Inference-time Motion Calibration for Video Generation

Zile Huang[1] and Ser-nam Lim[1*]

University of Central Florida, Orlando, United States
{zi057775, sernam}@ucf.edu

Abstract. Modern video generation models, despite high-quality frame synthesis, still struggle with motion fidelity and temporal consistency, causing artifacts and semantic drift; current solutions typically require extensive backbone retraining. We introduce a training-free framework for inference-time motion calibration that intervenes in the sampling trajectory of pretrained models to correct nascent inconsistencies efficiently. Our framework has two complementary components. First, **Dynamic Latent Control (DLC)** at selected steps estimates a clean latent, evaluates a VFM-based motion–semantic loss on proxy frames, and applies a corrective update to steer the trajectory. To control overhead, **Amortization Trajectory Rectification (ATR)** interleaves sparse calibration with fast amortized steps by caching a rectification field. Second, **Dynamic RoPE Control (DRC)** modulates temporal RoPE in attention blocks from the same motion signals, strengthening long-range coherence or sharpening short-range dynamics when flicker is detected. Extensive experiments on VBench and VBench 2.0 show that our method improves advanced motion fidelity metrics while incurring a small trade-off on superficial temporal smoothness for some backbones.

1 Introduction

Generative video models built on diffusion and flow-based generative modeling have recently achieved impressive progress, scaling from short, low-resolution clips to long, high-fidelity videos in both text-to-video (T2V) and image-to-video (I2V) settings [9, 13, 19, 24, 25, 37, 41, 44, 45, 52]. Despite architectural advances from 3D U-Nets to Diffusion Transformers (DiTs), these models remain largely optimized for per-frame visual quality rather than temporally coherent motion. The training objective focuses on frame-level reconstruction, and motion quality emerges only indirectly from temporal layers. Consequently, generated videos often exhibit high-frequency flicker, unnatural trajectories, and semantic drift over time. The cost of finetuning frontier video backbones is prohibitive for most practitioners, making inference-time calibration an attractive alternative.

However, extending these successes to dynamic video generation introduces two fundamental challenges. First, motion fidelity is hard to specify: language

* Corresponding author

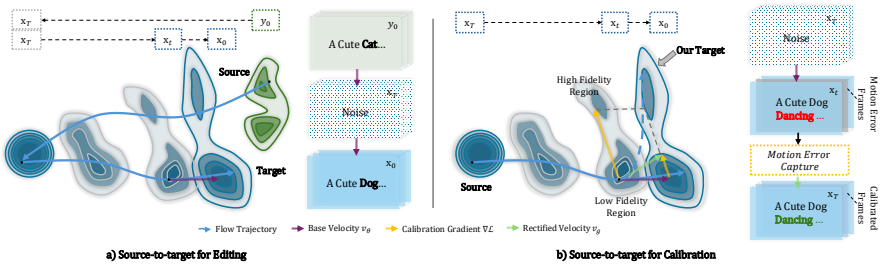

Fig. 1: Illustration of inference-time motion calibration. (a) Trajectory-shift view. In fixed-condition calibration (c_{src}=c_{tar}), the base trajectory drifts toward a low-fidelity region (red), while the calibrated trajectory (green) stays in a high-fidelity region. We do *not* invert real videos nor study prompt-based editing. **(b) Our calibration.** During sampling, if the base velocity v_θ drifts toward a low-fidelity region, DLC computes a VFM-based calibration gradient $\nabla\mathcal{L}$ on low-resolution proxy frames and yields a rectified velocity v^{DLC} that steers subsequent steps toward a high-fidelity region.

cannot capture subtle artifacts like high-frequency jitter, object flickering, or inconsistent trajectories, and backbones rarely have explicit motion objectives. Second, current inference-time scaling methods address quality improvement through *random search*: evolutionary or noise-trajectory methods maintain populations of candidates, scoring them with external verifiers and selecting high-reward samples [20, 27, 28, 40, 54, 58]. The verifier *ranks* samples but never induces an explicit motion-corrected distribution, so progress depends on random exploration rather than principled trajectory shaping. Moreover, search-based methods scale linearly with population size, making them expensive for large-scale deployment. Consequently, current models still produce unnatural trajectories, high-frequency flicker, and catastrophic semantic drift [16, 55, 60], as illustrated in Fig. 1. Search selects but cannot correct; using verifier feedback as *dense error signals* and feeding it back into the sampling dynamics is a natural alternative. However, even with trajectory-level correction, some failures persist due to fixed architectural inductive biases like temporal receptive field and positional encoding. This motivates an additional motion-aware *internal* controller that can adapt the temporal attention kernel online, complementing external trajectory steering. Our key insight is to convert language-free VFM feedback from a *ranking oracle* into a *differentiable energy* whose gradients can locally reshape the sampling dynamics. This leads to our central research question: *can we design a principled, inference-time motion calibration framework that directly steers the sampling trajectory toward a motion-corrected target distribution, while incurring only modest additional latency?* In this work, we answer this question affirmatively with a training-free framework that injects VFM-derived motion signals back into the sampling process. Our approach is complementary to backbone finetuning: it requires no weight updates and benefits from stronger VFMs as they become available, making it practical for deployment with frontier video generators.

Concretely, we instantiate this idea with an external latent controller (DLC) amortized by ATR, and an optional internal attention modulator (DRC), all without updating model weights. We cast motion calibration as a lightweight, inversion-free trajectory optimization at inference time. Instead of retraining the backbone, we reuse off-the-shelf Vision Foundation Models (VFMs) as language-free motion/semantic critics and backpropagate their signals to intermediate latents. The framework benefits from stronger VFMs as they become available, without requiring any retraining. We introduce *Dynamic Latent Control (DLC)*, an external correction loop that, at selected timesteps, estimates a clean latent, decodes low-resolution proxy frames, and computes a composite loss over motion incoherence and semantic deviation using optical flow, DINO, and CLIP features. The resulting gradient defines a latent correction that steers the trajectory toward regions of higher motion fidelity and semantic accuracy. Formally, the composite loss is $\mathcal{L} = w_1\mathcal{L}_{\text{flow}} + w_2\mathcal{L}_{\text{struct}} + w_3\mathcal{L}_{\text{sem}}$, where each term is standardized to zero mean and unit variance along the batch to ensure balanced gradients.

We focus on temporal RoPE because it is a low-dimensional, inference-time controllable knob that reshapes how attention logits vary with temporal lag, without modifying any weights. Despite trajectory-level calibration, some intrinsic architectural biases like fixed temporal receptive field and suboptimal positional encoding persist. Our key empirical observation is that these failures often arise from a mismatch between the fixed temporal coupling and the actual motion regime: fast motion is oversmoothed across distant frames, producing flicker, whereas slow motion lacks sufficient long-range coupling. We therefore introduce *Dynamic RoPE Control (DRC)*, a lightweight motion-aware internal controller that uses the same VFM-derived motion signals as DLC to dynamically modulate temporal Rotary Position Embeddings via a temporal dilation and a phase bias: jittery motion triggers a more local receptive field to suppress flicker; smooth motion allows stronger long-range coherence along the dominant motion direction.

While VFM-based gradients provide rich motion signals, applying DLC at every timestep introduces substantial overhead. We therefore introduce *Amortization Trajectory Rectification (ATR)*, which exploits the smoothness of the VFM-induced rectification field along the ODE trajectory. Low-resolution proxy decoding and frame aggregation encourage this; we also apply norm clipping for stability. ATR interleaves sparse *slow* steps (full DLC) with fast base steps that amortize the cached rectification field, achieving most of the calibration benefit at a fraction of the cost. We derive a Lipschitz-based error bound for this approximation and empirically characterize the quality–speed trade-off as the slow-step interval varies.

Unlike best-of-N search that improves quality by sampling many trajectories and ranking them, we perform *single-trajectory calibration* by feeding VFM errors back as *gradient-based* updates to the sampling dynamics. ATR amortizes expensive steps, enabling motion-aware control with only a small number of VFM backward evaluations [11, 35].

In summary, our contributions are three-fold:

1. We formulate *inference-time motion calibration* for pretrained video generators as a training-free, trajectory-level energy shaping problem, where VFM-derived signals provide gradients that steer sampling toward motion-consistent samples. This converts language-free VFM feedback from a ranking oracle into a differentiable energy whose gradients locally reshape the sampling dynamics.
2. We propose *Dynamic Latent Control* (DLC) together with *Amortization Trajectory Rectification* (ATR), which leverage VFM-based motion and semantic gradients to steer the sampling trajectory efficiently. We analyze the amortization error both theoretically (Lipschitz-based bound) and empirically, and characterize how calibration frequency trades off motion quality versus inference speed.
3. We introduce *Dynamic RoPE Control* (DRC), an internal temporal attention modulation mechanism driven by the same motion signals as DLC. DRC dynamically reshapes the temporal receptive field based on motion regime, complementing external trajectory steering. We demonstrate that combining DLC, ATR, and DRC consistently improves motion metrics and frame quality on both T2V and I2V benchmarks.

2 Related Works

2.1 Inference-time Scaling in Diffusion Models

Inference-time scaling uses extra compute during sampling to improve outputs without retraining [27, 28]. Classifier-free guidance trades diversity for fidelity [8, 14]; artifacts from high guidance are mitigated by adaptive schedules [20, 40, 54], while negative prompts suppress undesired content [2, 47]. Other controllers fix low-level mechanics or pursue multi-objective alignment and variational guidance [17,23,33]. A complementary view is search under *verifiers*: recent works scale this paradigm to image and video generation by performing evolutionary search over denoising trajectories or optimizing the noise trajectory via contextual-bandit-style policies [11,35]. While these methods can optimize arbitrary rewards, their cost scales linearly with the number of candidates and they treat the generator as a black box. In contrast, our approach views test-time scaling as *continuous trajectory calibration*: we construct a differentiable rectification field in latent space and directly modify the model's velocity field, amortizing this correction across many solver steps for fine-grained motion control at a fraction of the search cost.

2.2 Motion Modeling in Vision Generative Models

Motion modeling techniques are broadly categorized as implicit or explicit. In implicit approaches, spatiotemporal convolutions and attention learn motion as an emergent property [1,12,34,48]. Frame-Aware Video Diffusion [26] further assigns each frame an independent noise schedule for zero-shot I2V and interpolation.

Explicit methods use intermediate representations for greater control: MoCoGAN and TwoStreamVAN disentangle content and motion in separate subspaces [42,43], while optical flow serves as either a generation target [31] or a guidance signal [22, 30]. High-level control extends to skeletal poses and user-drawn trajectories [49,59]. All of these approaches bake motion priors into the architecture or training objective, and explicit guidance methods treat motion signals as *conditioning inputs* rather than online *feedback*. Our work is complementary: we use VFMs to measure motion and semantic errors during sampling and transform these signals into a calibration field that steers the latent trajectory and temporal attention at test time—requiring no retraining and no weight updates to the underlying generator.

3 Methodology

Parameterization bridge. We describe our method in a flow-ODE form, $\dot{z}_t = v_\theta(t, z_t \mid c)$. In diffusion samplers, each denoising step can be interpreted as an ODE-like update, and we use standard clean-latent estimators $\hat{z}_0(t)$ (e.g., DDIM) to instantiate the same controller.

3.1 Dynamic Latent Control

Overview. We cast generation as *trajectory optimization* for flow models whose dynamics follow $\frac{d}{dt} z_t = v_\theta(t, z_t \mid c)$. Given a suboptimal sampling path, we steer it toward a target distribution with higher temporal and semantic coherence. Unlike approaches that determine the target solely via conditioning, we *measure* motion/consistency errors with vision foundation models (VFMs) and *feed* their gradients back to intermediate latents along the trajectory. This performs *energy reweighting* under a fixed condition: the VFM signals act as a differentiable quality energy whose gradients locally reshape the sampling dynamics. The key advantage is that no retraining is required—the framework benefits from stronger VFMs as they become available.

Calibration Target. Let $p_\theta(z_0 \mid c)$ be the base model distribution with fixed c, and $\mathcal{L}(z_0)$ an externally evaluated quality energy such as temporal consistency, motion fidelity, or semantic stability. We define a Boltzmann-reweighted target:

$$p^\star(z_0 \mid c) \propto p_\theta(z_0 \mid c)\, \exp\big(-\lambda_{\mathrm{E}}\, \mathcal{L}(z_0)\big), \tag{1}$$

whose score is $\nabla_{z_0} \log p^\star = \nabla_{z_0} \log p_\theta - \lambda_{\mathrm{E}} \nabla_{z_0} \mathcal{L}$. We do *not* claim exact sampling from $p^\star$; DLC applies a small, stability-constrained perturbation that empirically shifts samples toward lower energy. The quality energy $\mathcal{L}$ is a composite of motion, structural, and semantic terms measured on proxy decoded frames, as detailed in Appendix. Let z_t denote the state at time $t \in [0, 1]$, with t=0 the clean data endpoint and t=1 the source/noise endpoint. We form a *terminal* estimate of the clean latent by a local Euler extrapolation to t=0:

$$\hat{z}_0(t) \;\approx\; z_t \;-\; s_\tau\, \tau_t\, v_\theta(t, z_t \mid c), \tag{2}$$

where τ_t denotes the remaining-time scale ($\tau_t \approx t$ for a uniform grid), and s_τ is a user-adjustable extrapolation scale ($s_\tau \approx 1.0$–1.25 empirically). This extrapolation approximates the clean latent at each intermediate step, enabling VFM evaluation on proxy decoded frames without waiting for full denoising. On Gaussian paths, map sample-space gradients to velocity space with the standard coefficient b_t. Additional details for diffusion pipeline are provided in Appendix.

Assumption 1 (Intermediate Semantic Coherence) *Let z_t be the latent state at an intermediate timestep $t > 0$, and let $\hat{z}_0(t)$ be the corresponding clean-latent estimate.*

1. ***Semantic Coherence:*** *The decoded proxy frame $x^{proxy} = \mathcal{D}(\hat{z}_0(t))$ retains sufficient structure, such as discernible objects and motion cues, so that $\mathcal{L}(x^{proxy})$ provides a non-trivial, informative error signal.*
2. ***Proxy-Target Correlation:*** *The proxy-loss gradient $\nabla_{\hat{z}_0}\mathcal{L}\big(\mathcal{D}(\hat{z}_0)\big)$ is sufficiently correlated with the intractable gradient on the fully denoised sample, so small updates at time t improve the final sample.*

Under Assumption 1, a continuous-time controller adds a *quality shaping* term to the velocity field.

Remark 1 (Sign convention). Sampling proceeds from t=1 to t=0; the clean endpoint is $\hat{z}_0 = z_t - \tau_t v_t$. Adding $+\beta_t b_t J_{\mathcal{D}}^{\top} \nabla_x \mathcal{L}$ to the velocity induces a displacement $-\tau_t \beta_t b_t J_{\mathcal{D}}^{\top} \nabla_x \mathcal{L}$ on the endpoint, so the plus sign in velocity space corresponds to descent in endpoint space.

This yields:

$$v^{\mathrm{DLC}}(z_t, t \mid c) = v_\theta(t, z_t \mid c) + \beta_t\, b_t\, J_{\mathcal{D}}(\hat{z}_0(t))^{\top}\, \nabla_x \mathcal{L}\Big(x^{\mathrm{proxy}}\Big), \tag{3}$$

where x^{proxy}=$\mathcal{D}(\hat{z}_0(t))$, b_t is the standard Gaussian-path coefficient mapping score-space to the velocity domain, and β_t is a user-chosen schedule controlling guidance strength. Equation (3) is the first-order continuous-time form; Algorithm implements it with K inner steps, momentum β_{mom}, and effective step size η_{eff} that absorbs the ratio $\beta_t b_t / \tau_t$. Momentum is a practical stability heuristic and is not part of the theoretical guarantee. For stability, the per-step correction is capped via gradient-norm clipping with γ=2.0; when the VFM signal is noisy or the decoder is ill-conditioned, this ensures that the strong base-model prior smoothly overrides the erroneous VFM gradient. In all experiments, we evaluate the energy on this terminal proxy; this relies on the proxy–target gradient correlation in Assumption 1. At early denoising steps ($t \to 1.0$), the latent is still dominated by noise and the proxy frames carry insufficient structure for informative VFM signals. We therefore activate DLC strictly from t_{start}=0.5 onward, after which the proxy latent carries sufficient low-frequency structure. A denoising-stage ablation in Appendix Table confirms that intervening at t=0.5 significantly outperforms both early (t=1.0, FVD 412.3) and late (t=0.2, FVD 358.4) starts. Equivalently, a small proxy latent step $\Delta z_0 = -\eta_t \nabla_{\hat{z}_0}\mathcal{L}$ re-encodes as a corrected velocity $v' \approx v_\theta + (\eta_t/\tau_t)\nabla_{\hat{z}_0}\mathcal{L}$. Increasing velocity along $+\nabla_{\hat{z}_0}\mathcal{L}$ makes $z_0 = z_t - \tau_t v$ move along $-\nabla_{\hat{z}_0}\mathcal{L}$, thereby decreasing $\mathcal{L}$.

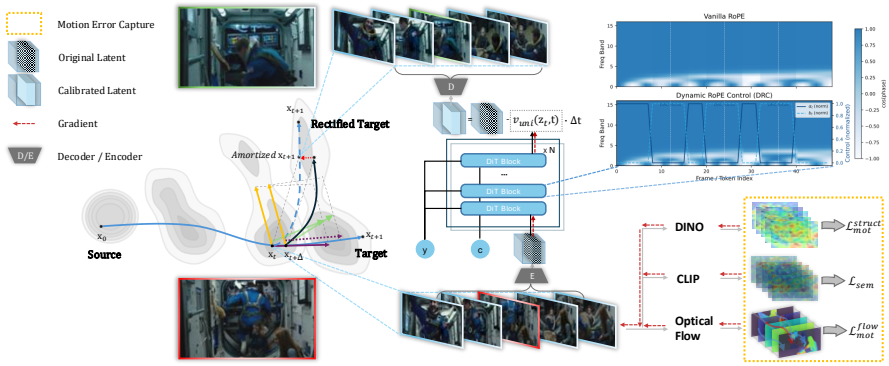

Fig. 2: Overview of inference-time motion calibration. During sampling, the base velocity field may drift toward low-fidelity regions, producing motion artifacts or semantic inconsistencies. At selected timesteps, *Dynamic Latent Control (DLC)* computes a clean latent estimate and evaluates motion and semantic deviations using vision foundation models. The resulting gradient defines a rectification field that adjusts the velocity direction, steering the trajectory to high-fidelity regions. To reduce computational cost, *Amortization Trajectory Rectification (ATR)* caches this rectification and reuses it across subsequent fast steps. Dynamic RoPE Control (DRC) reshapes the temporal attention kernel whenever motion errors are detected.

Conditional Shift (optional extension). In this paper, our main setting is fixed-condition calibration: the condition c remains constant throughout sampling. As an optional extension, our framework can be combined with conditional-mixing residuals when the target condition changes; we refer the reader to Appendix for the full derivation.

3.2 Dynamic RoPE Control

Our second component, **Dynamic RoPE Control (DRC)**, dynamically reshapes the temporal attention kernel via Rotary Position Embeddings (RoPE). Fast, jittery motion triggers a more localized temporal receptive field to suppress flicker, whereas slowly varying motion permits stronger long-range coupling.

Controllable Temporal Phase. Standard RoPE applies position-dependent rotations to query $\mathbf{q}_r$ and key $\mathbf{k}_s$ vectors at frame indices r and s. We introduce two online-controlled parameters: a temporal dilation $\rho_t > 0$ and a phase bias $\phi_t \in \mathbb{R}$. At each denoising step t, the relative phase gap for temporal lag $\Delta = r - s$ becomes:

$$\Delta\phi_i = \Big(\frac{r-s}{\rho_t} + 2\phi_t\Big)\omega_i. \tag{4}$$

Here, ρ_t controls the temporal receptive field: shrinking ρ_t sharpens attention on mitigating flicker, while enlarging it strengthens long-range coherence. The bias ϕ_t introduces an asymmetric shift in the relative-position kernel over positive and negative temporal lags, favoring either forward or backward temporal offsets.

Table 1: Quantitative comparison of T2V models on VBench 1.0 and VBench 2.0 motion metrics. For our method, "N" in row labels denotes the number of VFM-guided calibration steps (N_{VFM}), not search population size.

Method	Temporal Quality			Advanced Motion Fidelity			
	Motion Smoothness ↑	Temporal Flickering ↓	Dynamic Degree ↑	Motion Rationality ↑	Motion Order ↑	Human Anatomy ↑	Mechanics ↑
Sora [32]	–	–	–	34.48%	15.15%	86.45%	62.22%
Kling 1.6 [21]	–	–	–	38.51%	29.29%	86.99%	65.55%
HunyuanVideo [18]	98.70%	99.10%	57.00%	34.48%	26.94%	88.58%	76.09%
CogVideoX-1.5 [53]	96.92%	98.66%	70.97%	33.91%	26.60%	59.72%	80.80%
ModelScopeT2V [46]	95.79%	98.28%	66.39%	–	–	–	–
AnimateDiff-v2 [10]	94.63%	98.89%	68.89%	–	–	–	–
VideoCrafter-2.0 [6]	97.73%	98.41%	42.50%	–	–	–	–
Gen-2 [38]	99.58%	99.56%	18.89%	–	–	–	–
Latte-1 [29]	99.50%	99.74%	47.50%	–	–	–	–
Gen-3 [39]	97.73%	98.89%	59.86%	–	–	–	–
CogVideoX-1.5 [53] + EvoSearch	96.80%	98.72%	71.40%	34.00%	26.70%	59.80%	81.00%
CogVideoX-1.5 [53] + Noise Trajectory Search	97.05%	98.78%	71.80%	34.10%	26.80%	59.90%	81.10%
HunyuanVideo [18] + EvoSearch	98.55%	99.00%	57.50%	34.60%	27.10%	88.70%	76.40%
HunyuanVideo [18] + Noise Trajectory Search	98.50%	98.95%	57.70%	34.70%	27.20%	88.65%	76.50%
CogVideoX-1.5 [53] + *Ours (N=8)*	**97.40%**	**99.10%**	72.80%	**36.00%**	**29.00%**	61.20%	82.00%
CogVideoX-1.5 [53] + *Ours (N=16)*	97.20%	98.90%	**73.20%**	35.20%	28.50%	**61.40%**	**82.30%**
HunyuanVideo [18] + *Ours (N=1)*	96.00%	98.20%	66.50%	35.20%	27.60%	89.10%	77.10%
HunyuanVideo [18] + *Ours (N=4)*	96.40%	98.50%	67.80%	36.10%	28.40%	89.70%	78.00%
HunyuanVideo [18] + *Ours (N=8)*	**96.90%**	**98.80%**	**68.60%**	**37.20%**	29.50%	**90.30%**	78.80%
HunyuanVideo [18] + *Ours (N=16)*	96.70%	98.70%	68.20%	36.80%	**29.70%**	90.10%	**79.10%**

Adaptive Control from VFM Signals. The parameters (ρ_t, ϕ_t) are updated at runtime from motion signals extracted from the same proxy frame sequence used for DLC. We summarize physical motion with a robust velocity statistic u_i and a discrete temporal derivative j_i, both computed from optical flow magnitudes. In addition, we derive a temporal asymmetry score Δs_t from VFM inter-frame similarities. After normalization to $[0, 1]$, the parameters are updated: $\rho_t = \rho_0\, e^{-\kappa_u \hat{u}_i - \kappa_j \hat{j}_i}$ (clamped for stability) and $\phi_t = \beta\, \Delta s_t$ (constrained to $[\phi_{\min}, \phi_{\max}]$). Fast or jerky motion drives ρ_t toward its lower bound to suppress flicker, while slow motion allows long-range interactions.

3.3 Amortization Trajectory Rectification

Dynamic Latent Control (Sec. 3.1) introduces a test-time calibration policy that intermittently adjusts the sampling trajectory using VFM-derived gradients. While these corrections tend to improve motion fidelity and temporal consistency, computing the gradients at every correction step is computationally costly: each DLC step requires VFM forward/backward passes and proxy decoding. To reduce the overhead, we propose an amortized strategy, **Amortization Trajectory Rectification (ATR)** that interleaves sparse *slow* corrections with *fast* base steps within a single solver trajectory:

- **Calibration (*slow*) Step:** Executed sparsely every S_{ATR} steps at solver index k_s. This step performs the complete Dynamic Latent Control procedure to compute the calibrated *velocity prediction* v'_{k_s}.
- **Base (*fast*) Step:** Executed for the $S_{\text{ATR}}-1$ steps between slow steps. This step runs the standard, base-model *velocity field* without DLC calibrations.

Amortized Rectification. To prevent error accumulation in the *fast* steps, we use the information from the last *slow* step. At a *slow* step k_s, we compute two

velocity predictions: The standard model prediction, $v_\theta(z_{k_s}, t_{k_s})$, and the fully corrected prediction, v'_{k_s}, from our DLC algorithm. We define the **Rectification Field** $R(z,t)$ as the difference between the DLC-calibrated *velocity* that would be obtained if DLC were applied at (z,t), and the standard *velocity predictions* at any given state (z,t):

$$R(z,t) \triangleq v'_{\text{ideal}}(z,t) - v_\theta(z,t). \tag{5}$$

At the *slow* step k_s, we compute the exact value $R(z_{k_s}, t_{k_s}) = v'_{k_s} - v_\theta(z_{k_s}, t_{k_s})$ and cache it. For all subsequent *fast* steps k_f with $k_s < k_f < k_s + S_{\text{ATR}}$, we *approximate* the calibrated *velocity* v'_{approx} by adding this cached correction to the standard *velocity predictions*:

$$v'_{\text{approx}}(z_{k_f}, t_{k_f}) = v_\theta(z_{k_f}, t_{k_f}) + R(z_{k_s}, t_{k_s}). \tag{6}$$

This approach amortizes the cost of one expensive DLC computation over S_{ATR} fast, lightweight steps.

Theoretical Justification. This approximation is valid because the rectification field $R(z,t)$, which is derived from smooth VFMs, is assumed to be locally smooth. The error of our approximation is the difference between the cached correction and the true correction:

$$\mathcal{E}_{\text{approx}} = R(z_{k_s}, t_{k_s}) - R(z_{k_f}, t_{k_f}). \tag{7}$$

Because the solver step size is small, the state z_{k_f} is close to z_{k_s} and the time t_{k_f} is close to t_{k_s}. If R is smooth, this error $\mathcal{E}_{\text{approx}}$ will be small. This intuition is formalized in the following proposition, which guarantees that the approximation error is bounded by the state drift and time drift since the last slow step.

Proposition 1 (Error Bound for Amortized Rectification). *Suppose the rectification field $R(z,t)$ in Eq. 5 is L-Lipschitz continuous with respect to the state z and K-Lipschitz continuous with respect to time t. The approximation error $\mathcal{E}_{approx}$ at a fast step k_f, using the cached correction from the last slow step k_s, is bounded:*

$$\begin{aligned} \|\mathcal{E}_{approx}\|_2 &\le \|R(z_{k_s}, t_{k_s}) - R(z_{k_f}, t_{k_s})\|_2 \\ &\quad + \|R(z_{k_f}, t_{k_s}) - R(z_{k_f}, t_{k_f})\|_2 \\ &\le L \cdot \|z_{k_s} - z_{k_f}\|_2 + K \cdot |t_{k_s} - t_{k_f}|. \end{aligned} \tag{8}$$

Proposition 1 guarantees that the approximation error is bounded by the state drift and time drift since the last slow step. By performing the slow correction step frequently enough by keeping S_{ATR} small, we ensure that both terms remain small, thus keeping deviations bounded while delivering substantial speedup.

4 Experiment

4.1 Setup

Evaluation Settings. To comprehensively evaluate the quality of our generated videos, we employ a suite of evaluation metrics that combine visual fidelity, temporal consistency, and semantic alignment. For text-to-video (T2V) generation, we use the evaluation benchmarks from `VBench` [15] and `VBench-2.0` [60]. For image-to-video (I2V) evaluation, we adopt the image suite defined in `VBench-I2V` [15].

Backbones. We evaluate on four representative video generators spanning different architectures and parameterizations: CogVideoX-1.5 (DiT with ϵ-prediction), HunyuanVideo (DiT with flow matching), Wan 2.1 (DiT with flow matching), and FramePack (3D U-Net with ϵ-prediction). For each backbone, we use the public inference defaults and keep them fixed across all compared methods.

Compute Accounting. For fairness, we report wall-clock latency and align the expensive test-time budget across methods. Search baselines scale generator calls with population size N_{pop}, while our method uses a single trajectory with N_{VFM} sparse calibration steps and amortizes the remaining steps via ATR. All latency numbers are measured on a single H200-140GB GPU with batch size 1 under mixed-precision inference. More details are in Appendix.

4.2 Main Result

Comparison with Baselines on T2V Generation. Table 1 summarizes the results. Proprietary models such as Sora, Kling, and HunyuanVideo lead on advanced motion fidelity, with Kling at 38.51% `Motion Rationality` and HunyuanVideo at 88.58% `Human Anatomy`, while open-source models like VideoCrafter-2.0 excel in temporal stability (`Motion Smoothness` 97.73%), at the cost of lower dynamic expressiveness. Search-based baselines, EvoSearch and Noise Trajectory Search, yield only marginal improvements, $\leq$0.3% on all advanced motion metrics. This supports our design rationale: ranking-only feedback cannot induce explicit motion corrections, whereas our gradient-based calibration steers the trajectory toward lower energy regions. Search baselines use N_{pop}=8, matching our C_{gen} (generator forward evaluations); our method additionally spends C_{vfm} verifier evaluations, so gains are not from larger generator expenditure. Our framework consistently outperforms both base models and search baselines, validating that DLC and DRC jointly address motion artifacts that ranking cannot fix. On CogVideoX-1.5, we raise `Motion Rationality` from 33.91% to **36.00%**, `Motion Order` from 26.60% to **29.00%**, and `Dynamic Degree` to **73.20%** at N_{VFM}=16, while matching top open-source models on temporal stability. On HunyuanVideo, gains are even larger: `Human Anatomy` improves from 88.58% to **90.30%** and `Mechanics` from 76.09% to **79.10%**, all without any weight updates. These improvements reflect DLC's ability to correct nascent motion drift and DRC's adaptation of the temporal receptive field to the motion regime. Our framework

Table 2: Quantitative comparison of I2V models on VBench I2V metrics. For our method, "N" in row labels denotes the number of VFM-guided calibration steps (N_{VFM}), not search population size.

Method	Alignment Evaluation			Visual Quality Evaluation		
	i2v_subject ↑	i2v_background ↑	Camera Motion ↑	Dynamic Degree ↑	Motion Smoothness ↑	Aesthetic Quality ↑
Hunyuan I2V [18]	94.50%	94.00%	23.00%	17.74%	96.00%	62.04%
Wan 2.1 [45]	96.30%	95.60%	33.50%	46.02%	97.20%	63.12%
Skyreels-v2-I2V [4]	97.70%	97.00%	36.00%	51.16%	97.60%	58.94%
Skyreels-v2-DF [4]	97.30%	96.50%	35.00%	47.30%	97.40%	61.10%
FramePack [56]	93.50%	93.00%	21.50%	20.05%	96.50%	63.94%
FramePack F1 [56]	94.00%	93.50%	22.40%	24.42%	96.70%	63.10%
EasyAnimate [51]	96.00%	95.20%	33.00%	45.76%	97.00%	61.41%
DynamiCrafter-1024 [50]	96.71%	96.05%	35.44%	47.40%	97.38%	66.46%
SEINE-512x320 [7]	94.85%	94.02%	23.36%	34.31%	96.68%	58.42%
I2VGen-XL [57]	96.74%	95.44%	13.32%	24.96%	98.31%	65.33%
ConsistI2V [36]	94.69%	94.57%	33.60%	18.62%	97.38%	59.00%
VideoCrafter-I2V [5]	90.97%	90.51%	33.58%	22.60%	98.00%	60.78%
SVD-XT-1.1 [3]	97.51%	97.62%	–	43.17%	98.12%	60.23%
FramePack [56] + EvoSearch	93.60%	93.10%	21.60%	20.20%	96.60%	63.98%
FramePack [56] + Noise Trajectory Search	93.70%	93.10%	21.70%	20.15%	96.65%	63.96%
Wan 2.1 [45] + EvoSearch	96.40%	95.70%	33.60%	46.20%	97.35%	63.20%
Wan 2.1 [45] + Noise Trajectory Search	96.45%	95.75%	33.65%	46.25%	97.30%	63.18%
Hunyuan I2V [18] + EvoSearch	94.60%	94.10%	23.10%	17.85%	96.15%	62.10%
Hunyuan I2V [18] + Noise Trajectory Search	94.65%	94.10%	23.15%	17.80%	96.10%	62.08%
FramePack [56] + *Ours (N=8)*	94.40%	**93.80%**	**22.80%**	**21.10%**	**97.30%**	**64.32%**
FramePack [56] + *Ours (N=16)*	**94.80%**	93.20%	22.10%	20.60%	96.90%	64.18%
Wan 2.1 [45] + *Ours (N=1)*	96.90%	96.20%	34.00%	46.30%	97.60%	63.28%
Wan 2.1 [45] + *Ours (N=4)*	97.40%	96.70%	35.20%	46.80%	98.00%	63.45%
Wan 2.1 [45] + *Ours (N=8)*	**98.10%**	**97.40%**	**36.80%**	**47.70%**	**98.50%**	**63.72%**
Wan 2.1 [45] + *Ours (N=16)*	97.80%	97.10%	36.00%	47.20%	98.20%	63.58%
Hunyuan I2V [18] + *Ours (N=1)*	95.00%	94.40%	24.00%	17.95%	96.50%	62.18%
Hunyuan I2V [18] + *Ours (N=4)*	95.60%	95.00%	25.10%	**18.58%**	96.90%	62.26%
Hunyuan I2V [18] + *Ours (N=8)*	**96.50%**	**95.90%**	**26.50%**	18.55%	**97.50%**	**62.42%**
Hunyuan I2V [18] + *Ours (N=16)*	96.10%	95.50%	25.90%	18.38%	97.20%	62.34%

consistently outperforms both base models and search baselines, validating that DLC and DRC jointly address motion artifacts that ranking cannot fix. On CogVideoX-1.5, we raise `Motion Rationality` from 33.91% to **36.00%**, `Motion Order` from 26.60% to **29.00%**, and `Dynamic Degree` to **73.20%** at N_{VFM}=16, while matching top open-source models on temporal stability. On HunyuanVideo, gains are even larger: `Human Anatomy` improves from 88.58% to **90.30%** and `Mechanics` from 76.09% to **79.10%**, all without any weight updates. These improvements reflect DLC's ability to correct nascent motion drift and DRC's adaptation of the temporal receptive field to the motion regime.

Comparison with Baselines on I2V Generation. Table 2 shows that leading I2V models already achieve strong alignment, e.g., SVD-XT-1.1 at 97.51% / 97.62% `i2v_subject` / `i2v_background`, and substantial dynamics, Skyreels-v2-I2V `Dynamic Degree` 51.16%. EvoSearch and Noise Trajectory Search again provide negligible gains, <0.2% across all metrics, further demonstrating that motion-unaware search cannot improve camera control or temporal coherence. For I2V, these baselines use population size N_{pop}=8, matching or exceeding our VFM evaluation count. Our framework delivers consistent, architecture-agnostic improvements, confirming that the plug-in design generalizes across backbones. On FramePack, `Motion Smoothness` rises from 96.50% to **97.30%** and `Aesthetic Quality` from 63.94% to **64.32%**. On Wan 2.1, `i2v_subject` / `i2v_background` reach **98.10%** / **97.40%** with `Motion Smoothness` at **98.50%**.

Table 3: Gradient-baseline comparison on HunyuanVideo. Our ATR-amortized method achieves superior quality at a fraction of the latency of dense-gradient approaches.

Method	FVD↓	Motion↑	Latency(s)↓
Baseline	430.2	0.58	5.2
Universal Guidance [11]	385.4	0.63	48.6
Ours (N=8, ATR)	**328.5**	**0.70**	6.8

On Hunyuan I2V, `Camera Motion` improves most notably from 23.00% to **26.50%**. This gain directly reflects DLC's ability to correct camera trajectory drift during sampling, a dimension where search-based methods show weaker gains because they lack gradient feedback to steer the latent path. These results validate that our calibration framework is not limited to T2V generation but extends naturally to I2V tasks where temporal consistency with the input image is critical.

Comparison with Gradient-Based Baselines. To position our approach against prior gradient-based guidance methods, we compare with Universal Guidance [11] on HunyuanVideo (Table 3). Universal Guidance performs dense VFM backpropagation through 3D-UNet Jacobians at every denoising step, which incurs a 9.3× latency overhead (48.6 s vs. 5.2 s baseline). In contrast, our ATR-amortized method (N=8) achieves superior quality (FVD 328.5 vs. 385.4) with a mere 1.3× overhead (6.8 s). This validates our principled amortization: by caching the rectification field and reusing it across fast steps, calibration scales practically without the linear compute cost of dense-gradient methods. The key insight is that the rectification field varies smoothly along the trajectory, enabling efficient amortization while preserving calibration quality.

Scalability with Stronger Vision Foundation Models. Table 4 illustrates how Vision Foundation Models and Amortization Trajectory Rectification jointly affect performance and latency. We evaluate six VFM configbinations ranging from lightweight (CLIP ViT-B/32 with frame difference) to powerful (CLIP ViT-L/14 with RAFT optical flow). Stronger VFMs consistently improve `FVD`, `CLIP`, and `Motion` across configurations, validating that richer motion supervision yields more accurate rectification gradients. For example, replacing frame difference with RAFT for motion estimation reduces FVD from 405.7 to 354.8 (CLIP ViT-B/32) and from 392.3 to 329.7 (DINOv2 ViT-B/14). The computational cost of full DLC rises accordingly—from 28.5 s to 55.7 s— but ATR drastically reduces inference latency by amortizing corrections across fast steps, bringing it close to baseline levels (5.2 s) while preserving nearly identical performance. This demonstrates that our design scales with both VFM strength and the number of calibration steps N_{VFM}, offering an efficient quality–speed trade-off.

Table 4: Effect of VFM strength and amortization. We evaluate VFM backbones for the DLC module. Stronger models provide more accurate gradients and better performance but significantly increase inference latency, as shown in the full DLC rows. Applying Amortization Trajectory Rectification (ATR, Sec 3.3) drastically reduces latency to near-baseline levels while maintaining almost identical performance.

Control Modules			VFM Configuration		Performance Metrics			Cost
DRC	DLC	ATR	$\mathcal{L}_{\text{sem}}$	$\mathcal{L}_{\text{mot}}$	FVD ↓	CLIP ↑	Motion ↑	Latency (s) ↓
×	×	×	Baseline	Baseline	430.2	0.270	0.58	5.2
Full DLC Guidance (ATR Disabled)								
✓	✓	×	CLIP ViT-B/32	Frame Difference	405.7	0.279	0.61	28.5
✓	✓	×	CLIP ViT-L/14	Frame Difference	392.3	0.286	0.62	32.1
✓	✓	×	DINOv2 ViT-B/14	Frame Difference	386.5	0.283	0.64	30.7
✓	✓	×	CLIP ViT-B/32	RAFT	354.8	0.281	0.67	45.3
✓	✓	×	DINOv2 ViT-B/14	RAFT	329.7	0.287	0.70	48.8
✓	✓	×	CLIP ViT-L/14	DINOv2 Consistency	338.1	0.289	0.69	51.2
✓	✓	×	**CLIP ViT-L/14**	**RAFT**	**318.9**	**0.293**	**0.72**	55.7
Applying Amortization Trajectory Rectification (ATR)								
✓	✓	✓	CLIP ViT-B/32	Frame Difference	406.1	0.279	0.60	5.9
✓	✓	✓	DINOv2 ViT-B/14	RAFT	330.5	0.286	0.70	6.4
✓	✓	✓	**CLIP ViT-L/14**	**RAFT**	**319.2**	**0.292**	**0.72**	6.8

4.3 Ablation Study

Effect of VFM Strength. Table 4 shows that Dynamic Latent Control provides meaningful motion corrections even with lightweight error signals such as frame difference, demonstrating that the calibration mechanism is robust to verifier quality. Stronger VFMs yield substantially larger gains: replacing frame difference with RAFT for motion estimation and CLIP ViT-L/14 for semantic supervision consistently improves `FVD`, `Motion`, and `CLIP` across configurations. For example, with CLIP ViT-B/32, replacing frame difference with RAFT reduces FVD from 405.7 to 354.8 and improves Motion from 0.61 to 0.67. With DINOv2 ViT-B/14, the improvement is even larger: FVD drops from 386.5 to 329.7 and Motion from 0.64 to 0.70. This confirms that richer motion supervision translates into more accurate rectification gradients and better trajectory calibration, as predicted by our energy-based formulation. The best performance is achieved with CLIP ViT-L/14 and RAFT (FVD 318.9, Motion 0.72), but at the cost of 55.7 s latency. ATR reduces this to 6.8 s while preserving nearly identical performance, demonstrating the practical value of amortization.

Effect of Main Components. Table 5 decomposes our system on HunyuanVideo to isolate the contribution of each component. Starting from the baseline (FVD 430.2), adding DLC alone reduces FVD to 358.6 and improves Motion from 0.58 to 0.68, demonstrating that external trajectory calibration provides substantial gains. This improvement comes from the VFM-based gradient signal that steers the sampling trajectory toward regions of higher motion fidelity. Adding DRC alone yields a smaller improvement (FVD 401.3), as it only modulates internal attention without correcting the trajectory. However, combining DLC and DRC yields the best FVD (318.9), showing that external trajectory calibration and

Table 5: Component isolation study on HunyuanVideo. DLC and DRC provide complementary motion corrections; ATR amortizes the cost with minimal quality loss.

Configuration	FVD↓	Motion↑	Temp. Flicker↑
Baseline	430.2	0.58	98.70%
+DLC only	358.6	0.68	99.05%
+DRC only	401.3	0.62	99.20%
+DLC+DRC (Full, No ATR)	318.9	0.72	99.10%
Full + ATR (N=8)	328.5	0.70	98.80%

internal attention modulation are complementary: DLC corrects global trajectory drift while DRC adapts local temporal attention to the motion regime. Crucially, ATR slashes VFM forward/backward passes from 50 to 8 per trajectory (8.2× speedup) with minimal quality degradation (FVD 328.5 vs. 318.9), validating that the rectification field is smooth enough for efficient amortization. The small FVD increase ($< 3\%$) confirms that the cached correction remains accurate across fast steps.

The "DLC-only" row (without ATR) shows the upper bound of our calibration quality: with VFM evaluation at every step, we achieve FVD 318.9 and Motion 0.72. This represents the best possible trajectory correction given the current VFM signals. The gap between "DLC-only" and "Full + ATR" (328.5 vs. 318.9) quantifies the amortization cost, which is small compared to the gap between baseline and DLC-only (430.2 vs. 318.9). This validates our design choice: amortization provides a practical way to achieve near-optimal calibration quality at a fraction of the computational cost.

Hyperparameter Sensitivity. Figure 3 reports sensitivity to six key hyperparameters governing trajectory shaping, dynamic modulation, and amortization. We systematically vary each parameter while holding others fixed, using HunyuanVideo on the 100-prompt VBench split.

DLC parameters. The extrapolation scale s_τ exhibits a clear sweet spot around 1.0–1.25: values below 1.0 under-correct motion errors, while $s_\tau > 1.25$ over-compresses the trajectory and destabilizes the ODE solver. The Gaussian path coefficient b_t is more robust, with a mild optimum near b_t=1.0; moderate deviations preserve most performance, but large departures gradually degrade `CLIP` and motion alignment. The quality weight λ_{E} has the strongest effect: increasing it from zero to $\lambda_{\mathrm{E}}\approx$0.5–0.75 yields large drops in `FVD` (from 430 to ≈320) and consistent gains across metrics, while λ_{E}=1.0 begins to over-regularize and introduces a slight `FVD` trade-off. The DLC guidance strength β_t exhibits a similar sweet spot: β_t=0 reverts to the unguided baseline; moderate $\beta_t\approx$1.0–1.5 yields the best trade-off; excessive β_t slightly over-regularizes the trajectory.

DRC parameters. The decay strength κ governing how ρ_t and ϕ_t respond to motion signals exhibits an intermediate optimum around $\kappa\approx$1.0–1.5: very weak decay (κ=0) makes DRC unresponsive, while overly strong decay destabilizes

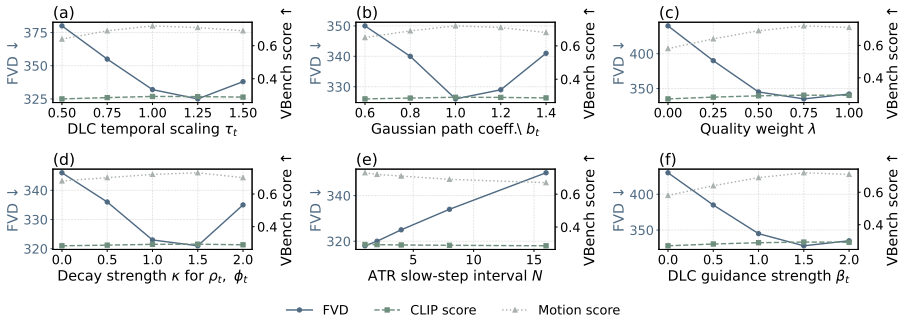

Fig. 3: Ablation on trajectory and modulation hyperparameters. (a) DLC temporal scaling τ_t; $\tau_t \approx 1.0$–1.25 yields the best trade-off. (b) Gaussian path coefficient b_t; performance is most stable around $b_t \approx 1.0$. (c) Quality weight λ_{E} on external VFM; mid-range values improve all metrics before slight over-regularization at λ_{E}=1.0. (d) Decay strength κ for DRC parameters ρ_t and ϕ_t; intermediate decay produces the lowest FVD and highest coherence. (e) ATR slow-step interval N; smaller N improves quality at higher latency, larger N trades quality for speed. (f) DLC guidance strength β_t; β_t=0 reverts to baseline, moderate $\beta_t \approx 1.0$–1.5 is optimal, excessive values over-regularize.

temporal attention. This confirms that DRC benefits from adaptive control rather than fixed settings.

ATR parameters. The slow-step interval N trades quality for latency: N=1 (full DLC every step) achieves the lowest `FVD` (318.9) and highest `CLIP`/`Motion`, while N=2 or N=4 yields only a modest `FVD` increase (+7 or +12) with substantial latency savings; N=8 or N=16 further reduces cost at the expense of gradual quality decline, supporting the quality–speed trade-off in Table 4.

We expose N as a continuous, user-controllable guidance-budget knob; the integer values shown are operating points, not the only legal choices. On HunyuanVideo, N=16 consistently improves all metrics; on CogVideoX-1.5 it acts as a low-pass filter that slightly smooths fine-grained compositional coherence while preserving Human Anatomy and Mechanics.

Latency Overhead. Full DLC without amortization requires 50 VFM passes (55.7 s, 10.7× overhead); ATR reduces this to 8 slow steps (6.8 s, 1.3× overhead) while preserving nearly identical performance (Table 4), measured on a single H200-140GB GPU with batch size 1.

5 Conclusion

We presented a training-free framework for inference-time motion calibration with two complementary components: DLC (external VFM-based correction loop) and DRC (internal temporal attention modulation), amortized by ATR for 1.3× baseline latency. Experiments on VBench show consistent motion fidelity improvements across T2V and I2V backbones.

References

1. Almog, G., Shamir, A., Fried, O.: REED-VAE: RE-Encode Decode training for iterative image editing with diffusion models. Computer Graphics Forum (Proceedings of Eurographics) (2025)
2. Ban, Y., Wang, R., Zhou, T., Cheng, M., Gong, B., Hsieh, C.J.: Understanding the impact of negative prompts: When and how do they take effect? arXiv preprint (2024), https://arxiv.org/abs/2406.02965
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Chen, G., Lin, D., Yang, J., Lin, C., Zhu, J., Fan, M., Zhang, H., Chen, S., Chen, Z., Ma, C., et al.: Skyreels-v2: Infinite-length film generative model. arXiv preprint arXiv:2504.13074 (2025)
5. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., Shan, Y.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023), referenced as VideoCrafter-1.0 in VBench++
6. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. arXiv preprint arXiv:2401.09047 (2024), referenced as VideoCrafter-2.0 in VBench++
7. Chen, X., Wang, Y., Zhang, L., Zhuang, S., Ma, X., Yu, J., Wang, Y., Lin, D., Qiao, Y., Liu, Z.: SEINE: Short-to-Long Video Diffusion Model for Generative Transition and Prediction. In: ICCV (2023)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems. vol. 34, pp. 8780–8794 (2021), https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html
9. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis (2024), https://arxiv.org/abs/2403.03206
10. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. arXiv preprint arXiv:2307.04725 (2023)
11. He, H., Liang, J., Wang, X., Wan, P., Zhang, D., Gai, K., Pan, L.: Scaling image and video generation via test-time evolutionary search (2025), https://arxiv.org/abs/2505.17618
12. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Video diffusion models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 28334–28347 (2022)
13. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
14. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv:2207.12598 (2022), https://arxiv.org/abs/2207.12598
15. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: VBench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)

16. Ji, P., Xiao, C., Tai, H., Huo, M.: T2vbench: Benchmarking temporal dynamics for text-to-video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) (2024), evGenFM Workshop
17. Kim, S., Kim, M., Park, D.: Test-time alignment of diffusion models without reward over-optimization. In: International Conference on Learning Representations (ICLR) (2025), https://openreview.net/forum?id=vi3DjUhFVm
18. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., et al.: Hunyuanvideo: A systematic framework for large video generative models (2024)
19. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: Hunyuanvideo: A systematic framework for large video generative models (2025), https://arxiv.org/abs/2412.03603
20. Koulischer, F., Handke, F., Deleu, J., Demeester, T., Ambrogioni, L.: Feedback guidance of diffusion models. arXiv preprint (2025), https://arxiv.org/abs/2506.06085
21. Kuaishou: Kling: Ai video generation model. https://klingai.kuaishou.com/ (2024), accessed: 2024-06-06
22. Liang, Y., Wang, Y., Liu, Y., Zhang, Z., Wang, T.: Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. arXiv preprint arXiv:2312.08537 (2024)
23. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024), https://openaccess.thecvf.com/content/WACV2024/papers/Lin_Common_Diffusion_Noise_Schedules_and_Sample_Steps_Are_Flawed_WACV_2024_paper.pdf
24. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
25. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
26. Liu, Y., Ren, Y., Artola, A., Liu, Y., Zeng, T., Chan, R.H., Morel, J.M.: Redefining temporal modeling in video diffusion: The vectorized timestep approach. arXiv preprint arXiv:2410.03160 (2024)
27. Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X., Xie, S.: Inference-time scaling for diffusion models beyond scaling denoising steps. arXiv preprint (2025), https://arxiv.org/abs/2501.09732
28. Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X., Xie, S.: Scaling inference time compute for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2025), https://openaccess.thecvf.com/content/CVPR2025/papers/Ma_Scaling_Inference_Time_Compute_for_Diffusion_Models_CVPR_2025_paper.pdf
29. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent Diffusion Transformer for Video Generation. arXiv preprint arXiv:2401.03048 (2024)
30. Nam, H., Kim, J., Lee, D., Ye, J.C.: Optical-flow guided prompt optimization for coherent video generation. arXiv preprint (2025)

31. Ni, S., Li, C., Liu, Z.: Conditional image-to-video generation with latent flow diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18451–18461 (2023)
32. OpenAI: Sora: Creating video from text. https://openai.com/index/sora/ (2024), accessed: 2024-02-15
33. Pandey, K., Sofian, F.M., Draxler, F., Karaletsos, T., Mandt, S.: Variational control for guidance in diffusion models. arXiv preprint (2025), https://arxiv.org/abs/2502.03686
34. Park, Y., Jung, H., Bae, S., Yun, S.Y.: Temporal alignment guidance: On-manifold sampling in diffusion models. arXiv preprint arXiv:2510.11057 (2025)
35. Ramesh, V., Mardani, M.: Test-time scaling of diffusion models via noise trajectory search (2025), https://arxiv.org/abs/2506.03164
36. Ren, W., Yang, H., Zhang, G., Wei, C., Du, X., Huang, W., Chen, W.: Consisti2v: Enhancing visual consistency for image-to-video generation (2024), https://arxiv.org/abs/2402.04324
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
38. Runway: Gen-2: Next-generation video generation. https://research.runwayml.com/gen2 (2023), accessed: 2023-09-25
39. Runway: Gen-3: Scaling video generation models. https://runwayml.com/research/introducing-gen-3-alpha (2024), accessed: 2024-06-17
40. Sadat, S., Hilliges, O., Weber, R.M.: Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In: International Conference on Learning Representations (ICLR) (2025), https://proceedings.iclr.cc/paper_files/paper/2025/hash/d1450d6c10c6b6cf1b80964357f5fa08-Abstract-Conference.html
41. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
42. Sun, X., Zhang, Z., Qiao, Y., Wang, G., Wang, H.: Twostreamvan: Improving motion modeling in video generation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2743–2751 (2020)
43. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: MoCoGAN: Decomposing motion and content for video generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1526–1535 (2018)
44. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025), https://arxiv.org/abs/2503.20314
45. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
46. Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., Zhang, S.: Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06071 (2023)
47. Wang, Y., et al.: Dynamic negative guidance of diffusion models. arXiv preprint (2024), https://arxiv.org/abs/2410.14398

48. Wu, S.C., et al.: Latentps: Image editing using latent representations in diffusion models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW). pp. 130–139 (2025)
49. Wu, W., Chen, A., Zhang, L., Wang, L., Wang, Z., Liu, H.: Mind the time: Temporally-controlled multi-event video generation. arXiv preprint (2025)
50. Xing, J., Xia, M., Zhang, Y., Chen, H., Wang, X., Wong, T.T., Shan, Y.: DynamiCrafter: Animating Open-domain Images with Video Diffusion Priors. In: ECCV (2024)
51. Xu, J., Zou, X., Huang, K., Chen, Y., Liu, B., Cheng, M., Shi, X., Huang, J.: Easyanimate: A high-performance long video generation method based on transformer architecture. arXiv preprint arXiv:2405.18991 (2024)
52. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Zhang, Y., Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer (2025), https://arxiv.org/abs/2408.06072
53. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
54. Yehezkel, S., Dahary, O., Voynov, A., Cohen-Or, D.: Navigating with annealing guidance scale in diffusion space. arXiv preprint (2025), https://arxiv.org/abs/2506.24108
55. Yuan, S., Huang, J., He, X., Ge, Y., Shi, Y., Chen, L., Luo, J., Yuan, L.: Identity-preserving text-to-video generation by frequency decomposition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12978–12988 (June 2025)
56. Zhang, L., Agrawala, M.: Packing input frame context in next-frame prediction models for video generation. arXiv preprint arXiv:2504.12626 (2025)
57. Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., Zhou, J.: I2VGen-XL: High-Quality Image-to-Video Synthesis via Cascaded Diffusion Models. arXiv preprint arXiv:2311.04145 (2023)
58. Zhang, X., Lin, H., Ye, H., Zou, J.Y., Ma, J., Liang, Y., Du, Y.: Inference-time scaling of diffusion models through classical search. arXiv preprint (2025), https://arxiv.org/abs/2505.23614
59. Zhao, Z., Wang, Y., Wu, J.B., Liu, Y., Liu, G., Liu, B.K., Liu, Y.: Motion prompting: Controlling video generation with motion trajectories. arXiv preprint arXiv:2404.03073 (2024)
60. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Zhang, Y., He, J., Zheng, W.S., Qiao, Y., Liu, Z.: VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025)