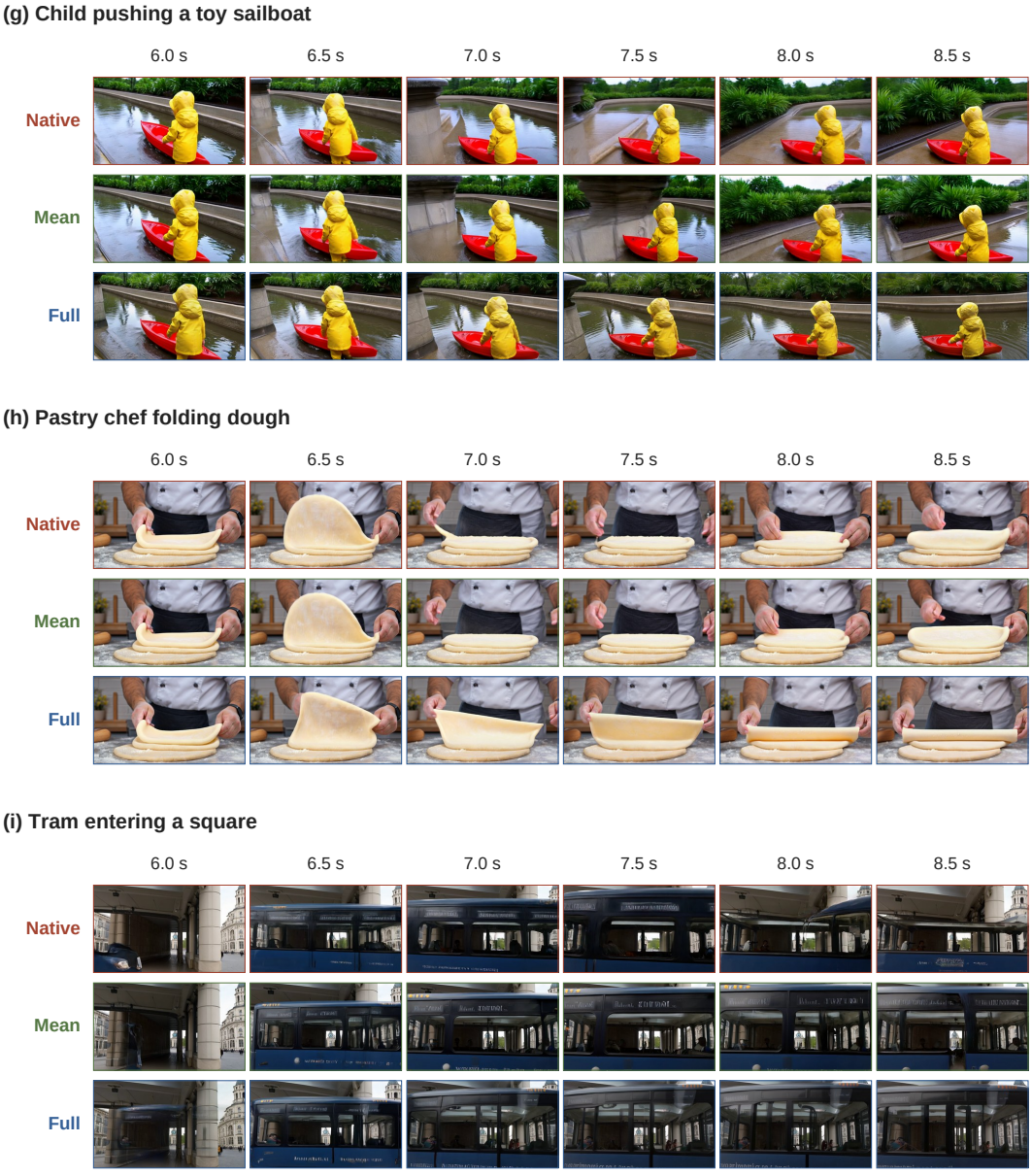



Figure 10: Additional motion-scene comparisons (continued). Panels (g)–(i) show a child pushing a toy sailboat, a pastry chef folding dough, and a tram entering a square. Rows are Native, Mean and Full ($\lambda = 1$). All nine videos use the same fixed 6.0–8.5 s display interval, sampled every 0.5 s. Within each scene, all arms use the unadapted frozen base, the same prompt and seed 9102701, and the paired generation RNG protocol. The existing Native/Full outputs come from the adaptation evaluations in Appendix G.1; the missing Mean outputs were generated with the same settings and an exact match to Native's first 21 unmodified latent frames. No trained adapter is enabled. Child and chef videos last 8.8125 s; tram videos last 59.8125 s, all at 16 fps. The video companion provides continuous, frame-aligned playback of each complete three-arm comparison. These three development examples are separate from the six-prompt pilot and independent cohort; the local frame sequences do not establish motion correctness or superiority.

The main cohort establishes the reported Full–Mean text–subject tradeoff. The retained developmental scores and temporal descriptions motivate further mechanism tests, but do not add independent confirmation to that comparison.

# E CORRECTION CONTROLS, IMPLEMENTATION COST, AND PUBLISHED CONTEXT

This appendix tests alternatives to the centered correction at a common write location and reports implementation cost under controlled hardware conditions. Published systems are listed separately because their evaluation protocols differ.

## E.1 CENTERED-DOSE CONTROLS AND SAME-POSITION ALIGNMENT

This study asks how much the correction geometry matters beyond mean control. It retains actual-writeback dose checks separately from quality comparisons.

**Development control protocol.** The separately registered study uses six development prompts, seed 0, 960 latent frames and 3,837 decoded frames at 16 fps. All arms use the same seven VBench dimensions, evaluator settings and scoring seed. Twenty non-overlapping windows are weighted by frame count ($19 \times 192 + 189$), then prompts receive equal weight. Prompts, settings and reporting rules were fixed before the comparisons finished. These remain development inputs; the additional controls do not create an independent test set.

**Simple statistical alignment.** The five arms are instrumented Mean and Full, Mean plus a random centered correction, Mean plus a uniform centered contraction, and mean/std alignment, for 30 assigned trajectories. Every arm intervenes on the same newly written value rows and uses the same early reference blocks 2–6. Statistical alignment matches the reference mean and population standard deviation independently in each flattened head/channel, with blend 1 and $\epsilon = 10^{-6}$. The reference standard deviation is the arithmetic mean of the per-block population standard deviations, not the standard deviation of pooled tokens. This is a controlled statistical baseline, not a reproduction of TetherCache's complete recall and alignment system.

**Dose measured after the actual writeback.** Let $Q$ denote bfloat16 conversion, $\mathcal{C}$ token centering, and $\Delta_M(X)$ and $\Delta_F(X)$ the Mean and Full corrections on input $X$. For each Full event, the donor target is

$$d_F = \|\mathcal{C}\left[Q(X_F - \Delta_F(X_F)) - Q(X_F - \Delta_M(X_F))\right]\|_F . \tag{16}$$

Subtraction and centering use analysis precision after conversion. Each alternative starts from its own Mean update and chooses a scale $g$ to match

$$d_C = \|\mathcal{C}\left[Q(X_C - \Delta_M(X_C) + gU_C) - Q(X_C - \Delta_M(X_C))\right]\|_F \tag{17}$$

to the corresponding donor target. The random direction is a private-seed, rank-16 centered additive direction independent of the cache spectrum. The uniform direction contracts incoming centered content without reversing it solely to reach a target. A fixed bounded search uses relative tolerance 1%, absolute tolerance $10^{-6}$ and at most 24 iterations. Zero targets, infeasible targets, mismatches and common-component leakage are retained.

Thus the matched magnitude is the centered increment beyond the Mean correction evaluated on each arm's own input. Later inputs $X_F$ and $X_C$ can diverge, and quantization can add a common increment. Matching this norm therefore does not isolate direction as the only causal difference. We retain all six prompt-level differences, their mean and median, and all seven dimensions for Full–Mean, Full–random, Full–uniform, mean/std–Mean and mean/std–Full. No additional significance tests or composite score are used.

**Completed quality comparisons.** All 30 assigned trajectories were generated, decoded and scored, yielding 4,200 metric-window records. Tables 27 and 28 report all five fixed contrasts and all seven dimensions. The anonymous supplement retains all 210 prompt-level differences and the raw and normalized window scores with their frame weights. Imaging quality is divided by 100

once; other dimensions use the evaluator's stated scale. Differences are in normalized score units, not percentage points.

Mean/std alignment raises mean aesthetic, imaging and text scores above Mean but remains below Full. Full also has higher means on these dimensions than the random and uniform controls. These gains have costs: Full lowers mean subject consistency and smoothness relative to the random control. Mean and median differences can have opposite signs, as for Full–Mean subject consistency; the averages do not describe every prompt. Dynamic degree measures activity, so an increase is not necessarily an improvement.

The six prompts and seed are shared with the decoded-video pilot. Mean and Full are separately instrumented runs in this correction-control study. The study extends the comparison on these development inputs and provides neither an independent replication nor evidence that complex control is generally necessary.

Table 27: Mean paired differences across all six development prompts. $F$: Full; $M$: Mean; $R$: Mean plus matched random correction; $U$: Mean plus dose-targeted uniform contraction; $S$: mean/std alignment. The 28 infeasible uniform targets remain included. Each trajectory first averages its 20 windows by actual frame count, then prompts receive equal weight. All contrasts are descriptive.

| Dimension | $F-M$ | $F-R$ | $F-U$ | $S-M$ | $S-F$ |
|---|---|---|---|---|---|
| Subject | -0.0084 | -0.0058 | +0.0355 | -0.0072 | +0.0012 |
| Background | -0.0015 | +0.0024 | +0.0074 | -0.0052 | -0.0037 |
| Smoothness | -0.0011 | -0.0013 | +0.0048 | +0.0011 | +0.0022 |
| Dynamic | +0.1833 | +0.2333 | +0.1250 | -0.0165 | -0.1999 |
| Aesthetic | +0.0769 | +0.0679 | +0.0981 | +0.0262 | -0.0507 |
| Imaging | +0.0339 | +0.0351 | +0.2040 | +0.0239 | -0.0100 |
| Text | +0.0563 | +0.0337 | +0.0659 | +0.0402 | -0.0161 |

Table 28: Median paired differences for the same six prompts and all fixed contrasts in Table 27. Windows are not treated as independent samples. No additional tests or composite score are introduced.

| Dimension | $F-M$ | $F-R$ | $F-U$ | $S-M$ | $S-F$ |
|---|---|---|---|---|---|
| Subject | +0.0017 | +0.0049 | +0.0382 | -0.0008 | -0.0035 |
| Background | +0.0012 | +0.0051 | +0.0105 | -0.0040 | -0.0054 |
| Smoothness | +0.0008 | +0.0013 | +0.0040 | +0.0004 | +0.0004 |
| Dynamic | +0.0000 | +0.0000 | +0.0000 | +0.0500 | -0.1001 |
| Aesthetic | +0.0501 | +0.0447 | +0.0971 | +0.0123 | -0.0448 |
| Imaging | +0.0275 | +0.0292 | +0.2143 | +0.0218 | -0.0019 |
| Text | +0.0588 | +0.0380 | +0.0563 | +0.0363 | -0.0054 |

**Actual writeback dose and retained exceptions.** The complete logs contain $281{,}700$ intervention events. Random and uniform controls each attempt $56{,}340$ donor-dose matches. Every random event meets the fixed tolerance; its maximum relative mismatch is $0.228\%$. Uniform contraction meets tolerance on $56{,}312$ events. The remaining 28, all on one prompt, have certified infeasible targets under the registered non-reversing contraction; all events and their videos are retained. The largest uniform relative mismatch, including those exceptions, is $30.66\%$. No donor target is zero.

Post-cast increments also have a small nonzero common component. Across all events, Full, random and uniform increments have mean common norms $1.346$, $1.417$ and $1.220$, respectively, and mean centered norms $492.641$, $492.680$ and $492.612$. These are means of event norms, not norms of a pooled increment. All per-trajectory norm summaries and mismatch counts accompany the quality data. This common component and the divergent later inputs limit interpretation of the comparison as a test of direction alone.

### E.2 Controlled generation cost at the main duration

The cost study uses the same A100-SXM4-80GB for all eight runs: two fixed prompts (p20 and p21, seed 0) crossed with Native, Mean, Full and Sink. Method order is reversed on the second prompt. Each run uses bfloat16 and produces 960 latent frames, equivalent to 3,837 output frames at 16 fps. Measurements include synchronized generation wall time, per-chunk CUDA-event time and peak allocated/reserved memory. Model loading, text encoding, VAE decoding, scalar probes and controller diagnostic recording are excluded from the time measurements. Peak memory is the absolute PyTorch allocated/reserved peak, including resident model allocations, all persistent caches and inference temporaries; it is not an incremental cache-only measurement or driver-level memory total. Each run starts with an empty device process list; process, memory, clock, temperature and power traces are sampled every two seconds. Any detected foreign GPU process invalidates the run. The resulting measurements apply to this implementation and duration. Shared-device timings elsewhere record resource use and are excluded from the overhead comparison.

Table 29 includes all eight completed runs: time and throughput are averaged across the two prompts; each time ratio is computed against Native within prompt, then averaged. Memory is the larger of the two reserved-memory peaks. One interrupted Mean attempt has no timing result. Its conservative budget charge remains in resource accounting and is excluded from these measurements.

Table 29: Controlled generation cost for 3,837 nominal output frames (239.8 s) on one A100-SXM4-80GB. All eight fixed prompt–method runs passed the sampled exclusivity check. Frames/s excludes VAE decoding; memory is the absolute allocator peak, not incremental cache memory.

| Method | Seconds | Ratio to Native | Frames/s | Peak reserved GiB |
|---|---|---|---|---|
| Native | 470.27 | 1.000 | 8.16 | 9.62 |
| Mean | 478.33 | 1.017 | 8.02 | 9.62 |
| Full | 736.43 | 1.566 | 5.21 | 9.86 |
| Sink (4) | 469.98 | 0.999 | 8.16 | 9.54 |

The paired timing study quantifies the practical price of the centered extension; it supplies no additional evidence of quality or transfer.

### E.3 Historical correction magnitudes and short-duration cost

Table 30: Relative correction magnitude $\|\alpha\Delta\|_F/\|X\|_F$ at 240 s unless marked otherwise. "Shrink" is the mean selected-direction scaling factor: 1 means no centered shrinkage, not proportional contraction. "Silent" counts chunk-layer updates with zero total correction, requiring both mean and centered updates to vanish.

| Configuration | mean | p95 | shrink | silent |
|---|---|---|---|---|
| low-rank $r$=16, $\lambda$=1 | 43.4% | 0.841 | 0.585 | 0.0000% |
| low-rank $r$=16, $\lambda$=4 | 26.8% | 0.585 | 0.952 | 0.0053% |
| low-rank $r$=16, $\lambda$=4, $b$=4$\tau$ (60 s) | 3.9% | — | 0.986 | 26.4% |
| reference-directed mean | 19.5% | 0.259 | 1.000 | 0.0000% |

**Wall-clock overhead.** Shared-GPU measurements vary from 12 to 152 minutes for the same arm, and the fastest open-loop run is slower than the fastest closed-loop run. Such measurements cannot isolate controller cost from resource contention. A separate timing check used an otherwise idle A100-80GB, one job on the device, 240 latent frames per run and prompts paired between arms. The open loop takes 162.5 s (median, $n = 2$); the mean trust region alone takes 167.4 s, +4.8% paired; the full operator at $r = 16$ takes 256.8 s, +57.9% paired (per-prompt +60.0% and +55.7%). The full operator computes a per-chunk, per-layer eigendecomposition of $\Sigma - G_{\mathrm{ref}}$. This small timing sample estimates implementation cost, not a hardware-independent constant. Mean-only control has lower measured cost; its output effects are evaluated separately from timing.

### E.4 Recorded alternative controllers and numerical implementations

Two completed development checks test alternatives to the implemented method. They do not change the frozen Mean/Full configurations or add evidence to the main quality comparison.

**A ratio-gated reference alternative.** A single p21, seed-0 screen generates 240 latent frames (957 decoded frames, 59.8125 s) and reuses the matching Mean trajectory's prefix. The alternative uses the same new-$V$ write site and chunks 2–6, but triggers only above twice the layer's maximum reference $\rho$. It averages the current and reference means, then radially clips the result if needed to satisfy that ratio cap. This is a separate controller, not Mean in Section 5.1. Conditioning, initial noise and pre-generation RNG are matched; the first 21 unmodified latent frames agree exactly.

Only $27/2{,}190$ eligible layer–chunk events trigger ($1.233\%$), all with actual bfloat16 changes and no recorded numerical anomaly. Triggered-event mean $\rho$ changes from $0.33022$ to $0.19171$. Every post-cast event meets the layer's cap. However, the largest triggered FP32 post-update ratio is only $65.53\%$ of its cap: the mean blend already lies inside the ball, so this screen does not exercise a binding radial clip.

Table 31: Single-example ratio-controller screen. Energies and $\rho$ are means across all 2,190 eligible events, before / after actual writeback. The two feedback trajectories diverge; these are not same-input operator comparisons or independent event samples.

| Quantity | Mean | Ratio-gated alternative |
|---|---|---|
| $\rho$ | $0.213695/0.166559$ | $0.247987/0.246279$ |
| $\|\mu\|^2$ | $294.7970/231.7465$ | $314.9854/312.8117$ |
| $\operatorname{tr}\Sigma$ | $1452.075848/1452.071352$ | $1343.845297/1343.845264$ |
| Mean actual relative correction | $14.5877\%$ | $0.2275\%$ |
| Maximum actual relative correction | $24.1770\%$ | $20.1294\%$ |

The first trigger is at block 46, after approximately 35.06 s. Recorded AI-assisted fixed-frame and late-clip inspection finds repeated trails already at 30 s, before this trigger, and extensive late stripes and scene fragmentation. Mean also has artifacts, but no visual benefit is established for the alternative. This inspection is neither a blinded human comparison nor a computed quality or motion score. Triggered events have mean relative correction $18.456\%$, so a smaller all-event average does not mean every active correction is gentler. No VBench scoring was performed. The single new generation takes 119.258 s; total allocated GPU time including checks and decode is 259.291 s. The reused Mean run has a different full duration, precluding an overhead comparison.

**A factorized numerical implementation.** One paired p20, seed-0 check compares the original and a factorized Full implementation for 30 latent frames (7.5 s), with 90 active corrections per arm. The first 24 frames agree exactly, but the complete outputs differ: relative $L_2$ error $0.0138553660$, maximum absolute error $2.1484375$, and final-frame relative $L_2$ error $0.0532890521$. Both are finite and their conditioning, initial noise and RNG records match. Generation times are $15.4396/15.2716$ s, with reserved-memory peaks $10{,}206{,}838{,}784/10{,}208{,}935{,}936$ bytes. This short, startup-heavy pair has no decoded quality measurement and does not demonstrate lower long-video overhead or numerical equivalence. Neither variant replaces the original controller in the paper.

### E.5 Published systems under their original protocols

**Separate published protocols.** Table 32 reproduces the authors' VBench-Long scores alongside each paper's own Self-Forcing baseline. FreqForcing uses the first 128 MovieGen prompts; TeaCache samples 128 prompts for 30/60 s and 32 for 240 s. Both refine prompts with Qwen2.5-7B-Instruct. These cohorts differ from our registered 50-prompt, two-seed protocol; matching seeds, evaluator revisions and temporal aggregation are not established. The table supplies external context, without ranking these methods against ours or pooling scores across protocols.

The local controls establish descriptive differences at matched writeback magnitude; the published tables provide context, not a common-protocol ranking.

**(a) FreqForcing: original Table 1** (Li et al., 2026)

| Length | Method | DD | MS | OC | IQ | AQ | SC | BC |
|---|---|---|---|---|---|---|---|---|
| 60 s | Self-Forcing | 33.29 | 98.45 | 18.81 | 66.90 | 56.77 | 96.53 | 96.32 |
| 60 s | FreqForcing | 59.58 | 98.07 | 20.94 | 69.52 | 60.91 | 97.27 | 96.47 |
| 120 s | Self-Forcing | 27.42 | 98.34 | 16.17 | 61.40 | 51.22 | 97.02 | 96.29 |
| 120 s | FreqForcing | 58.97 | 98.22 | 20.98 | 68.89 | 60.68 | 97.38 | 96.47 |

**(b) TetherCache: original Table 1** (Meng et al., 2026)

| Length | Method | DD | IQ | CC | OC | Drift | Total | Quality | Semantic |
|---|---|---|---|---|---|---|---|---|---|
| 30 s | Self-Forcing | 24.29 | 69.79 | 68.27 | 26.11 | 0.95 | 80.86 | 82.03 | 76.16 |
| 30 s | TetherCache | 48.57 | 70.00 | 89.74 | 25.71 | -0.51 | 82.70 | 83.40 | 79.94 |
| 60 s | Self-Forcing | 34.48 | 67.06 | 67.56 | 23.49 | 11.34 | 78.82 | 81.53 | 67.97 |
| 60 s | TetherCache | 43.45 | 69.84 | 95.01 | 26.07 | 0.33 | 82.11 | 82.51 | 80.49 |
| 240 s | Self-Forcing | 28.99 | 58.51 | 78.78 | 14.98 | 7.84 | 71.22 | 77.95 | 44.34 |
| 240 s | TetherCache | 35.50 | 68.58 | 98.96 | 25.70 | 1.33 | 80.17 | 81.50 | 74.85 |

Table 32: **Author-reported results under separate evaluation protocols.** All numbers retain their source table's scale and aggregation; no rescaling or averaging across papers is applied. DD: dynamic degree; MS: motion smoothness; OC: overall consistency; IQ: imaging quality; AQ: aesthetic quality; SC: subject consistency; BC: background consistency; CC: color consistency. DD measures motion activity and is not uniformly better when higher. Other scores are higher-is-better except TetherCache's Drift, the first-to-last-clip drop in imaging quality. Its Total, Quality and Semantic columns retain the source's aggregates. The different column sets and all values come from the original papers, not our experiments.

# F TRANSFER DIAGNOSTICS AND QUALITY ACROSS GENERATORS

The transfer studies ask whether the diagnosed feature change and the controller's output effects persist outside the main generator. MAGI supplies a distinct architecture; Causal Forcing tests another trained generator in the Wan family. Complete, exploratory and incomplete comparisons remain separate.

## F.1 SECOND-BACKBONE ACTIVE-PROBE INTEGRITY CHECK

We tested whether active scalar probes alter the full-controller trajectory on MAGI-1 4.5B base. One prompt and seed were run with probes disabled and enabled over 288 video frames (12 s). Each run recorded 374 clean-cache events and 136 nonzero interventions. Artifact comparison found identical conditioning, initial noise, generated latents, decoded tensors, encoded video, generation RNG records and controller decisions under the two observation settings. Both runs completed. Peak allocated and reserved memory with probes enabled were 14.67 and 16.40 GiB. The two-run allocation consumed 3.020 GPU-hours; this resource total is not a controlled estimate of method overhead.

The comparison repeats one trajectory under two observation settings. It supports probe equivalence for this case, without adding an independent sample or comparing native and full generation. Concentration dynamics, mean/full differences, matched-random controls and perceptual quality are separate questions, addressed by the development comparison in Appendix D.10. Generalization to independent prompts remains unestablished.

## F.2 SIX-ARM SHORT MAGI DEVELOPMENT COMPARISON

The MAGI-1 4.5B base comparison uses one sailboat prompt, seed 73109 and 288 frames at 24 fps (12 s). Its six arms are native, Top-SVD, mean, full, random-Top and random-mean. The native/full generation comparison uses these same runs and adds no independent sample. Conditioning, initial noise and generation RNG are paired across arms.

Each randomized arm replays the absolute correction norm of its corresponding donor at all 136 active events. The maximum relative mismatch between actually stored corrections is $3.85 \times 10^{-6}$ for random-Top and $4.23 \times 10^{-6}$ for random-mean. This matches the applied norm at an event; it does not match the norm relative to an evolving cache or the subsequent feedback trajectory.

Table 33: Complete MAGI development comparison: one prompt/seed, 12 seconds. Random controls match donor per-event absolute corrections after cache storage. These are descriptive dimension scores, not an official total or efficacy test. Imaging quality is normalized by100 once; all dynamic-degree scores are zero.

| Dimension | Native | Top-SVD | Rand-Top | Rand-Mean | Mean | Full |
|---|---|---|---|---|---|---|
| Subject consistency | 0.93810 | 0.93808 | 0.93807 | 0.93827 | 0.93821 | 0.93806 |
| Background consistency | 0.95610 | 0.95270 | 0.95626 | 0.95659 | 0.95802 | 0.95544 |
| Motion smoothness | 0.99562 | 0.99557 | 0.99565 | 0.99565 | 0.99560 | 0.99567 |
| Dynamic degree | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Aesthetic quality | 0.67990 | 0.67724 | 0.68005 | 0.67929 | 0.67526 | 0.67957 |
| Imaging quality | 0.73677 | 0.73734 | 0.73478 | 0.73710 | 0.73338 | 0.73586 |
| Text consistency | 0.25557 | 0.25510 | 0.25588 | 0.25568 | 0.25559 | 0.25534 |

Among the six quality/text dimensions, mean scores above random-mean only on background consistency, and Top-SVD scores above random-Top only on subject consistency and imaging quality. Full–mean differences are mixed; full scores below native on five dimensions. All six dynamic-degree scores are zero. One short example cannot establish cross-backbone perceptual superiority or a general effect of dominant-direction suppression.

Native cache trends depend on the measurement interval. From chunk 0 to 1, the median layer-wise common-energy ratio is 1.482 and the uncentered participation-ratio ratio is 0.466 (later/earlier). Comparing late chunks 7–10 with reference chunks 2–6 reverses both trends: the common-energy ratio is 0.910 and the uncentered participation-ratio ratio is 1.151. The centered-energy and centered participation-ratio ratios are 1.049 and 1.029, respectively. The startup concentration increase therefore does not establish persistent growth after the reference interval.

### F.3 Four completed longer MAGI Native diagnostics

We completed all four separately registered MAGI-1 4.5B base Native trajectories at 24 s and seed 73109, covering camera motion, scene change, a stable scene and subject motion. Each has 576 frames at 24 fps and 782 layer–chunk records (34 layers and 23 clean chunks), for 3,128 records in total. Prompt selection followed the registered identifier rule; no prompt was replaced after inspecting its outcome. These are exploratory cache diagnostics, not an independent quality test.

For each metric and layer, we divide its late temporal mean (chunks 7–22) by its reference temporal mean (chunks 2–6), then report the median across all 34 layers. The common-fraction difference is also computed within each layer before taking its median. All raw chunk curves are retained, including startup and decreases. Layers and chunks are repeated measurements; identical Native pre/post records are counted once.

Table 34: All four registered 24 s Native diagnostics. The first five numeric columns are median layer-wise late/reference ratios; the last is the median absolute change in common fraction. PR denotes participation ratio; $E_\mu$ and $E_c$ are mean and centered energy, and $f$ is common fraction. No controller or video-quality effect is estimated.

| Scene | $E_\mu$ | $E_c$ | $f$ | $P(M)$ | $P(\Sigma)$ | $\Delta f$ |
|---|---|---|---|---|---|---|
| Camera motion | 1.1140 | 1.0352 | 1.0587 | 0.9352 | 1.0155 | +0.01074 |
| Scene change | 1.2332 | 0.9947 | 1.2085 | 0.8221 | 0.9932 | +0.02624 |
| Stable scene | 1.1309 | 1.0608 | 1.0490 | 0.9717 | 1.0505 | +0.00809 |
| Subject motion | 1.0495 | 0.9926 | 1.0664 | 0.9181 | 0.9483 | +0.00684 |

Mean energy and common fraction increase in each per-prompt summary, while uncentered PR decreases. Centered PR rises for camera motion and the stable scene and falls for scene change

and subject motion. Thus centered and uncentered concentration do not have a uniform shared trend. These reference-relative summaries do not imply monotonic growth, harmful drift, a precursor to perceptual failure, or controller efficacy. The earlier 12 s sailboat trace had the opposite post-reference mean-energy and uncentered-PR trends; differing prompts and durations prevent identifying a horizon effect. The completed two-prompt controller scans below reuse the camera-motion and scene-change Native videos. They do not complete the separate four-scene, three-arm registration.

### F.4 Complete MAGI Mean and Full development scans

We retain two separately specified scans on the camera-motion and scene-change prompts, both previously observed during development. They use MAGI-1 4.5B base, seed 73109, 576 frames at 24 fps, $720 \times 720$ resolution, 64 sampling steps, guidance 3, a four-chunk window and reference chunks 2–6 in all 34 layers. Each video is evaluated in two non-overlapping 288-frame windows using the seven VBench dimensions. Windows and prompts are averaged equally; imaging scores are divided by 100 once. All settings, both prompts and unfavorable outcomes are included. These are development comparisons with one seed, not independent significance tests.

The Mean scan contains Native and four $(\alpha, k)$ settings, where $k$ multiplies the reference-derived mean trust-region radius. It reuses four Native/default-Mean videos and generates six more, giving ten videos and 140 scores. The Full scan fixes the mean component at $(0.5, 2)$ and crosses centered tolerance $\lambda \in \{2, 4\}$ with rank $r \in \{4, 16\}$. It adds eight videos and reuses Native and Mean $(0.5, 2)$ for each prompt, giving 168 scores. The scans therefore contain 18 video assignments and 252 distinct scores, with four shared comparators. The Mean anchor was selected after the preceding development outcomes; its selection is not held-out validation.

Table 35: Complete exploratory MAGI Mean scan: two previously observed development prompts, one seed, 24 s per video. All ten videos and 140 raw window scores are included. Mean settings are $(\alpha, k)$, where $k$ is the trust-region radius multiplier. Each entry averages two fixed 12 s windows; imaging is divided by 100 once. All dynamic-degree scores are zero.

| Method | Subject | Background | Smooth | Aesthetic | Imaging | Text |
|---|---|---|---|---|---|---|
| Camera-motion prompt | | | | | | |
| Native | 0.99373 | 0.97395 | 0.99688 | 0.65829 | 0.78036 | 0.25401 |
| Mean $(1, 1)$ | 0.99409 | 0.97655 | 0.99711 | 0.66267 | 0.78230 | 0.25495 |
| Mean $(0.25, 1)$ | 0.99387 | 0.97755 | 0.99696 | 0.66001 | 0.78180 | 0.25470 |
| Mean $(0.5, 1)$ | 0.99415 | 0.97811 | 0.99706 | 0.66393 | 0.78208 | 0.25532 |
| Mean $(0.5, 2)$ | 0.99392 | 0.97826 | 0.99703 | 0.66156 | 0.78186 | 0.25419 |
| Scene-change prompt | | | | | | |
| Native | 0.99455 | 0.98571 | 0.99723 | 0.55128 | 0.74182 | 0.18250 |
| Mean $(1, 1)$ | 0.98612 | 0.98240 | 0.99697 | 0.53750 | 0.75316 | 0.18186 |
| Mean $(0.25, 1)$ | 0.99375 | 0.98810 | 0.99724 | 0.54980 | 0.74731 | 0.18007 |
| Mean $(0.5, 1)$ | 0.99134 | 0.98586 | 0.99710 | 0.54040 | 0.75059 | 0.18143 |
| Mean $(0.5, 2)$ | 0.99452 | 0.98746 | 0.99727 | 0.55118 | 0.74382 | 0.18156 |

Table 36: Mean–Native differences averaged equally over the two MAGI development prompts. These descriptive score differences do not imply significance. Each configuration passes the per-prompt 0.01 subject/background/smoothness tolerance, but none has nonnegative mean differences in all three aesthetic/imaging/text targets.

| $(\alpha, k)$ | Subject | Background | Smooth | Aesthetic | Imaging | Text |
|---|---|---|---|---|---|---|
| $(1, 1)$ | -0.00403 | -0.00035 | -0.00001 | -0.00470 | +0.00663 | +0.00015 |
| $(0.25, 1)$ | -0.00033 | +0.00300 | +0.00005 | +0.00011 | +0.00346 | -0.00088 |
| $(0.5, 1)$ | -0.00139 | +0.00216 | +0.00002 | -0.00262 | +0.00524 | +0.00012 |
| $(0.5, 2)$ | +0.00008 | +0.00303 | +0.00010 | +0.00158 | +0.00175 | -0.00038 |

**Local gains and cross-prompt tradeoffs.** All four Mean configurations improve subject/background consistency, smoothness, aesthetic, imaging and text scores on the camera-motion

prompt. For example, $(0.5, 1)$ improves aesthetic, imaging and text by $0.00564$, $0.00172$ and $0.00131$. On the scene-change prompt, all four lower aesthetic and text scores. The fixed selection rule requires each prompt's subject/background/smoothness difference to be at least $-0.01$ and pooled aesthetic/imaging/text differences to be nonnegative. All Mean settings pass the tolerance but none passes all three targets. This engineering tolerance is not a statistical non-inferiority margin validated for perception.

Table 37: Complete MAGI Full development scan: eight new videos and four reused comparators, 168 window scores. Full settings denote (centered tolerance, rank); its mean component uses $(\alpha, k) = (0.5, 2)$. Entries average two fixed 12 s windows. All dynamic-degree scores are zero.

| Method | Subject | Background | Smooth | Aesthetic | Imaging | Text |
|---|---|---|---|---|---|---|
| Camera-motion prompt | | | | | | |
| Native | 0.99373 | 0.97395 | 0.99688 | 0.65829 | 0.78036 | 0.25401 |
| Mean $(0.5, 2)$ | 0.99392 | 0.97826 | 0.99703 | 0.66156 | 0.78186 | 0.25419 |
| Full $(2, 4)$ | 0.99396 | 0.97848 | 0.99703 | 0.66073 | 0.78194 | 0.25438 |
| Full $(2, 16)$ | 0.99398 | 0.97883 | 0.99703 | 0.66102 | 0.78172 | 0.25451 |
| Full $(4, 4)$ | 0.99392 | 0.97826 | 0.99703 | 0.66156 | 0.78186 | 0.25419 |
| Full $(4, 16)$ | 0.99392 | 0.97826 | 0.99703 | 0.66156 | 0.78186 | 0.25419 |
| Scene-change prompt | | | | | | |
| Native | 0.99455 | 0.98571 | 0.99723 | 0.55128 | 0.74182 | 0.18250 |
| Mean $(0.5, 2)$ | 0.99452 | 0.98746 | 0.99727 | 0.55118 | 0.74382 | 0.18156 |
| Full $(2, 4)$ | 0.99445 | 0.98797 | 0.99722 | 0.55250 | 0.74337 | 0.18146 |
| Full $(2, 16)$ | 0.99446 | 0.98831 | 0.99722 | 0.55114 | 0.74365 | 0.18166 |
| Full $(4, 4)$ | 0.99451 | 0.98753 | 0.99727 | 0.55103 | 0.74395 | 0.18136 |
| Full $(4, 16)$ | 0.99451 | 0.98753 | 0.99727 | 0.55103 | 0.74395 | 0.18136 |

Table 38: Full differences averaged equally over both MAGI prompts. All candidates pass the fixed per-prompt subject/background/smoothness tolerance of $-0.01$ against each comparator. None improves or preserves all three pooled aesthetic/imaging/text targets against both comparators. No significance is inferred from this development scan.

| $(\lambda, r)$ | Subject | Background | Smooth | Aesthetic | Imaging | Text |
|---|---|---|---|---|---|---|
| Differences from Native | | | | | | |
| $(2, 4)$ | +0.000066 | +0.003398 | +0.000072 | +0.001828 | +0.001561 | -0.000334 |
| $(2, 16)$ | +0.000080 | +0.003744 | +0.000071 | +0.001288 | +0.001594 | -0.000176 |
| $(4, 4)$ | +0.000073 | +0.003069 | +0.000094 | +0.001508 | +0.001812 | -0.000485 |
| $(4, 16)$ | +0.000073 | +0.003069 | +0.000094 | +0.001508 | +0.001812 | -0.000485 |
| Differences from Mean $(0.5, 2)$ | | | | | | |
| $(2, 4)$ | -0.000013 | +0.000367 | -0.000026 | +0.000248 | -0.000187 | +0.000047 |
| $(2, 16)$ | +0.000001 | +0.000713 | -0.000027 | -0.000292 | -0.000154 | +0.000205 |
| $(4, 4)$ | -0.000006 | +0.000038 | -0.000004 | -0.000072 | +0.000063 | -0.000103 |
| $(4, 16)$ | -0.000006 | +0.000038 | -0.000004 | -0.000072 | +0.000063 | -0.000103 |

All four Full settings likewise improve the six camera-example scores over Native. Pooled subject, background, smoothness, aesthetic and imaging means exceed Native, while text is lower by $0.000176$–$0.000485$. No Full setting passes the joint targets against both Native and the Mean anchor, although all pass the per-prompt tolerances. Full $(\lambda, r) = (2, 4)$ gains $0.000248$ aesthetic and $0.000047$ text over Mean but loses $0.000187$ imaging; Full $(\lambda, r) = (2, 16)$ gains $0.000205$ text but loses $0.000292$ aesthetic and $0.000154$ imaging. These small, selected-development differences do not establish an additional Full benefit.

**Duplicate outputs, motion, and cost.** At tolerance $4$, ranks $4$ and $16$ produce byte-identical complete MP4s within each prompt. On the camera prompt they also match the Mean-anchor MP4. These assignments remain in the table, but are not independent replications or evidence for a centered-only contribution. A large tolerance can weaken centered shrinkage; the hashes alone do not identify its internal activation. All 36 distinct windows have zero dynamic-degree scores. AI-assisted inspection of fixed frames at 4, 12, 20 and 23 s does not establish the requested full camera

orbit or completed train departure; it is neither a human study nor continuous-video motion measurement. Thus smoother or more consistent frames cannot be interpreted as better motion realization. New generation and scoring consumed $19.5024$ GPU-hours for the Mean scan and $26.2791$ for Full; reused videos are not charged again. These shared-resource totals are accounting, not controlled overhead estimates.

### F.5 MAGI Full–Mean confirmation: incomplete evidence at the cutoff

A separate fixed confirmation assigns $16$ newly authored prompts, one seed ($92317$), and Mean/Full at $24$ s, for $32$ videos. Both arms use $(\alpha, k) = (0.5, 2)$; Full adds $(\lambda, r) = (2, 16)$. The text-consistency primary endpoint and these parameters were selected from the earlier development results and fixed before the new outcomes. There is no Native arm, so a completed comparison would address an increment over Mean, not benefit over the native model.

The original attempt terminated with $31/32$ complete generations. One Mean assignment was preempted after $13/24$ chunks, while its paired Full video completed. This was a recorded scheduler preemption, not an established numerical or out-of-memory failure. Under that attempt's failure policy, all $16$ scoring groups remained unsubmitted, leaving $0/448$ planned scores. Its reviewed generation accounting was $101.397846$ GPU-hours, including the cancellation tail, with zero scoring cost. This original terminal record is preserved in `evidence/magi_confirmation_status.json`.

A subsequently authorized execution amendment reused all $31$ complete videos and permitted recovery of the missing Mean assignment with unchanged scientific parameters. The archived observation of September 23, 2026, 06:37 UTC recorded the new attempt as running and all scoring groups as unsubmitted. Its prior cumulative accounting, including the first failed recovery, was $101.632013$ GPU-hours before the amended attempt's additional charge. This historical amendment and observation remain preserved in `evidence/magi_recovery_snapshot.json`.

The subsequent terminal review (September 23, 2026, 08:05 UTC) records the second authorized recovery as preempted and incomplete, with $31/32$ complete generations and $0/448$ planned quality scores. Cumulative generation accounting is $103.087290$ GPU-hours, including both recovery attempts. The terminal record reports no active known owners and no automatic retry; it is preserved in `evidence/magi_confirmation_terminal_20260923.json`.

No complete, independently reviewed quality result from this confirmation is available at the edition's cutoff, so it contributes no effect size, confidence interval or significance claim. This is incomplete evidence, not a completed nonsignificant comparison. No partial-pair inference or imputation is used. Resource charges describe shared-resource accounting, not controller overhead.

### F.6 Complete Causal Forcing long-video comparison

We evaluate the official Causal Forcing long-video EMA model (Zhu et al., 2026). It shares the Wan2.1-1.3B backbone family with Self-Forcing; this is a transfer test across trained generators, not a second independent architecture. Eight fixed prompts (two per subject-motion, camera-motion, scene-change and stable-scene stratum). Prompts are the first eight in the sealed Self-Forcing list and are not globally unseen. All assignments complete generation, continuous decoding and scoring; none is replaced. Each uses $960$ latent frames, $3{,}837$ decoded frames at $16$ fps ($239.81$ s), $480 \times 832$ resolution and bfloat16, with the native $24$-latent-frame cache and three-frame sink.

**Controller integration and measurement.** The controller edits only newly finalized clean $V$ once per chunk in all $30$ layers, after the final denoising step. Three unfinished startup writes are skipped. Mean uses $\alpha = 1$, radius multiplier $1$ and reference chunks 2–6; Full adds rank $16$ and centered tolerance $1$. The original controller and precision are retained. A native hook/no-hook check gives bitwise-identical latents and RNG state, and paired arms share initial-noise hashes. All $320$ chunks have basic measurements; exact concentration terms use fixed layers $0$, $14$ and $29$. Early and late intervals are chunks 7–53 and 273–319. We average time within layer, layers within trajectory, then the two seeds within prompt. Layer/chunk observations are repeated measurements, not statistical replicates.

Quality uses all 6,720 records from 20 non-overlapping windows per video ($19 \times 192 + 189$ frames), the same seven dimensions, frame-weighted window means and equal seed means. Eight prompts are the paired statistical units. The fixed four primary tests form one two-sided paired-$t$ Holm family; 95% intervals are unadjusted. All other comparisons are descriptive.

Table 39: Causal Forcing long-model evaluation: all 48 videos, eight prompts and two seeds per arm. Scores use frame-weighted windows and equal seeds. Full–Mean intervals are unadjusted prompt-paired 95% $t$ intervals; they are not equivalence or non-inferiority tests.

| Metric | Native | Mean | Full | Full–Mean | 95% CI |
|---|---|---|---|---|---|
| Subject | 0.90869 | 0.90485 | 0.90277 | $-0.00208$ | $[-0.01351, +0.00936]$ |
| Background | 0.92336 | 0.92611 | 0.92116 | $-0.00495$ | $[-0.00847, -0.00143]$ |
| Smoothness | 0.97217 | 0.97288 | 0.97228 | $-0.00060$ | $[-0.00267, +0.00147]$ |
| Dynamic degree | 0.96252 | 0.92812 | 0.96882 | $+0.04071$ | $[-0.03001, +0.11142]$ |
| Aesthetic | 0.58360 | 0.58028 | 0.57294 | $-0.00734$ | $[-0.02020, +0.00552]$ |
| Imaging | 0.71956 | 0.70483 | 0.71604 | $+0.01121$ | $[-0.00369, +0.02611]$ |
| Text | 0.23895 | 0.23568 | 0.23898 | $+0.00330$ | $[-0.00593, +0.01252]$ |

Table 40: All four fixed Causal Forcing primary comparisons form one Holm family. None is significant. These eight prompt pairs do not establish an absence of effects.

| Comparison / metric | Difference | 95% CI | Raw $p$ | Holm $p$ |
|---|---|---|---|---|
| Mean–Native / Aesthetic | $-0.00332$ | $[-0.01268, +0.00604]$ | 0.4292 | 1.0000 |
| Mean–Native / Text | $-0.00327$ | $[-0.01325, +0.00670]$ | 0.4634 | 1.0000 |
| Full–Mean / Subject | $-0.00208$ | $[-0.01351, +0.00936]$ | 0.6808 | 1.0000 |
| Full–Mean / Text | $+0.00330$ | $[-0.00593, +0.01252]$ | 0.4265 | 1.0000 |

**Representation effects without confirmed quality transfer.** Native common fraction is 0.20308 early and 0.20294 late, with change $-0.00014$ (95% CI $[-0.00483, 0.00454]$); four of eight prompt changes are positive and none has late fraction above 0.5. Common-energy change is 0.45443 ($[-6.67571, 7.58456]$), and centered-energy change is $-0.11200$ ($[-24.57759, 24.35359]$). The concentration change in the three fixed layers is 0.00321 ($[-0.00400, 0.01043]$), decomposing into common 0.00392, alignment 0.00111 and centered $-0.00182$. Dividing the common change by this small signed total gives 122.07%, which is neither a total-energy fraction nor reliable evidence of dominant growth.

Both controllers reduce late mean-reference distance $\|\mu - \mu_{\text{ref}}\|_2$ in all eight prompt averages: Mean–Native is $-4.61128$ ($[-5.75039, -3.47218]$), and Full–Native is $-4.61257$ ($[-5.75071, -3.47443]$). Mean lowers late common fraction by 0.03452, while Full changes it by $-0.00081$ ($[-0.00906, 0.00744]$); shrinking centered energy also changes the fraction's denominator. Full's recorded centered shrink multiplier is below one at every intervention layer/chunk, demonstrating execution rather than benefit. All four quality tests have Holm $p = 1$. Full–Mean text is positive on six prompts, but its interval includes zero; imaging increases descriptively and aesthetic decreases. The unadjusted negative background interval in Table 39 is a secondary observation, not a separate multiplicity-corrected harm claim. The complete study consumed 22.8333 allocated GPU-hours including checks, generation, decoding and scoring; it does not estimate exclusive-GPU overhead.

### F.7 Native Causal Forcing follow-up: energy and spectral direction

A separate fixed Native-only study uses the next eight prompts in the sealed list and new seeds 20270413, 20270429. They do not overlap the preceding Causal Forcing prompts, but are reused Self-Forcing prompts. All 16 960-latent-frame trajectories complete with 67,200 observer records; only one preselected trajectory is continuously decoded for visualization. There are no new controlled-arm or quality measurements. A 72-latent-frame native hook/no-hook check preserves bitwise latents and RNG. The full protocol, all layer/head summaries and input hashes are released.

**Objects and fixed tests.** At 28 fixed autoregressive positions, measurements cover all 30 layers and 12 heads: newly finalized clean values, the complete stored clean-value cache, and actual terminal-denoising attention. The last object uses the first $4{,}680$ query outputs belonging to the chunk being completed. For values, $f = \|\mu\|^2/(\|\mu\|^2 + \operatorname{tr}\Sigma)$; for actual attention output $O$, $g = \|O\hat{\mu}\|_2^2/\|O\|_F^2$ measures energy projected onto the actual attended value pool's unit mean direction $\hat{\mu}$. Heads are concatenated before these layer-level fractions are computed. Actual pools include anchors, recent clean tokens and current noisy tokens; they are distinct from newly written or stored clean $V$. Startup, early and late intervals are blocks 0–6, 7–53 and 273–319. Within an interval we average sampled times, layers and seeds equally before paired analysis across eight prompts. The four fixed tests share one Holm family. The predeclared post-startup dominance criterion requires both a positive change with Holm $p < 0.05$ and a late-level $95\%$ interval lower bound above $0.5$, for the same object.

Table 41: Native-only Causal Forcing follow-up: all 16 generated trajectories, eight prompts and two seeds. Four two-sided paired tests share one Holm family. $V$ entries measure common-energy fraction $f$; attention entries measure projected output energy $g$. Each interval is an unadjusted $95\%$ prompt-level interval.

| Object / contrast | Earlier | Later | Difference | $95\%$ CI | Holm $p$ |
|---|---|---|---|---|---|
| New $V$: early–startup | 0.16300 | 0.20077 | $+0.03777$ | $[+0.02639, +0.04915]$ | 0.00041 |
| New $V$: late–early | 0.20077 | 0.20925 | $+0.00848$ | $[-0.00190, +0.01886]$ | 0.20126 |
| Stored $V$: late–early | 0.17298 | 0.18156 | $+0.00858$ | $[-0.00079, +0.01795]$ | 0.20126 |
| Attention $g$: late–early | 0.28251 | 0.28673 | $+0.00422$ | $[-0.01174, +0.02018]$ | 0.55167 |

The startup increase is supported, but none of the three post-startup objects meets the joint criterion. New-$V$ common fraction rises from $0.16300$ at startup to $0.20077$ early and $0.20925$ late. This establishes neither continued majority-energy growth nor its universal absence. The terminal cache flush changes query/pool shapes. A descriptive sensitivity restricted to the maximum query shape gives attention-$g$ change $0.00495$ ($[-0.01052, 0.02041]$) and actual-pool-$f$ change $0.00200$ ($[-0.00353, 0.00753]$), retaining this limitation. All comparisons keep the denoising phase fixed: they concern autoregressive video position, not successive denoising steps within a chunk.

**A strong spectral direction need not hold majority energy.** An exploratory description uses all 36 heads in fixed layers 0, 14 and 29, with exact 128-dimensional spectra of the actual attention value pool. Let $u_1(M)$ be the leading eigenvector of its uncentered second moment. Squared mean-direction alignment rises from $0.7269$ to $0.7759$ on average, and $\|\mu\|^2/\lambda_1(\Sigma)$ rises from $1.1398$ to $1.4565$. The latter compares rank-one common strength with the strongest centered mode, not with total energy. Thus it can exceed one while $f < 0.5$.

Table 42: Exploratory spectral description of all 36 heads in three fixed layers of the same Native follow-up trajectories. $u_1(M)$ is the leading uncentered eigenvector. Prompt-level $95\%$ intervals are unadjusted; these were not additional confirmatory tests. Neither quantity measures total-energy majority.

| Quantity | Early | Late | Difference | $95\%$ CI |
|---|---|---|---|---|
| $\cos^2(\mu, u_1(M))$ | 0.7269 | 0.7759 | $+0.0490$ | $[+0.03365, +0.06443]$ |
| $\|\mu\|^2/\lambda_1(\Sigma)$ | 1.1398 | 1.4565 | $+0.3167$ | $[+0.21933, +0.41403]$ |

Of 36 head-level prompt means, 32 have alignment above $0.5$ in both intervals; 21 early and 29 late have the strength ratio above one. All heads are retained. Layer 14 has a small alignment decrease ($0.8372$ to $0.8288$), so the aggregate increase is not universal across layers. Centered top eigenvalues also decline ($30.013$ to $27.346$), and pool composition changes near the terminal flush; no separate flush-excluded spectral sensitivity was performed. These unadjusted exploratory intervals do not replace the fixed primary tests or establish causal quality effects. The study used $2.1411$ allocated GPU-hours including checks and its one decode, without additional quality scoring. These studies bound the transfer claim: a measurable change in cache statistics can occur without a confirmed perceptual gain, and an incomplete confirmation cannot be interpreted as either a positive or a null result.

# G EXPLORATORY ADAPTATION TO CORRECTED HISTORY

These optional training studies ask whether exposure to corrected history helps the generator use Full. They are separate from the training-free method and its main comparison, and retain every completed configuration and the incomplete gradient-path attempt.

## G.1 EXPLORATORY ADAPTATION TO CONTROLLED HISTORY

**Training comparison.** We test whether exposing a model to controlled history during adaptation improves its subsequent use of Full. Both training arms start from the same Self-Forcing/Wan2.1-T2V-1.3B EMA generator and a zero-output LoRA initialization. The generator base is frozen; rank-8 LoRA on the query, value and output projections of all 30 attention layers contains $2{,}211{,}840$ trainable parameters. Conventional adaptation uses uncontrolled history, while controller-aware adaptation applies Full during the clean context refresh. The initial batches complete either 2 generator and 10 critic updates or 10 generator and 50 critic updates per arm. A later registered continuation reaches 50 generator and 250 critic updates per arm. Within each batch, the arms match the fixed four-prompt training schedule, seeds, prefix lengths, update counts and trainable generator size. They do not match GPU time.

The objective is DMD with generator/critic learning rates $2 \times 10^{-6}/4 \times 10^{-7}$, batch size 1 and bfloat16 computation. The real-score teacher is the available Wan2.1-T2V-1.3B, a resource adaptation from the original recipe's 14B teacher. Prefixes of 8 or 16 chunks are generated without gradients. The teacher-supported clip contains 21 latent frames, with the generator loss masked to the final 3 frames of the current suffix. Historical KV states and reference statistics are detached; the controller is not differentiated and there is no additional spectral loss. Full retains rank 16, zero-based reference indices $[2, 7)$, and intervention on newly written values at the no-gradient clean-context refresh. This tests training with exposure to corrected history, while leaving the history-writing gradient path outside the completed quality comparison.

**Completed evaluation and sampling units.** Each panel below contains the frozen model, conventional adapter and controller-aware adapter, each evaluated with inference control Off or Full. All videos use seed 9102701, $832 \times 480$ pixels and 16 fps. The short evaluations contain 141 frames (8.8125 s) and use different prompts: one describes a child moving a toy sailboat while the camera tracks sideways, and the other describes a pastry chef folding dough. Both long evaluations contain 957 frames (59.8125 s) and use a prompt describing a tram emerging from an underpass into a square. Note that these prompts have been used for development checks; therefore, they do not confer an unseen-test guarantee.

The short panels each contain 42 raw metric records. Each long panel contains 210: six videos, five temporal windows and seven dimensions. Window lengths are $(192, 192, 192, 192, 189)$ frames, and their frame-weighted mean gives one value per trajectory and dimension. The windows are repeated observations of one prompt and seed. The two long batches also use different generation hosts; differences between batches are not estimates of the effect of increasing training from two to ten updates. All comparisons reported here are descriptive and within their own batch, without confidence intervals, significance tests, a custom overall score or human-preference claims.

**Observed quality tradeoffs.** Neither the short nor the long evaluations establish that controller-aware adaptation with Full improves quality consistently over both core baselines. In the ten-update long-video batch, controller-aware training with Full increases 2 of the six quality/text dimensions and decreases 4 relative to frozen EMA with Full; against conventional training with Full, 3 increase and 3 decrease. For example, imaging quality changes by $-0.039466$ and text/video consistency by $-0.022307$ relative to frozen EMA with Full, while the corresponding differences against conventional training with Full are $+0.005550$ and $+0.022507$. These are descriptive differences in normalized score units, not percent improvements or statistical effect estimates. All three Full configurations have dynamic degree 1 in every panel. For the two-update long batch, controller-aware inference with control Off has dynamic degree $0.200627$, versus 1 with Full; its higher aesthetic and imaging scores must therefore be read alongside this change in activity. The corresponding Off value in the ten-update long batch is $0.802508$. These measurements do not by themselves establish visual freezing or whether the motion follows the prompt.

Table 43: Complete short-video adaptation results: six configurations per panel, 141 frames at 16 fps (8.8125 s), one prompt and seed per panel. The panels use different prompts and are not a training curve. Imaging quality is divided by 100 once; all other scores retain their raw scale. Dynamic degree measures activity, not motion correctness.

2 *generator*/10 *critic updates per arm; exploratory* `test_00`

| | Frozen EMA | | Conventional | | Controller-aware | |
|---|---|---|---|---|---|---|
| Dimension | Off | Full | Off | Full | Off | Full |
| Subject consistency | 0.901003 | 0.904740 | 0.903564 | 0.907546 | 0.892840 | 0.898227 |
| Background consistency | 0.890247 | 0.897518 | 0.909344 | 0.905636 | 0.899679 | 0.901045 |
| Motion smoothness | 0.960869 | 0.961181 | 0.963055 | 0.962362 | 0.960974 | 0.958210 |
| Dynamic degree | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Aesthetic quality | 0.600703 | 0.603562 | 0.594354 | 0.584867 | 0.585418 | 0.589463 |
| Imaging quality | 0.766201 | 0.769698 | 0.772909 | 0.772911 | 0.764801 | 0.764012 |
| Text/video consistency | 0.282874 | 0.275954 | 0.275035 | 0.274059 | 0.291125 | 0.282801 |

10 *generator*/50 *critic updates per arm; development* `dev_00`

| | Frozen EMA | | Conventional | | Controller-aware | |
|---|---|---|---|---|---|---|
| Dimension | Off | Full | Off | Full | Off | Full |
| Subject consistency | 0.848646 | 0.838289 | 0.851016 | 0.857726 | 0.848495 | 0.843915 |
| Background consistency | 0.941534 | 0.939617 | 0.943713 | 0.943682 | 0.936977 | 0.935936 |
| Motion smoothness | 0.983470 | 0.984401 | 0.984036 | 0.983906 | 0.986844 | 0.986856 |
| Dynamic degree | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 | 1.000000 |
| Aesthetic quality | 0.635366 | 0.627195 | 0.629539 | 0.636858 | 0.631591 | 0.635500 |
| Imaging quality | 0.729987 | 0.737848 | 0.732064 | 0.723156 | 0.732590 | 0.733452 |
| Text/video consistency | 0.236837 | 0.237456 | 0.242911 | 0.243077 | 0.236620 | 0.235567 |

Table 44: Complete long-video adaptation results: six configurations per panel on the tram prompt, seed 9102701, with 957 frames at 16 fps (59.8125 s). Each entry weights five windows by their frame counts (192, 192, 192, 192, 189). Score scales match Table 43. The two batches are separate development checks, not a causal comparison of two versus ten training updates.

2 *generator*/10 *critic updates per arm; first long-video batch*

| | Frozen EMA | | Conventional | | Controller-aware | |
|---|---|---|---|---|---|---|
| Dimension | Off | Full | Off | Full | Off | Full |
| Subject consistency | 0.895917 | 0.816475 | 0.883569 | 0.861775 | 0.898152 | 0.785342 |
| Background consistency | 0.910492 | 0.897464 | 0.909814 | 0.895264 | 0.917897 | 0.880173 |
| Motion smoothness | 0.979018 | 0.982834 | 0.979471 | 0.985691 | 0.988743 | 0.982628 |
| Dynamic degree | 0.802508 | 1.000000 | 1.000000 | 1.000000 | 0.200627 | 1.000000 |
| Aesthetic quality | 0.389062 | 0.517673 | 0.441384 | 0.488080 | 0.569604 | 0.519762 |
| Imaging quality | 0.693433 | 0.680611 | 0.636722 | 0.693677 | 0.758825 | 0.620313 |
| Text/video consistency | 0.123525 | 0.225810 | 0.160369 | 0.196792 | 0.207086 | 0.231021 |

10 *generator*/50 *critic updates per arm; subsequent long-video batch*

| | Frozen EMA | | Conventional | | Controller-aware | |
|---|---|---|---|---|---|---|
| Dimension | Off | Full | Off | Full | Off | Full |
| Subject consistency | 0.895917 | 0.816475 | 0.903263 | 0.836253 | 0.901335 | 0.817502 |
| Background consistency | 0.910492 | 0.897464 | 0.901452 | 0.899638 | 0.924995 | 0.890337 |
| Motion smoothness | 0.979018 | 0.982834 | 0.982381 | 0.981173 | 0.986923 | 0.982768 |
| Dynamic degree | 0.802508 | 1.000000 | 1.000000 | 1.000000 | 0.802508 | 1.000000 |
| Aesthetic quality | 0.389062 | 0.517673 | 0.510460 | 0.539539 | 0.495973 | 0.522471 |
| Imaging quality | 0.693433 | 0.680611 | 0.737611 | 0.635594 | 0.734549 | 0.641144 |
| Text/video consistency | 0.123525 | 0.225810 | 0.145127 | 0.180996 | 0.188873 | 0.203504 |

**Execution evidence and the gradient boundary.** All three Full configurations in each long batch record 2,190 corrected layer-write events with actual bfloat16 changes and 72 chunks reading controlled history across 30 layers (2,160 layer–chunk reads). These records establish that the adapted model uses the modified state during long generation; the quality tables provide the separate assessment of its effects. Ideal pre-cast correction diagnostics are not treated as actual post-cast norm doses. The first long batch's frozen-model/Off trajectory retains missing complete cache, RNG and per-configuration timing records; its previously verified video is included without reconstructing those logs.

A separate local-history differentiation attempt verified one actual generator update through two recent clean-cache writers: all 120 pre/post gradient hooks across 30 layers were finite and nonzero, with actual bfloat16 changes and matched first-step clip/RNG records. That attempt was stopped before its second generator update because of predicted memory-limit pressure. It supplies a single-step gradient-path check and no completed quality comparison. All rows in Tables 43 and 44 use detached-history adaptation. That single-step check contributes no quality result to these tables. The complete longer-training continuation is reported separately below.

**Reproducibility.** A CPU-only script verifies the frame-weighted aggregation and reproduces Tables 43 and 44 from the released scalar records. Imaging quality is divided by 100 exactly once; other dimensions retain their native score scales.

**Complete fifty-update continuation.** A separately fixed evaluation uses four prompts authored after the earlier results, held out from the four-prompt training schedule, with seeds 951501 and 951502. It retains all 48 assigned videos: six configurations for each of the eight prompt–seed pairs. Each video has 957 frames at 16 fps; five non-overlapping windows provide 1,680 raw metric records in total. The final checkpoint is fixed at 50 generator and 250 critic updates per arm, with both optimizers and the trained critic restored from the ten-update continuation. No checkpoint is selected using quality. The teacher, LoRA, losses and detached-history treatment remain as above.

Table 45: Complete 50-update developmental evaluation: all six configurations on four prompts and two seeds at 59.8125 s. Scores average frame-weighted windows, then seeds within prompts and prompts equally. Imaging is divided by 100 once. Dynamic degree measures activity; larger is not uniformly better. These descriptive means have no hypothesis tests or confidence intervals and are separate from the independent cohort.

| Model / control | Subject | Background | Smooth | Dynamic | Aesthetic | Imaging | Text |
|---|---|---|---|---|---|---|---|
| Frozen / Off | 0.9062 | 0.9343 | 0.9855 | 0.5263 | 0.6184 | 0.6291 | 0.2537 |
| Frozen / Full | 0.9177 | 0.9409 | 0.9837 | 0.4498 | 0.5801 | 0.5615 | 0.2789 |
| Conventional / Off | 0.9180 | 0.9371 | 0.9857 | 0.5004 | 0.6064 | 0.6088 | 0.2463 |
| Conventional / Full | 0.9152 | 0.9370 | 0.9827 | 0.5000 | 0.5847 | 0.5837 | 0.2800 |
| Controller-aware / Off | 0.9256 | 0.9465 | 0.9881 | 0.4506 | 0.6170 | 0.6411 | 0.2524 |
| Controller-aware / Full | 0.9142 | 0.9349 | 0.9812 | 0.5748 | 0.5818 | 0.5617 | 0.2793 |

Controller-aware adaptation with Full has small positive differences in mean aesthetic, imaging and text scores relative to frozen-model Full ($+0.00169$, $+0.00026$, $+0.00043$), alongside lower subject consistency, background consistency and smoothness, and higher dynamic degree. Against conventional-adaptation Full, its aesthetic, imaging and text scores are lower ($-0.00296$, $-0.02195$, $-0.00071$) as retained in Table 45. The completed adaptation results are mixed; a larger learned benefit is not needed for, or established by, the present cache-controller evidence.