arXiv:2503.23951v3 [cs.CV] 24 Nov 2025

JointTuner: Appearance-Motion Adaptive Joint Training for Customized Video Generation

Fangda Chen[1], Shanshan Zhao[2], Chuanfu Xu[1], Long Lan[1]*
[1]College of Computer Science and Technology, National University of Defense Technology
[2]Alibaba International Digital Commerce
{fdchen.nudt, xuchuanfu, long.lan}@nudt.edu.cn
sshan.zhao00@gmail.com

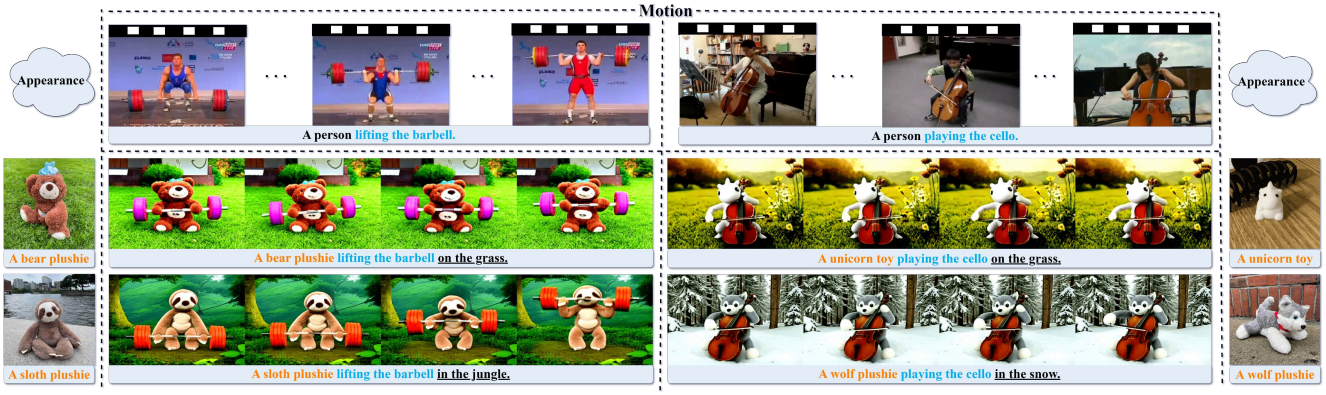


Figure 1. **Results of customized video generation using JointTuner.** Given paired appearance and motion inputs, it produces videos that reflect both the desired subject appearance and motion patterns through adaptive joint training.

Abstract

Recent advancements in customized video generation have led to significant improvements in the simultaneous adaptation of appearance and motion. Typically, decoupling the appearance and motion training, prior methods often introduce concept interference, resulting in inaccurate rendering of appearance features or motion patterns. In addition, these methods often suffer from appearance contamination, in which background and foreground elements from reference videos distort the customized video. This paper aims to alleviate these issues by proposing ***JointTuner****. The core motivation of our JointTuner is to enable joint optimization of both appearance and motion components, upon which two key innovations are developed, i.e.,* ***Gated Low-Rank Adaptation (GLoRA)*** *and* ***Appearance-independent Temporal Loss (AiT Loss)****. Specifically, GLoRA uses a context-aware activation layer, analogous to a gating regulator, to dynamically steer LoRA modules toward learning either appearance or motion while maintaining spatio-temporal consistency. Moreover, with the finding that channel-temporal shift noise suppresses appearance-related low-frequencies while enhancing motion-related high-frequencies, we designed the AiT Loss. This loss adds the same shift to the diffusion model's predicted noise during fine-tuning, forcing the model to prioritize learning motion patterns. JointTuner's architecture-agnostic design supports both UNet (e.g., ZeroScope) and Diffusion Transformer (e.g., CogVideoX) backbones, ensuring its customization capabilities scale with the evolution of foundational video models. Furthermore, we present a systematic evaluation framework for appearance-motion combined customization, covering 90 combinations evaluated along four critical dimensions: semantic alignment, motion dynamism, temporal consistency, and perceptual quality. Our project homepage is available online.*

1. Introduction

Recent advances in text-to-video generation have defined three core tasks in customized video generation (CVG): appearance customization using images to preserve subject identity [22, 28, 46, 49]; motion customization by transferring motion from reference videos [27, 48, 54]; and appearance-motion combined customization, which simultaneously controls both aspects [9, 20, 25, 34, 47, 56]. Appearance customization typically focuses on static attributes, while motion customization emphasizes dynamic behavior. For complex motions or fine-grained appearances beyond textual description, combining images for identity and videos for motion enables more precise results. In this

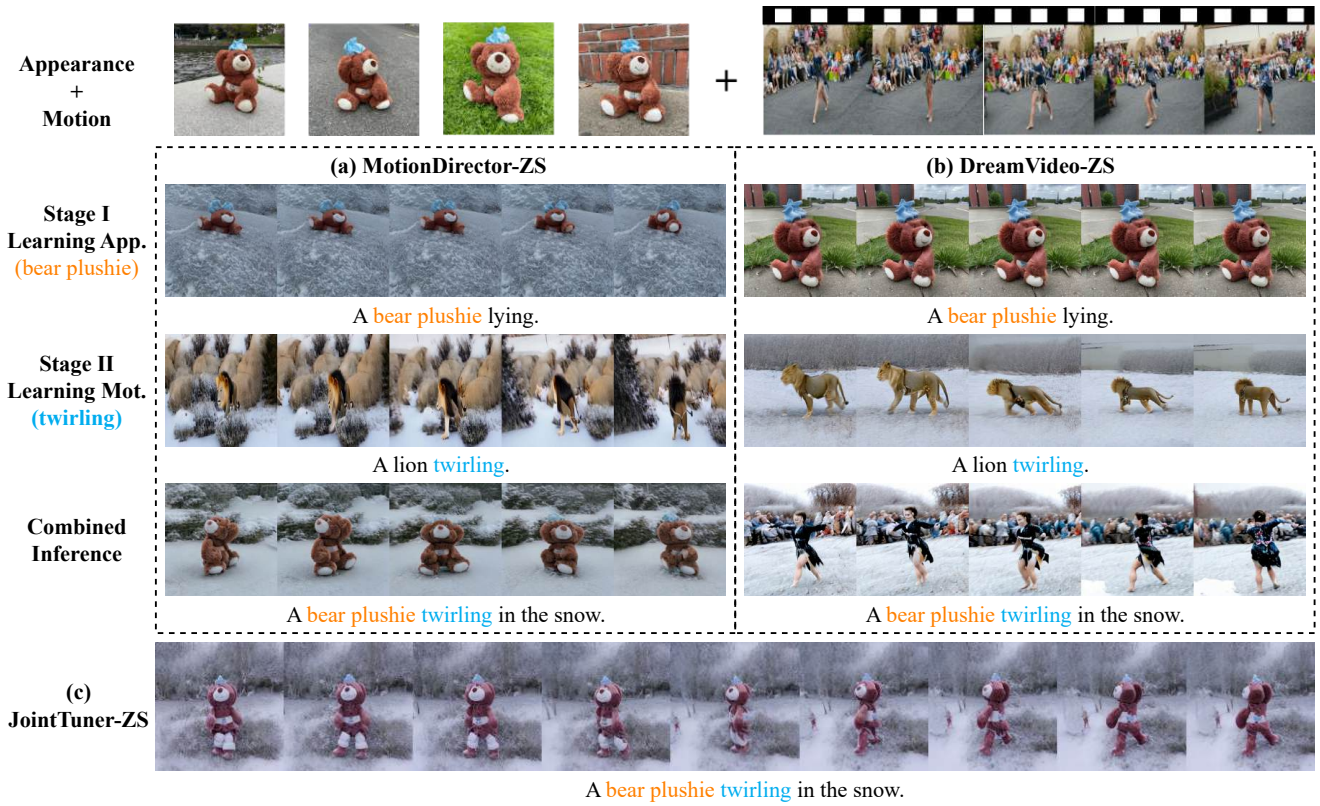


Figure 2. Illustration of a failure case from advanced customized video generation methods. (a) and (b) show videos generated by MotionDirector-ZS [56] and DreamVideo-ZS [47] across three stages: appearance learning (bear plushie), motion learning (twirling), and combined inference. (c) presents results from JointTuner-ZS, which improves inference by jointly learning both appearance and motion. Note that "-ZS" denotes methods based on ZeroScope [40].

work, we focus on this setting to customize both appearance and motion, aiming to enhance overall generation quality.

Existing appearance-motion combined customization methods typically adopt stage-wise training, separately optimizing appearance and motion components [47, 56]. Such a decoupled process often leads to *concept interference*: as shown in Fig. 2(a), models struggle to generate desired motion due to poor spatial-temporal fusion. Another issue is *appearance contamination*, where unwanted elements from reference videos leak into the output (Fig. 2(b)). Recently, Diffusion Transformer (DiT)-based text-to-video models [11, 43] have enabled higher-quality and longer video generation. Yet, simply inserting learnable parameters into DiT, training it stage-wise, and fusing the modules for inference still leads to *concept interference* [45].

To overcome these challenges, we introduce JointTuner, an adaptive joint training framework designed to concurrently learn both appearance and motion. Our approach introduces two key innovations: **(1) Gated Low-Rank Adaptation (GLoRA)**: To resolve conflicts between appearance and motion, we introduce GLoRA, an architecture that dynamically adjusts the activation of modality-specific expert modules. This is achieved through a gating regulator that modulates each LoRA based on the input modality. The gating regulator learns to activate the LoRAs focused on either appearance or motion during training, enabling dynamic composition of specialized modules at inference, suppressing concept interference, and allowing coherent joint modeling. **(2) Appearance-independent Temporal Loss (AiT Loss)**: As shown in Fig. 3, channel- and temporal-wise shifts disrupt appearance yet preserve motion, while spatial shifts mainly change color tones. Spectral analysis (Fig. 4) confirms that these shifts suppress low-frequency appearance signals and enhance high-frequency motion components [29, 55]. Based on this, AiT Loss injects channel–temporal shift noise during training, guiding the model to learn motion patterns independent of appearance and suppress contamination from the reference video's appearance. Consequently, as shown in Fig. 2(c), our method produces more complete motion and significantly reduces interference from reference video appearance.

While most existing appearance-motion customization methods [2, 45, 56] rely on curated datasets and appearance-centric metrics like CLIP-Image or Subject Consistency, they often neglect motion evaluation due to the lack of a unified benchmark. To bridge this gap and enable balanced assessment, we introduce a comprehensive benchmark comprising 90 subject-motion combinations from

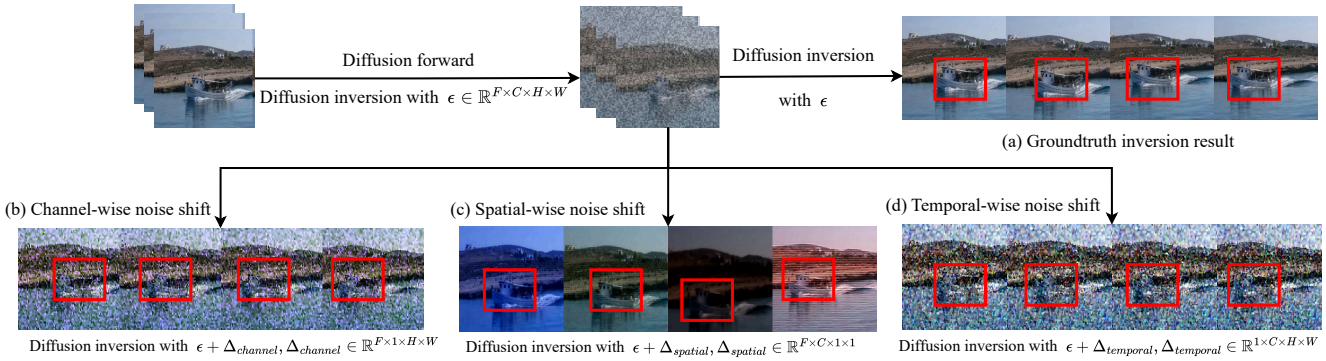


Figure 3. **Illustration of the impact of noise shift on diffusion inversion in latent space with CogVideoX [11].** Starting from a clean video, random noise $\epsilon \in \mathbb{R}^{F\times C\times H\times W}$ is added during the diffusion forward process, followed by diffusion inversion to recover the original video. Red bounding boxes highlight regions of interest.

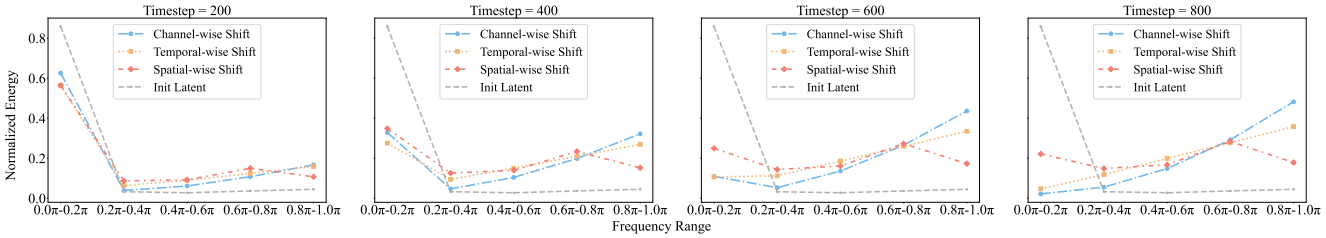


Figure 4. **Frequency distribution of latent and shifted latent signals across shift types and time steps.** These results show normalized energy for three shift types: Channel-wise, Spatial-wise, and Temporal-wise at time steps 200, 400, 600, and 800.

public datasets [21, 32, 35, 39, 56]. Our accompanying evaluation framework includes 10 metrics across four dimensions, semantic alignment, motion dynamism, temporal consistency, and perceptual quality, along with three composite scores, establishing a standardized protocol for appearance-motion combined customization.

Our contributions are as follows:

- We propose **JointTuner**, an adaptive joint training framework for appearance-motion combined customization, compatible with both UNet- and DiT-based architectures.
- We introduce **GLoRA**, a gating adaptation mechanism that reduces concept interference and enables seamless integration of appearance and motion representations.
- We find that injecting shift noise along channel and temporal dimensions suppresses appearance-related low-frequency components while enhancing motion-related high-frequency ones, leading to the design of **AiT Loss** to reduce appearance contamination.
- We establish a comprehensive evaluation framework for appearance-motion combined customization, featuring diverse subject-motion combinations and metrics assessing semantic alignment, motion fidelity, temporal coherence, and perceptual quality.

2. Related Work

2.1. Text-to-Video Generation

Text-to-video generation has progressed from early GAN-based [1, 42] and autoregressive models [8, 30] to diffusion-based methods [4, 17, 37], which now dominate in quality and controllability. Two diffusion architectures prevail. UNet-based models such as AnimateDiff [10] introduce motion modules into text-to-image pipelines; ModelScopeT2V [44], ZeroScope [40], and Stable Video Diffusion [3] enhance fidelity and realism via spatiotemporal attention and multi-stage training. Diffusion Transformer-based models like CogVideoX [11, 52], Wan [43], and HunyuanVideo [19] leverage hierarchical training, adaptive fusion, and scalable pretraining with 3D VAE designs to achieve high-resolution, long-range, and semantically aligned generation. These developments provide a foundation for customized video synthesis.

2.2. Customized Video Generation

Customized video generation involves appearance control, motion control, or both. Appearance methods preserve identity from a few images [28, 46, 49, 50], while motion methods transfer dynamics from reference videos [36, 48, 54]. Combined approaches often use LoRA or Adapters to separately tune spatial and temporal layers [34, 56], sometimes enhanced with text inversion [47]. However, stage-wise training suffers from concept interference. CustomTTT [2] uses test-time tuning to reduce domain shift but risks error accumulation. DualReal [45] jointly trains appearance and motion adapters in CogVideoX [11] via a controller that fuses text and visual features during training. In inference, only text is available, causing the controller to assign higher weights to appearance, which often

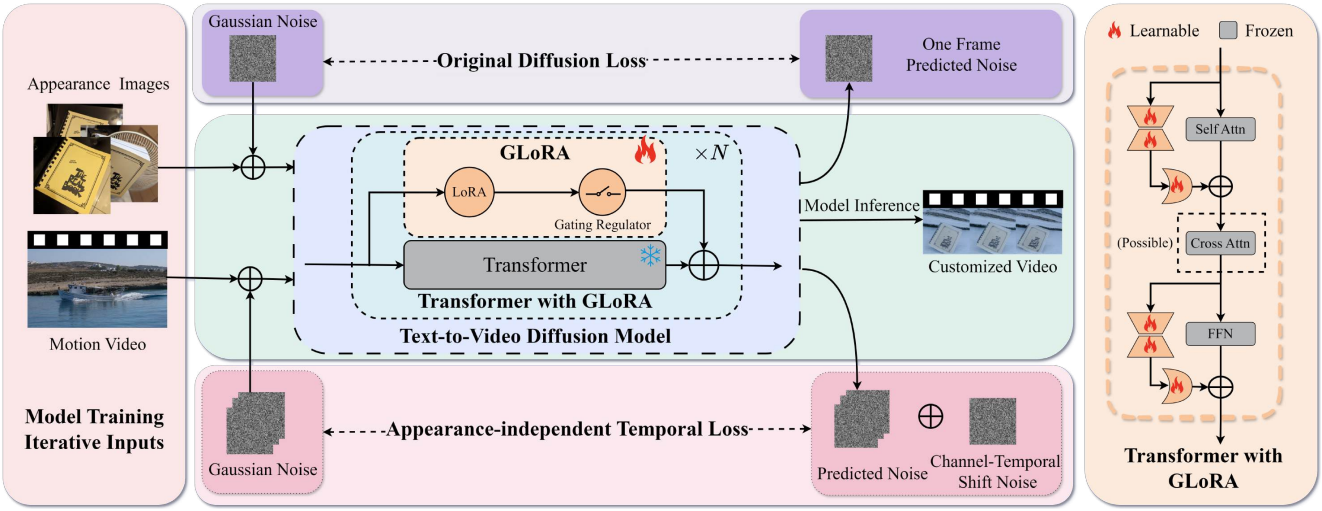


Figure 5. **Architecture of JointTuner, an adaptive joint training framework with two main steps**: (1) integrating GLoRA into the transformer blocks for efficient fine-tuning, and (2) optimizing GLoRA with two complementary losses. The original diffusion loss leverages reference images to preserve appearance details, and the AiT Loss utilizes reference videos to focus on motion patterns. The pre-trained text-to-video model remains frozen throughout training; only GLoRA parameters are updated. During inference, trained GLoRA weights are loaded, and customized videos are generated conditioned solely on the input prompt.

leads to weak motion customization. Moreover, to mitigate appearance leakage, some methods swap appearance components at inference [34, 56] or condition on an appearance image [47], yet robust disentanglement remains unresolved.

3. Method

In this section, we introduce the concepts of text-to-video diffusion models and low-rank adaptation, followed by the proposed JointTuner framework.

3.1. Preliminaries

Text-to-Video Diffusion Models Following previous customized video generation methods [47, 56], we adopt a diffusion-based framework where the denoising network ϵ_θ is trained to predict additive Gaussian noise from a noisy video latent z_t, conditioned on a text prompt y. We explore two architectural backbones: (1) *UNet Backbone.* A 3D UNet [6] adopts a U-shaped encoder and decoder structure that combines convolutional blocks and Transformer blocks. It integrates spatial Transformers for per-frame semantics and temporal Transformers for inter-frame dynamics. (2) *DiT Backbone.* Diffusion Transformer (DiT) [31] replaces all the convolutional layers with a fully attention-based design. It allows long-range dependency modeling, improves capacity on high-resolution content, and naturally extends to video by applying spatiotemporal attention. The network (ϵ_θ) is optimized with a text encoder (τ_θ) via:

$$\mathcal{L}_{z_0,y,\epsilon,t}\left[\epsilon, \epsilon_\theta(z_t, t, \tau_\theta(y))\right], \tag{1}$$

where $z_0 \in \mathbb{R}^{F\times C\times H\times W}$ denotes the latent code of the video, and F, C, H, W denote the number of frames, channels, height, and width of the latent code, respectively. y is the text prompt, $\epsilon \sim \mathcal{N}(0, I)$ represents Gaussian noise, and $t \sim \mathcal{U}(0, T)$ corresponds to the time steps. A typical $\mathcal{L}(\cdot)$ is Mean Squared Error (MSE) Loss, formally as follows:

$$\mathbb{E}_{z_0,y,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right]. \tag{2}$$

Low-Rank Adaptation (LoRA) LoRA [13] enables parameter-efficient adaptation by decomposing weight updates as $W = W_0 + BA$, where $W_0 \in \mathbb{R}^{d\times k}$ is frozen, and $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$ (with $r \ll d, k$) are trainable low-rank matrices. This design significantly reduces the number of trainable parameters while preserving model capacity. The forward pass is given by

$$x' = W_0 x + \alpha \cdot BAx, \tag{3}$$

where x is the input feature and x' is the adapted output feature, and α is a fixed scaling factor.

3.2. Gated Low-Rank Adaptation

Concept interference arises in stage-wise paradigms because fine-tuned parameters learned in separate stages are not effectively fused during inference. To resolve this, a unified joint training framework is needed. However, applying conventional LoRA to joint fine-tuning leads to sub-optimal performance due to conflicts between shared LoRA modules adapting to distinct modalities [47, 56].

To mitigate such conflicts, each LoRA can be trained as a modality-specific expert module, dedicated either to appearance or motion. We propose **Gated Low-Rank Adaptation (GLoRA)**, a novel architecture that introduces a context-aware linear activation layer, termed the gating regulator,

Models (on CogVideoX [12])	*Composite Metrics*			*Semantic Alignment*			*Motion Dynamism*		*Temporal Consistency*		*Perceptual Quality*		
	ARS ↓	NAS ↑	AAS ↑	CLIP-T ↑	CLIP-I ↑	Mot-Fid ↑	Mot-Smth ↑	Dyn-Deg ↑	Subj-Con ↑	Bkgd-Con ↑	Pick-Sc ↑	Aesth-Q ↑	Img-Q ↑
CogVideoX-5B [12]	3.10	91.82	65.70	33.03	61.18	34.76	98.52	93.81	93.22	95.60	20.94	62.32	63.59
+ LoRA [12]	3.70	86.76	61.46	31.39	65.93	39.74	98.78	48.57	94.70	95.97	20.25	58.32	60.90
DualReal-Cog5B [45]	1.90	92.04	64.95	33.36	71.30	42.75	98.30	56.67	95.70	95.74	20.97	63.63	71.10
JointTuner-Cog2B	2.90	93.28	66.35	31.32	67.71	70.16	98.98	64.05	95.41	95.17	20.53	59.58	60.65
JointTuner-Cog5B	3.40	94.66	67.75	31.16	68.55	62.15	97.73	80.22	94.80	95.09	20.35	59.33	68.11

Table 1. Quantitative comparison of DiT-based methods, built on CogVideoX. We highlight the best and second-best values for each metric.

Models (on CogVideoX [12])	Example: *A clock driving in the snow.*				
	AAS ↑	CLIP-T ↑	CLIP-I ↑	Mot-Fid ↑	Img-Q ↑
CogVideoX-5B [12]	68.52	33.87	58.87	34.11	75.61
+LoRA [12]	57.96	31.13	68.75	41.63	59.28
DualReal-Cog5B [45]	70.64	31.15	63.90	54.70	72.83
JointTuner-Cog2B	71.53	31.81	56.61	67.43	74.65
JointTuner-Cog5B	72.25	30.17	56.28	91.35	73.88

Table 2. Partial quantitative results corresponding to the customized video in Fig. 6. We highlight the best and second-best values for each metric.

to dynamically modulate the activation state of each LoRA based on input modality. The forward propagation formulation of GLoRA is defined as follows:

$$x' = W_0 x + \underbrace{\text{Mean}(\sigma(Gx))}_{\text{gating regulation}} \cdot BAx, \tag{4}$$

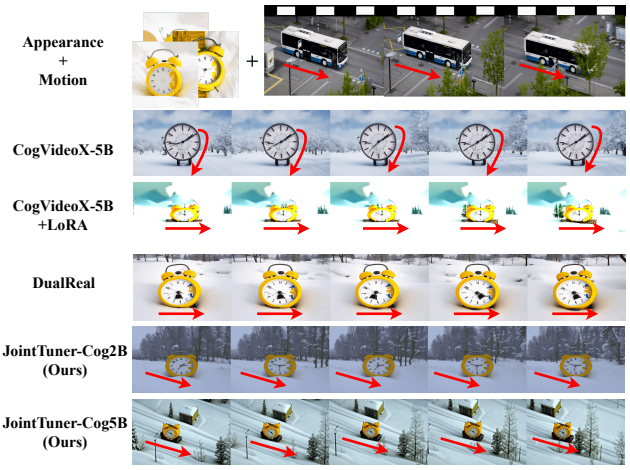
where G is a learnable projection, σ is the sigmoid function, and $\text{Mean}(\cdot)$ ensures stability across frames and batches.

The gating regulator receives the input to the associated layer, which carries rich spatial and temporal semantics, serving as a natural context signal. During training, by alternating between appearance-focused and motion-focused reconstruction tasks, the gating regulators learn to selectively activate or inhibit their corresponding LoRA according to whether they specialize in appearance or motion. At inference time, this gating mechanism enables dynamic composition of semantically specialized LoRAs, effectively suppressing concept interference and enabling coherent fusion.

3.3. Appearance-independent Temporal Loss

In customized video generation, motion reference videos often intertwine spatial appearance with temporal motion, leading to appearance leakage during fine-tuning. To address this, we propose a shift perturbation strategy that introduces structured noise along specific feature dimensions to disentangle motion and appearance representations. As illustrated in Fig. 3, we apply three types of perturbations: channel-wise $(F, C{=}1, H, W)$, spatial-wise $(F, C, H{=}1, W{=}1)$, and temporal-wise $(F{=}1, C, H, W)$. Channel-wise and temporal-wise shifts effectively disrupt spatial correlations while preserving coherent motion trajectories, whereas spatial-wise shifts mainly modify color tones and degrade motion consistency.

Given prior findings that motion is encoded in high-frequency components, while appearance resides in low-



Prompt: A clock gliding through a snowy landscape, its hands frozen in time as it appears to drive across a serene, wintery countryside.

Figure 6. Qualitative results of appearance-motion combined customization on CogVideoX.

frequency components [29, 55], we further conduct a spectral analysis to explain the phenomena. Specifically, we perform a 3D Fast Fourier Transform on the denoised latents and normalize the energy distribution across frequency bands. The quantitative results, presented in Fig. 4, show that channel-wise and temporal-wise shifts suppress low-frequency components and amplify high-frequency signals, confirming their effectiveness in isolating motion features. In contrast, spatial-wise shift loses effectiveness at higher timesteps, indicating limited ability to disentangle motion.

Building upon these insights, we propose the **Appearance-independent Temporal Loss (AiT Loss)** to enforce motion-focused learning. During fine-tuning, we inject shift noise $\Delta_{noise} \in \mathbb{R}^{1\times1\times H\times W}$ that remains constant across both temporal and channel dimensions into the predicted noise. This structured perturbation disrupts appearance-related cues such as texture, color, and identity, compelling the diffusion model to rely solely on motion dynamics for reconstruction. Consequently, AiT Loss enhances motion fidelity and mitigates appearance contamination during customized video generation. The AiT Loss is formally defined as follows:

$$\mathcal{L} z_0, y, \epsilon, t \left[\epsilon, (\epsilon_{\theta,\pi}(z_t, t, \tau_\theta(y)) + \Delta_{noise})\right], \tag{5}$$

where π denotes the injected GLoRA parameters.

Models (on ZeroScope [40])	*Composite Metrics*			*Semantic Alignment*			*Motion Dynamism*		*Temporal Consistency*		*Perceptual Quality*		
	ARS ↓	NAS ↑	AAS ↑	CLIP-T ↑	CLIP-I ↑	Mot-Fid ↑	Mot-Smth ↑	Dyn-Deg ↑	Subj-Con ↑	Bkgd-Con ↑	Pick-Sc ↑	Aesth-Q ↑	Img-Q ↑
MotionDirector-ZS [56]	2.00	92.78	62.71	28.56	70.73	70.12	96.66	34.22	94.12	94.49	20.37	53.27	64.62
DreamVideo-ZS [47]	2.50	92.27	62.22	27.39	51.70	72.60	96.94	61.56	89.92	94.10	20.01	48.66	59.33
JointTuner-ZS	1.50	96.14	64.15	32.14	62.44	72.61	97.78	46.89	91.92	93.83	21.19	57.14	65.53

Table 3. Quantitative comparison of UNet-based methods, built on ZeroScope. We highlight the best and second-best values for each metric.

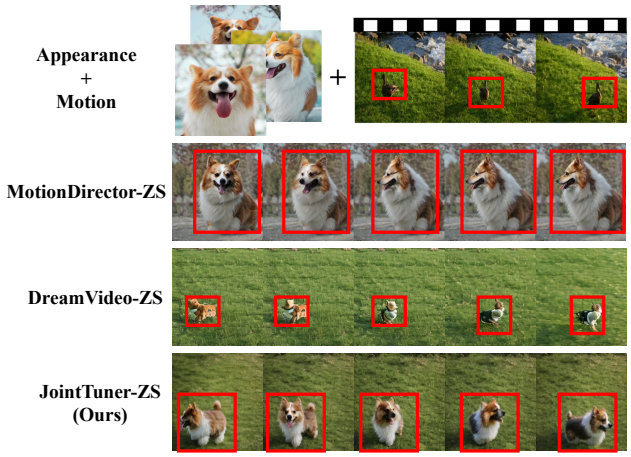


Figure 7. Qualitative results of appearance-motion combined customization on ZeroScope.

3.4. Training and Inference

The model is jointly trained with GLoRA by alternating between image and video inputs. Images are treated as single-frame videos, while videos are uniformly sampled into N frames. Image inputs use the original denoising loss (Eq. 1) to preserve appearance, whereas video inputs adopt the AiT Loss (Eq. 5) to focus on motion patterns and suppress appearance contamination.

During inference, the trained GLoRA parameters are integrated into the pre-trained text-to-video diffusion model, enabling the generation of customized videos conditioned solely on the textual prompt.

We provide the pseudocode for JointTuner and the stage-wise strategy approach in Appendix A, and compare the differences between their training and inference procedures.

4. Experiments

This section presents the proposed benchmark, comparisons with state-of-the-art methods, ablation studies, and failure cases. More results, including hyperparameter analysis and further visualization results, are provided in Appendix D.

4.1. Benchmark

Most existing appearance-motion customization methods [2, 45, 56] rely on researcher-curated datasets and emphasize appearance-oriented metrics (e.g., CLIP-Image, Subject Consistency), while often overlooking motion-specific evaluation. To address this, we construct a comprehensive benchmark. Below is a brief introduction of the benchmark, and more details can be found in Appendix B.

Dataset. We construct a comprehensive dataset of 90 appearance-motion combinations using public sources, comprising 15 subjects from CustomDiffusion [21], DreamBooth [35], and MotionDirector [56], along with nine motion types from DAVIS2016 [32] and three composite motions (each composed of five video clips) from UCF101 [39]. The dataset includes six distinct subject-motion pairing strategies and is evaluated under five specific background settings: grassland, jungle, snowscape, beach, and cobblestone street. This yields a benchmark of 450 customized videos covering diverse combinations of subjects, motions, and backgrounds.

Metrics. We evaluate generation quality across four dimensions: (1) *Semantic alignment* includes CLIP-Text (CLIP-T) for text-video alignment, CLIP-Image (CLIP-I) for image-video correspondence, and Motion-Fidelity (Mot-Fid), which measures motion pattern consistency using trajectory tracking from CoTracker3 [15]. (2) *Motion dynamism* is quantified by two metrics: Motion-Smoothness (Mot-Smth), which assesses temporal coherence via AMT-based interpolation priors, and Dynamic-Degree (Dyn-Deg), which measures motion intensity using optical flow magnitude from RAFT [41]. Both are implemented in VBench [14]. (3) *Temporal consistency* includes Subject-Consistency (Subj-Con) and Background-Consistency (Bkgd-Con), evaluated using DINO and CLIP feature similarity across frames. (4) *Perceptual quality* aggregates Pick-Score (Pick-Sc), Aesthetic-Quality (Aesth-Q), and Imaging-Quality (Img-Q) at the frame level. To holistically compare methods, we report three **composite metrics**: the Average Rank Score (ARS), which represents the mean rank of a model across all individual metrics where a lower rank indicates better performance; the Normalized Average Score (NAS), in which each metric is normalized so that the best-performing method receives a score of 100 and NAS is the average of these normalized values; and the Absolute Average Score (AAS), defined as the arithmetic mean of raw metric values.

4.2. Main Results

We compare JointTuner with state-of-the-art methods. For clarity, "-ZS", "-Cog2B", and "-Cog5B" indicate variants built on ZeroScope [40], CogVideoX-2B [12], and

Models (on CogVideoX [12])	*Composite Metrics*			*Semantic Alignment*			*Motion Dynamism*		*Temporal Consistency*		*Perceptual Quality*		
	ARS ↓	NAS ↑	AAS ↑	CLIP-T ↑	CLIP-I ↑	Mot-Fid ↑	Mot-Smth ↑	Dyn-Deg ↑	Subj-Con ↑	Bkgd-Con ↑	Pick-Sc ↑	Aesth-Q ↑	Img-Q ↑
CogVideoX-5B	2.30	93.49	65.70	33.03	61.18	34.76	98.52	93.81	93.22	95.60	20.94	62.32	63.59
Cog5B + LoRA	2.60	88.52	61.46	31.39	65.93	39.74	98.78	48.57	94.70	95.97	20.25	58.32	60.90
Cog5B + GLoRA	2.80	91.05	62.90	30.48	69.28	55.06	97.80	40.00	95.13	95.43	20.13	57.76	67.91
Cog5B + GLoRA + AiT (Ours)	2.30	96.89	67.75	31.16	68.55	62.15	97.73	80.22	94.80	95.09	20.35	59.33	68.11

Table 4. Quantitative ablation results for each core component of the proposed JointTuner, built on CogVideoX-5B. We highlight the best and second-best values for each metric.



Figure 8. Qualitative ablation results for each component of JointTuner based on CogVideoX-5B.

CogVideoX-5B [12], respectively. Implementation details of JointTuner and experimental settings for the comparison methods are provided in Appendix C.

Quantitative results. As shown in Tab. 1, CogVideoX-5B exhibits poor semantic alignment, and naive joint tuning via typical LoRA yields only marginal improvements. This indicates that simply injecting LoRA for joint fine-tuning also suffers from concept interference, leading to poor performance in combined customization. DualReal achieves good performance in appearance-related scores (high CLIP-I), but the controller that fuses appearance and motion suffers from a gap between training and inference, resulting in poor motion scores (low Mot-Fid and Dyn-Deg). In contrast, JointTuner-Cog2B achieves good performance on both CLIP-I and Mot-Fid scores, striking a better balance. JointTuner-Cog5B further improves CLIP-I and Img-Q, benefiting from the stronger backbone. On ZeroScope (Tab. 3), JointTuner leads in perceptual quality and consistently ranks within the top-2 across most metrics. While MotionDirector excels in static fidelity, it underperforms in motion capture; conversely, DreamVideo achieves high motion scores but suffers from degraded appearance. Additional experiments on ModelScope [44] corroborate these trends and are reported in Appendix D. This indicates that stage-wise paradigms, due to concept interference, fail to achieve strong performance in both appearance and motion.

Qualitative results. As illustrated in Fig. 6, CogVideoX-5B, without any fine-tuning, exhibits no customization in either appearance or motion. Simply injecting LoRA for joint fine-tuning suffers from concept interference, achieving only appearance customization. DualReal fails to generate the desired motion trajectories due to the gap between training and inference. In contrast, JointTuner, through a carefully designed joint training strategy, successfully achieves effective integration of both appearance and motion. Similarly, in Fig. 7, JointTuner produces semantically aligned outputs, whereas MotionDirector yields visually sharp but static frames, and DreamVideo captures dynamic motion at the cost of identity consistency.

Metrics analysis. To assess the effectiveness of our comprehensive evaluation metric, we conducted a case study using the example in Fig. 6, with quantitative results in Tab. 2. Based on prior appearance-related metrics, both direct LoRA injection with joint fine-tuning and DualReal achieve high scores. However, as shown in Fig. 6, neither method successfully generates the desired motion trajectories. Therefore, it is necessary to introduce motion-related metrics to quantitatively demonstrate this limitation. Furthermore, visual results show that our JointTuner effectively integrates appearance and motion customization and achieves the best overall evaluation scores, reflecting the consistency between our proposed comprehensive evaluation metrics and the qualitative results.

4.3. Ablation Studies

We perform ablation studies on two video generation backbones, CogVideoX-5B (DiT-based) and ZeroScope (UNet-based). Additional results, including the effect of the gating regulator in GLoRA and qualitative and quantitative analyses based on ZeroScope, are provided in Appendix D.

Quantitative ablations on CogVideoX-5B. We conduct quantitative ablation experiments on CogVideoX-5B to assess the impact of GLoRA and AiT Loss within the DiT backbone. As shown in Tab. 4, integrating GLoRA leads to consistent improvements over standard LoRA across all key metrics, including NAS, AAS, CLIP-I, and Mot-Fid. The gains are particularly pronounced in appearance consistency and motion realism, demonstrating the effectiveness of the gating regulator. Furthermore, when AiT Loss is incorporated during training, we observe additional performance gains, especially in text-to-video alignment and video quality, reflected in higher CLIP-I and improved perceptual quality metrics. These results confirm that AiT Loss effectively mitigates appearance contamination while preserving target visual semantics.

Models (on ZeroScope [40])	*Composite Metrics*			*Semantic Alignment*			*Motion Dynamism*		*Temporal Consistency*		*Perceptual Quality*		
	ARS ↓	NAS ↑	AAS ↑	CLIP-T ↑	CLIP-I ↑	Mot-Fid ↑	Mot-Smth ↑	Dyn-Deg ↑	Subj-Con ↑	Bkgd-Con ↑	Pick-Sc ↑	Aesth-Q ↑	Img-Q ↑
Spatial Shift	2.40	93.28	60.69	31.13	59.80	59.50	98.06	34.89	92.23	94.17	20.80	52.82	63.54
Full Shift	2.10	99.11	64.06	31.19	62.07	75.11	97.54	47.56	91.70	93.54	21.05	56.45	64.43
Channel-Temp Shift (Ours)	1.50	99.43	64.15	32.14	62.44	72.61	97.78	46.89	91.92	93.83	21.19	57.14	65.53

Table 5. Quantitative results of different noise shift designs on ZeroScope. We highlight the best and second-best values for each metric.

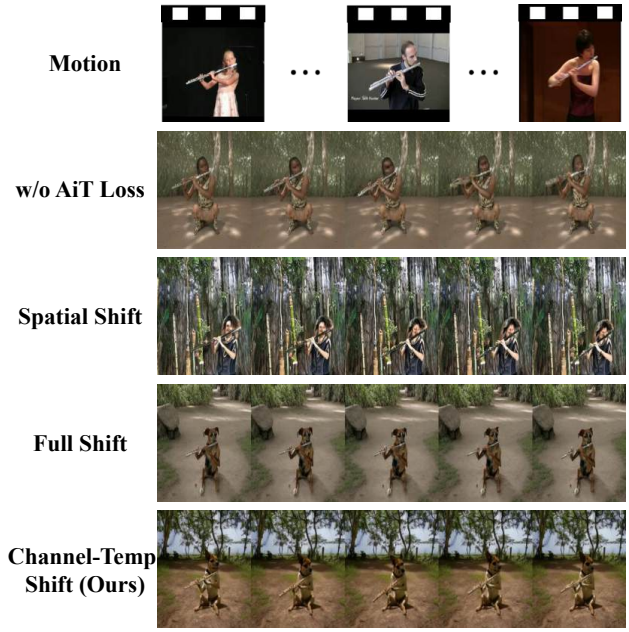


Figure 9. Motion training outcomes using three noise shift strategies in the AiT Loss based on ZeroScope.

Qualitative ablations on CogVideoX-5B. Illustrated in Fig. 8, models equipped with GLoRA generate videos with significantly enhanced appearance and motion fidelity. During training, the model learns to dynamically activate appearance- or motion-specialized LoRA modules based on input semantics; at inference time, this enables coherent fusion of both features, preserving subject identity while producing natural motion. Moreover, the inclusion of AiT Loss further refines output quality, particularly in background rendering and adherence to textual prompts, without compromising motion dynamics.

Impact of noise shift designs. We compare three noise shift strategies for AiT Loss in ZeroScope-based implementation: Spatial Shift $(F, C, H{=}1, W{=}1)$, Full Shift (F, C, H, W), and Channel-Temp Shift $(F{=}1, C{=}1, H, W)$. Tab. 5 shows that Channel-Temp Shift achieves the best performance in semantic alignment and perceptual quality. Visuals in Fig. 9 confirm it effectively suppresses appearance interference while preserving motion dynamics. In contrast, Spatial Shift suffers from severe appearance contamination, leading to distorted subject identities. Full Shift mitigates this issue to some extent, but still struggles with degraded background quality.

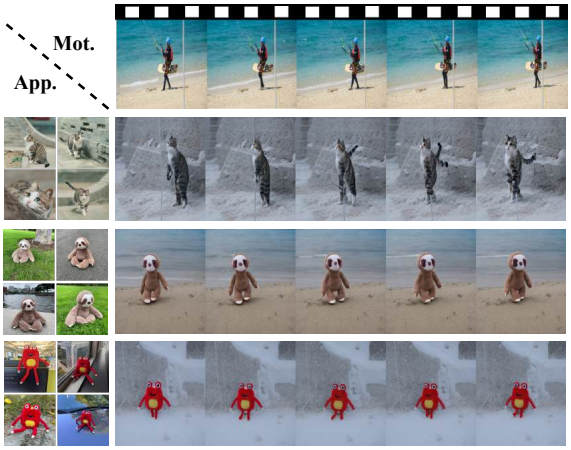


Figure 10. Failure cases of our JointTuner based on ZeroScope, illustrating the results on different subjects: a cat, a sloth plushie, and a monster toy walking.

4.4. Failure Cases and Limitations

Despite the overall strong performance, our JointTuner may fail under extreme views. As illustrated in Fig. 10, the system occasionally encounters issues when the reference and target viewpoints differ significantly. For example, the side-facing cat aligns well with motion, while the sloth plushie and monster toy fail to maintain structural integrity due to the lack of side-view references. These cases highlight the limitations of appearance-motion composition when training data does not sufficiently cover certain viewpoints.

5. Conclusion

In this work, we propose JointTuner, an adaptive joint training framework for appearance-motion combined customization. By integrating Gated LoRA and AiT Loss, JointTuner effectively mitigates concept interference that arises in stage-wise tuning and alleviates appearance contamination from reference videos. It is compatible with both UNet and DiT architectures, and demonstrates a strong ability to balance alignment between appearance and motion. To enable fair comparisons, we also introduce a comprehensive evaluation benchmark for combined customization. The benchmark includes 90 appearance-motion combinations and uses 10 automatic metrics in four dimensions, providing a foundation for assessing combined customization. Future work will explore strategies to augment reference priors and integrate 3D-aware representations to further improve robustness and generalization in appearance-motion combined customization.

References

[1] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1995–2001, 2019. 3

[2] Xiuli Bi, Jian Lu, Bo Liu, Xiaodong Cun, Yong Zhang, Weisheng Li, and Bin Xiao. Customttt: Motion and appearance customized video generation via test-time training. *arXiv preprint arXiv:2412.15646*, 2024. 2, 3, 6

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 4

[7] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3677–3686, 2020. 4

[8] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022. 3

[9] Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7621–7630, 2024. 1

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024. 3

[11] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2, 3, 4

[12] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. 2022. 5, 6, 7

[13] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. 2021. 4

[14] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 6, 3, 4

[15] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proc. arXiv:2410.11831*, 2024. 6, 3

[16] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 4

[17] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 3

[18] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 36652–36663, 2023. 4

[19] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3

[20] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 1

[21] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. 2023. 3, 6, 2

[22] Gihyun Kwon and Jong Chul Ye. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation. *arXiv preprint arXiv:2410.05591*, 2024. 1

[23] LAION-AI. aesthetic-predictor, 2022. https://github.com/LAION-AI/aesthetic-predictor. 4

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[25] Xiaomin Li, Xu Jia, Qinghe Wang, Haiwen Diao, Mengmeng Ge, Pengxiang Li, You He, and Huchuan Lu. Motrans:

Customized motion transfer with text-driven video diffusion models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3421–3430, 2024. 1

[26] Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023. 3

[27] Huijie Liu, Jingyun Wang, Shuai Ma, Jie Hu, Xiaoming Wei, and Guoliang Kang. Separate motion from appearance: Customizing motion via customizing text-to-video diffusion models. *arXiv preprint arXiv:2501.16714*, 2025. 1

[28] Jia-Wei Liu, Yan-Pei Cao, Jay Zhangjie Wu, Weijia Mao, Yuchao Gu, Rui Zhao, Jussi Keppo, Ying Shan, and Mike Zheng Shou. Dynvideo-e: Harnessing dynamic nerf for large-scale motion-and view-change human-centric video editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7664–7674, 2024. 1, 3

[29] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *Advances in Neural Information Processing Systems*, 37:131434–131455, 2024. 2, 5

[30] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Ccvs: context-aware controllable video synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 14042–14055, 2021. 3

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 4

[32] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Computer Vision and Pattern Recognition*, 2016. 3, 6, 2

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4

[34] Yixuan Ren, Yang Zhou, Jimei Yang, Jing Shi, Difan Liu, Feng Liu, Mingi Kwon, and Abhinav Shrivastava. Customize-a-video: One-shot motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 332–349. Springer, 2024. 1, 3, 4

[35] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3, 6, 2

[36] Qingyu Shi, Jianzong Wu, Jinbin Bai, Jiangning Zhang, Lu Qi, Yunhai Tong, and Xiangtai Li. Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10995–11005, 2025. 3

[37] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. 2022. 3

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[39] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3, 6, 2

[40] Spencer Sterling. Zeroscope, 2023. https://huggingface.co/cerspense/zeroscope_v2_576w. 2, 3, 6, 8, 4

[41] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 6, 3

[42] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1526–1535, 2018. 3

[43] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3

[44] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 3, 7, 4

[45] Wenchuan Wang, Mengqi Huang, Yijing Tu, and Zhendong Mao. Dualreal: Adaptive joint training for lossless identity-motion fusion in video customization. *arXiv preprint arXiv:2505.02192*, 2025. 2, 3, 5, 6, 4

[46] Zhao Wang, Aoxue Li, Lingting Zhu, Yong Guo, Qi Dou, and Zhenguo Li. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*, 2024. 1, 3

[47] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024. 1, 2, 3, 4, 6

[48] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7589–7599. IEEE Computer Society, 2023. 1, 3

[49] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. *arXiv preprint arXiv:2408.13239*, 2024. 1, 3

[50] Xuancheng Xu, Ming Tao, and Bing-Kun Bao. Clgc: Continuous layout guidance for consistent text-to-video editing. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2025. 3

[51] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 3

[52] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[53] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8476, 2024. 3

[54] Hidir Yesiltepe, Tuna Han Salih Meral, Connor Dunlop, and Pinar Yanardag. Motionshop: Zero-shot motion transfer in video diffusion models with mixture of score guidance. *arXiv preprint arXiv:2412.05355*, 2024. 1, 3

[55] Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers. *arXiv preprint arXiv:2502.15894*, 2025. 2, 5

[56] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290. Springer, 2024. 1, 2, 3, 4, 6

JointTuner: Appearance-Motion Adaptive Joint Training for Customized Video Generation

Supplementary Material

This supplementary material is organized into the following sections:

- **Algorithmic Comparisons**: Presents pseudocode and detailed training and inference procedures for JointTuner and the stage-wise training paradigm, highlighting differences in parameter handling and optimization flow.
- **Benchmark Details**: Describes the construction of our proposed benchmark, including the selection of subject and motion, data preprocessing strategies for different model architectures, and the design of fine-grained evaluation metrics across multiple quality dimensions.
- **Experiment Settings**: Provides hyperparameters, training configurations, and hardware settings for JointTuner and compared methods.
- **Extended Analysis and Additional Results**: Includes extended quantitative and qualitative results, ablation studies, analysis of the gating regulator in our GLoRA, sensitivity to key hyperparameters, and additional customized videos that demonstrate the appearance-motion combined customization capabilities of JointTuner in comparison with state-of-the-art approaches.

A. Algorithmic Comparisons

We present the detailed procedures of JointTuner and the stage-wise training paradigm in Algorithms 1 and 2, respectively. The stage-wise approach consists of two distinct training phases: first, spatial LoRA parameters are optimized on image data to adapt the model's appearance representation; subsequently, temporal LoRA parameters are trained separately on video data to refine motion dynamics. During inference, both sets of parameters are applied concurrently. In contrast, JointTuner trains on mixed batches comprising both images and videos, employing modality-specific loss functions while updating a unified set of GLoRA parameters throughout the training process.

A fundamental limitation of the stage-wise strategy lies in its decoupling of appearance and motion optimization into disjoint phases. Because spatial and temporal LoRA parameters are learned independently, their respective adaptations may become misaligned or even contradictory. When jointly applied during inference, such incompatibilities can lead to concept interference. For instance, the appearance-adapted module may emphasize fine-grained spatial details, whereas the motion-adapted module may inadvertently modify spatial structures to satisfy temporal consistency. This conceptual conflict often compromises subject fidelity, resulting in incomplete or inconsistent motion patterns.

JointTuner addresses this challenge through a joint op-

Algorithm 1: JointTuner

Input : Training data (images and videos), pretrained model weights
Output: Trained model, generated video

Initialize models: vae, τ_θ, ϵ_θ;
Load pretrained model weights;
Add GLoRA parameters π to ϵ_θ to obtain $\epsilon_{\theta,\pi}$;
Adaptive Joint Training:
for *each epoch* **do**
  for *each batch in data loader* **do**
    Encode batch into latent space z_0 using vae;
    Sample noise ϵ and add to z_0 to obtain z_t;
    Encode text embeddings $\tau_\theta(y)$;
    if *batch is image* **then**
      Compute original diffusion loss with predicted noise:

$$\mathcal{L}_{z_0,y,\epsilon,t}\left[\epsilon, \epsilon_{\theta,\pi}(z_t, t, \tau_\theta(y))\right];$$

    else if *batch is video* **then**
      Compute AiT loss with predicted noise:

$$\mathcal{L}_{z_0,y,\epsilon,t}\left[\epsilon, (\epsilon_{\theta,\pi}(z_t, t, \tau_\theta(y)) + \Delta_{noise})\right];$$

    Update GLoRA parameters π;
Save trained GLoRA checkpoints;
Inference:
Initialize pipeline with pretrained models;
Load trained GLoRA parameters π;
for *each prompt* **do**
  Prepare input latents (random noise) $\hat{z}_t$;
  Perform the denoising process step by step, ultimately yielding the denoised latent $\hat{z}_0$:

$$\hat{z}_{t-1} = \hat{z}_t - \hat{\epsilon}_{\theta,\pi}(\hat{z}_t, t, \tau_\theta(y));$$

  Decode denoised latent $\hat{z}_0$ into video using vae;

Algorithm 2: Stage-wise Training Method

Input : Training data (images and videos), pretrained model weights
Output: Trained model, generated video

Initialize models: vae, τ_θ, ϵ_θ;
Load pretrained model weights;
Stage I – Appearance Training:
Add finetuning parameters π_s to ϵ_θ to obtain ϵ_{θ,π_s};
for *each epoch* **do**
  for *each image in image data loader* **do**
    Encode image to latent space z_0 using vae;
    Sample noise ϵ and add to z_0 to obtain z_t;
    Encode text embeddings $\tau_\theta(y)$;
    Compute diffusion loss with predicted noise:

$$\mathcal{L}_{z_0,y,\epsilon_i,t}\left[\epsilon, \epsilon_{\theta,\pi_s}(z_t, t, \tau_\theta(y))\right];$$

    Update spatial tuning parameters π_s;
Save trained spatial tuning checkpoints.
Stage II – Motion Training:
Add finetuning parameters π_t to ϵ_θ to obtain ϵ_{θ,π_t};
for *each epoch* **do**
  for *each video in video data loader* **do**
    Encode video to latent space z_0 using vae;
    Sample noise ϵ and add to z_0 to obtain z_t;
    Encode text embeddings $\tau_\theta(y)$;
    Compute diffusion loss with predicted noise:

$$\mathcal{L}_{z_0,y,\epsilon_v,t}\left[\epsilon, \epsilon_{\theta,\pi_t}(z_t, t, \tau_\theta(y))\right];$$

    Update temporal tuning parameters π_t;
Save trained temporal tuning checkpoints;
Combined Inference:
Initialize pipeline with pretrained models;
Load trained spatial finetuning parameters π_s;
Load trained temporal finetuning parameters π_t;
for *each prompt* **do**
  Prepare input latents (random noise) $\hat{z}_t$;
  Perform the denoising process step by step, ultimately yielding the denoised latent $\hat{z}_0$:

$$\hat{z}_{t-1} = \hat{z}_t - \hat{\epsilon}_{\theta,\pi_s,\pi_t}(\hat{z}_t, t, \tau_\theta(y));$$

  Decode denoised latent z_0 into video using vae;

timization framework that trains a single set of parameters under simultaneous supervision from both appearance (derived from static images) and motion (extracted from videos). By alternately optimizing appearance-focused and motion-focused reconstruction objectives, the gating regulators within GLoRA learn to selectively activate or suppress their associated modules based on whether they specialize in modeling appearance or motion. During inference, this gating mechanism facilitates the dynamic composition of specialized GLoRAs, enabling adaptive integration of appearance and motion features. This joint training paradigm effectively mitigates inter-parameter conflicts and ensures coherent alignment between the customized subject identity and its intended motion, thereby producing more reliable and temporally consistent video generation.

B. Benchmark Details

B.1. Datasets

We construct a comprehensive dataset comprising 90 appearance-motion combinations using public sources [21, 32, 35, 39, 56]. As summarized in Tab.A1, the dataset integrates 15 subjects from CustomDiffusion [21], DreamBooth [35], and MotionDirector [56], along with nine motion types from DAVIS2016 [32] and three composite motions (each comprising 5 video clips) from UCF101 [39]. We design six distinct subject-motion pairing configurations, as shown in Tab. A2. To further assess generalization, we adopt five diverse background settings, including natural scenes (grassland, jungle, snowscape, and beach) and an urban environment (cobblestone street). This results in a comprehensive benchmark dataset of 450 customized videos, covering diverse combinations of subjects, motions, and backgrounds.

To standardize the motion input format for different backbones, we employ two preprocessing strategies: a 16-frame sequence for UNet-based methods and a 49-frame sequence for DiT-based methods. For UNet-based methods pre-trained on short video segments, each video is first processed to extract a 16-frame sequence. Specifically, for the DAVIS2016 Dataset [32], frames are sampled every two frames, whereas for the UCF101 Dataset [39], frames are sampled every six frames. This results in 16-frame clips with a frame rate of 8 FPS and a duration of 2 seconds, while maintaining the original video resolution. In addition to the 16-frame setting, we create a 49-frame input setting by directly extracting the first 49 frames from each original video, without temporal downsampling. These longer clips are intended for DiT-based models, such as CogVideoX [11], which are pre-trained on longer video sequences and benefit from the extended temporal context. All 49-frame videos are preprocessed to a fixed resolution of 720×480.

Category	Description	Samples
Rigid Subject	Rigid objects with no articulation or deformable parts.	guitar (CustomDiffusion/instrument_music2), book (CustomDiffusion/things_book), car (CustomDiffusion/transport_car6), backpack (DreamBooth/backpack_dog), clock (DreamBooth/clock)
Non-Humanoid Subject	Subjects lacking human-like limb structure.	cat (CustomDiffusion/pet_cat5), wolf_plushie (DreamBooth/wolf_plushie), tortoise_plushie (CustomDiffusion/plushie_tortoise), unicorn_toy (CustomDiffusion/toy_unicorn), dog (DreamBooth/dog)
Humanoid Subject	Subjects with articulated limbs capable of standing posture.	pink_plushie (CustomDiffusion/plushie_pink), bear_plushie (DreamBooth/bear_plushie), sloth_plushie (DreamBooth/grey_sloth_plushie), monster_toy (DreamBooth/monster_toy), terracotta_warrior (MotionDirector/Terracotta_Warrior)
Rigid Motion	Global movement of rigid objects without internal deformation.	boat_sailing (DAVIS2016/boat), bus_traveling (DAVIS2016/bus), train_turning (DAVIS2016/train)
Non-Human Motion	Motion patterns typical of non-human entities.	bear_walking (DAVIS2016/bear), duck_walking (DAVIS2016/mallard-fly), dog_walking (DAVIS2016/dog)
Human Motion (Without Props)	Body-only motion without interaction with external objects.	person_dancing (DAVIS2016/breakdance-flare), person_twirling (DAVIS2016/dance-twirl), person_walking (DAVIS2016/kite-walk)
Human Motion (With Props)	Motion involving interaction with props (5 clips per motion).	person_lifting_barbell (UCF101/CleanAndJerk), person_playing_cello (UCF101/PlayingCello), person_playing_flute (UCF101/PlayingFlute)

Table A1. Overview of the subject and motion categories in our benchmark dataset. Samples are annotated with their instance names and respective dataset/model origins.

Subject Category	Compatible Motion Categories
Rigid Subject	Rigid Motion
Non-Humanoid Subject	Non-Human Motion Human Motion (Without Props) Human Motion (With Props)
Humanoid Subject	Human Motion (Without Props) Human Motion (With Props)

Table A2. Allowed subject-motion category combinations used for generating customized videos.

Furthermore, we annotate each video clip using BLIP-2 [24] to generate descriptive captions that serve as textual prompts during training. To further improve the quality of prompts for DiT-based methods, we employ Qwen3 [51] to enrich the captions. For example, a simple caption such as "*A bear plushie playing the flute on the grass*" is expanded into a richer and more vivid prompt: "*A bear plushie sitting on a lush green meadow, serenely playing a wooden flute under a gentle, golden sunset*".

B.2. Metrics

We establish a comprehensive evaluation framework across four dimensions: semantic alignment, motion dynamism, temporal consistency, and perceptual quality. For overall assessment, we report three composite metrics. The specific metrics are described as follows.

Composite Metrics

- **Average Rank Score (ARS):** Mean model ranking score across all ten metrics in the same table.
- **Normalized Average Score (NAS):** Scores are normalized against the top-performing method (scaled to 100) within a single table, and then averaged across all evaluation metrics for each model's performance.
- **Absolute Average Score (AAS):** Arithmetic mean of raw metric scores.

Semantic Alignment

- **CLIP-Text (CLIP-T):** Measures text-video alignment via frame-wise cosine similarity between CLIP [33] embeddings of input prompts and generated frames.
- **CLIP-Image (CLIP-I):** Evaluates visual-semantic correspondence using CLIP [33] image encoder to compare reference images and generated frames.
- **Motion-Fidelity (Mot-Fid):** Assesses motion pattern consistency using CoTracker3 [15] from diffusion-motion-transfer [53].

Motion Dynamism

- **Motion-Smoothness (Mot-Smth):** Evaluates temporal coherence using video interpolation priors from AMT [26] via VBench [14].
- **Dynamic-Degree (Dyn-Deg):** Quantifies motion intensity using optical flow estimation via RAFT [41] within VBench [14].

Temporal Consistency

- **Subject-Consistency (Subj-Con):** Tracks object persis-

Models (on ModelScope [44])	*Composite Metrics*			*Semantic Alignment*			*Motion Dynamism*		*Temporal Consistency*		*Perceptual Quality*		
	ARS ↓	NAS ↑	AAS ↑	CLIP-T ↑	CLIP-I ↑	Mot-Fid ↑	Mot-Smth ↑	Dyn-Deg ↑	Subj-Con ↑	Bkgd-Con ↑	Pick-Sc ↑	Aesth-Q ↑	Img-Q ↑
MotionDirector-MS	2.20	92.98	61.81	28.91	68.00	67.42	96.01	37.56	92.40	93.41	20.34	51.04	63.05
DreamVideo-MS	2.30	93.76	62.31	27.69	54.44	78.44	96.16	51.11	91.22	94.85	20.10	48.69	60.41
JointTuner-MS (Ours)	1.50	97.42	63.88	32.08	63.31	69.65	97.50	48.44	91.47	93.43	21.10	56.50	65.33

Table A3. Quantitative comparison of UNet-based methods, built on ModelScope. We highlight the best and second-best values for each metric.

tence using DINO [5] feature similarity via VBench [14].
- **Background-Consistency (Bkgd-Con):** Evaluates background stability using CLIP [33] feature similarity across frames via VBench [14].

Perceptual Quality
- **Pick-Score (Pick-Sc):** Predicts human preference scores using PickScore [18] with frame-level averaging.
- **Aesthetic-Quality (Aesth-Q):** Measures artistic merit using LAION aesthetic predictor [23] via VBench [14].
- **Imaging-Quality (Img-Q):** Evaluates technical quality using MUSIQ [16] trained on SPAQ [7] via VBench [14].

C. Experiment Settings

C.1. Implementation Details

Our JointTuner is compatible with both UNet-based and DiT-based diffusion models. The implementation details are provided below.
- **UNet-based JointTuner.** The joint training framework uses GLoRA parameters in transformers of UNet, including ZeroScope [40] and ModelScope [44], trained for 2000 iterations with a learning rate of 5×10^{-4}. Training requires 17GB of GPU memory and takes approximately 30 minutes on an RTX 4090. Inference utilizes DDIM [38] with 30 sampling steps to generate 16-frame videos at 8 FPS with a resolution of 384×384.
- **DiT-based JointTuner.** The joint training framework uses GLoRA parameters in transformers of DiT, including CogVideoX [11], trained for 1000 iterations with a learning rate of 1×10^{-4}. Training requires 20GB of GPU memory for CogVideoX-2B and 38GB for CogVideoX-5B. The training duration is approximately 60 minutes for a task involving five images and a single clip video, and around 120 minutes for a task with five images and five clip videos, based on the A100 (40 GB). Inference is performed with 50 sampling steps to generate 49-frame videos at 8 FPS and a resolution of 720×480.

C.2. Comparison Methods

We compare JointTuner with several open-source methods under consistent evaluation settings without external controls. For stage-wise training, we use ZeroScope [40] and ModelScope [44] as base models and compare against **MotionDirector** [56] and **DreamVideo** [47]. For joint training, CogVideoX [11] serves as the baseline. We further compare JointTuner (implemented on both CogVideoX-2B and CogVideoX-5B) with **CogVideo-5B**, its **LoRA** variant, and **DualReal** [45]. Below is an overview of the state-of-the-art methods along with the experimental setup.
- **MotionDirector [56]**: MotionDirector decouples appearance and motion learning via dual-path LoRA modules and introduces an appearance-debiased temporal loss to improve motion generalization across diverse subjects. We use the settings from the official MotionDirector implementation. For appearance learning, the spatial LoRA is trained for 1,000 iterations at a learning rate of 5.0×10^{-4}. For motion learning, the temporal LoRA is trained for 3,000 iterations at 5.0×10^{-4} for multi-reference motions and 300 iterations for single-reference motions. During inference, 16-frame videos at 8 FPS are generated using DDIM with 30 sampling steps.
- **DreamVideo [47]**: DreamVideo separates training into subject and motion stages, using identity and motion adapters with appearance-guided motion learning to enable flexible subject-motion composition. We follow the settings from the official DreamVideo implementation. For appearance learning, the text identity is optimized for 3,000 iterations at 1.0×10^{-4}, and the identity adapter is trained for 1,000 iterations at 1.0×10^{-5}. For motion learning, the motion adapter is trained for 1,000 iterations at 1.0×10^{-5} for multi-reference motions and 3,000 iterations for single-reference motions. During inference, 16-frame videos at 8 FPS are generated using DDIM with 50 sampling steps and classifier-free guidance.
- **DualReal [45]**: DualReal employs joint training with alternating optimization phases and a controller to achieve lossless fusion of identity and motion without cross-dimensional conflict. We follow the settings from the official implementation of DualReal. DualReal is trained for 1000 iterations with a learning rate of 1×10^{-3} based on CogVideoX-5B [11]. During inference, 49-frame videos at 8 FPS are generated with 50 sampling steps.

D. Extended Analysis and Additional Results

D.1. Overall comparison on ModelScope

We conducted experiments on another UNet-based model, ModelScope [44], and the results are shown in Tab. A3. JointTuner leads in perceptual quality and consistently ranks in the top-2 across most metrics. While MotionDirector excels in static fidelity (high CLIP-I, Subj-Con), it falls short in motion capture. On the other hand, DreamVideo

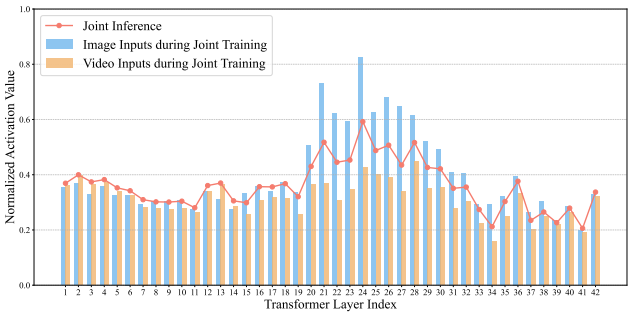
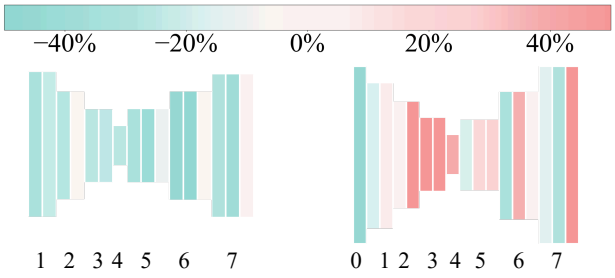


Figure A1. Activation values of gating regulators across different Transformer layers in CogVideoX-5B during both training and inference.



(a) Spatial Transformer Layers (b) Temporal Transformer Layers

Figure A2. Changes in the values of gating regulators across different Transformer layers when transitioning from image input to video input, as observed in ZeroScope.

captures motion well (high CLIP-T, Mot-Fid) but suffers from degradation in appearance. In contrast, JointTuner generates videos with better overall appearance and motion.

D.2. Analysis of Gated Modulation in GLoRA

This section presents a comprehensive analysis of how GLoRA's gated modulation mechanism enables adaptive feature learning during joint training and feature fusion during joint inference. We select 30 appearance-motion pairs for the experiment.

Gated Regulator Analysis on CogVideoX-5B. We record the activation magnitudes of GLoRA across all Transformer layers under three settings: (i) joint training with image inputs, (ii) joint training with video inputs, and (iii) joint inference. The layer-wise activation patterns reveal a clear modality-driven specialization. Specifically, in the early layers (1–14), video inputs mostly elicit stronger activations than image inputs, suggesting a heightened sensitivity to motion-related features. In contrast, in the deeper layers (15–42), image inputs dominate, reflecting a greater emphasis on static appearance features. This stratified behavior indicates that distinct GLoRA modules naturally specialize according to input modality during joint training. Crucially, during inference, the gating mechanism dynamically refines this specialization in a context-aware manner. In motion-sensitive early layers, the unified activation values often exceed those observed during either image-only or

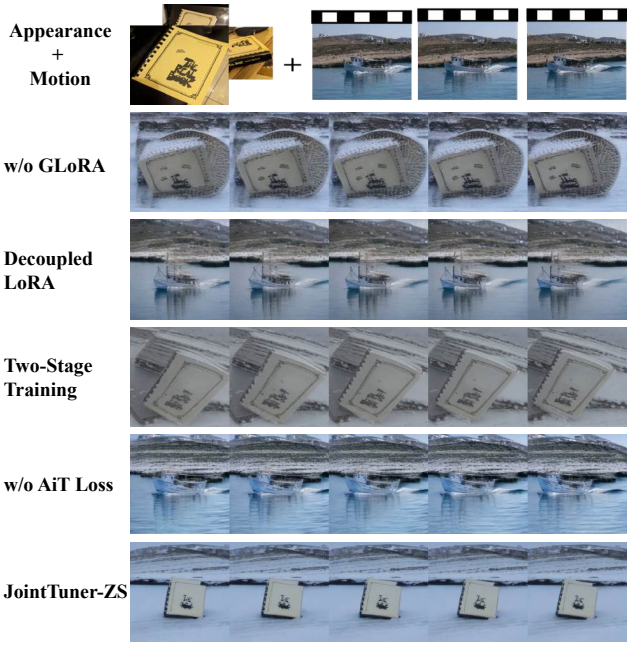


Figure A3. Qualitative ablation results for each component of JointTuner based on ZeroScope.

video-only training, implying an intentional amplification of motion signals at test time. Conversely, in appearance-biased later layers, inference activations typically lie between the image- and video-only training values, frequently approximating their average, suggesting a deliberate attenuation of overly dominant appearance responses. Through these complementary adjustments, GLoRA achieves a balanced and adaptive composition of modality-specific experts, suppressing potentially conflicting ones, thereby enabling coherent joint modeling of appearance and motion and effectively mitigating concept interference.

Gated Regulator Analysis on ZeroScope. Leveraging the UNet architecture's explicit separation of spatial and temporal Transformer blocks, we directly monitor GLoRA activation patterns when switching the input modality from images to videos during joint training. As shown in Fig. A2, the activations in spatial Transformer layers drop significantly upon transitioning to video inputs, while those in temporal Transformer layers rise markedly. This shift confirms that GLoRA employs learnable modules that dynamically modulate the contributions of spatial and temporal LoRA pathways based on the input modality.

D.3. Ablation Studies on ZeroScope

We further conduct ablation studies on the UNet-based model ZeroScope, evaluating four key variants: **(i) w/o GLoRA**, which employs standard LoRA without gating regulation; **(ii) Decoupled LoRA**, which alternately

Models (on ZeroScope [40])	*Composite Metrics*			*Semantic Alignment*			*Motion Dynamism*		*Temporal Consistency*		*Perceptual Quality*		
	ARS ↓	NAS ↑	AAS ↑	CLIP-T ↑	CLIP-I ↑	Mot-Fid ↑	Mot-Smth ↑	Dyn-Deg ↑	Subj-Con ↑	Bkgd-Con ↑	Pick-Sc ↑	Aesth-Q ↑	Img-Q ↑
w/o GLoRA	3.60	94.26	62.49	28.94	60.18	72.03	96.20	46.89	91.14	93.31	20.46	52.23	63.54
Decoupled LoRA	3.20	95.20	63.05	28.95	59.89	73.83	96.01	48.89	91.21	93.21	20.35	54.62	63.59
Two-Stage Training	3.50	92.94	62.26	28.37	69.08	73.24	95.54	34.67	93.11	94.01	20.29	51.57	62.68
w/o AiT Loss	2.80	96.53	63.86	29.68	60.30	80.08	96.71	49.33	90.82	93.11	20.83	55.04	62.64
JointTuner-ZS (Ours)	1.80	97.46	64.15	32.14	62.44	72.61	97.78	46.89	91.92	93.83	21.19	57.14	65.53

Table A4. Quantitative ablation results for each core component of the proposed JointTuner, on ZeroScope. We highlight the best and second-best values for each metric.

optimizes spatial and temporal typical LoRA modules; **(iii) Two-Stage Training**, which separates appearance and motion learning into distinct training phases; and **(iv) w/o AiT Loss**, which uses the original diffusion loss instead of AiT loss when processing video inputs during joint training.

Quantitative results. As shown in Tab. A4, JointTuner-ZS significantly outperforms the *w/o GLoRA* variant, reducing ARS from 3.6 to 1.8 and improving NAS (+3.2) and AAS (+1.66). Moreover, *w/o GLoRA* performs worse than *Decoupled LoRA*, further demonstrating that gating regulation enhances the effectiveness of unified LoRA. JointTuner-ZS also surpasses both *Decoupled LoRA* and *Two-Stage Training*. Specifically, *Decoupled LoRA* yields degraded CLIP-T and CLIP-I scores, while *Two-Stage Training* further reduces perceptual quality across Pick-Sc, Aesth-Q, and Img-Q. These results highlight the superiority of joint optimization in aligning spatial and temporal features. Additionally, *w/o AiT Loss* exhibits notable drops in CLIP-T and CLIP-I, indicating that the AiT loss is crucial for disentangling motion from appearance while preserving the target subject's visual identity.

Qualitative results. As shown in Fig. A3, removing gating regulation (*w/o GLoRA*) causes motion artifacts and appearance degradation. The *Decoupled LoRA* variant produces incomplete appearance reconstruction, while *Two-Stage Training* further compromises motion coherence. In contrast, *w/o AiT Loss* suffers from appearance corruption, underscoring the importance of the AiT loss in maintaining visual fidelity.

D.4. Hyperparameter Sensitivity

Since the loss functions for appearance learning and motion learning differ, we also experimented with adjusting the weight of the AiT Loss, denoted by λ, to evaluate its effect on motion quality in customized video generation. We tested various values of λ and analyzed five key metrics, with results presented in Fig. A4. Notably, Dyn-Deg and Mot-Fid improve by 28.68% and 8.86%, respectively, when $\lambda = 10$, while CLIP-I decreases by 13.54%. In contrast, CLIP-T and Pick-Sc exhibit minimal variation, indicating lower sensitivity. While $\lambda = 10$ provides the optimal performance, the results also show that using $\lambda = 1$ does not lead to significant degradation. In practice, if a setting with $\lambda = 1$ produces unsatisfactory results, increasing it to $\lambda = 10$ can often yield better motion quality, making $\lambda = 10$ a reliable reference for future customizations.

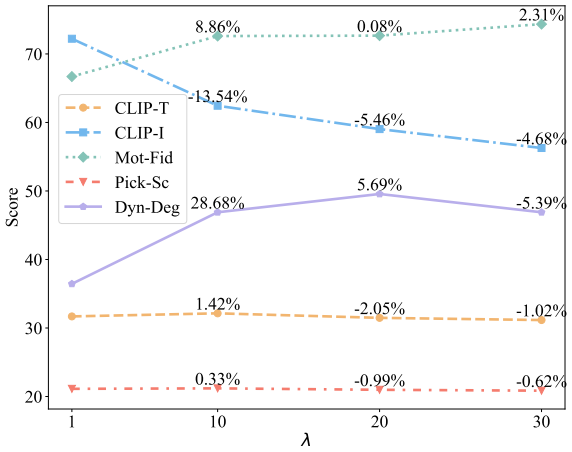


Figure A4. Quantitative analysis of the performance of our JointTuner based on ZeroScope, with varying AiT loss weight, λ.

D.5. More Customized Videos

In this section, we showcase more appearance-motion combined customized videos generated by JointTuner and provide a qualitative comparison with current state-of-the-art methods. The results are as follows.



(a) A curious cat flying across a sunlit meadow, chasing after floating dandelion seeds, with soft wind gently swaying the tall grass around.

(b) A joyful dog flying across a sunlit meadow, chasing butterflies in a serene countryside setting.

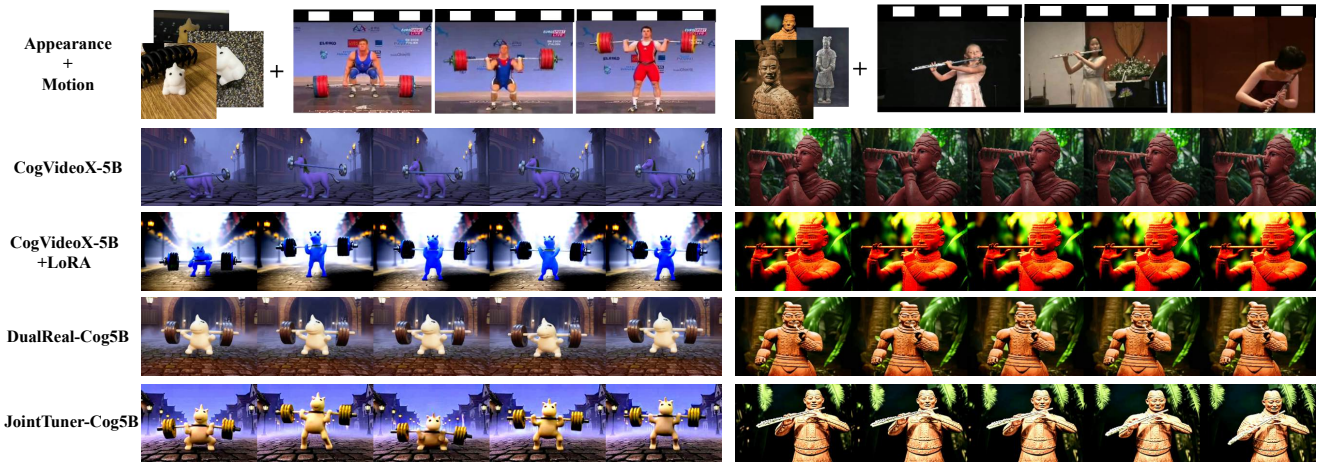
Figure A5. Qualitative results of appearance-motion combined customization on CogVideoX-5B.



(a) A curious cat walking cautiously through a snowy landscape, leaving delicate paw prints behind, surrounded by silent trees dusted with fresh snow.

(b) A curious cat playing the flute in a serene snowy landscape, surrounded by silent trees and soft, falling snowflakes.

Figure A6. Qualitative results of appearance-motion combined customization on CogVideoX-5B.



(a) A whimsical unicorn toy lifting a gleaming barbell on a misty cobblestone street, surrounded by old-world architecture and flickering gas lamps.

(b) A terracotta warrior playing a delicate flute, surrounded by lush greenery in an ancient jungle.

Figure A7. Qualitative results of appearance-motion combined customization on CogVideoX-5B.