TOWARDS REALISTIC TEMPORAL DYNAMICS IN VIDEO DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-video diffusion models have achieved remarkable progress in visual fidelity and semantic alignment, but still struggle to reproduce the temporal dynamics observed in real videos. We characterize this discrepancy along two complementary dimensions: *motion magnitude*, which measures how much temporal change occurs, and *temporal pacing*, which describes how that change unfolds over time. Across three video generators, we observe similar temporal miscalibration patterns, while the required calibration varies across models and individual samples. Based on these observations, we introduce *Adaptive Temporal Coordinate Calibration*, an inference-time method that reparameterizes the temporal RoPE coordinates of a pretrained video diffusion model. During denoising, we predict temporal statistics from intermediate representations and calibrate them toward real video dynamics. We evaluate our method on LTX-Video-13B, Wan2.1-T2V-1.3B, and CogVideoX-5B using held-out SSv2 and Panda-70M videos with external temporal benchmarks, including ChronoMagic-Bench and Ref4D-VideoBench. On the held-out dataset, our method reduces M-Gap by 17.2% and Prog-W_1 by 19.0% on average across the three evaluated models, while largely preserving visual quality. The method requires no additional denoising steps and adds less than 0.7% overhead across the evaluated models.

1 INTRODUCTION

Text-to-video (T2V) diffusion models have advanced rapidly in visual fidelity, semantic alignment, and overall generation quality (Wan et al., 2025; Yang et al., 2025; HaCohen et al., 2024). However, reproducing the temporal dynamics observed in real videos remains challenging. As shown in Fig. 1, generated videos can contain visually plausible frames while exhibiting insufficient motion or unrealistic temporal progression. Recent studies have reported related limitations, including incorrect physical time scales, temporal stagnation and abrupt transitions, and underdeveloped irreversible processes (Gao et al., 2026; Metzer et al., 2026; Xu et al., 2026).

We study this discrepancy through two complementary properties: *motion magnitude*, which measures how much temporal change occurs, and *temporal pacing*, which describes how that change is distributed over time. Across LTX-Video-13B (HaCohen et al., 2024), Wan2.1-T2V-1.3B (Wan et al., 2025), and CogVideoX-5B (Yang et al., 2025), generated videos consistently exhibit lower motion magnitude, although the degree of mismatch varies across models and samples. Temporal pacing, in contrast, shows weak correspondence between real and generated videos, with the mismatch appearing mainly in the overall distribution. These observations motivate different calibration strategies for the two properties, while effective correction also depends on when each property can still be reliably predicted and controlled during generation.

We therefore examine how the predictability and controllability of these temporal statistics change throughout denoising. Intermediate representations become increasingly informative about the final temporal statistics, while the remaining influence of a temporal coordinate intervention generally decreases. This creates a read–write trade-off between prediction accuracy and control authority. Effective calibration must therefore determine not only what correction is needed, but also when to apply it along the denoising trajectory.

Based on these observations, we introduce *Adaptive Temporal Coordinate Calibration*, an inference-time framework that calibrates temporal dynamics by reparameterizing the temporal RoPE coordi-

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.

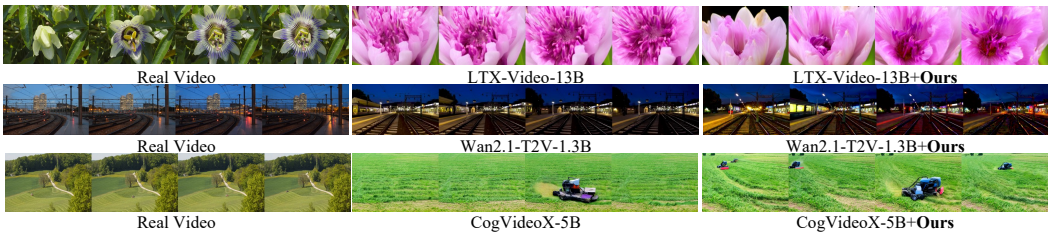


Figure 1: **Real–generated temporal dynamics across video backbones.** For each backbone, we compare uniformly sampled frames from a real reference video, the corresponding baseline generation, and our calibrated result under the same text condition. Baseline generations can exhibit insufficient or differently allocated temporal change relative to the real video, whereas our method adjusts both motion magnitude and temporal pacing toward the observed real-video dynamics.

nates of a pretrained video diffusion model. We use two low-dimensional controls: α primarily adjusts motion magnitude, while γ redistributes temporal change over time. During denoising, intermediate representations are used to predict the temporal statistics, and model-specific intervention steps are selected based on their predictability and remaining control authority. Real-video calibration statistics then provide the targets for calibrating the two temporal properties. The resulting procedure calibrates motion magnitude and temporal pacing without modifying the pretrained denoiser or adding denoising steps.

We evaluate our framework on three T2V backbones: LTX-Video-13B, Wan2.1-T2V-1.3B, and CogVideoX-5B. We use held-out videos from Something-Something v2 (Goyal et al., 2017) and Panda-70M (Chen et al., 2024), together with external temporal benchmarks including ChronoMagic-Bench (Yuan et al., 2024), Ref4D-VideoBench (Wei et al., 2026), and VBench (Huang et al., 2024). On the held-out SSv2 and Panda-70M set, our method reduces M-Gap by 17.2% and Prog-W_1 by 19.0% on average across the three models while largely preserving visual quality, and further improves MT and Prog-W_1 for all three models on ChronoMagic-Bench.

Our contributions are summarized as follows:

- We analyze the mismatch between real and generated temporal dynamics through motion magnitude and temporal pacing. Across three T2V generators, we observe consistent under-motion, distinct pacing discrepancies, and a read–write trade-off between prediction accuracy and control authority during denoising.
- We introduce *Adaptive Temporal Coordinate Calibration*, which predicts temporal statistics from intermediate representations and calibrates motion magnitude and temporal pacing through temporal RoPE reparameterization using real-video statistics.
- We demonstrate improved temporal dynamics across three video diffusion models on held-out real-video data and external temporal benchmarks, while preserving visual quality and requiring no additional denoising steps.

2 Related Work

2.1 Video Generation and Temporal Positional Control

Modern text-to-video diffusion models represent temporal positions using positional encodings such as rotary positional embeddings (RoPE) (Wan et al., 2025; Yang et al., 2025; HaCohen et al., 2024). Recent methods use these coordinates for inference-time control. RIFLEx (Zhao et al., 2025) modifies RoPE frequencies for temporal length extrapolation, while RoPECraft (Gökmen et al., 2025) and Infinity-RoPE (Yesiltepe et al., 2026) reparameterize temporal coordinates for motion control and long-horizon generation. Our work instead uses temporal coordinates to calibrate generated dynamics toward real-video statistics based on the predicted temporal statistics of each generation.

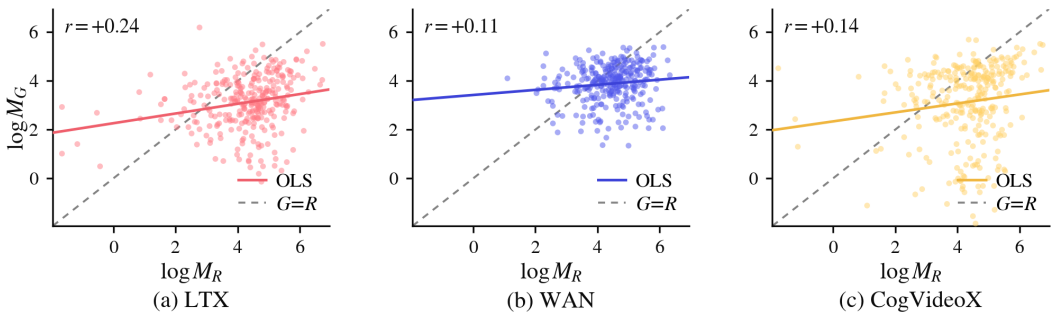


Figure 2: **Paired real–generated motion magnitude.** Each point compares $\log M$ of a real video (R) and its caption-matched generation (G) for LTX-Video, Wan2.1, and CogVideoX (left to right). The dashed line denotes perfect calibration and the solid line the ordinary least-squares fit.

2.2 Temporal Dynamics in Video Diffusion Models

Recent work has increasingly examined temporal dynamics in video generation. The Pulse of Motion (Gao et al., 2026) studies incorrect physical time scales, ReTime (Metzer et al., 2026) analyzes nonlinear temporal progression, and related work investigates failures in irreversible processes (Xu et al., 2026). FlowMo (Shaulov et al., 2025) improves motion coherence by extracting temporal structure from model predictions and using it for inference-time guidance. Other studies show that intermediate representations contain physical and motion-related information (Esmati et al., 2026; Ding et al., 2026). Our work instead characterizes temporal dynamics through motion magnitude and temporal pacing, and uses their intermediate predictions to calibrate generation toward real-video statistics.

2.3 Evaluating Temporal Dynamics in Generated Videos

Video-generation benchmarks evaluate complementary aspects of temporal quality. VBench (Huang et al., 2024) measures motion activity, smoothness, and temporal consistency, and ChronoMagic-Bench (Yuan et al., 2024) focuses on temporal progression in time-lapse generation. Ref4D-VideoBench (Wei et al., 2026) further provides reference-based evaluation against real videos. Together, these benchmarks cover complementary properties including motion magnitude, temporal consistency, and event progression. Whereas these benchmarks are designed for post-hoc evaluation, we use real-video temporal statistics as reference information for inference-time calibration.

3 Analysis of Temporal Miscalibration

Generated videos can appear visually plausible while their temporal dynamics differ from those of real videos. We characterize this discrepancy through two complementary properties: *motion magnitude*, which measures how much temporal change occurs, and *temporal pacing*, which describes how that change is distributed over time. We then examine how the predictability and controllability of these statistics change during denoising.

3.1 Motion Magnitude Miscalibration

Given a video $\mathbf{x} = \{I_t\}_{t=0}^{T-1}$, we define the frame-rate-normalized temporal change and its average as

$$m_t = \overline{|I_{t+1} - I_t|} \cdot \text{fps}, \qquad M(\mathbf{x}) = \frac{1}{T-1} \sum_{t=0}^{T-2} m_t.$$

Here, M summarizes the overall amount of temporal change in a video.

As shown in Fig. 2, real and generated motion magnitudes are weakly correlated. The paired correlations are 0.24, 0.11, and 0.14 for LTX-Video-13B (HaCohen et al., 2024), Wan2.1-T2V-1.3B (Wan et al., 2025), and CogVideoX-5B (Yang et al., 2025), respectively. All three models also exhibit

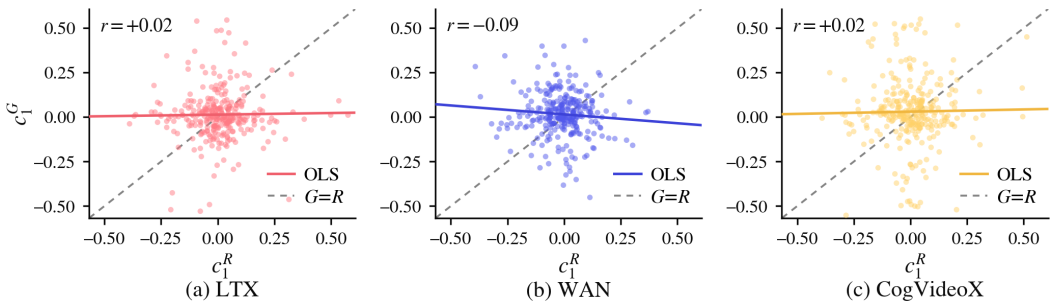


Figure 3: **Paired real–generated temporal pacing.** We compare c_1 between caption-matched real and generated videos for LTX-Video, Wan2.1, and CogVideoX. Pairwise correlations remain close to zero across all three models.

under-motion, with generated-to-real geometric-mean ratios of 0.32, 0.55, and 0.33. Thus, the direction of the magnitude mismatch is shared across models, while its severity and required correction vary across models and samples. Detailed statistics are provided in Appendix C.1.

3.2 Temporal Pacing Miscalibration

Motion magnitude does not capture when temporal change occurs. We therefore define *temporal pacing* from the cumulative temporal change profile. After resampling m_t to L bins, let

$$u_i = \frac{i + \frac{1}{2}}{L}, \qquad C_i = \frac{\sum_{j=0}^{i} \tilde{m}_j}{\sum_{j=0}^{L-1} \tilde{m}_j}.$$

We measure the deviation from uniform progression using

$$c_1 = \frac{\sum_{i=0}^{L-1} (C_i - u_i) \sin(\pi u_i)}{\sum_{i=0}^{L-1} \sin^2(\pi u_i)}. \tag{1}$$

Positive c_1 indicates earlier accumulation of temporal change, negative c_1 indicates later accumulation, and values near zero indicate a more uniform progression.

Pooling SSv2 and Panda-70M, temporal pacing shows little sample-wise correspondence between real and generated videos (Fig. 3), with paired correlations of 0.02, −0.09, and 0.02. Its mismatch instead appears mainly in the overall distribution. On SSv2 (Goyal et al., 2017), generated pacing is more dispersed than real pacing, with standard-deviation ratios of 1.74×, 1.46×, and 2.34×. On Panda-70M (Chen et al., 2024), the direction reverses, with ratios of 0.66×, 0.84×, and 0.85×. Thus, the pacing discrepancy depends on the video domain rather than following a single global bias. This motivates calibrating pacing toward real-video distributions rather than directly matching each generation to its paired real video.

3.3 Read–Write Trade-off During Denoising

Inference-time calibration depends on both whether a temporal statistic can be predicted and whether it can still be modified. We measure *readability* by the cross-validated correlation between predictions from intermediate hidden representations and the final $\log M$ or c_1, and *control authority* by the effect of applying the corresponding temporal-coordinate control from step s onward. Write responses are normalized by the response at the first evaluated step.

Figure 4 shows that readability and control authority evolve differently across models and temporal properties. At the start of denoising, motion magnitude is already moderately predictable (0.36, 0.49, and 0.60 for LTX-Video, Wan2.1, and CogVideoX), whereas pacing remains largely unreadable (−0.09, −0.09, and −0.05). Pacing becomes informative later, particularly for LTX-Video.

Meanwhile, control authority generally decreases over denoising, with model-dependent intervention windows. LTX-Video rapidly loses magnitude control after $s = 6$ while retaining later pacing control, CogVideoX favors earlier intervention, and Wan2.1 exhibits a broader magnitude-control

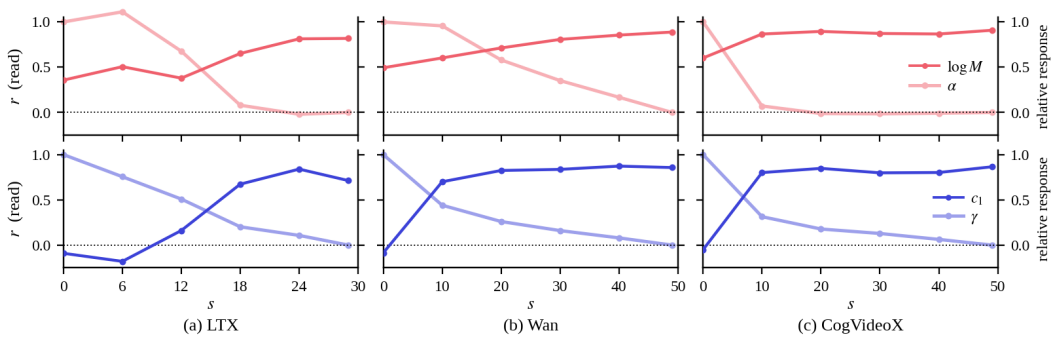


Figure 4: **Read–write trade-off across denoising.** Dark curves show prediction correlation for $\log M$ (top) and c_1 (bottom), while light curves show the remaining response to α and γ, respectively. Columns correspond to LTX-Video-13B, Wan2.1-T2V-1.3B, and CogVideoX-5B.

window. Thus, later representations can improve prediction while reducing the remaining effect of temporal-coordinate control, motivating the model-specific intervention schedules used in Sec. 4.

We interpret these curves as diagnostic response profiles rather than a direct rule for selecting intervention steps. The final operating points additionally account for model-specific control responses and validated operating ranges. Consequently, an earlier intervention can be preferred when delaying it substantially reduces write authority, even if later representations are more predictive.

4 Adaptive Temporal Coordinate Calibration

We introduce Adaptive Temporal Coordinate Calibration, a two-component inference-time framework for aligning motion magnitude and temporal pacing with real-video dynamics. Stage I denotes magnitude calibration and Stage II pacing calibration, and their chronological order is model dependent. Using the read–write profiles in Sec. 3.3, we select model-specific intervention points that balance prediction accuracy and remaining control authority. The pretrained generator remains unchanged, with no additional denoising steps.

Let $\ell \in \{0, \dots, T_\tau - 1\}$ denote the native temporal RoPE index and $u = \ell/(T_\tau - 1)$. We reparameterize the temporal coordinate as

$$\tau_{\alpha,\gamma}(\ell) = \alpha\ell + \gamma(T_\tau - 1)\sin(\pi u), \tag{2}$$

where α primarily controls motion magnitude and γ redistributes temporal change over time. The identity setting $(\alpha, \gamma) = (1, 0)$ recovers the pretrained model.

We require $\alpha > \pi|\gamma|$ to preserve temporal ordering and restrict both controls to empirically validated ranges. We use $(s_M, s_{c_1}) = (6, 17)$ for LTX-Video, $(5, 1)$ for Wan2.1, and $(3, 5)$ for CogVideoX. These points need not maximize predictability, since delaying an intervention can reduce its remaining control authority. Because α and γ act on the same temporal coordinate, their responses are characterized separately for each model.

4.1 Offline Real–Generated Calibration Bank

For each model, we construct separate magnitude and pacing banks:

$$\mathcal{B}_M = \left\{(q_{M,n}, \log M_n^G, \log M_n^R)\right\}_{n=1}^{N_B},$$

$$\mathcal{B}_{c_1} = \left\{(q_{c_1,n}, c_{1,n}^G, c_{1,n}^R)\right\}_{n=1}^{N_B}.$$

Here, $q_{M,n}$ and $q_{c_1,n}$ are compact signatures extracted at the corresponding intervention steps, with each bank constructed under the corresponding generation trajectory.

The signature summarizes temporal differences from three intermediate transformer blocks using the log mean magnitude, an eight-bin temporal profile, and the first sine mode of cumulative temporal deviation, yielding a 30-dimensional representation. For a query q, we retrieve its K nearest

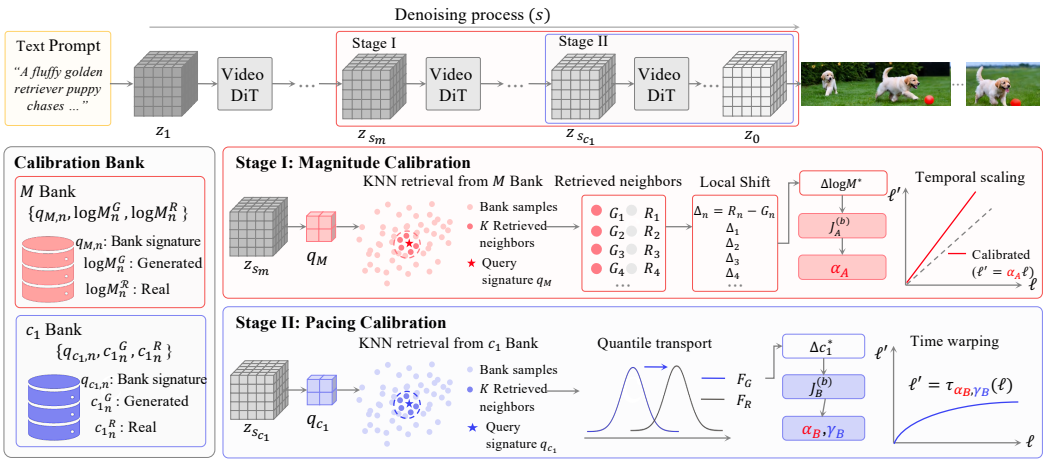


Figure 5: **Overview of Adaptive Temporal Coordinate Calibration.** Stage I calibrates motion magnitude and Stage II temporal pacing using intermediate representations and real–generated calibration banks. Their chronological order is model dependent, and both operate within the original denoising trajectory through temporal RoPE reparameterization.

generated examples with weights

$$w_n \propto \exp\left(-d(q, q_n)^2/\sigma^2\right).$$

A support factor further attenuates corrections for queries far from the calibration bank.

4.2 Stage I: Motion Magnitude Calibration

At $s_M^{(b)}$, the current signature retrieves a neighborhood $\mathcal{N}_M$, from which we compute

$$\Delta \log M^\star = g_M \sum_{n \in \mathcal{N}_M} w_n \left(\log M_n^R - \log M_n^G\right),$$

where $g_M \in [0, 1]$ is the support factor.

Writing $a = \log \alpha$, we obtain

$$a_A = \Pi_{\mathcal{A}_A^{(b)}}\left(\frac{\Delta \log M^\star}{J_A^{(b)}}\right), \qquad \alpha_A = \exp(a_A),$$

where $J_A^{(b)}$ is the measured response of $\log M$ to $\log \alpha$, and $\Pi_{\mathcal{A}_A^{(b)}}$ restricts the update to the validated range. If pacing calibration has already been applied, as in Wan2.1, the magnitude update is computed in the presence of the active γ control.

4.3 Stage II: Temporal Pacing Calibration

At $s_{c_1}^{(b)}$, we predict the current pacing statistic $\hat{c}_1$ and retrieve a local neighborhood $\mathcal{N}_{c_1}$. Because paired real and generated pacing values show weak correspondence, we construct a distributional target using quantile transport:

$$c_1^\star = \left(F_R^{\mathcal{N}_{c_1}}\right)^{-1}\left(F_G^{\mathcal{N}_{c_1}}(\hat{c}_1)\right),$$

with

$$\Delta c_1^\star = g_{c_1}(c_1^\star - \hat{c}_1).$$

Let α_{act} denote the temporal scale active at the pacing intervention. If magnitude calibration has already occurred,

$$\alpha_{\text{act}} = \min\left(\alpha_A, \bar{\alpha}_{c_1}^{(b)}\right),$$

Table 1: **In-domain temporal calibration on the held-out SSv2 and Panda-70M set.** The evaluation uses 100 real–prompt pairs, with 50 videos from each dataset and three random seeds per prompt. M-Gap and c_1-W_1 measure discrepancies in the temporal statistics used by our controller, while Flow-M Gap and Prog-W_1 provide independent measures of motion magnitude and temporal progression. VBench Dynamic Degree and Imaging Quality evaluate overall motion activity and visual quality.

Method	M-Gap ↓	Flow-M Gap ↓	c_1-W_1 ↓	Prog-W_1 ↓	Dynamic Degree ↑	Imaging Quality ↑
LTX-Video-13B	1.4110	1.3960	0.0226	0.0436	0.2933	0.6091
+ Ours	**1.1140**	**1.0400**	**0.0172**	**0.0314**	**0.3867**	**0.6108**
Wan2.1-T2V-1.3B	1.0690	1.4380	0.0200	0.0385	0.6100	**0.6919**
+ Ours	**0.8330**	**1.3060**	0.0200	**0.0342**	**0.6970**	0.6650
CogVideoX-5B	1.2590	**0.5845**	0.0484	0.0576	0.5600	**0.5778**
+ Ours	**1.1509**	0.5945	**0.0367**	**0.0474**	**0.5767**	0.5740

Table 2: **Results on external temporal benchmarks.** Ref4D evaluates reference-based motion dynamics, while ChronoMagic evaluates metamorphic progression. We additionally report Prog-W_1 to measure temporal progression, together with VBench Dynamic Degree and Imaging Quality.

	Ref4D			**ChronoMagic**			
Method	Motion ↑	Flow-M Gap ↓	Prog-W_1 ↓	MT ↑	Prog-W_1 ↓	Dynamic Degree ↑	Imaging Quality ↑
LTX-Video-13B	64.8300	1.9754	**0.0210**	2.3250	0.0253	0.1800	0.6079
+ Ours	**65.5200**	**1.6680**	0.0278	**2.4120**	**0.0176**	**0.2233**	**0.6125**
Wan2.1-T2V-1.3B	66.2500	1.7820	**0.0261**	3.1130	0.0643	0.4570	**0.6351**
+ Ours	**67.0500**	**1.6190**	0.0341	**3.4500**	**0.0567**	**0.5200**	0.6281
CogVideoX-5B	**69.8500**	**0.3569**	0.1165	2.8890	0.0324	0.4567	0.5540
+ Ours	68.8300	0.4332	**0.1041**	**2.9860**	**0.0207**	**0.4633**	**0.5575**

whereas $\alpha_{\text{act}} = 1$ when pacing is applied first.

Because the pacing response can be nonlinear and asymmetric, we use the measured response curve $R_{c_1}^{(b)}(\gamma; \alpha_{\text{act}})$:

$$\gamma_B = \Pi_{\Gamma^{(b)}} \left[\left(R_{c_1}^{(b)} \right)^{-1} (\Delta c_1^\star; \alpha_{\text{act}}) \right].$$

The resulting γ_B remains active during subsequent denoising. For Wan2.1, pacing is applied first and the later magnitude update adjusts α while retaining γ_B. Additional details on signature construction, intervention steps, support factors, response measurements, and operating ranges are provided in Appendix B.

5 Experiments

Experimental setup. We evaluate Adaptive Temporal Coordinate Calibration on LTX-Video-13B, Wan2.1-T2V-1.3B, and CogVideoX-5B. For the held-out evaluation, we use 100 real–prompt pairs, consisting of 50 videos from SSv2 and 50 from Panda-70M, with three random seeds per prompt. For external evaluation, we use ChronoMagic-Bench (Yuan et al., 2024) and Ref4D-VideoBench (Wei et al., 2026) with two random seeds per prompt, together with VBench Dynamic Degree and Imaging Quality (Huang et al., 2024). The same prompt–seed pairs are shared between the base and calibrated models to enable paired comparison.

All intervention steps, operating ranges, response mappings, and controller hyperparameters are determined using the calibration split only. No held-out evaluation samples are used for model selection or controller tuning. We report M-Gap as the mean paired absolute discrepancy in $\log M$ and c_1-W_1 as the Wasserstein-1 distance between the real and generated pacing distributions. Flow-M Gap and Prog-W_1 provide independent evaluations based on optical flow and semantic progression, respectively.

Table 3: **Component-wise ablation of magnitude calibration (MC) and pacing calibration (PC) on LTX-Video-13B.** All results use the same 100-video held-out subset with one random seed per prompt.

Exp.	MC	PC	Flow-M Gap ↓	Prog-W_1 ↓	M-Gap ↓	c_1-W_1 ↓
1			1.396	0.0484	1.329	0.0273
2		✓	1.463	**0.0282**	1.381	<u>0.0192</u>
3	✓		<u>1.087</u>	0.0414	**1.023**	0.0210
4	✓	✓	**1.048**	<u>0.0319</u>	<u>1.049</u>	**0.0127**

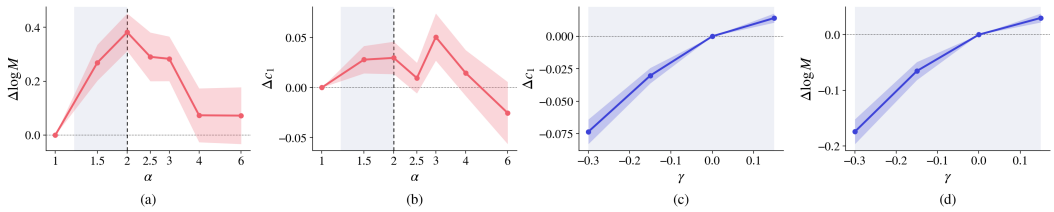


Figure 6: **Effects of temporal coordinate controls on LTX-Video-13B.** From left to right, we show $\alpha \to \Delta \log M$, $\alpha \to \Delta c_1$, $\gamma \to \Delta c_1$, and $\gamma \to \Delta \log M$. Shaded bands denote mean $\pm$ standard error. Gray regions indicate the 5th–95th percentile operating range.

5.1 Main Results

Table 1 evaluates temporal alignment on the pooled held-out SSv2 and Panda-70M set using both controller-space and independent temporal metrics. Calibration reduces M-Gap and improves Prog-W_1 and Dynamic Degree across all three models. For LTX-Video, the improvement extends to all reported temporal metrics, with Flow-M Gap decreasing from 1.396 to 1.040 and Prog-W_1 from 0.0436 to 0.0314. Wan2.1 similarly improves both magnitude metrics and Prog-W_1, while CogVideoX improves M-Gap and both pacing metrics despite a small increase in Flow-M Gap. Imaging Quality remains largely comparable, with modest model-dependent changes.

Table 2 evaluates transfer to external temporal benchmarks. On Ref4D, LTX-Video and Wan2.1 improve reference-based Motion and Flow-M Gap, whereas CogVideoX improves Prog-W_1 while showing a higher Flow-M Gap. On ChronoMagic, all three models improve MT, Prog-W_1, and Dynamic Degree after calibration. Imaging Quality changes only slightly across models. Overall, these results show that temporal calibration transfers across different generators and evaluation settings, although its effects on motion magnitude and temporal progression remain model and benchmark dependent.

5.2 Ablations and Analysis

Component Ablation. We ablate magnitude calibration (MC) and pacing calibration (PC). M-Gap and c_1-W_1 measure the statistics directly targeted by the controller, while Flow-M Gap and Prog-W_1 provide independent measures of motion magnitude and temporal progression.

Table 3 shows complementary effects of MC and PC. MC primarily reduces magnitude discrepancies, while PC substantially improves temporal progression (Prog-W_1: $0.0484 \to 0.0282$). Their combination achieves the lowest Flow-M Gap and c_1-W_1 while remaining competitive on the other metrics. This behavior reflects the coupling between the two controls, which modify the same temporal coordinate system.

Effects of Temporal Coordinate Controls. Figure 6 shows that α and γ have distinct but coupled effects. Increasing α primarily raises motion magnitude, while its effect on c_1 remains modest within the operating range. The magnitude response weakens outside this range, supporting a bounded model-specific scale rather than unrestricted temporal scaling. In contrast, γ changes temporal pacing nonlinearly and asymmetrically, while also introducing substantial magnitude cross-talk. These interactions motivate the measured Stage-II response and the restricted operating range used by our controller. Table 4 further supports the interaction between the two controls. Freezing α during

Table 4: **Effect of Stage-II magnitude adjustment on LTX-Video-13B.** All results use the same 100-video held-out subset with one random seed per prompt.

Method	Flow-M Gap ↓	M-Gap ↓	Prog-W_1 ↓	c_1-W_1 ↓
Baseline	1.396	1.329	0.0484	0.0273
Full (freeze α)	<u>1.107</u>	**1.038**	<u>0.0385</u>	<u>0.0200</u>
Full (Ours)	**1.048**	<u>1.049</u>	**0.0319**	**0.0127**

Table 5: **Inference efficiency.** Runtime comparisons use identical generation settings across models.

Model	**GPU**	Base (s)	Ours (s)	Overhead
LTX-Video-13B	A6000	117.96	117.97	0.008%
Wan2.1-T2V-1.3B	B200	39.11	39.35	0.614%
CogVideoX-5B	A6000	392.68	392.96	0.071%

Stage II slightly improves M-Gap, but worsens Flow-M Gap, Prog-W_1, and c_1-W_1, supporting the joint Stage-II adjustment.

Effect of Pacing Target Construction. We compare quantile transport with a simpler mean-shift target. Mean shifting gives the lowest pooled Prog-W_1 (0.0260 vs. 0.0319), whereas quantile transport consistently achieves lower c_1-W_1 across SSv2, Panda-70M, and the pooled set, and also gives the best Panda-70M Prog-W_1. We therefore use quantile transport as a distributional calibration rule, rather than as a method that uniformly minimizes every progression metric.

When Does Adaptive Targeting Help? A global magnitude shift already captures much of the systematic under-motion when most samples require correction in the same direction. Conditional targeting becomes more useful when the required correction varies across samples. On ChronoMagic-Bench, the local target tracks sample-specific oracle corrections with a Pearson correlation of $r = 0.394$, whereas a global target is constant across all samples. For the subset requiring reduced motion, local targeting lowers the target error from 2.125 to 1.402, a 34.0% reduction. This behavior is also reflected in the generated videos, where local targeting reduces Flow-M Gap from 0.808 to 0.576 relative to the global target. In contrast, on Ref4D-VideoBench, where the correction is predominantly one-sided, the global target remains competitive and achieves a lower Flow-M Gap. Thus, sample conditioning is not uniformly superior, but provides additional flexibility when a single dataset-level correction cannot represent heterogeneous temporal mismatches. Additional ablations are provided in Appendix D.

5.3 Inference Efficiency

Table 5 reports end-to-end generation time under matched inference settings. Our method introduces no additional denoiser evaluations and adds less than 0.7% overhead across all three models. The additional computation consists of intermediate-representation processing, calibration-bank retrieval, and lightweight coordinate updates within the original denoising trajectory. Because the models are evaluated on different GPUs, absolute runtimes are not intended for cross-model comparison; the reported overhead is measured relative to each corresponding baseline.

6 Conclusion

We study temporal miscalibration in text-to-video diffusion models by separating temporal dynamics into motion magnitude and temporal pacing. Our analysis shows that the discrepancy between generated and real-video dynamics varies across models and video domains, and reveals a read–write trade-off between predicting temporal statistics and retaining control authority during denoising. Motivated by these findings, we introduce Adaptive Temporal Coordinate Calibration, which predicts temporal statistics from intermediate representations and adjusts temporal coordinates without modifying the pretrained generator or adding denoising steps. Across multiple models and benchmarks, our method improves temporal alignment under independent motion and progression metrics while largely preserving visual quality. These results highlight that realistic temporal dynamics depend not only on how much motion occurs, but also on how that change unfolds over time.

AI USE STATEMENT

We used generative AI tools to assist with literature discovery, research ideation and experimental planning, drafting portions of the manuscript, and language editing and polishing. AI tools were also used to assist with code drafting and debugging for some experimental and visualization pipelines. All AI-assisted outputs were reviewed by the authors. References suggested by AI tools were manually verified against the original sources, and experimental procedures, code, numerical results, and scientific claims were independently checked by the authors before inclusion in the paper. Generative AI was not used to generate the evaluation datasets or to establish mathematical proofs. The authors take full responsibility for the final content of the paper, including all text, analyses, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies temporal calibration and evaluation of pretrained text-to-video generation models using existing video datasets and publicly released model checkpoints. The study does not involve human subjects, the collection of new personal data, or deployment of the proposed method on individuals. We use the datasets and pretrained models in accordance with their stated terms and intended research use. Improving the temporal realism of generated videos may also increase the plausibility of synthetic visual content and could therefore contribute to misuse in applications such as deceptive or misleading media. Our work is intended for research on the analysis, evaluation, and improvement of generative video models rather than for such applications. We encourage responsible use of the proposed techniques together with appropriate provenance, disclosure, and safeguards for generated content.

REPRODUCIBILITY STATEMENT

We provide the implementation and experimental details necessary to reproduce our results in the main paper and supplementary material. The method section specifies the temporal statistics, calibration procedure, intervention schedule, and temporal-coordinate reparameterization. The appendix reports the exact pretrained checkpoints, inference configurations, datasets, evaluation benchmarks, calibration-bank construction, control-response measurements, and ablation settings used in our experiments. We additionally report the number of evaluated samples, generation seeds, hyperparameters, and metric definitions where applicable. All comparisons between the base and calibrated models use matched prompts and generation seeds. Code for the calibration procedure, analysis scripts, and evaluation pipeline is provided in the anonymous supplementary material.

REFERENCES

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13320–13331. IEEE, 2024.

Tongsheng Ding, Zhen Luo, Yixuan Yang, Boyu Wang, Luyang Xie, Jinyu Yang, and Feng Zheng. Dreamtraj: Generating 6-dof object trajectories by reading unrendered video diffusion latents. *arXiv preprint arXiv:2608.00486*, 2026.

Parsa Esmati, Somjit Nath, Katja Hofmann, Derek Nowrouzezahrai, Samira Ebrahimi Kahou, and Majid Mirmehdi. The invisible hand of physics: When video diffusion models know more than they show. *arXiv preprint arXiv:2606.05328*, 2026.

Xiangbo Gao, Mingyang Wu, Siyuan Yang, Jiongze Yu, Pardis Taghavi, Fangzhou Lin, and Zhengzhong Tu. The pulse of motion: Measuring physical frame rate from visual dynamics. *arXiv preprint arXiv:2603.14375*, 2026.

Ahmet Berke Gökmen, Yiğit Ekin, Bahri Batuhan Bilecen, and Aysegul Dundar. Ropecraft: Training-free motion transfer with trajectory-guided rope optimization on diffusion transformers. *Advances in Neural Information Processing Systems*, 38:21459–21483, 2025.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 5843–5851. IEEE, 2017.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

Gal Metzer, Sagi Polaczek, Ali Mahdavi-Amiri, Raja Giryes, and Daniel Cohen-Or. Video analysis and generation via a semantic progress function. *arXiv preprint arXiv:2604.22554*, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Ariel Shaulov, Itay Hazan, Lior Wolf, and Hila Chefer. Flowmo: Variance-based flow guidance for coherent motion in video generation. *Advances in Neural Information Processing Systems*, 38: 82725–82754, 2025.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiajia Wei, Yujia He, Yuhan Hou, Hang Qi, Sihua Wang, Jincheng Shi, Kwok Fung Li, Zibin Zheng, and Weibin Wu. Ref4d-videobench: Four-dimensional reference-based evaluation of text-to-video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7719–7729, 2026.

Jian Xu, Yanning Wu, Delu Zeng, John Paisley, and Qibin Zhao. Diagnosing under-development of irreversible processes in video generation. *arXiv preprint arXiv:2608.00617*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Hidir Yesiltepe, Tuna Meral, Adil Kaan Akan, Kaan Oktay, and Pinar Yanardag. Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40256–40265, 2026.

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *Advances in Neural Information Processing Systems*, 37:21236–21270, 2024.

Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers. *arXiv preprint arXiv:2502.15894*, 2025.

A Experimental Details

A.1 Evaluation Benchmarks

ChronoMagic-Bench (Yuan et al., 2024). ChronoMagic-Bench evaluates temporal and metamorphic understanding in time-lapse video generation. It contains 1,649 prompts with real-video references and covers four broad categories, including biological processes, human creations, meteorological phenomena, and physical phenomena. We evaluate the same 150-prompt subset across models and variants and report MTScore, which evaluates metamorphic change over time following the benchmark protocol. The same prompts and random seeds are used for the corresponding base and calibrated generations.

VBench (Huang et al., 2024). VBench is a comprehensive benchmark for evaluating the quality and text–video consistency of T2V models across multiple dimensions. We report Dynamic Degree and Imaging Quality to examine whether temporal calibration changes the overall motion magnitude or degrades visual quality. Dynamic Degree measures the amount of motion in the generated video, while Imaging Quality evaluates frame-level visual fidelity and low-level artifacts.

Ref4D-VideoBench (Wei et al., 2026). Ref4D-VideoBench is a reference-based benchmark that evaluates generated videos against real-video references across semantic, motion, event-temporal, and world-knowledge dimensions. We primarily use its *Motion Dynamics* metric, which evaluates the consistency of motion with respect to the real reference. Since generating the full set of 600 videos for every model and variant is computationally expensive, we randomly sample 150 prompts using a fixed seed and use the same subset across all models and variants.

A.2 Temporal Evaluation Metrics

We report both controller-space metrics, which directly use the temporal statistics defined in Sec. 3, and independent metrics based on optical flow and semantic progression.

A.2.1 Controller Space Metrics

M-Gap. For caption-matched real and generated videos, we measure the paired log-magnitude discrepancy as

$$M\text{-Gap} = \frac{1}{N}\sum_{i=1}^{N}\left|\log M(\mathbf{x}_i^G) - \log M(\mathbf{x}_i^R)\right|.$$

Lower values indicate closer agreement with the motion magnitude of the corresponding real videos.

c_1-W_1. We measure the distributional discrepancy in temporal pacing using the Wasserstein-1 distance between the real and generated c_1 distributions:

$$c_1\text{-}W_1 = W_1\left(P_{\mathcal{R}}(c_1), P_{\mathcal{G}}(c_1)\right).$$

Lower values indicate closer agreement with the real-video pacing distribution.

A.2.2 Independent Temporal Evaluation Metrics

Our controller uses pixel-space motion magnitude M and pacing statistic c_1 to construct temporal corrections. To avoid evaluating the method solely with the same quantities used for calibration, we additionally use two independent temporal metrics.

Flow-M Gap. We estimate optical flow between consecutive frames using RAFT (Teed & Deng, 2020) and compute the mean flow magnitude over forward–backward consistent pixels. For a set of real and generated videos, Flow-M Gap is defined as

$$\text{Flow-}M\text{ Gap} = \left|\mathbb{E}_{\mathbf{x}\sim\mathcal{G}}[\log M_{\text{flow}}(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim\mathcal{R}}[\log M_{\text{flow}}(\mathbf{x})]\right|.$$

Lower values indicate closer agreement with real-video motion magnitude.

Prog-W_1. To evaluate temporal progression independently of the pixel-based pacing statistic c_1, we extract DINOv2 (Oquab et al., 2023) features from each frame and measure adjacent-frame semantic change. Let $\xi_q(\mathbf{x}) \in [0, 1]$ denote the normalized temporal location at which a fraction q of the cumulative semantic change has occurred. We evaluate $q \in \{0.25, 0.50, 0.75\}$ and compute

$$\text{Prog-}W_1 = \frac{1}{3} \sum_{q \in \{0.25, 0.50, 0.75\}} W_1 \left(P_{\mathcal{R}}(\xi_q), P_{\mathcal{G}}(\xi_q) \right),$$

where $P_{\mathcal{R}}(\xi_q)$ and $P_{\mathcal{G}}(\xi_q)$ denote the distributions of the progression location across real and generated videos, respectively. Lower values indicate closer agreement with real-video temporal progression.

A.3 Baseline Models

LTX-Video-13B (HaCohen et al., 2024). We use the publicly released LTX-Video-13B checkpoint (`ltxv-13b-0.9.8-dev`) as one of our base T2V models. LTX-Video represents temporal positions through rotary positional embeddings, providing a direct interface for temporal-coordinate reparameterization. Unless otherwise specified, we follow the official inference configuration and use 30 denoising steps. Our method is applied at inference time without updating the pretrained model parameters.

Wan2.1-T2V-1.3B (Wan et al., 2025). We use the publicly released Wan2.1-T2V-1.3B checkpoint and follow its official inference configuration with 50 denoising steps. Our temporal intervention is applied at inference time without modifying the pretrained model parameters or other model components.

CogVideoX-5B (Yang et al., 2025). We use the publicly released CogVideoX-5B checkpoint and follow its official inference configuration with 50 denoising steps. CogVideoX employs 3D rotary positional embeddings for spatiotemporal position encoding. We apply the same temporal-coordinate reparameterization to its temporal rotary coordinates while leaving all pretrained model parameters unchanged.

B Additional Method Details

This section provides implementation details of temporal statistics estimation, calibration-bank retrieval, empirical response measurement, and the complete calibration procedure. The calibration split is disjoint from all held-out evaluation samples. Intervention steps, control-response measurements, operating ranges, retrieval hyperparameters, and all other controller settings are selected using this calibration split only. No held-out evaluation samples are used for model selection or controller tuning.

B.1 Temporal Statistics Estimation and Stage Selection

Let $s \in \{0, \ldots, S-1\}$ denote the denoising-step index. At each candidate step s, we extract intermediate representations from three transformer blocks. For a representation

$$H_s \in \mathbb{R}^{T_h \times P_h \times D_h},$$

T_h, P_h, and D_h denote the temporal, spatial, and feature dimensions, respectively. For temporal position $\ell = 0, \ldots, T_h - 2$, we compute

$$m_\ell^H = \sqrt{\text{mean}_{p,d} \left[H_s(\ell+1, p, d) - H_s(\ell, p, d) \right]^2}.$$

For each block, we summarize this temporal sequence using its log mean magnitude, an eight-bin temporal profile, and the first sine-mode coefficient of its cumulative temporal deviation. Concatenating the summaries from the three blocks yields a 30-dimensional hidden-state signature q.

For stage selection, we fit lightweight ridge predictors from q to the final $\log M$ and c_1 and evaluate their predictive correlation across denoising steps. Predictability alone is insufficient for selecting

Table 6: **Model-specific intervention steps.** s_M and s_{c_1} denote the denoising steps used for motion-magnitude and temporal-pacing calibration, respectively.

Model	S	s_M	s_{c_1}
LTX-Video-13B	30	6	17
Wan2.1-T2V-1.3B	50	5	1
CogVideoX-5B	50	3	5

an intervention point, because a later hidden representation can be easier to read after much of the remaining control authority has already been lost. We therefore consider prediction accuracy together with the empirical response of the corresponding temporal-coordinate control. The resulting model-specific intervention steps are summarized below.

The selected steps need not follow the same chronological order across models. For LTX-Video and CogVideoX, magnitude calibration is applied before pacing calibration, whereas Wan2.1 uses the pacing intervention first. We therefore treat s_M and s_{c_1} as property-specific intervention points rather than imposing a fixed ordering between the two components. The magnitude predictor is used only for stage selection; online magnitude calibration directly retrieves the bank using the current hidden-state signature. The pacing predictor estimates $\hat{c}_1$ from the trajectory actually present at s_{c_1} under the model-specific intervention order.

B.2 Calibration Banks and Local Retrieval

The calibration data consist of caption-matched real and generated videos from subsamples of SSv2 (Goyal et al., 2017) and Panda-70M (Chen et al., 2024) and are disjoint from the held-out evaluation set. For each model, we construct separate magnitude and pacing banks:

$$\mathcal{B}_M = \left\{ \left(q_{M,n}, \log M_n^G, \log M_n^R \right) \right\}_{n=1}^{N_B},$$

and

$$\mathcal{B}_{c_1} = \left\{ \left(q_{c_1,n}, c_{1,n}^G, c_{1,n}^R \right) \right\}_{n=1}^{N_B},$$

where n indexes calibration-bank samples.

Each bank is constructed to match the control history encountered at its corresponding intervention step. For LTX-Video and CogVideoX, magnitude calibration precedes pacing calibration, so the pacing bank and pacing predictor are constructed from trajectories with the magnitude correction already active. For Wan2.1, where $s_{c_1} < s_M$, pacing calibration is encountered first and its bank and predictor are constructed from the corresponding pre-magnitude trajectory. This ensures that online queries are compared with calibration signatures obtained under the same preceding interventions.

At inference time, the current query signature q_M or q_{c_1} is standardized using the corresponding bank statistics and retrieves its K nearest generated samples. For a query q, let

$$d_n = d(q, q_n)$$

denote the standardized distance to bank sample n. The retrieved samples receive normalized Gaussian weights

$$w_n = \frac{\exp(-d_n^2/\sigma^2)}{\sum_{r \in \mathcal{N}} \exp(-d_r^2/\sigma^2)}.$$

We additionally attenuate corrections for queries that are poorly supported by the calibration bank. Let

$$\bar{d}(q) = \frac{1}{K} \sum_{n \in \mathcal{N}} d_n$$

denote the mean distance to the retrieved neighbors. For each stage, we compare $\bar{d}(q)$ with the distribution of within-bank K-NN distances. The support factor is

$$g(\bar{d}) = \begin{cases} 1, & \bar{d} \le d_{50}, \\ \dfrac{d_{95} - \bar{d}}{d_{95} - d_{50}}, & d_{50} < \bar{d} < d_{95}, \\ 0, & \bar{d} \ge d_{95}, \end{cases}$$

Table 7: **Empirical temporal-coordinate sensitivities.** J_A is measured from the magnitude intervention point, whereas J_B is measured from the pacing intervention point.

Model	J_A	$J_{B,11}$	$J_{B,12}$	$J_{B,21}$	$J_{B,22}$
LTX-Video	0.533	0.129	0.369	0.018	0.250
Wan2.1-T2V-1.3B	1.006	1.104	-0.610	0.017	1.251
CogVideoX-5B	0.891	0.900	-0.566	0.020	0.912

where d_{50} and d_{95} denote the median and 95th percentile of the corresponding bank-distance distribution. We denote the stage-specific support factors by g_M and g_{c_1}.

B.3 Empirical Control Responses

The effects of α and γ depend on when they are introduced during denoising. We therefore estimate response models at the two model-specific intervention points of each generator.

At Stage I, we measure the magnitude sensitivity

$$J_A^{(b)} = \frac{\partial \log M}{\partial \log \alpha}$$

using temporal-scale perturbations introduced at $s_M^{(b)}$.

At the pacing intervention, we measure the coupled response

$$J_B^{(b)} = \begin{bmatrix} \partial \log M/\partial \log \alpha & \partial \log M/\partial \gamma \\ \partial c_1/\partial \log \alpha & \partial c_1/\partial \gamma \end{bmatrix}$$

using α and γ perturbations introduced at $s_{c_1}^{(b)}$.

Although $J_A^{(b)}$ and $J_{B,11}^{(b)}$ describe the response of $\log M$ to the same coordinate parameter, they are measured at different denoising steps and therefore correspond to different remaining response horizons.

These measurements show that the two controls are not orthogonal. In particular, γ can produce substantial magnitude cross-talk, whose strength and direction depend on the generator and intervention step. We therefore use the response coefficients only as coarse first-order sensitivities within empirically validated control ranges.

For LTX-Video, the deployed pacing controller additionally uses the measured pacing response over the validated operating range because the $\gamma \to c_1$ response is visibly nonlinear and asymmetric. For Wan2.1, the applied $|\gamma|$ is small and the pacing update uses the local $J_{B,22}$ sensitivity; magnitude calibration is applied later and compensates for the magnitude cross-talk already induced by γ. The sensitivity matrix is therefore interpreted as a local response summary, not as a universal linear model of the full control range.

B.4 Model-Specific Control Range

As in the main paper, let $\ell \in \{0, \dots, T_\tau - 1\}$ denote the native temporal RoPE index. The temporal coordinate is reparameterized as

$$\tau_{\alpha,\gamma}(\ell) = \alpha\ell + \gamma(T_\tau - 1)\sin\left(\frac{\pi\ell}{T_\tau - 1}\right).$$

Its derivative with respect to the temporal coordinate is

$$\frac{\partial \tau_{\alpha,\gamma}}{\partial \ell} = \alpha + \pi\gamma\cos\left(\frac{\pi\ell}{T_\tau - 1}\right).$$

Hence, monotonic temporal coordinates require

$$\alpha > \pi|\gamma|.$$

Table 8: **Model-specific temporal-scale ranges.** The deployed cap denotes an additional bound used by the final controller.

Model	α range	Deployed cap
LTX-Video-13B	[0.5, 2.0]	1.2 at s_{c_1}
Wan2.1-T2V-1.3B	[0.714, 1.4]	1.4
CogVideoX-5B	[0.5, 1.4]	1.2

Table 9: **Shared controller hyperparameters.** Bank sizes and retrieval neighborhood sizes are model specific.

Parameter	Value
Ridge coefficient λ	10^{-3}
Support lower percentile	50
Support upper percentile	95
Monotonicity margin	0.95

In implementation, we use the slightly stricter constraint

$$|\gamma| \leq 0.95\,\frac{\alpha}{\pi}.$$

The admissible range of α is determined separately for each model using offline response sweeps. LTX-Video uses $\alpha \in [0.5, 2.0]$ for magnitude calibration and restricts the active scale to at most 1.2 when pacing calibration is applied. Wan2.1 uses $\alpha \in [1/1.4, 1.4] \approx [0.714, 1.4]$. CogVideoX uses a sweep range of $[0.5, 1.4]$, with a deployed upper bound of 1.2. These ranges reflect model-specific response behavior rather than a shared universal operating interval.

The shared controller hyperparameters are summarized in Table 9.

B.5 Complete Calibration Procedure

The labels Stage I and Stage II refer to the magnitude and pacing objectives, respectively; their chronological order can differ across models. At each intervention point, the query signature retrieves a calibration neighborhood constructed under the same preceding control history.

Magnitude calibration. The query signature q_M retrieves a magnitude neighborhood $\mathcal{N}_M$. The target correction is

$$\Delta \log M^\star = g_M \sum_{n \in \mathcal{N}_M} w_n \left(\log M_n^R - \log M_n^G\right).$$

Writing $a = \log \alpha$, a provisional update is obtained from the measured magnitude sensitivity $J_A^{(b)}$ and projected to the model-specific operating range. When no pacing control is active, this reduces to

$$a_A = \Pi_{\mathcal{A}^{(b)}}\left(\frac{\Delta \log M^\star}{J_A^{(b)}}\right), \qquad \alpha_A = \exp(a_A).$$

Pacing calibration. At $s_{c_1}^{(b)}$, the pacing predictor estimates $\hat{c}_1$ and the query signature retrieves a pacing neighborhood $\mathcal{N}_{c_1}$. We construct the distributional target

$$c_1^\star = \left(F_R^{\mathcal{N}_{c_1}}\right)^{-1}\left(F_G^{\mathcal{N}_{c_1}}(\hat{c}_1)\right), \qquad \Delta c_1^\star = g_{c_1}(c_1^\star - \hat{c}_1).$$

The corresponding γ update is restricted by the validated response range and the temporal-coordinate monotonicity constraint.

Magnitude-first models. For LTX-Video and CogVideoX, $s_M < s_{c_1}$. Magnitude calibration is therefore applied first and pacing calibration is conditioned on the already modified trajectory.

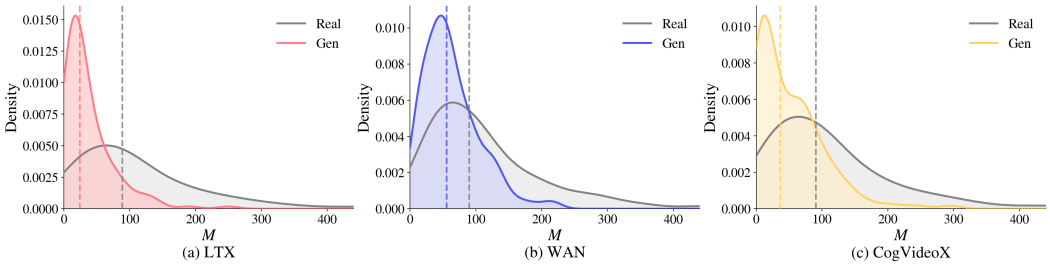


Figure 7: **Motion-magnitude distributions for real and generated videos.** All three generators exhibit systematic under-motion. The generated-to-real geometric-mean magnitude ratios are $M_G/M_R = 0.32$, 0.55, and 0.33 for LTX-Video, Wan2.1, and CogVideoX, respectively.

At the pacing intervention, the active magnitude scale is restricted to its deployed range before solving for γ. For LTX-Video, where the pacing response is nonlinear and asymmetric, the controller numerically inverts the measured response curve over the validated range. Within a local linear approximation, the same pacing-priority update can be written as

$$\gamma_B = \frac{\Delta c_1^\star - J_{B,21}^{(b)} \Delta a_{\text{app}}}{J_{B,22}^{(b)}},$$

where Δa_{app} is the magnitude-scale change actually active at the pacing intervention. This prevents γ from compensating for magnitude corrections that cannot be realized within the admissible α range.

Pacing-first Wan2.1. Wan2.1 uses $s_{c_1} = 1 < s_M = 5$. At the pacing intervention, no magnitude correction is yet active, so the controller computes

$$\gamma_B = \Pi_{\Gamma^{(b)}} \left(\frac{\Delta c_1^\star}{J_{B,22}^{(b)}} \right).$$

When magnitude calibration is subsequently applied, it compensates for the magnitude change already induced by γ_B:

$$a_A = \Pi_{\mathcal{A}^{(b)}} \left(\frac{\Delta \log M^\star - J_{B,12}^{(b)} \gamma_B}{J_A^{(b)}} \right), \qquad \alpha_A = \exp(a_A).$$

Thus, the same two temporal properties are calibrated across all three models without requiring a fixed chronological ordering of the two interventions.

All updates remain within the original denoising trajectory. The procedure requires no trajectory restart or additional denoiser evaluations, and the final video is decoded only after completing the original sampling schedule.

C Additional Analysis

C.1 Detailed Motion-Magnitude Statistics

We provide additional statistics for the motion-magnitude mismatch discussed in Sec. 3.1. Figure 7 compares the distributions of motion magnitude between caption-matched real and generated videos for all three models. Real and generated magnitudes exhibit only modest positive correspondence, with paired correlations of 0.24, 0.11, and 0.14 for LTX-Video, Wan2.1, and CogVideoX, respectively.

All three models exhibit lower motion magnitude than the corresponding real videos, although the severity of the mismatch differs across generators. LTX-Video and CogVideoX produce roughly one third of the real-video magnitude on average, whereas the deficit is less severe for Wan2.1. The modest paired correlations additionally indicate substantial sample-level variation. A global magnitude target is compared with state-conditioned local retrieval in Sec. D.1.

Table 10: **Motion-magnitude statistics.** r denotes the paired real–generated correlation, M_G/M_R the geometric-mean magnitude ratio, and d the standardized difference in log-magnitude space.

Model	N_{eval}	r	M_G/M_R	d
LTX-Video	300	0.24	0.32	-0.94
Wan2.1	300	0.11	0.55	-0.64
CogVideoX	299	0.14	0.33	-0.76

Table 11: **Robustness to the motion measure.** Values are reported in log-magnitude space.

Measure	Real	Base	Ours
Pixel difference	4.46	3.24	3.65
Residual flow	2.86	1.36	1.96

C.2 Robustness to Motion Measurement

Our primary motion magnitude M uses spatially averaged frame differences. To test whether the observed behavior is specific to this measure, we repeat the analysis using RAFT optical flow after compensating for dominant global affine motion. We denote the resulting residual-flow magnitude by M_{flow}.

Across real videos, the Pearson correlation between $\log M$ and $\log M_{\text{flow}}$ is 0.59 (Spearman $\rho_{\text{S}} = 0.49$), while for generated videos it is 0.57 ($\rho_{\text{S}} = 0.55$). The two measures therefore capture related but distinct temporal variation.

The real–generated discrepancy remains pronounced under residual flow. Moreover, W_1 decreases from 0.965 to 0.449 on SSv2, from 2.031 to 1.434 on Panda-70M, and from 1.496 to 0.895 overall. Thus, the observed calibration improvement is not specific to the frame-difference definition of M.

Residual flow does not completely remove camera-motion effects. Its correlation with estimated camera motion decreases from 0.54 to 0.42 for generated videos, but remains similar for real videos (0.51 versus 0.52). We therefore interpret it as an alternative temporal-change measure rather than a complete separation of object and camera motion.

C.3 Distributional Structure of Temporal Pacing

Unlike motion magnitude, temporal pacing exhibits almost no sample-wise correspondence between caption-matched real and generated videos. Across both domains and all three models, $|r| \leq 0.12$.

The mismatch appears primarily at the distribution level. On SSv2, generated pacing is more dispersed than real pacing for all three models, whereas the direction reverses on Panda-70M.

The near-zero paired correlations make a particular caption-matched real c_1 an unreliable sample-wise point target. In contrast to the shared under-motion observed for magnitude, the direction of the pacing spread mismatch depends on the video domain. We therefore formulate pacing calibration at the distribution level. Conditional quantile transport is our default because it can accommodate changes in both location and spread; Sec. D.1 compares it with a simpler mean-shift target.

C.4 Read–Write Trade-off Across Denoising

The main paper analyzes not only when the final temporal statistics become predictable from intermediate hidden representations, but also how long a coordinate intervention retains influence over the final video. For a denoising step s, we measure readability using the cross-validated ridge correlation between the hidden-state signature $q(H_s)$ and the final $\log M$ or c_1.

For magnitude control, we apply $\alpha = 2.0$ from step s onward and measure the resulting $\Delta \log M$ relative to an identity run with the same prompt and random seed. For pacing control, we use the

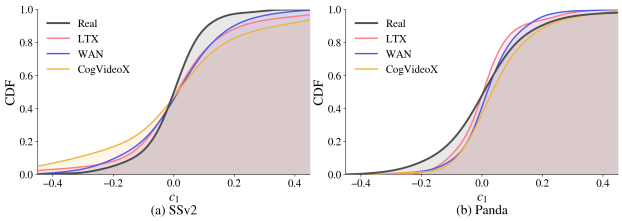


Figure 8: **Distributional mismatch of temporal pacing.** Empirical cumulative distributions of c_1 for real and generated videos. Generated pacing is broader on SSv2 but narrower on Panda-70M.

Table 12: **Temporal-pacing distribution statistics.**

Domain	Model	σ_R	σ_G	σ_G/σ_R	W_1	r
SSv2	LTX-Video	0.101	0.176	1.74×	0.0465	+0.07
	Wan2.1	0.102	0.148	1.46×	0.0405	−0.12
	CogVideoX	0.101	0.236	2.34×	0.0956	+0.01
Panda	LTX-Video	0.149	0.099	0.66×	0.0383	−0.04
	Wan2.1	0.114	0.095	0.84×	0.0375	−0.04
	CogVideoX	0.143	0.122	0.85×	0.0435	+0.01

symmetric perturbation $\gamma = \pm 0.15$ and compute

$$\frac{c_1(+\gamma) - c_1(-\gamma)}{2\gamma}.$$

Each write measurement uses ten SSv2 clips with matched seeds. Because the absolute response depends on the model and perturbation magnitude, we normalize each write curve by its earliest-step response and use the normalized value only to compare how control authority decays over denoising.

The resulting curves exhibit a clear read–write trade-off. At the beginning of generation, $\log M$ is already moderately predictable for all three models, whereas c_1 is not: the initial correlations are $0.357/-0.091$ for LTX-Video, $0.494/-0.086$ for Wan2.1, and $0.602/-0.051$ for CogVideoX, for $\log M/c_1$ respectively. As denoising proceeds, readability generally improves, but the remaining write response decreases. The decay rate is highly model dependent. At approximately one fifth of the trajectory, the relative α response is 1.112 for LTX-Video, 0.957 for Wan2.1, and only 0.071 for CogVideoX. Thus, CogVideoX loses most of its magnitude-control authority much earlier, whereas Wan2.1 retains a comparatively broad interval in which magnitude is both readable and controllable.

The two control channels also decay differently. In CogVideoX, the relative α response is already near zero by step 10, while the relative γ response remains 0.314 at step 10 and 0.178 at step 20. This indicates that pacing can remain adjustable after substantial magnitude control has been lost.

The α perturbation used in this diagnostic is intentionally shared across models to compare relative response decay and should not be interpreted as the deployed operating range of each controller. The earliest-step normalization can also make later relative responses appear smaller when very early interventions substantially alter sample identity. We therefore interpret the curves as diagnostic response profiles rather than as a rule that directly determines the deployed intervention steps. The final controller additionally considers model-specific control behavior, validated operating ranges, and the interaction between the two coordinates. Consequently, the deployed points can differ from those suggested by predictability alone, as in Wan2.1 and CogVideoX.

D Additional Ablations and Controls

Unless otherwise specified, all ablation experiments are conducted on LTX-Video-13B using the same 100 held-out videos with one random seed per prompt. The same seed is shared across compared variants to enable paired evaluation.

Table 13: **Sensitivity to magnitude-target construction on the held-out set.**

Magnitude target	Flow-M Gap ↓	M-Gap ↓
Baseline	1.396	1.329
Local KNN	1.087	1.023
Global	**1.080**	**1.016**

Table 14: **Sensitivity to pacing-target construction.** Mean shift more strongly reduces the independent Prog-W_1 metric, whereas quantile transport gives better alignment of the pacing distribution measured by c_1-W_1.

	SSv2		**Panda**		**All**	
Pacing target	Prog-W_1 ↓	c_1-W_1 ↓	Prog-W_1 ↓	c_1-W_1 ↓	Prog-W_1 ↓	c_1-W_1 ↓
Baseline	0.0551	0.0424	0.0482	0.0445	0.0484	0.0273
Mean shift	**0.0366**	0.0306	0.0267	0.0309	**0.0260**	0.0146
Quantile transport (Ours)	0.0470	**0.0297**	**0.0234**	**0.0281**	0.0319	**0.0127**

D.1 Alternative Target Constructions

Magnitude target construction. We compare the local KNN magnitude target used by our method with a single global correction shared across samples.

Local and global targets perform similarly on the held-out evaluation set, indicating that a substantial part of the magnitude mismatch is a systematic scale bias. Local retrieval is therefore not uniformly superior in aggregate magnitude error. Its role is to condition the requested correction on the current generation, which becomes more relevant when different samples require different correction directions. We examine this behavior on external benchmarks in Sec. D.4.

Pacing target construction. We compare the conditional quantile transport used by our method with a simpler local mean-shift correction. Because the two procedures optimize different aspects of the distribution, we report both the independent progression metric and the controller-space c_1-W_1 discrepancy.

The two target constructions exhibit a clear trade-off. Mean shifting achieves the lowest pooled Prog-W_1 and gives a particularly large improvement on SSv2. Quantile transport, in contrast, consistently produces the lower c_1-W_1 discrepancy on SSv2, Panda-70M, and the pooled evaluation set, and also obtains the best Panda-70M Prog-W_1.

This distinction follows from their objectives. A mean shift adjusts only the local first moment and can directly move the average progression toward the real distribution. Quantile transport instead preserves the sample's relative position within the generated distribution and maps it to the corresponding position in the real distribution, allowing differences in both location and spread to be corrected. Since our pacing analysis reveals a distributional mismatch with weak sample-wise correspondence, we use quantile transport as the default distributional calibration rule. We do not claim that it uniformly minimizes every external progression metric.

D.2 Saturation-Aware Pacing Allocation

For models in which magnitude calibration precedes pacing calibration, the pacing-stage response couples magnitude and pacing. A naive implementation can first solve an unconstrained two-dimensional problem and then independently clamp α. Wan2.1 is not included in this comparison because its pacing intervention precedes magnitude calibration. When the proposed magnitude update exceeds its admissible range, however, the corresponding magnitude residual is unattainable through α. Retaining this residual causes γ to compensate for a control demand that cannot actually be realized.

Table 15: **Stage-II control magnitude under alternative saturation handling.** Values report average $|\gamma|$ over 100 held-out clips.

Stage-II allocation	CogVideoX	LTX-Video
Post-hoc clamp	0.129	0.179
Unprojected 2-row LS	0.208	0.551
Pacing priority	0.040	0.122
Projected 2-row LS	0.018	0.038

We therefore condition the pacing correction on the projected magnitude update, as described in Sec. 4.3. We also evaluate an alternative projected two-row least-squares solution. After setting

$$a_B = \Pi_{\mathcal{A}^{(b)}}(a_A + \Delta a),$$

the reachable target removes the saturated magnitude residual. The resulting two-row solution is

$$\gamma_{B,\mathrm{2D}} = \frac{J_{B,22}^{(b)}\left[\Delta c_1^\star - J_{B,21}^{(b)}(a_B - a_A)\right]}{\left(J_{B,12}^{(b)}\right)^2 + \left(J_{B,22}^{(b)}\right)^2 + \lambda}.$$

Thus, the projected two-row solution is a shrunk version of the pacing-priority solution, with shrinkage determined by the magnitude cross-talk $J_{B,12}$.

Without projection, two-row least squares can substantially increase $|\gamma|$ because the pacing control attempts to compensate for the unreachable magnitude residual. Our pacing-priority rule removes this failure mode while retaining direct authority over the pacing objective. The projected two-row variant is more conservative and is included as an ablation.

D.3 Detailed Control-Response Statistics

The response curves in Fig. 6 show that the two coordinates are coupled but asymmetric. For LTX-Video, increasing α within the stable regime strongly changes motion magnitude while producing comparatively little change in c_1. The response weakens and becomes non-monotonic outside this regime.

The pacing coordinate γ directly changes c_1 but can also modify the pixel-space magnitude. For LTX-Video, $\gamma = -0.30$ produces $\Delta c_1 = -0.074$ and $\Delta \log M = -0.174$ at the deployed Stage-II scale, illustrating both the asymmetry of the pacing response and its magnitude cross-talk. Restricting the active Stage-II scale to $\alpha = 1.2$ reduces this interference relative to retaining the Stage-I scale. Because the response differs across models, admissible control ranges and response mappings are selected separately rather than shared across models.

D.4 When Does Adaptive Targeting Help?

We further examine whether the local KNN target captures sample-specific variation in the required magnitude correction. This diagnostic uses the same 50 ChronoMagic and 50 Ref4D clips employed in our local and global target analysis. These subsets are used only for the sample-level diagnostic and are distinct from the larger benchmark sets used for the main evaluation. For each clip, let

$$\delta_i = \log M_i^R - \log M_i^{\mathrm{base}}$$

denote the oracle correction indicated by its real-video reference, and let $\hat{\delta}_i$ denote the correction target produced by the controller. A global target assigns the same $\hat{\delta}$ to every clip, whereas local retrieval can vary the target according to the current generation.

Table 16 shows that the local target exhibits positive correspondence with the clip-specific oracle correction in all three settings. The association is strongest on ChronoMagic, where correction demands are more heterogeneous. On the held-out set, where the mismatch is predominantly one-sided, the correlation is weaker ($r = 0.187$, $p = 0.063$). These results suggest that sample conditioning becomes more informative when the required correction varies substantially across clips.

Table 16: **Sample-level variation captured by local magnitude targeting.** Pearson r measures the correspondence between the local target $\hat{\delta}$ and the oracle correction δ.

Domain	N	$\mathrm{mean}(\delta)$	$\mathrm{sd}(\delta)$	Increase required	Pearson r
Held-out	100	1.222	1.077	90%	0.187
Ref4D	50	1.334	1.023	93%	0.260
ChronoMagic	50	0.393	1.375	64%	**0.394**

Table 17: **Sample-level target error on ChronoMagic.** Errors are reported separately for clips requiring increased ($\delta > 0$) and reduced ($\delta \leq 0$) motion.

Target	$\mathrm{sd}(\hat{\delta})$	Pearson r	$\delta > 0$ Error	$\delta \leq 0$ Error
Local KNN	0.557	0.394	0.808	**1.402**
Global	0.000	–	**0.663**	2.125

To examine this behavior more directly, we separate ChronoMagic clips according to the sign of the oracle correction. The local target has nonzero sample-to-sample variation ($\mathrm{sd}(\hat{\delta}) = 0.557$), whereas the global target is constant by construction.

The global target is slightly more accurate for the majority subset requiring increased motion. However, because it applies the same positive shift to every clip, it performs poorly on samples requiring the opposite correction. For the 18 ChronoMagic clips with $\delta \leq 0$, local targeting reduces target error from 2.125 to 1.402, corresponding to a 34.0% reduction. Thus, the advantage of local targeting is not a uniformly lower average error, but its ability to represent correction variation that a single global shift cannot express.

We finally compare the two targets after video generation.

The generated-video results follow the same pattern. ChronoMagic contains substantial variation in the required correction, and local targeting gives a lower Flow-M Gap than the global target. Ref4D is much more strongly dominated by increased-motion corrections, for which the shared global shift is more effective. Local targeting also gives a lower pooled Prog-W_1 on the external clips (0.0158 versus 0.0208).

These results do not imply that local targeting is universally superior. Rather, a global correction is effective when the temporal mismatch is largely one-sided, whereas local conditioning provides additional flexibility when different clips require substantially different corrections.

E Hyperparameter Sensitivity

The principal control hyperparameters are determined from offline response measurements rather than tuned to optimize the held-out evaluation metrics. The temporal-scale range is selected independently for each model, and γ is additionally restricted by temporal-coordinate monotonicity. Support thresholds and the shared ridge coefficient are summarized in Table 9.

The response measurements also determine how pacing corrections are realized. LTX-Video exhibits a nonlinear and asymmetric $\gamma \to c_1$ response, so the final controller uses the measured response over its validated range. Wan2.1 operates with much smaller $|\gamma|$ and uses the local first-order pacing sensitivity, with the later magnitude intervention compensating for γ-induced magnitude cross-talk. These model-specific choices follow the measured control behavior rather than assuming that one response model is appropriate for all generators.

Table 18: **Local vs. global magnitude targeting on external benchmarks.**

Benchmark	Increase required (%)	Local Flow-M Gap ↓	Global Flow-M Gap ↓
ChronoMagic	64	**0.576**	0.808
Ref4D	93	1.634	**1.473**



Figure 9: **Additional qualitative results for LTX-Video-13B.** For each example, we compare the base generation with our calibrated output under the same prompt and random seed. Frames are sampled uniformly over time; full videos are provided in the supplementary material.

F Additional Qualitative Results

F.1 Comparison with Base Models

We provide additional qualitative comparisons between the base generators and their temporally calibrated outputs. For each example, the base and calibrated videos use the same text condition and random seed. We uniformly sample representative frames over the generated sequence to visualize differences in motion progression. The accompanying supplementary videos provide the full temporal sequences.

F.2 Prompts for Qualitative Examples

Table 19 lists the prompts used for the qualitative examples of LTX-Video-13B, Wan2.1-T2V-1.3B, and CogVideoX-5B, ordered from top to bottom within each figure. Table 20 similarly lists the prompts used for the motivation examples in Fig. 1.

F.3 Failure Cases and Limitations

Our temporal reparameterization changes how the pretrained generator allocates its existing temporal dynamics; it does not introduce new motion primitives. Consequently, when real-video motion lies substantially beyond the reachable response of the pretrained model, as observed for high-motion SSv2 examples, calibration may reduce but not eliminate the magnitude gap.

A second limitation arises from coupling between the two temporal controls. Although α acts primarily on the pixel-space magnitude statistic, γ can also change motion magnitude while adjusting temporal pacing. The controller therefore represents a compromise between the two temporal objectives and need not outperform the corresponding single-component controller on every metric individually.

The relation between the controller statistic and an independent motion measure can also be model dependent. For CogVideoX, increasing α can raise the pixel-difference magnitude while reducing RAFT flow magnitude, so improvement in the calibrated pixel-space statistic does not necessarily



Figure 10: **Additional qualitative results for Wan2.1-T2V-1.3B.**

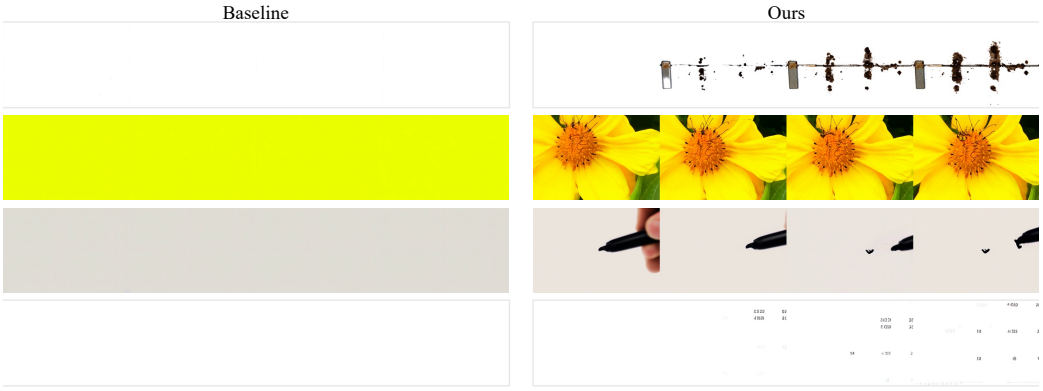


Figure 11: **Additional qualitative results for CogVideoX-5B.**

imply improvement in Flow-M Gap. This actuator mismatch is one reason we report independent temporal metrics in addition to the quantities used by the controller.

Finally, the calibration statistics depend on the temporal regimes represented in the calibration bank. Local retrieval and distributional pacing calibration provide a unified way to construct state-conditioned corrections, but simpler global corrections can be competitive in some regimes. Moreover, when a test sample lies outside the support of the calibration bank, the retrieved neighborhood may provide an unreliable reference correction. Our support factor attenuates such updates based on the distance to the calibration data, but it cannot provide reliable targets for temporal regimes that are absent from the bank. Expanding the calibration data to broader video domains is therefore an important direction for improving robustness beyond the temporal regimes considered in this work.

Table 19: **Text prompts for qualitative examples.** Prompts are listed in the same top-to-bottom order as the corresponding qualitative examples for each backbone.

Backbone	Example	Prompt
LTX-Video-13B	1	Time-lapse of a marbled cake baking process: starting with freshly poured batter in pans, signs of rising by initial heat exposure, uniform rising and solidifying midway, significant structured marbling and defined surface formation, and finally a fully baked, solid cake ready to be removed from the oven.
	2	Time-lapse of a white hibiscus blooming: starting as a tightly closed bud with an adjacent wilting flower, gradually opening its petals, and showing stamen and vibrant colors as it reaches full bloom, with the second flower continuing to wilt. The main flower exhibits beginning signs of petal wilting by the end of the video.
	3	Time-lapse of a lettuce seed germinating and growing into a mature plant. Initially, a seedling emerges from the soil, followed by leaves appearing and growing larger. The plant continues to develop, with some leaves turning reddish as it matures, culminating in a fully-grown plant with large green and red leaves.
	4	Time-lapse of a cityscape at dawn, showing the progression of the sunrise. The sequence captures the transition from pre-dawn light to a fully risen sun, with increasing brightness and visibility of the city as the sun ascends and illuminates the buildings. The scene evolves from darkness to a fully illuminated early morning cityscape.
Wan2.1-T2V-1.3B	1	Putting trivet underneath candle holder.
	2	Snow-covered evergreen trees stand prominently in the foreground, their branches heavy with fresh snow. Behind them, a dense white cloud layer blankets the lower slopes of a majestic, snow-capped mountain range under a bright blue sky dotted with scattered clouds. Snowflakes gently fall from above, adding to the serene winter landscape and enhancing the pristine whiteness of the scene. The overall lighting is soft yet clear, highlighting the crisp textures of the snow on both the trees and the distant mountains.
	3	Time-lapse of mushrooms growing using an organic grow kit: starting with no visible growth, slight darkening and initial signs by early stages, followed by noticeable darkening, small clusters forming, larger clusters developing, and significant growth, culminating in fully grown mushrooms ready for harvesting.
	4	Time-lapse of a massive explosion impact on a planetary surface: beginning with a bright burst, progressing to an expanding, glowing impact region with rising debris. A shockwave disperses debris, forming a prominent crater surrounded by molten, bright areas. The crater measures 840 km and gradually cools, with the impact zone darkening, marking the end of major explosive events.
CogVideoX-5B	1	Time-lapse of metallic objects undergoing significant corrosion: starting clean, early signs of dark spots appear, followed by spreading rust and increased oxidation, leading to severe degradation and discoloration by the end.
	2	A butterfly is pollinating squash flowers in a garden.
	3	Putting a black pen on a surface.
	4	Covering calendar page with calendar page.

Table 20: **Text prompts for the motivation examples.** Prompts are listed in the same top-to-bottom order as the examples shown in Fig. 1.

Backbone	Example	Prompt
LTX-Video-13B	1	Time-lapse of a passionflower blooming: Starting from a tightly closed bud, it gradually swells and begins to open. The outer petals peel back as the inner petals and central structures unfold. The flower fully opens, displaying intricate stamen and vibrant purple filaments, reaching peak bloom by the video's end.
Wan2.1-T2V-1.3B	2	Time-lapse of a railway station transitioning from early evening to night. A train arrives and departs as daylight fades. City lights turn on at dusk, and artificial lighting takes over, fully illuminating the station by night. The video concludes with consistent nighttime conditions and minimal changes in lighting.
CogVideoX-5B	3	Time-lapse of a lawn mower cutting grass in a scenic grassland, starting from the lower right, moving leftward while creating a visible path of mowed grass. The mower continues progressing through the field, reaching the middle, and finally completing its task at the far left edge with the entire area mowed.