# When Distillation Breaks Motion Control: Restoring Generative Trajectories for Fast Video Generators

Jintao Rong[1,2,*], Xin Xie[2,*], Xinyi Yu[1], Linlin Ou[1,†], Xinyu Zhang[3], Chunhua Shen[1,4], and Dong Gong[2,†]

[1]Zhejiang University of Technology, China [2]UNSW Sydney, Australia
[3]University of Auckland, New Zealand [4]Zhejiang University, China

**Abstract.** Training-free motion customization imposes motion patterns from reference videos onto video generators through test-time computation. Most existing methods target full diffusion models, requiring many denoising steps and high computational cost. With the rise of efficient distilled models, a natural question arises: *can test-time motion customization be applied directly to distilled generators with their accelerated sampling and efficiency gains?* However, our analysis reveals that existing training-free techniques fail on distilled models. Distillation fundamentally alters the denoising dynamics that prior test-time guidance relies on, and the large denoising steps of distilled generators discard the dense intermediate states that score guidance requires, rendering existing motion control strategies incompatible with fast generation. To address this limitation, we propose `MotionEcho`, a novel training-free *test-time distillation* framework that enables motion customization for distilled video generators. The key idea is to correct the student model's sampling trajectory with restricted usage of a high-quality diffusion teacher at inference time. Teacher supervises the student's denoising by re-noising the student's endpoint onto its dense trajectory to form a motion-aligned clean endpoint, then interpolating it with the student's, while an adaptive scheduling mechanism determines when and how much teacher guidance is needed. As a result, `MotionEcho` restores generative trajectories for distilled video generators via lightweight, adaptive test-time teacher guidance, enabling accurate motion control without compromising generation efficiency. Extensive experiments on multiple distilled video generation models demonstrate that our method significantly improves motion fidelity and visual quality while retaining the efficiency advantages of distilled generation. Project page: MotionEcho.

## 1 Introduction

Text-to-Video (T2V) diffusion models [6,17,24,35,45,50,61] have recently achieved remarkable progress in conditional video generation, enabling the synthesis of high-quality and temporally coherent videos from textual prompts. Beyond text

* Equal contribution. † Corresponding authors. The work was done during Jintao Rong's Research Practicum visit at UNSW Sydney.

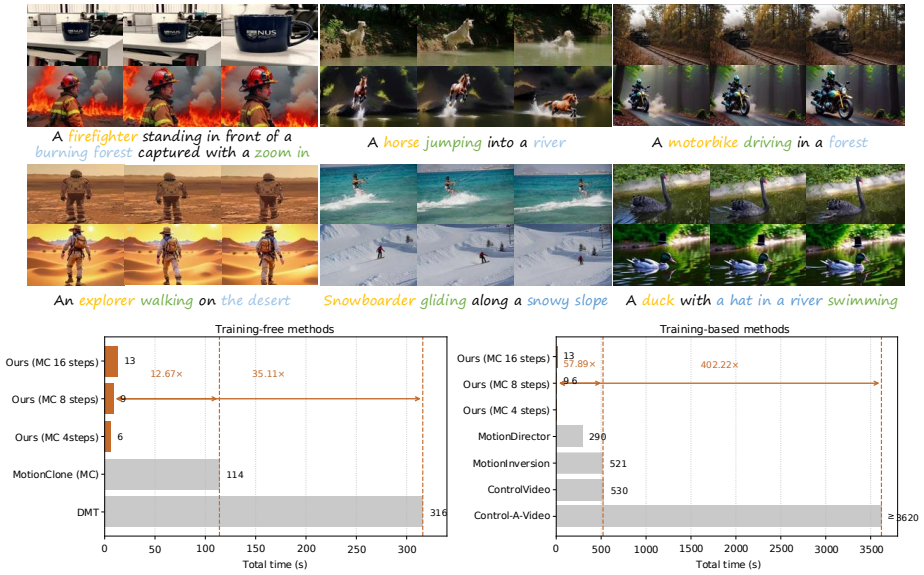


**Fig. 1:** `MotionEcho` is a training-free motion customization framework that enables effective motion control on distilled video generators through adaptive, teacher-guided test-time distillation. It achieves high-quality motion transfer across diverse motion types (top) while preserving the efficiency advantages of distilled generation (bottom).

guidance, a growing body of works has focused on motion customization [8, 19, 22, 31, 33, 37, 52, 56, 62, 68–70], which aims to generate videos that follow motion patterns extracted from reference videos while synthesizing novel visual content from text descriptions.

Existing motion customization approaches can be broadly categorized into training-based and training-free methods. Training-based methods incorporate motion information into T2V models via LoRA adapter [70], temporal attention adaptation [19], motion embeddings [52], and ControlNet-style condition branches [8, 68, 69]. While effective, such approaches require costly optimization for each motion condition. To enable more flexible motion control, training-free methods [22, 31, 37, 56, 62] inject motion patterns via test-time computation through optimization or regularization-based guidance, making them adaptable across different models and motion inputs. However, despite their flexibility, most existing training-free motion customization methods are designed for full diffusion models, whose iterative sampling process involves dozens of denoising steps. The training-free methods introduce additional computations on top of it.

With the development of distilled T2V generators [11, 24, 25, 29, 32, 53], which compress the multi-step denoising process of a teacher diffusion model into a few-step student model via distillation, inference is significantly accelerated while maintaining competitive generation quality. *This motivates a key question: can training-free test-time motion customization be adapted for distilled T2V models*

*with altered generative trajectories?* To investigate, we conducted preliminary experiments by directly applying representative training-free motion customization methods, including MotionClone [31] and DiTFlow [37], to distilled T2V models including T2V-Turbo-V2 [25] and Video-Blade [11]. Fig. 2 shows the generated videos exhibit significant motion degradation, *i.e.*, the motion cannot be exactly transferred from reference videos to targeted ones. We attribute these failures to fundamental differences between standard diffusion models and distilled generators: i) distilled models collapse multiple denoising steps into a small number of large-step updates, making the progressive interventions used in existing training-free methods *overly coarse and ineffective*; ii) the denoising dynamics of distilled models, trained to *approximate multi-step trajectories*, differ substantially from those of the teacher models with *dense* timesteps, rendering prior motion guidance strategies incompatible. The large denoising steps of distilled generators discard the dense intermediate states that score guidance requires.

To address this, we propose `MotionEcho`, a novel training-free framework that restores generative trajectories for distilled T2V models to enable effective test-time motion customization, while preserving the efficiency of the distilled fast generator. The key idea is to correct the student model's disrupted sampling trajectory using selective teacher guidance during inference, as a form of complementary test-time distillation, which bridges the dynamic gap without replacing the efficient distilled backbone. Specifically, the distilled model serves as the primary inference backbone to maintain efficient generation, while a high-quality teacher (full) diffusion model provides auxiliary supervision during selected reverse timesteps. Through endpoint prediction and interpolation, the teacher guidance acts as an *echo* that corrects the student's sampling trajectory and enables accurate motion transfer. To preserve efficiency, we further introduce an *adaptive acceleration strategy* that dynamically determines when teacher supervision is necessary and decides how to adjust the optimization budget at each denoising step. This adaptive scheduling balances motion fidelity and computational efficiency during inference. Quantitative and qualitative experiments demonstrate that our `MotionEcho` enables motion customization success on distilled models. As illustrated in Fig. 1, different motion patterns extracted from reference videos can be accurately transferred to targeted generated videos, while keeping high inference speed. Our contributions are as follows:

- We first identify the limitations of applying existing training-free motion customization methods to fast distilled T2V models and provide empirical and analytical evidence of the resulting motion degradation.
- We propose `MotionEcho`, a novel training-free test-time distillation framework that enables motion customization for distilled video diffusion models via teacher-guided test-time distillation during inference.
- We introduce an adaptive acceleration strategy that dynamically schedules when and how much teacher supervision is applied at each denoising step, effectively improving motion fidelity without sacrificing efficiency.

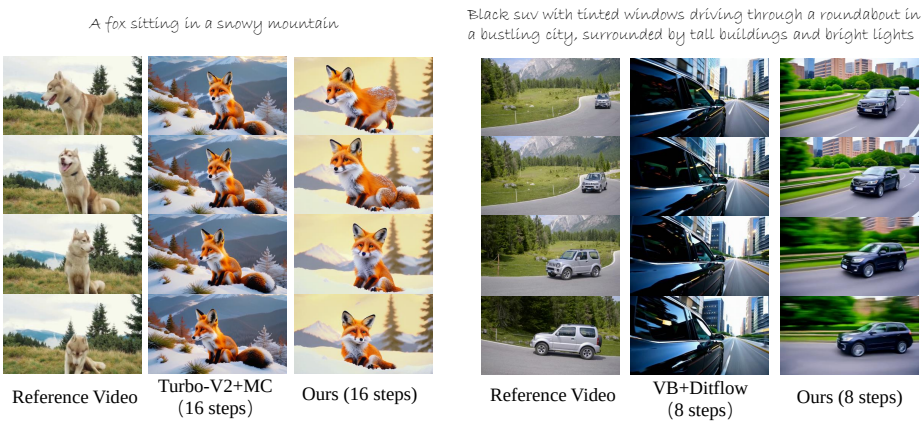


**Fig. 2: Visual comparison of motion customization on fast distilled video models.** Directly applying existing training-free methods to distilled models leads to severe degradation. For instance, the baseline fails to mimic the reference object motion (left, the fox fails to raise its head) and suffers from catastrophic structural degradation under dynamic camera motion (right, the background and car shape become corrupted). In contrast, our proposed `MotionEcho` successfully transfers the reference motion while maintaining high visual quality and strict text alignment during a highly accelerated few-step inference regime.

– We demonstrate through extensive experiments on multiple distilled video diffusion models that `MotionEcho` achieves superior motion fidelity, visual quality, and efficiency compared to other methods across two benchmarks.

## 2 Related Work

### 2.1 Text-to-Video Diffusion Models

With the advent of diffusion models [15, 47, 49], remarkable advancements in content creation have been witnessed in the text-to-image field [3,10,20,23,30,34, 36,43,44,57,58,66]. Following this success, recent text-to-video diffusion models, such as Sora [35], Gen-3 Alpha [42], and Veo [45], have achieved high-quality video generation but remain unavailable to the public. To break this limitation, the seminal work VDM [17] leverages a 3D UNet [9], factorized over space and time, for unconditional video generation. Subsequently, Imagen Video [14] and Make-A-Video [46] adopt cascade frameworks to enable high-resolution text-conditioned video generation in pixel space. To reduce the computational cost, LVDM [13], VideoLDM [4], and MagicVideo [71] extend diffusion models to a 3D latent space, facilitating training on large-scale video datasets [2, 59].

Rather than training from scratch, AnimateDiff [12] trains a domain adapter for quality enhancement and a temporal module for motion prior learning via

lightweight LoRAs [18]. VideoCrafter2 [7] constructs a UNet with temporal attention layers and uses low-quality videos for motion learning while using high-quality images for appearance learning. CogVideoX [61] uses a Diffusion Transformer (DiT) with expert transformers and multiple resolution training for coherent videos. Wan [50] advances large-scale video generation by using a diffusion transformer framework. T2V-Turbo [24] and T2V-Turbo-V2 [25] adopt consistency models [48] to distill from VideoCrafter2 [7] for accelerating the sampling process. Video-Blade [11] jointly trains adaptive block-sparse attention with sparsity-aware step distillation, enabling data-free distillation. OPAD [60] explores training-based concept personalization for one-step generation.

### 2.2 Video Motion Customization

Motion customization aims to enable the generated videos to follow specific motion patterns from reference videos while remaining semantically aligned with textual descriptions [8, 19, 26–28, 31, 37, 52, 54, 56, 67, 69, 70]. The pioneering work Tune-A-Video [54] introduces one-shot video generation by tuning SD [43] on a single reference video. ControlVideo [69] and Text2Video-Zero [22] inherit priors from ControlNet [65] and leverage cross-frame interactions to enable zero-shot controllable video synthesis. Furthermore, Control-A-Video [8] incorporates trainable motion layers into UNet and ControlNet [65] for temporal modeling, achieving motion control conditioned on depth, sketch, or motion information from reference videos. Despite achieving high-quality customization, these methods still suffer from motion misalignment.

To address this, existing motion-specific tuning methods, such as VMC [19], MotionDirector [70], FlexiMMT [27], and MotionInversion [52], are designed to decouple the learning of appearance and motion, enabling the generalization of distinct motion patterns to diverse textual prompts and scenes. Additionally, DMT [62] and MotionClone [31] employ training-free strategies, extracting motion priors from the latent representations of reference videos and constructing an energy function to guide motion customization during inference. MOFT [56] is a training-free approach that uses Principal Component Analysis (PCA) [1] to derive motion subspaces from the video diffusion model to control video motion. Zhang *et al.* [67] propose a motion consistency loss combined with inverted reference noise to enhance temporal coherence and motion accuracy. DiTFlow [37] transfers motion in DiT-based T2V models by extracting Attention Motion Flow (AMF) from cross-frame attention and performing latent optimization for motion guidance. FreeNoise [40] and FreeTraj [39] control video motion by rescheduling the initial noise. Furthermore, Video-MSG [26] generates a video motion sketch through multimodal planning and injects it into T2V generators via noise inversion.

## 3 Methods

We first establish a diagnostic baseline by directly applying existing training-free motion customization methods to distilled models, observing significant vi-

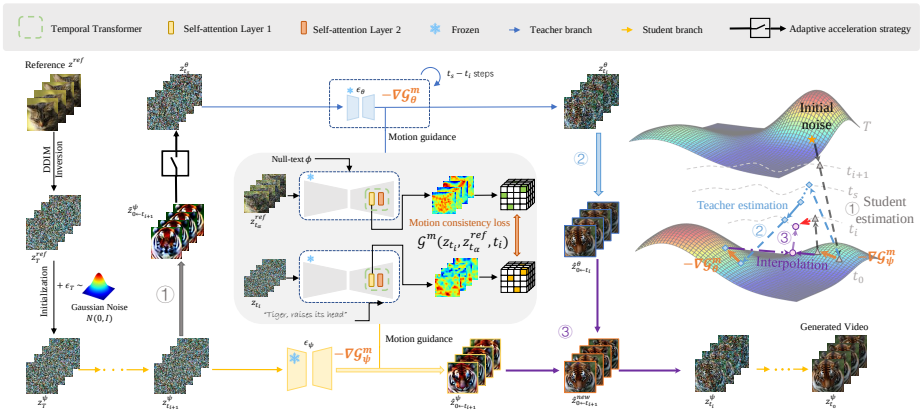


**Fig. 3: Pipeline of MotionEcho.** Given a reference video, motion priors are extracted to initialize the student model with a motion-preserving noisy latent. During inference, the teacher (blue path ②) and student (gray path ①) models perform motion customization using motion loss gradients. Teacher guidance is applied via prediction interpolation (purple path ③) at sub-interval endpoints. The student then generates the final video in a few steps with high motion fidelity.

sual artifacts and temporal inconsistencies arising from coarse denoising and mismatched sampling dynamics. To address this, we propose `MotionEcho`, a training-free framework that restores generative trajectories to enable effective motion customization in fast video generators. By strictly maintaining the distilled student as the primary inference backbone, `MotionEcho` leverages a fine-grained teacher diffusion model as an on-demand trajectory corrector via score distillation. To maintain efficiency, we further develop an adaptive acceleration strategy that triggers the teacher's guidance and truncates its denoising iterations. The whole pipeline of `MotionEcho` is demonstrated in Fig. 3.

### 3.1 Preliminaries

*Video Diffusion Models and Motion Customization.* Video diffusion models [3,4, 7, 12, 50, 61] generate videos by progressively denoising random noise through a learned reverse diffusion process. Given an input video $x$, a pre-trained encoder $\mathcal{E}$ maps it into the latent space $z_0 = \mathcal{E}(x)$, which is gradually corrupted into $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$ represents Gaussian noise and $\bar{\alpha}_t$ is a hyperparameter for controlling the diffusion process. To capture the data distribution, a neural network $\epsilon_\theta$ is trained to estimate the added noise from the noisy input $z_t$, which is formulated as follows:

$$\min_{\theta} \mathbb{E}_{\mathcal{E}(x),\, \epsilon_t \sim \mathcal{N}(0,\mathbf{I}),\, t \sim \mathcal{U}(1,T)} \left[ \left\| \epsilon_t - \epsilon_\theta(z_t, c, t) \right\|_2^2 \right], \tag{1}$$

where $c$ is the condition such as the textual prompt. During inference, the sampling process begins with standard Gaussian noise. To enhance controllability,

we typically introduce classifier-free guidance [16] or an additional customized energy function $\mathcal{G}(z_t, t)$ [31, 56, 62–64] during sampling:

$$\hat{\epsilon}_\theta = \epsilon_\theta(z_t, c, t) + \omega\big(\epsilon_\theta(z_t, c, t) - \epsilon_\theta(z_t, \phi, t)\big) - \eta\nabla_{z_t}\mathcal{G}(z_t, t), \tag{2}$$

where $\phi$ denotes the unconditional null prompt, while $\omega$ and $\eta$ are scaling factors that determine the strength of classifier-free guidance and additional guidance, respectively. From this viewpoint, motion customization can be instantiated by defining $\mathcal{G}$ as a motion energy that aligns the generated video with a reference motion pattern. While training-based methods [8, 52, 54, 68–70] adapt the model via fine-tuning or LoRA [18], training-free approaches [31,37,56,62,67] inject motion priors at test time by optimizing the sampling trajectory through motion-conditioned guidance. Both paradigms target the same objective of minimizing the mismatch between motion representations extracted from noisy latents, $\mathbb{E}_{z_0, z^{\text{ref}}, t, \epsilon_t}\left[\left\|\mathcal{M}(z_t^{\text{ref}}) - \mathcal{M}(z_t^g)\right\|_2^2\right]$, where $\mathcal{M}(\cdot)$ is a motion feature extractor, $z_t^{\text{ref}}$ is the noisy reference latent, and $z_t^g$ is the generated latent.

*Distillation for Accelerated Video Diffusion Models.* The inherently iterative nature of video diffusion models results in slow inference and substantial computational overhead. To alleviate this, recent efforts leverage knowledge distillation [11, 24, 25, 29, 32, 51] to compress generation process. Our framework is compatible with two primary distillation paradigms. For UNet-based models (e.g., T2V-Turbo-V2 [25]), methods typically adopt consistency distillation [48]. A student model $\epsilon_\psi$ is trained to approximate the macroscopic denoising trajectories of the teacher $\epsilon_\theta$ under a coarser timestep schedule $\mathcal{U}_\psi(1, T)$, formulated as $\min_\psi \mathbb{E}_{z_0, c, t\sim\mathcal{U}_\psi}\left[\|\epsilon_\psi(z_t, c, t) - \epsilon_\theta(z_t, c, t)\|_2^2\right]$. Conversely, for DiT-based architectures (e.g., Video-Blade [11]), the generation process is accelerated via trajectory distribution distillation [32]. The student model $\epsilon_\psi$ is trained to match, in a single large step, the result of multiple teacher steps over an interval $[t_{i-1}, t_i]$. Formally, the objective is $\min_\psi \mathbb{E}_{z_{t_i}, c, i}\left[\|\epsilon_\psi(z_{t_i}, c, t_i) - \tilde{\epsilon}_\theta(z_{t_i}, c, t_i \to t_{i-1})\|_2^2\right]$, where $\tilde{\epsilon}_\theta$ is the equivalent one-step noise prediction derived from the teacher's multi-step ODE solver output $\Phi_\theta(z_{t_i}, t_i \to t_{i-1})$.

### 3.2 Test-time Motion Customization for Fast Video Generators

To enable training-free motion customization on distilled, fast T2V models and preserving efficiency, we explore an intuitive solution that directly integrates motion control [31, 37, 56, 62] into the sampling process of the distilled video diffusion model $\epsilon_\psi$. This solution retains the efficiency of the distilled student model while injecting motion representations without any training. For each large denoising step from $t_{i+1}$ to $t_i$ in the distilled model, the predicted noise is modified as:

$$\hat{\epsilon}_\psi(z_{t_{i+1}}^\psi, c, t_{i+1}) = \tilde{\epsilon}_\psi(z_{t_{i+1}}^\psi, c, t_{i+1}) - \eta\nabla_{z_{t_{i+1}}^\psi}\mathcal{G}^m(z_{t_{i+1}}^\psi, z_{t_\alpha}^{\text{ref}}, t_{i+1}) \tag{3}$$


**Fig. 4: Qualitative comparisons on VideoCrafter2-based methods.** Our method enables unified object, camera, and hybrid motion transfer with high motion fidelity and low inference time.

where $\tilde{\epsilon}_\psi(z^\psi_{t_{i+1}}, c, t_{i+1}) = (1+\omega_\psi)\epsilon_\psi(z^\psi_{t_{i+1}}, c, t_{i+1}) - \omega_\psi\epsilon_\psi(z^\psi_{t_{i+1}}, \phi, t_{i+1})$ denotes classifier-free guidance [16]. The proposed `MotionEcho` is compatible with diverse inference-time guidance strategies and generative models. As a representative case, we use MotionClone [31] on diffusion models to illustrate the design of `MotionEcho`. Extensions to other guidance types and to flow-based models are provided in Section 3.5 and the supplementary material.

Following MotionClone [31], $\mathcal{G}^m(z^\psi_{t_{i+1}}, z^{\text{ref}}_{t_\alpha}, t_{i+1}) = ||M^{\text{ref}}_{t_\alpha} \cdot (\mathcal{A}(z^{\text{ref}}_{t_\alpha}) - \mathcal{A}(z^\psi_{t_{i+1}}))||_2^2$ is the motion loss function, where $\mathcal{A}(\cdot)$ represents the temporal attention map and $M^{\text{ref}}_{t_\alpha}$ is a temporal mask derived from $\mathcal{A}(z^{\text{ref}}_{t_\alpha})$ at a fixed timestep $t_\alpha$ to enforce focused motion supervision. We only apply motion guidance during the initial $\tau$ percent of the sampling process. Then, $z^\psi_{t_i}$ is obtained by re-noising the predicted clean latent:

$$z^\psi_{t_i} = \sqrt{\bar{\alpha}_{t_i}}\hat{z}^\psi_{0\leftarrow t_{i+1}} + \sqrt{1-\bar{\alpha}_{t_i}}\epsilon_{t_i}, \tag{4}$$

where $\hat{z}^\psi_{0\leftarrow t_{i+1}} = \frac{z^\psi_{t_{i+1}} - \sqrt{1-\bar{\alpha}_{t_{i+1}}}\hat{\epsilon}_\psi(z^\psi_{t_{i+1}}, c, t_{i+1})}{\sqrt{\bar{\alpha}_{t_{i+1}}}}$ is the one-step prediction result. The distilled model generates results in a few denoising steps, during which motion customization is applied.

However, empirical evidence reveals its fundamental inadequacy for motion customization quality and effectiveness. As shown in Fig. 2, the generated videos exhibit significant motion degradation, particularly in scenarios with complex object layouts or rapid movements. The degradation is further corroborated quan-


**Fig. 5: Qualitative comparisons on Wan2.1-1.3B-based methods.** Our method achieves consistent high-fidelity motion preservation across diverse scenarios while maintaining faster inference than baseline methods.

titatively in Table 3. The cause of these failures lies in the distinct generative dynamics of distilled models. Specifically, their sparse timestep schedule forces motion gradients to act across much larger denoising intervals. This loss of control granularity transforms what should be progressive guidance into abrupt and unstable drifts in the latent space. Furthermore, distillation changes the sampling paradigm. While teacher models refine latents through a dense, incremental path, student models are trained to take discrete leaps toward macroscopic endpoints. Injecting localized motion priors into this highly compressed, large-step generation process creates a severe dynamic mismatch, rendering existing motion guidance mechanisms ineffective.

### 3.3 Restoring Trajectories via Teacher-Guided Distillation

*Teacher-Guided Trajectory Correction.* To address these limitations, we introduce a selective test-time distillation framework that uses a fine-grained teacher to correct the student. We treat the student's predicted clean latent with motion guidance, $\tilde{z}^{\psi}_{0\leftarrow t_{i+1}} = \frac{z^{\psi}_{t_{i+1}} - \sqrt{1-\bar{\alpha}_{t_{i+1}}}\,\tilde{\epsilon}_{\psi}(z^{\psi}_{t_{i+1}}, c, t_{i+1})}{\sqrt{\bar{\alpha}_{t_{i+1}}}}$, as an anchor. To bridge the student's large step with the teacher's dense schedule, we renoise this anchor to an intermediate timestep $t_s \in [t_i, t_{i+1}]$ on the teacher's trajectory. Then, the teacher model follows Eq. (3) and Eq. (4) to perform iterative motion customization: $z^{\theta}_{t_{s-1}} \leftarrow z^{\theta}_{t_s} + \tilde{\epsilon}_{\theta}(z^{\theta}_{t_s}, c, t_s) - \eta \nabla_{z^{\theta}_{t_s}} \mathcal{G}^m(z^{\theta}_{t_s}, z^{\text{ref}}_{t_\alpha}, t_s)$. This multi-step process yields a highly accurate, motion-aligned prediction $\hat{z}^{\theta}_{0\leftarrow t_i}$.

*Score Interpolation as an Echo.* To preserve the efficiency of the distilled model, we formulate the trajectory correction as a decomposed optimization problem. To transfer the teacher's corrected trajectory back to the fast sampling process, we define a score distillation sampling (SDS) objective that aims to align the student's estimate with the teacher's refined prediction: $\ell_{\text{distill}} = ||\tilde{z}^{\psi}_{0\leftarrow t_{i+1}} - \text{sg}\left[\hat{z}^{\theta}_{0\leftarrow t_i}\right]||_2^2$. During inference, directly minimizing this objective w.r.t. the intermediate noisy state $z^{\psi}_{t_{i+1}}$ requires computing the Jacobian $\frac{\partial \tilde{\epsilon}_{\psi}}{\partial z^{\psi}_{t_{i+1}}}$. This requires backpropagating through the entire network, which is fundamentally intractable. To solve this bottleneck, we decouple the prediction from the network by treating the student's predicted clean latent $\tilde{z}^{\psi}_{0\leftarrow t_{i+1}}$ as an independent optimization variable. This formulation shifts the gradient update from the complex, high-curvature noisy score space directly to the clean data manifold. By evaluating the gradient step exclusively on the denoised estimate, the network Jacobian is completely bypassed. With the target fixed, the analytical gradient is straightforwardly $\nabla_{\tilde{z}^{\psi}_{0\leftarrow t_{i+1}}} \ell_{\text{distill}} = 2(\tilde{z}^{\psi}_{0\leftarrow t_{i+1}} - \hat{z}^{\theta}_{0\leftarrow t_i})$. The gradient descent step elegantly reduces to a closed-form linear interpolation:

$$\hat{z}^{\text{new}}_{0\leftarrow t_{i+1}} = \tilde{z}^{\psi}_{0\leftarrow t_{i+1}} - \frac{\lambda}{2}\nabla_{\tilde{z}^{\psi}_{0\leftarrow t_{i+1}}} \ell_{\text{distill}} = (1-\lambda)\tilde{z}^{\psi}_{0\leftarrow t_{i+1}} + \lambda \hat{z}^{\theta}_{0\leftarrow t_i}, \tag{5}$$

where $\lambda$ controls the step size of the gradient descent, effectively acting as the strength of the teacher's restorative echo, and $\hat{z}^{\text{new}}_{0\leftarrow t_{i+1}}$ is the corrected student's predicted endpoint. Finally, the student model proceeds to the next noisy manifold $t_i$ using this interpolated clean latent:

$$z^{\psi}_{t_i} = \sqrt{\bar{\alpha}_{t_i}}\hat{z}^{\text{new}}_{0\leftarrow t_{i+1}} + \sqrt{1-\bar{\alpha}_{t_i}}\epsilon_{t_i}. \tag{6}$$

### 3.4 Adaptive Acceleration Strategy for Efficiency

Applying uniform teacher supervision across the initial $\tau$ percent of sampling is both computationally prohibitive and counterproductive due to varying denoising dynamics. To maintain efficiency, we introduce an adaptive acceleration strategy comprising two core mechanisms. i) *Step-wise Guidance Activation.* The decision to apply teacher guidance at each step becomes critical, which significantly affects the effectiveness of motion transfer. To make this decision more robust, we evaluate the necessity of distillation by computing the average motion loss over a moving window of previous steps, $\mathcal{G}^{\psi}_{t_i} = \frac{1}{W}\sum_{j=0}^{W-1}\mathcal{G}^m(z^{\psi}_{t_{i+j}}, z^{\text{ref}}_{t_\alpha}, t_{i+j})$, where $t_i$ denotes the current timestep and $W$ is the window size. If the average motion loss $\mathcal{G}^{\psi}_{t_i}$ exceeds a predefined threshold $\delta_1$, we activate teacher guidance for this step. ii) *Dynamic Truncation.* Once guidance is triggered, we control the internal guidance process of the teacher model via a dynamic truncation based on the smoothness of the motion loss. Specifically, during the interval from $t_s$ to $t_i$, we monitor the motion loss $\mathcal{G}^m(z^{\theta}_{t_s}, z^{\text{ref}}_{t_\alpha}, t_s)$. If the loss falls below a confidence threshold $\delta_2$, we terminate the motion transfer early and directly denoise the current latent to the latent of $t_i$.

### 3.5 MotionEcho for DiT-based Video Generators

Since adaptive distillation operates at the trajectory level, `MotionEcho` accommodates different backbones and motion extractors $\mathcal{M}(\cdot)$. DiT-based video generation methods achieve superior performance by jointly processing spatiotemporal information through attention mechanisms. Our method integrates both distilled DiT student models and teacher models with two key adaptations. Unlike UNet-based methods (*e.g.*, MotionClone [31]) that assume separable temporal attention, DiTs employ full spatiotemporal attention where motion and content are entangled. Therefore, we adopt generalized motion extractors (MOFT [56], DMT [62]) and DiT-specific methods (DiTFlow [37]). Once reference motion representations are extracted from DiT, we follow the test-time distillation paradigm with teacher motion guidance at specific steps to optimize $z_{t_i}^{\psi}$. Following DiTFlow [37], we inject motion priors by processing $z_0^{\text{ref}}$ through the DiT and injecting the resulting keys $K$ and values $V$ into the target attention mechanism.

## 4 Experiments

We evaluate `MotionEcho` on a wide range of distilled video generators, comparing against state-of-the-art training-based and training-free motion customization methods. We further analyze each component's contribution via ablation studies.

### 4.1 Experimental Setting

*Implementation Details.* Without requiring additional training, our method can be seamlessly integrated with various generative video diffusion models. To facilitate discussion, we denote the models by their abbreviations in parentheses. During inference, VideoCrafter2 [7] (VC2) is employed as the teacher model to provide guidance for the distilled model, T2V-Turbo-V2 [25] (TurboV2). On TurboV2, we use MotionClone [31] (MC) to align reference motion for the teacher and student models. To further demonstrate the scalability, `MotionEcho` is also generalized to the DiT distilled model Video-Blade [11] (VB), using DiT Wan2.1 with 1.3 billion (Wan2.1-1.3B) parameter variants as the teacher model. All our experiments are conducted using an NVIDIA A100-40G GPU. Experimental configurations and hyperparameters are detailed in the supplementary material.

*Datasets.* We evaluate our method on two distinct benchmarks. For methods based on TurboV2 and VC2, we remain consistent with prior works [52,62], using source videos obtained from the DAVIS dataset [38], WebVID dataset [2], and online video resources. We refer to the combined validation dataset as TurboBench, consisting of 66 video-edit text pairs derived from 22 unique videos. This benchmark comprehensively spans diverse real-world scenarios, encompassing multiple object categories, varied scene configurations, and rich motion patterns ranging from simple translational movements to complex non-rigid deformations. For methods based on Wan2.1-1.3B and VB, we extracted 34 real-world videos from

**Table 1:** Quantitative comparison of VideoCrafter2-based methods on TurboBench.

| Method | Backbone | TC (↑) | TA (↑) | MF (↑) | FID (↓) | ITC (↓) | TTC (↓) |
|---|---|---|---|---|---|---|---|
| ControlVideo [69] | ControlNet | 0.9420 | 0.251 | 0.964 | 379.37 | 80s | 450s |
| Control-A-Video [8] | ControlNet | 0.925 | 0.256 | 0.858 | 383.95 | 20s | hours |
| MotionDirector [70] | VideoCrafter2 | 0.922 | 0.329 | 0.851 | 373.21 | 10s | 280s |
| MotionInversion [52] | VideoCrafter2 | 0.951 | 0.321 | 0.885 | 351.72 | 32s | 489s |
| DMT [62] | VideoCrafter2 | 0.961 | 0.259 | 0.909 | 367.18 | 316s | - |
| MotionClone [31] | VideoCrafter2 | **0.978** | 0.335 | 0.876 | 369.49 | 114s | - |
| Ours (MC 16 steps) | TurboV2 | <u>0.976</u> | **0.348** | **0.933** | **322.97** | 13s | - |
| Ours (MC 8 steps) | TurboV2 | 0.967 | <u>0.338</u> | <u>0.931</u> | <u>335.65</u> | 9s | - |
| Ours (MC 4 steps) | TurboV2 | 0.956 | 0.323 | 0.927 | 347.91 | 6s | - |

the DAVIS dataset [38] for evaluation. Each video is paired with text prompts at three difficulty levels (easy, medium, hard). We refer to this comprehensive evaluation set as DavisBench. We conduct additional evaluation across different benchmarks [55] in the supplementary material.

*Metrics.* To ensure a comprehensive assessment, we perform an evaluation from multiple perspectives. (1) **Text Alignment (TA)**: The metric measures how well the generated video content corresponds to the input text prompt by calculating the average cosine similarity between the embeddings of all video frames and the text embeddings, using the CLIP model [41]. (2) **Temporal Consistency (TC)**: The metric evaluates the visual smoothness and coherence of generated videos. We compute cosine similarities between the CLIP image embeddings [41] across frames and aggregate results into a stability score, which reflects flicker artifacts and semantic discontinuities while allowing motion-driven appearance changes. (3) **Motion Fidelity (MF)**: We assess the effectiveness of motion transfer using the Motion Fidelity Score [62]. The metric relies on motion trajectories extracted via Co-Tracker [21] and computes the geometric consistency between the motion patterns in the source and generated videos. (4) **Fréchet Inception Distance (FID)**: We construct a reference image set from selected source videos and evaluate the quality of our generated videos using FID [5]. (5) **Time Cost**: This measures the total time required to complete motion transfer from a reference video, termed Inference Time Cost (ITC). To ensure a fair comparison, additional fine-tuning or optimization steps are taken into account, termed Training Time Cost (TTC).

### 4.2 Comparison with Existing Methods

To verify the effectiveness of our approach, we compare `MotionEcho` with training-based and training-free methods. The former comprises ControlVideo [69], Control-A-Video [8], MotionDirector [70], and MotionInversion [52]. The latter includes MotionClone [31], MOFT [56], DMT [62], and DiTFlow [37].

*Quantitative Analysis.* To objectively evaluate the performance of our method, we conduct quantitative comparisons on two benchmarks using different base

**Table 2:** Quantitative comparison of Wan2.1-1.3B-based methods on DavisBench.

| Method | Backbone | TC (↑) | TA (↑) | MF (↑) | FID (↓) | ITC | TTC (↓) |
|---|---|---|---|---|---|---|---|
| MotionDirector* | Wan2.1-1.3B | 0.924 | 0.278 | 0.627 | 135.571 | 20s | 3629s |
| MOFT [56] | Wan2.1-1.3B | 0.961 | **0.326** | 0.505 | 131.98 | 116s | - |
| DMT [62] | Wan2.1-1.3B | 0.963 | 0.321 | 0.622 | **129.49** | 117s | - |
| DiTFlow [37] | Wan2.1-1.3B | <u>0.969</u> | <u>0.324</u> | 0.605 | 134.42 | 120s | - |
| Ours (MOFT 8 steps) | VB | 0.962 | 0.312 | 0.731 | <u>130.01</u> | 50s | - |
| Ours (DMT 8 steps) | VB | 0.963 | 0.313 | <u>0.732</u> | 132.52 | 49s | - |
| Ours (DitFlow 8 steps) | VB | **0.971** | 0.311 | **0.733** | 131.82 | 54s | - |

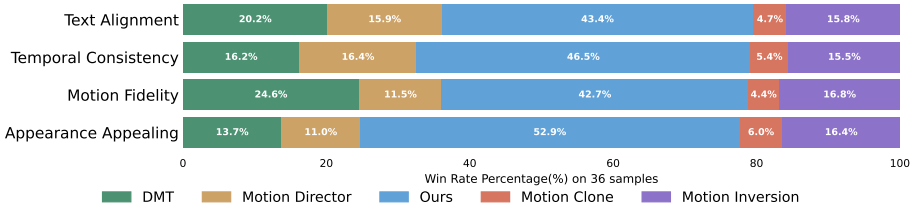


**Fig. 6:** User study results.

models. As shown in Table 1, our method (MC 16 steps) achieves the best performance across text alignment, motion fidelity, and FID score, while maintaining competitive temporal consistency within just 13 seconds. Compared to prior methods, Control-A-Video [8] and MotionDirector [70] show similar or higher inference times but significantly lower scores in key quality metrics and require costly training. Moreover, our method can reduce the inference steps while remaining highly competitive. With only 8 steps, it outperforms most baselines across all metrics in just 9 seconds. Even at 4 steps, it achieves a FID of 347.91 and maintains solid temporal consistency, reducing inference time to only 6 seconds. Additionally, we apply our method to the Wan2.1-1.3B-based distilled video model (VB), as shown in Table 2. Our approach consistently outperforms the training-free baselines (*e.g.*, MOFT, DMT, DiTFlow) on MF, while accelerating the inference process by more than 2×. Furthermore, compared to the training-based baseline MotionDirector* (reproduced on the Wan2.1-1.3B), our method completely bypasses the 3629s training time cost and still delivers substantially better temporal consistency, text alignment, and motion fidelity.

*Qualitative Evaluation.* We separately evaluate our method through qualitative comparisons against baselines based on VideoCrafter2 and Wan2.1-1.3B frameworks. In Fig. 4, compared to the over-smoothing and structural collapse observed in prior motion customization models, our method preserves spatial details while achieving high motion fidelity, particularly under challenging scenarios such as zoom out, snowy forests, and complex multi-object scenes. Fig. 5 demonstrates our method achieves consistent improvements on Wan2.1-1.3B-based motion guidance methods. Compared to MOFT, DMT, and DiTFlow,

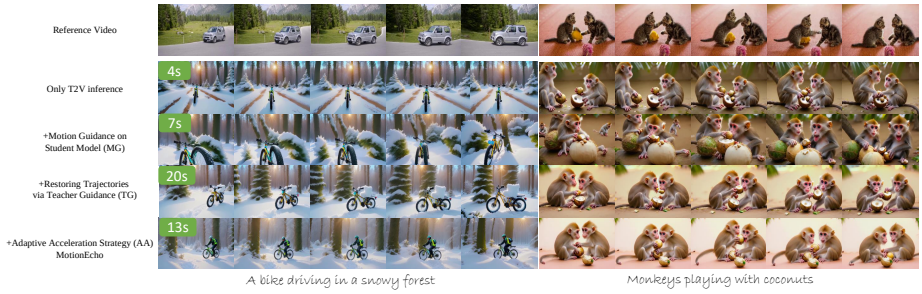


**Fig. 7:** Visualization of the ablation study results.

**Table 3:** Ablation study of the key components in our method. **MG**: Motion Guidance, **TG**: Teacher Guidance, **AA**: Adaptive Acceleration.

| Base Model | Backbone | MG | TG | AA | TC (↑) | TA (↑) | MF (↑) | TTC | ITC (↓) |
|---|---|---|---|---|---|---|---|---|---|
| MotionDirector* | TurboV2 | - | - | - | 0.965 | 0.327 | 0.806 | 300s | 4s |
| Ours | TurboV2 | × | × | × | 0.973 | 0.348 | 0.562 | - | 4s |
| | TurboV2 | ✓ | × | × | 0.967 | 0.324 | 0.833 | - | 7s |
| | TurboV2 | ✓ | ✓ | × | 0.965 | 0.336 | 0.912 | - | 20s |
| | TurboV2 | ✓ | ✓ | ✓ | **0.976** | **0.348** | **0.933** | - | 13s |

which struggle with motion preservation, our method maintains high fidelity and better visual quality across various motions. These visual results further validate our method is architecture-agnostic and effective in achieving unified and high-fidelity motion transfer, while maintaining efficiency.

*User Study.* We perform a user study to subjectively assess our method in comparison with VideoCrafter2-based methods, including DMT [62], MotionClone [31], MotionInversion [52], and MotionDirector [70]. We randomly sample 36 video-edit text pairs from the TurboBench dataset and synthesize motion videos with the aforementioned methods. We invite 50 participants and ask them to compare all generated videos from four different aspects. Fig. 6 reports win rates across four perceptual dimensions. MotionEcho achieves the highest preference in all categories, with particularly strong margins in appearance appeal (52.9%) and temporal consistency (46.5%), suggesting that teacher-guided trajectory correction produces videos that are not only more motion-accurate but also more visually coherent to human observers.

### 4.3 Ablation Study

We ablate each component in `MotionEcho` to achieve comprehensive assessment based on TurboV2 with TurboBench. The results are detailed in Table 3 and Fig. 7. First, directly applying Motion Guidance (**MG**) from MotionClone [31] injects motion priors and lifts MF from 0.562 to 0.833, but this coarse gradient intervention slightly degrades temporal consistency and text alignment. Then,

adding Teacher Guidance (**TG**) acts as a restorative echo that corrects trajectory deviations, further boosting MF to 0.912 and recovering TA to 0.336, though at the cost of ITC rising to 20s. Finally, the Adaptive Acceleration (**AA**) strategy selectively activates and truncates the teacher's denoising iterations, cutting ITC back to 13s while reaching the best overall performance. Notably, our full `MotionEcho` pipeline outperforms the training-based baseline MotionDirector (reproduced on TurboV2) in both quality and motion fidelity while eliminating its 300s training cost. Additional ablations are in the **supplementary material**.

## 5 Conclusion

We present `MotionEcho`, a training-free framework for motion customization in fast distilled video generators. We identify a fundamental incompatibility between prior guidance methods and distilled models, and address it via teacher-guided test-time distillation, which corrects the student's sampling trajectory through score interpolation. An adaptive acceleration strategy ensures selective and efficient teacher involvement. Experiments across multiple distilled backbones show that `MotionEcho` achieves state-of-the-art motion fidelity and visual quality while retaining the speed of distilled generation. In future work, we will explore unpaired teacher models with unshared latent spaces, a challenging open problem in diffusion model distillation, and investigate self-assessment mechanisms enabling more efficient and selective speculative guidance during inference.

## 6 Acknowledgments

The contribution of L. Ou was supported by the Baima Lake Laboratory Joint Funds of the Zhejiang Provincial Natural Science Foundation (No. LBMHD24F0 300023) and by the National Natural Science Foundation of China (No. 62373329). The contribution of X. Yu was supported by the Zhejiang Provincial Natural Science Foundation of China (No. LZ25F030003).

## References

1. Abdi, H., Williams, L.J.: Principal component analysis. Wiley interdisciplinary reviews: computational statistics **2**(4), 433–459 (2010)
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1728–1738 (2021)
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)

5. Bynagari, N.B.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Asian J. Appl. Sci. Eng **8**(1), 25–34 (2019)
6. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023)
7. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320 (2024)
8. Chen, W., Ji, Y., Wu, J., Wu, H., Xie, P., Li, J., Xia, X., Xiao, X., Lin, L.: Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. arXiv preprint arXiv:2305.13840 (2023)
9. Ciccek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19. pp. 424–432. Springer (2016)
10. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first International Conference on Machine Learning (2024)
11. Gu, Y., Li, X., Hu, Y., Zhuang, B.: Video-blade: Block-sparse attention meets step distillation for efficient video generation. arXiv preprint arXiv:2508.10774 (2025)
12. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
13. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)
14. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., Salimans, T.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
16. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
17. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in Neural Information Processing Systems **35**, 8633–8646 (2022)
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
19. Jeong, H., Park, G.Y., Ye, J.C.: Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9212–9221 (2024)
20. Jha, S., Yang, S., Ishii, M., Zhao, M., Mirza, M.J., Gong, D., Yao, L., Takahashi, S., Mitsufuji, Y., et al.: Mining your own secrets: Diffusion classifier scores for continual personalization of text-to-image diffusion models. In: International Conference on Learning Representations. vol. 2025, pp. 102294–102323 (2025)
21. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. In: European Conference on Computer Vision. pp. 18–35. Springer (2024)

22. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15954–15964 (2023)
23. Labs, B.F.: Flux.1-schnell (2024), `https://huggingface.co/black-forest-labs/FLUX.1-schnell`, accessed: 2024-08-17
24. Li, J., Feng, W., Fu, T.J., Wang, X., Basu, S., Chen, W., Wang, W.Y.: T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. arXiv preprint arXiv:2405.18750 (2024)
25. Li, J., Long, Q., Zheng, J., Gao, X., Piramuthu, R., Chen, W., Wang, W.Y.: T2v-turbo-v2: Enhancing video generation model post-training through data, reward, and conditional guidance design. arXiv preprint arXiv:2410.05677 (2024)
26. Li, J., Yu, S., Lin, H., Cho, J., Yoon, J., Bansal, M.: Training-free guidance in text-to-video generation via multimodal planning and structured noise initialization. arXiv preprint arXiv:2504.08641 (2025)
27. Li, Y., Gong, D., Cao, X., Yuan, J., Li, D., Zhou, L., Koh, Y.S., Yan, C., Zhang, X.: Let your image move with your motion!–implicit multi-object multi-motion transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11207–11217 (2026)
28. Liao, M., Lu, H., Zhang, X., Wan, F., Wang, T., Zhao, Y., Zuo, W., Ye, Q., Wang, J.: Evaluation of text-to-video generation models: A dynamics perspective. Advances in Neural Information Processing Systems **37**, 109790–109816 (2024)
29. Lin, S., Yang, X.: Animatediff-lightning: Cross-model diffusion distillation. arXiv preprint arXiv:2403.12706 (2024)
30. Lin, Z., Wang, Z., Liu, Q., Zhang, X., Liu, J.: Narratology meets text-to-image: a survey of consistency in ai generated storybook illustrations. Artificial Intelligence Review (2026)
31. Ling, P., Bu, J., Zhang, P., Dong, X., Zang, Y., Wu, T., Chen, H., Wang, J., Jin, Y.: Motionclone: Training-free motion cloning for controllable video generation. arXiv preprint arXiv:2406.05338 (2024)
32. Luo, Y., Hu, T., Sun, J., Cai, Y., Tang, J.: Learning few-step diffusion models by trajectory distribution matching. arXiv preprint arXiv:2503.06674 (2025)
33. Ma, Y., Liu, Y., Zhu, Q., Yang, X., Feng, K., Zhang, X., Yan, Z., Li, Z., Han, S., Qi, C., Chen, Q.: Effivmt: Video motion transfer via efficient spatial-temporal decoupled finetuning. In: International Conference on Learning Representations (2026)
34. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
35. OpenAI: Video generation models as world simulators (2024), `https://openai.com/sora/`, sora technical report
36. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
37. Pondaven, A., Siarohin, A., Tulyakov, S., Torr, P., Pizzati, F.: Video motion transfer with diffusion transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22911–22921 (June 2025)
38. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)

39. Qiu, H., Chen, Z., Wang, Z., He, Y., Xia, M., Liu, Z.: Freetraj: Tuning-free trajectory control in video diffusion models. arXiv preprint arXiv:2406.16863 (2024)
40. Qiu, H., Xia, M., Zhang, Y., He, Y., Wang, X., Shan, Y., Liu, Z.: Freenoise: Tuning-free longer video diffusion via noise rescheduling. arXiv preprint arXiv:2310.15169 (2023)
41. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
42. Research, R.: Introducing gen-3 alpha: A new frontier for video generation (2024), `https://runwayml.com/research/introducing-gen-3-alpha`, runway Gen-3 technical report
43. 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)
44. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
45. Sharma, A., Yu, A., Razavi, A., Toor, A., Pierson, A., Gupta, A., Waters, A., van den Oord, A., Tanis, D., Erhan, D., Lau, E., Shaw, E., Barth-Maron, G., Shaw, G., Zhang, H., Nandwani, H., Moraldo, H., Kim, H., Blok, I., Bauer, J., Donahue, J., Chung, J., Mathewson, K., David, K., Espeholt, L., van Zee, M., McGill, M., Narasimhan, M., Wang, M., Bińkowski, M., Babaeizadeh, M., Saffar, M.T., de Freitas, N., Pezzotti, N., Kindermans, P.J., Rane, P., Hornung, R., Riachi, R., Villegas, R., Qian, R., Dieleman, S., Zhang, S., Cabi, S., Luo, S., Fruchter, S., Nørly, S., Srinivasan, S., Pfaff, T., Hume, T., Verma, V., Hua, W., Zhu, W., Yan, X., Wang, X., Kim, Y., Du, Y., Chen, Y.: Veo. Placeholder Journal (2024), `https://deepmind.google/technologies/veo/`
46. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
47. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
48. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models (2023)
49. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
50. 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)
51. Wang, F.Y., Huang, Z., Bian, W., Shi, X., Sun, K., Song, G., Liu, Y., Li, H.: Animatelcm: Computation-efficient personalized style video generation without personalized video data. In: SIGGRAPH Asia 2024 Technical Communications, pp. 1–5 (2024)
52. Wang, L., Mai, Z., Shen, G., Liang, Y., Tao, X., Wan, P., Zhang, D., Li, Y., Chen, Y.: Motion inversion for video customization. arXiv preprint arXiv:2403.20193 (2024)
53. Wang, X., Zhang, S., Zhang, H., Liu, Y., Zhang, Y., Gao, C., Sang, N.: Videolcm: Video latent consistency model. arXiv preprint arXiv:2312.09109 (2023)

54. Wu, J.Z., Ge, Y., Wang, X., Lei, S.W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., Shou, M.Z.: Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7623–7633 (2023)
55. Wu, J.Z., Li, X., Gao, D., Dong, Z., Bai, J., Singh, A., Xiang, X., Li, Y., Huang, Z., Sun, Y., et al.: Cvpr 2023 text guided video editing competition. arXiv preprint arXiv:2310.16003 (2023)
56. Xiao, Z., Zhou, Y., Yang, S., Pan, X.: Video diffusion models are training-free motion interpreter and controller. vol. 37, pp. 7594–7611 (2024)
57. Xie, X., Gong, D.: Dymo: Training-free diffusion model alignment with dynamic multi-objective scheduling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13220–13230 (2025)
58. Xie, X., Guo, J., Gong, D.: Hyperalign: Hypernetwork for efficient test-time alignment of diffusion models. arXiv preprint arXiv:2601.15968 (2026)
59. Xue, H., Hang, T., Zeng, Y., Sun, Y., Liu, B., Yang, H., Fu, J., Guo, B.: Advancing high-resolution video-language representation with large-scale video transcriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5036–5045 (2022)
60. Yang, Y., Wu, T., Li, S., Yang, S., Wang, Y., van de Weijer, J., Wang, K.: Adversarial concept distillation for one-step diffusion personalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4321–4333 (2026)
61. 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., Bin, X., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) International Conference on Representation Learning. vol. 2025, pp. 83048–83077 (2025)
62. Yatim, D., Fridman, R., Bar-Tal, O., Kasten, Y., Dekel, T.: Space-time diffusion features for zero-shot text-driven motion transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8466–8476 (2024)
63. Ye, H., Lin, H., Han, J., Xu, M., Liu, S., Liang, Y., Ma, J., Zou, J., Ermon, S.: Tfg: Unified training-free guidance for diffusion models. arXiv preprint arXiv:2409.15761 (2024)
64. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23174–23184 (2023)
65. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
66. Zhang, X., Yang, L., Li, G., Cai, Y., Xie, J., Tang, Y., Yang, Y., Wang, M., Cui, B.: Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. arXiv preprint arXiv:2410.07171 (2024)
67. Zhang, X., Duan, Z., Gong, D., Liu, L.: Training-free motion-guided video generation with enhanced temporal consistency using motion consistency loss. arXiv preprint arXiv:2501.07563 (2025)
68. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arXiv preprint arXiv:2305.13077 (2023)

69. Zhao, M., Wang, R., Bao, F., Li, C., Zhu, J.: Controlvideo: conditional control for one-shot text-driven video editing and beyond. Science China Information Sciences **68**(3), 132107 (2025)
70. Zhao, R., Gu, Y., Wu, J.Z., Zhang, D.J., Liu, J.W., Wu, W., Keppo, J., Shou, M.Z.: Motiondirector: Motion customization of text-to-video diffusion models. In: European Conference on Computer Vision. pp. 273–290. Springer (2024)
71. Zhou, D., Wang, W., Yan, H., Lv, W., Zhu, Y., Feng, J.: Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018 (2022)