# MotionFlow: Attention-Driven Motion Transfer in Video Diffusion Models

**Tuna Han Salih Meral, Hidir Yesiltepe, Connor Dunlop, Pinar Yanardag**

Virginia Tech
tmeral@vt.edu, hidir@vt.edu, cdunlop@vt.edu, pinary@vt.edu

## Abstract

Text-to-video models have demonstrated impressive capabilities in producing diverse video content, yet often lack fine-grained control over motion. We address the problem of motion transfer: given a source video and a target text prompt, generate a new video that preserves the source motion while matching the target semantics and allowing large changes in appearance and scene layout. We introduce MotionFlow, a training-free framework that performs test-time latent optimization guided by attention-derived motion cues. MotionFlow first extracts cross-attention maps from a pre-trained video diffusion model and converts them into spatio-temporal motion masks for the source subject. During generation, it optimizes the target latents so that their evolving attention patterns align with these masks, while the target text controls appearance. This avoids direct attention-map replacement and any model-specific fine-tuning, reducing artifacts and improving flexibility. Qualitative and quantitative experiments, including a user study, show that MotionFlow outperforms existing methods in motion fidelity, temporal consistency, and versatility, even under drastic scene changes.

**Code** — https://github.com/tunahansalih/motionflow
**Extended version** — https://arxiv.org/abs/2412.05275

## Introduction

Generating realistic and controllable video remains a significant challenge in computer vision. While recent advances in diffusion models have enabled the creation of impressive text-to-video (T2V) models (Guo et al. 2023; Singer et al. 2022; Wang et al. 2023; Yu et al. 2024; Chen et al. 2023b; Hong et al. 2022; Yang et al. 2024; Zhou et al. 2022; OpenAI 2024), these models often struggle to provide fine-grained control over the motion of generated objects and characters—a critical requirement for a wide range of creative applications. Current T2V models excel at generating diverse content based on textual descriptions, but precisely dictating the way things move within the generated scene remains a significant limitation. We focus on motion transfer: given a source video and a target text prompt, the goal is to generate a new video that preserves the source motion while following the target semantics.



Imagine a filmmaker planning a new scene, eager to explore various motion styles for a character before committing to the labor-intensive process of shooting or animating the actual footage. Using MotionFlow, this filmmaker can repurpose video clips (e.g., a dog jumping into a lake) and transfer these motions directly into new settings (e.g., a rabbit jumping into a river surrounded by blooming flowers, as shown in Fig. 1). This capability allows for rapid prototyping of different motion effects, enabling filmmakers and animators to quickly visualize and iterate on their creative ideas, significantly reducing the time and resources required for pre-production and experimentation.

However, existing motion transfer methods have notable limitations. Many approaches struggle to maintain motion fidelity without unintentionally transferring unwanted appearance details or scene elements from the source video (Yatim et al. 2024; Zhao et al. 2025). Others depend on extensive training (Wu et al. 2022; Wang et al. 2024a; Li et al. 2024) or fine-tuning tailored to specific motion patterns (Zhao et al. 2025; Ren et al. 2024; Zhang et al. 2023a; Wang et al. 2024b), making them inflexible and impractical for real-world use where speed and adaptability are crucial. Unlike these methods, MotionFlow operates entirely at test time using pre-trained video diffusion models, offering a streamlined and versatile alternative.

To address these challenges, we propose MotionFlow, a training-free test-time approach that leverages the inherent capabilities of pre-trained video diffusion models without requiring additional training. MotionFlow introduces a strategy for motion transfer in which, instead of directly manipulating attention maps or retraining model components, it guides the generation process by optimizing the latent representations at each diffusion timestep. This optimization aims to align the evolving attention patterns of the target video with attention-derived motion masks from the source video, thereby transferring the desired dynamics while remaining independent of the source's appearance and scene composition. As illustrated in Fig. 2, by visualizing cross-attention maps, we observe how MotionFlow effectively transfers motion dynamics by ensuring the generated subject's attention aligns with the original motion patterns, all while adhering to the new edit prompt. Our contributions include:

- We introduce MotionFlow, a training-free test-time latent optimization method for attention-guided motion trans-

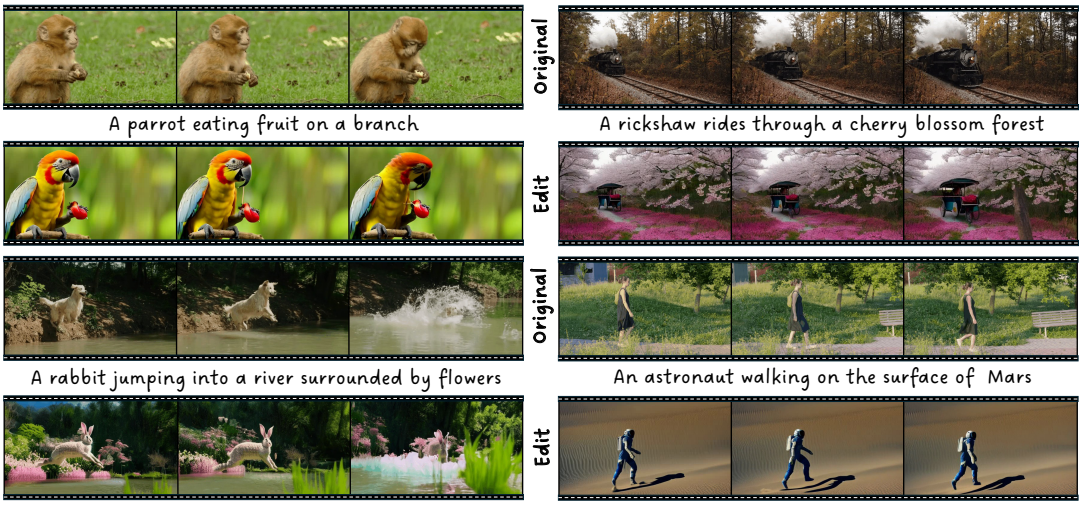


Figure 1: MotionFlow is a training-free method that utilizes cross-attention maps to capture and manipulate spatial and temporal dynamics for motion transfer. Shown are examples where motion from a source video (top row) is transferred to new subjects and scenes (bottom row), while preserving motion patterns and allowing large scene changes.

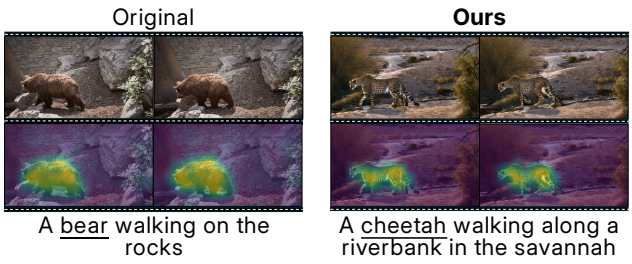


Figure 2: **Motivation.** Visualization of cross-attention maps for the subject tokens, showing how MotionFlow captures and transfers motion dynamics from the original video, ensuring accurate subject motion while adhering to new edit prompts.

fer. Unlike direct attention replacement or training-based approaches, it offers greater flexibility, artifact reduction, and generalization without per-video fine-tuning.

- We demonstrate, through extensive experiments and a user study, that MotionFlow achieves accurate motion transfer and generalizes to complex motions and robust scene alterations, providing a favorable trade-off compared to state-of-the-art methods.
- We will publicly release our code to promote further research and practical applications in video creation and editing.

## Related Work

**Text-to-Video Diffusion Models** Building on the success of Text-to-Image (T2I) models (Rombach et al. 2022; Saharia et al. 2022; Ramesh et al. 2022; Nichol et al. 2021), Text-to-Video (T2V) generation has advanced by extending 2D diffusion architectures with temporal layers (Guo et al. 2023; Singer et al. 2022; Wang et al. 2023; Yu et al. 2024; Chen et al. 2023b; Zhou et al. 2022; Yuan et al. 2024). Many works focus on fine-tuning these augmented models for video generation or introducing conditioning inputs such as depth maps for more precise control (Zhang et al. 2023b; Yin et al. 2023; Lin et al. 2023; Lian et al. 2023; Chen et al. 2023a). Attention-based guidance, crucial for T2I controllability (Chefer et al. 2023; Meral et al. 2023, 2024; Dahary et al. 2024; Helbling et al. 2025), is also foundational for T2V control, especially for temporal consistency. These techniques, which often involve manipulating attention maps or optimizing latent features, motivate the use of attention as a controllable interface in video diffusion models.

**Video Motion Editing** While T2V models can control motion via prompts, complex motions remain challenging to specify and preserve. Some methods use additional cues such as bounding boxes and depth maps (Li et al. 2024; Wang et al. 2024a; Jain et al. 2024; Ma, Lewis, and Kleijn 2023). Another approach is motion transfer from a reference video. Fine-tuning methods embed motion into model weights through training-intensive processes, using techniques like dedicated spatiotemporal layers or separate branches for motion and appearance (Wu et al. 2022; Zhao et al. 2025; Wei et al. 2024; Ren et al. 2024; Wang et al. 2024b). Inversion-based methods utilize DDIM inversion (Song, Meng, and Ermon 2020) and feature losses for guidance but often fail when there are significant geometric or appearance differences between source and target videos (Hertz et al. 2022; Tumanyan et al. 2023; Yatim et al. 2024; Bai et al. 2024; Jeong, Park, and Ye 2024; Yang et al. 2023; Kara et al. 2024).

Several recent works use attention mechanisms for editing. FateZero (Qi et al. 2023) and Video-P2P (Liu et al. 2024) perform direct attention map fusion or replacement. MotionClone (Ling et al. 2024) and DiTFlow (Pondaven

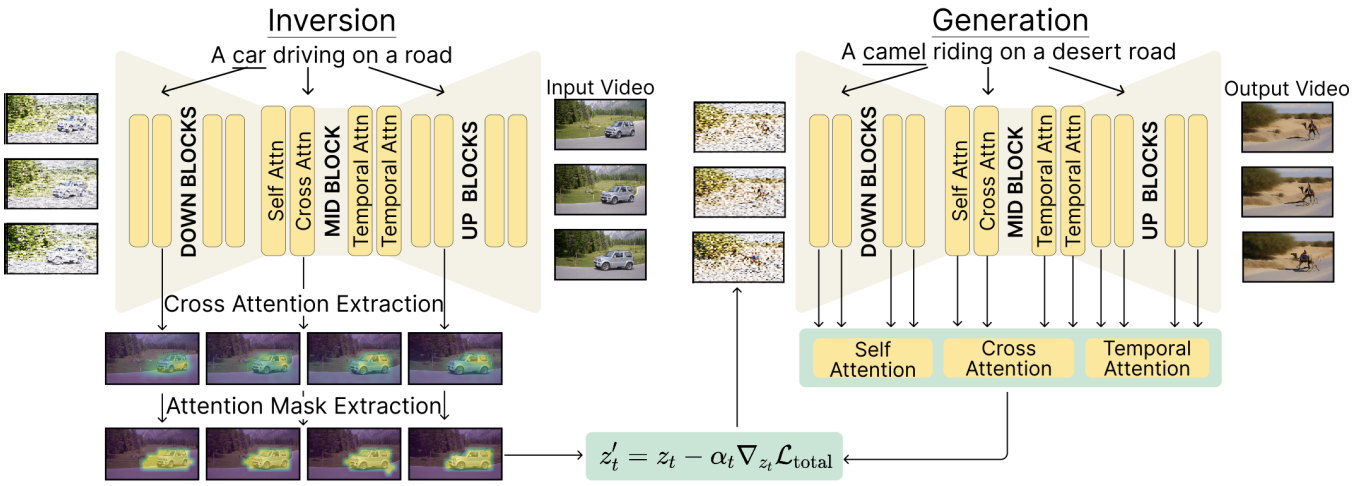


Figure 3: **Overview of MotionFlow framework.** The process consists of two stages: (1) Inversion, where latent representations and cross-attention maps are extracted from the source video to generate motion-guiding masks; (2) Guided Generation, where these masks and a target text prompt direct the creation of a new video that replicates the source's motion while adhering to the prompt's semantics.

et al. 2025) further explore training-free motion cloning and motion transfer in diffusion transformers via attention manipulation. While these methods offer control, direct manipulation of attention maps or architectural specialization can disrupt the generative process and cause artifacts, especially with large edits or substantial content changes. In contrast, MotionFlow avoids direct attention replacement by optimizing the latent representation to achieve the desired attention patterns, rather than forcing them, and can be applied to both UNet-based and Transformer-based video diffusion models. This yields greater flexibility and reduces artifacts while remaining training-free and applicable to pre-trained video diffusion models.

## Background

Diffusion models iteratively transform a noise sample $x_T \sim \mathcal{N}(0, \mathbf{I})$ into a structured output $x_0$ via a learned denoising process. Latent diffusion models (LDMs) (Rombach et al. 2022) operate in the latent space of a pre-trained autoencoder $(\mathcal{E}, \mathcal{D})$, where an input $x$ is encoded to $z = \mathcal{E}(x)$. This compression reduces computational cost while preserving high-fidelity generation.

Video LDMs augment T2I frameworks with temporal layers (e.g., convolution, attention) for inter-frame dependencies. We use the ZeroScope T2V model (cerspense 2023), an extension of Stable Diffusion with these mechanisms, ideal for modeling motion dynamics in our approach.

ZeroScope's UNet features temporal (cross-frame), self-attention (intra-frame), and cross-attention (within-frame, spatial-text). While we extract cross-attention ($A_{cross} \in \mathbb{R}^{F \times N \times K}$) features from the source subject ($N = h \times w$ latent dimensions, $K$ tokens) as motion cues, our optimization, described in the Methodology section, uses losses on all three attention types to guide target generation. This maintains spatio-temporal consistency, with self-attention ($A_{self} \in \mathbb{R}^{F \times N \times N}$) ensuring intra-frame coherence and temporal attention ($A_{temp} \in \mathbb{R}^{N \times F \times F}$) providing smooth inter-frame motion.

## Methodology

We propose a framework for generating a new video $\hat{V}$ by transferring the motion dynamics of a specific subject from a source video $V$, while ensuring compliance with a target text prompt $P$. Our method operates at test time without any training or fine-tuning. The core idea is to extract cross-attention maps from the source video during a DDIM inversion process and then use these maps to guide the generation of the new video via an optimization process applied to the latent representation. This optimization encourages the latent representations to evolve in a way that produces attention maps of the generated video aligned with those of the source video, thereby transferring the motion while allowing the scene content to be driven by the target text prompt $P$. Our method consists of two main stages: (1) Source Video Inversion and Attention Extraction, and (2) Guided Video Generation. These stages are illustrated in Fig. 3.

**Source Video Inversion.** We encode the source video $V \in \mathbb{R}^{F \times H \times W \times C}$ (where $F$ is the number of frames, and $H$, $W$, and $C$ are height, width, and channels) into a latent space using the pre-trained encoder $\mathcal{E}$. Each frame $V_f$ is encoded as $z_f = \mathcal{E}(V_f)$, where $z_f \in \mathbb{R}^{h \times w \times d}$, with $h \ll H$ and $w \ll W$. We then apply DDIM inversion (Song, Meng, and Ermon 2020) to the sequence of initial latents $\{z^s_{f,0}\}^{F-1}_{f=0}$ to obtain a corresponding sequence of noisy latents $\{z^s_t\}^{T}_{t=0}$, spanning the diffusion timesteps $T$. These inverted noisy latents $\{z^s_t\}$ allow for faithful reconstruction of the source video and serve as the foundation for extracting attention information. They are also stored to initialize the target video generation process.

**Attention-Based Mask Extraction.** To capture the spatio-temporal characteristics of the moving subject in the source video, we extract and process its cross-attention maps. First, we manually identify a set of key tokens $S^* = \{s_0, s_1, \ldots, s_k\}$ from the source prompt $P_s = \{p_1, p_2, \ldots, p_{L_s}\}$. These tokens (e.g., 'dog', 'running') typically describe the primary subject and its action. During the DDIM inversion process (or via a separate forward pass through the UNet using $\{z_t^s\}$ and $P_s$), we extract raw cross-attention maps from specific UNet layers known to capture rich semantic information (middle block, last down-sampling block, first up-sampling block (Meral et al. 2023; Chefer et al. 2023; Hertz et al. 2022)). For a given timestep $t$, frame $f$, and key token $s \in S^*$, the raw attention map (typically of shape $N_{\text{heads}} \times (h \cdot w) \times L_{\text{source}}$ from the UNet attention mechanism) is averaged across attention heads and spatially reshaped to yield $A_{cross}^{s,f,t} \in \mathbb{R}^{h \times w}$. This map $A_{cross}^{s,f,t}$ reflects the spatial regions the UNet attends to for token $s$ in frame $f$ at diffusion step $t$.

Finally, we convert these processed cross-attention maps into binary masks $M^{s,f,t} \in \{0,1\}^{h \times w}$ using an adaptive thresholding technique (Tang et al. 2022):

$$M^{s,f,t}[x,y] = \mathbb{I}\left(A_{cross}^{s,f,t}[x,y] > \tau \cdot \max_{i,j} A_{cross}^{s,f,t}[i,j]\right), \tag{1}$$

where $\mathbb{I}$ is the indicator function, $[x, y]$ are spatial coordinates, and $\tau$ is a threshold parameter. These binary masks $M^{s,f,t}$ highlight regions of highest attention for the key tokens, effectively capturing the subject's evolving spatial location and motion trajectory. These masks are the primary guidance for the subsequent video generation stage.

**Guided Generation.** With the source motion masks $M^{s,f,t}$ prepared, the second stage generates the target video $\hat{V}$ by iteratively denoising and optimizing a sequence of latent variables $\hat{z}_t$, guided by the target prompt $P_t$ and the source masks. This process is illustrated in Fig. 3 (right).

We initialize the target video's noisiest latent $\hat{z}_T$ using the corresponding inverted latent from the source video, i.e., $\hat{z}_T = z_T^s$. Optional low-frequency Gaussian noise can be added to $z_T^s$ to introduce content variation if desired. The generation then proceeds by iteratively applying a modified DDIM denoising process from $t = T$ down to $t = 1$. For a predefined number of initial diffusion timesteps, $N_{\text{update\_steps}}$ (e.g., the first 20 steps), each DDIM step is augmented with a latent optimization phase. This phase, repeated for $N_i$ iterations (the 'optimization iterations per step'), refines the current latent $\hat{z}_t$. During each optimization iteration, a forward pass is performed through the UNet using $\hat{z}_t$ and the embeddings of $P_t$. This yields the target video's current cross-attention maps $\hat{A}_{s,f,t}^{\text{cross}}$ (for each token $s$ in $P_t$ corresponding to a source key token $s'$), self-attention maps $\hat{A}_{f,t}^{\text{self}}$, and temporal attention maps $\hat{A}_{n,t}^{\text{temp}}$. A total loss, $L_{\text{total}} = \lambda_{CA}L_{CA} + \lambda_{SA}L_{SA} + \lambda_{TA}L_{TA}$, guides the latent update. The weights $\lambda_{CA}, \lambda_{SA}, \lambda_{TA}$ (empirically set to 1.0; see the Experiments section) balance three components, all guided by the source masks $M^{s',f,t}$.

The cross-attention loss $L_{CA}$ ensures that the subject of the target prompt $P_t$ (identified by token $s$, corresponding to source key token $s'$) spatially aligns with the motion trajectory defined by $M^{s',f,t}$. It aims to maximize the target cross-attention $\hat{A}_{s,f,t}^{\text{cross}}$ within these masked regions:

$$L_{CA} = \sum_{s \leftrightarrow s', f} \left(1 - \frac{\sum_n M^{s',f,t}[n] \cdot \hat{A}_{s,f,t}^{\text{cross}}[n]}{\sum_n \hat{A}_{s,f,t}^{\text{cross}}[n] + \epsilon}\right). \tag{2}$$

Here, $n$ indexes spatial locations, and the normalization by total target attention for token $s$ emphasizes the masked regions proportionally. The summation $\sum_{s \leftrightarrow s'}$ implies summing over pairs of semantically corresponding key tokens from $P_t$ and $S^*$. For example, if $P_s$ features a 'car' ($s'$) and $P_t$ a 'camel' ($s$), this loss aligns the 'camel's' attention with the 'car's' motion mask.

The self-attention loss $L_{SA}$ promotes intra-frame spatial consistency for the generated subject. It encourages strong self-attention connections $\hat{A}_{f,t}^{\text{self}}[n, n']$ between pairs of spatial locations $(n, n')$ that both fall within the source mask $M^{s',f,t}$. This mask $M^{s',f,t}$, derived from a source key token, delineates the target subject's expected spatial extent. Maximizing self-attention within this region promotes internal coherence (e.g., a generated camel's legs attend to its body if both fall within the area masked by a source car's motion). Let $\hat{A}_{\text{masked}}^{s',f,t}$ be the sum of self-attention values within the masked region:

$$\hat{A}_{\text{masked}}^{s',f,t} = \sum_{n,n'} \hat{A}_{f,t}^{\text{self}}[n, n'] \cdot M^{s',f,t}[n] \cdot M^{s',f,t}[n']. \tag{3}$$

The self-attention loss is then defined as:

$$L_{SA} = \sum_{s',f} \left(1 - \frac{\hat{A}_{\text{masked}}^{s',f,t}}{\sum_{n,n'} \hat{A}_{f,t}^{\text{self}}[n, n'] + \epsilon}\right). \tag{4}$$

The temporal attention loss $L_{TA}$ encourages smooth and consistent motion across frames. For each spatial location $n$ within the subject's mask $M^{s',f_1,t}$ in frame $f_1$, $L_{TA}$ promotes high temporal attention scores $\hat{A}_{n,t}^{\text{temp}}[f_1, f_2]$ between frame $f_1$ and other frames $f_2$. This maintains temporal correspondence for the moving subject, leading to smoother trajectories (e.g., a leg's pixels in one frame strongly attend to corresponding leg pixels in subsequent frames):

$$L_{TA} = \sum_{s'} \left(1 - \frac{\sum_{n,f_1,f_2} \hat{A}_{n,t}^{\text{temp}}[f_1, f_2] \cdot M^{s',f_1,t}[n]}{\sum_{n,f_1,f_2} \hat{A}_{n,t}^{\text{temp}}[f_1, f_2] + \epsilon}\right). \tag{5}$$

These individual loss components are then combined into a total loss $L_{\text{total}}$, guiding the latent update:

$$L_{\text{total}} = \lambda_{CA}L_{CA} + \lambda_{SA}L_{SA} + \lambda_{TA}L_{TA}. \tag{6}$$

The weights $\lambda_{CA}, \lambda_{SA}, \lambda_{TA}$ balance the influence of each attention modality. After computing $L_{\text{total}}$, the target latent $\hat{z}_t$ is updated via gradient descent: $\hat{z}_t \leftarrow \hat{z}_t - \alpha_t \nabla_{\hat{z}_t} L_{\text{total}}$, where $\alpha_t$ is a learning rate. Once $N_i$ iterations of this optimization are complete, the refined latent $\hat{z}_t$ undergoes a standard DDIM denoising step conditioned on $P_t$ to produce $\hat{z}_{t-1}$. For diffusion timesteps beyond the initial $N_{\text{update\_steps}}$,

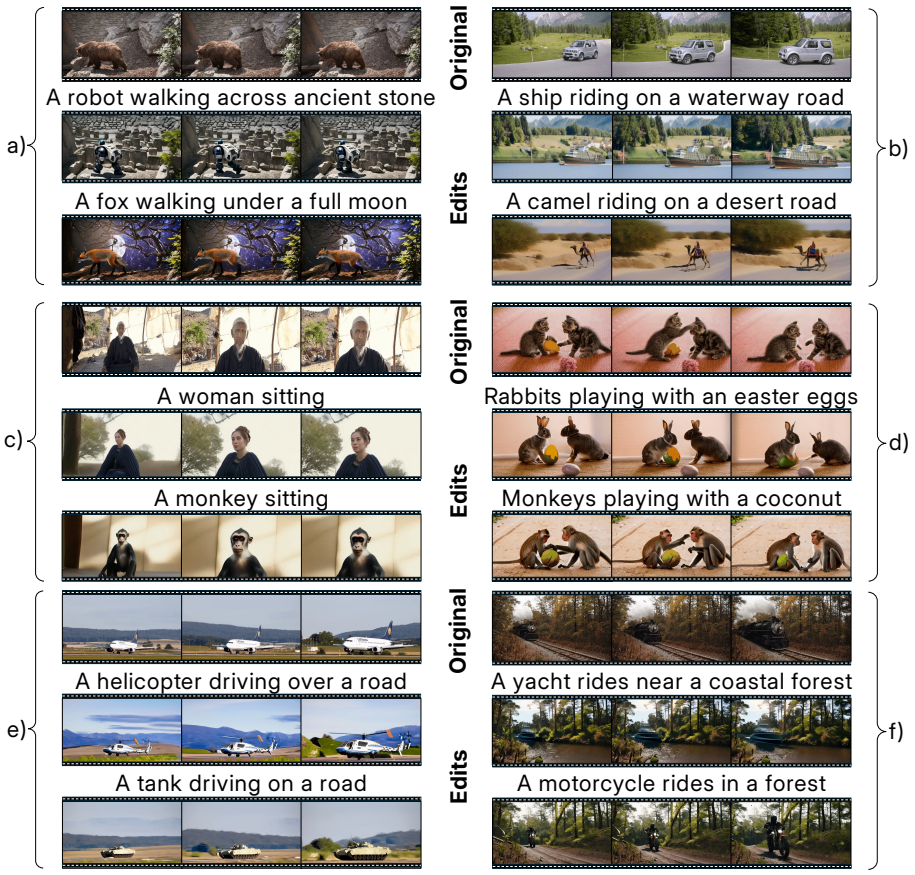


Figure 4: **Qualitative Results.** MotionFlow can successfully transfer a wide variety of motion types. Additionally, it can significantly alter scene layout based on the user-provided text prompt. Please see Supplementary Material for full videos.

only standard DDIM denoising is applied. This training-free process concludes by decoding $\hat{z}_0$ using the VAE decoder $\mathcal{D}$ to obtain the final video $\hat{V}$.

Crucially, the guidance via source masks $M^{s',f,t}$ is intentionally flexible. These masks define *where* consistency should be enforced but do not rigidly dictate *how* target spatial or temporal attention relationships must form. This allows MotionFlow to adapt source motion to new target geometries and appearances (e.g., transferring a car's motion to a camel retains trajectory and speed via mask guidance, while the camel's articulation and gait are formed coherently via $L_{SA}$ and $L_{TA}$, respectively), ensuring fidelity to the source motion's intent without forcing unnatural constraints. While this three-part loss is designed for UNet architectures, our framework's core principle of using attention-derived masks to guide latent optimization is highly adaptable. Our attention-derived mask guidance is architecture-agnostic and extends to diffusion transformers (e.g., CogVideoX; see Supplementary).

## Experiments

**Experimental Setup.** We perform DDIM inversion (Song, Meng, and Ermon 2020) on the source video to obtain the initial noise latent, following the process outlined in the Methodology section. For mask generation (Eq. 1), we set $\tau = 0.4$, empirically determined to balance motion capture and noise suppression based on preliminary experiments. During guided generation, we optimize the latent using $\alpha = 5.0$, selected for stable convergence, with equal loss weights $\lambda_{CA} = \lambda_{SA} = \lambda_{TA} = 1.0$. All experiments are conducted on NVIDIA L40 GPUs. In the generation stage, we perform latent updates for 20 out of 50 backward diffusion steps, with 20 optimization iterations per step, adapting a similar strategy to (Yatim et al. 2024) for iterative latent refinement. We limit updates to 20 steps to balance motion accuracy and computational efficiency, halting optimization to allow DDIM denoising to refine the output without artifacts. For experiments, we used a range of videos from the DAVIS (Pont-Tuset et al. 2017) dataset. This dataset provides a robust foundation for evaluating the effectiveness of MotionFlow in diverse motion transfer scenarios.

**Baselines and Metrics.** We compare MotionFlow to four recent motion-transfer methods: DMT (Yatim et al. 2024), VMC (Jeong, Park, and Ye 2024), MotionDirector (MD) (Zhao et al. 2025), and Motion Inversion (MI) (Wang et al. 2024b). Each method is evaluated on 100 video–prompt pairs sampled from DAVIS (Pont-Tuset et al. 2017). We re-

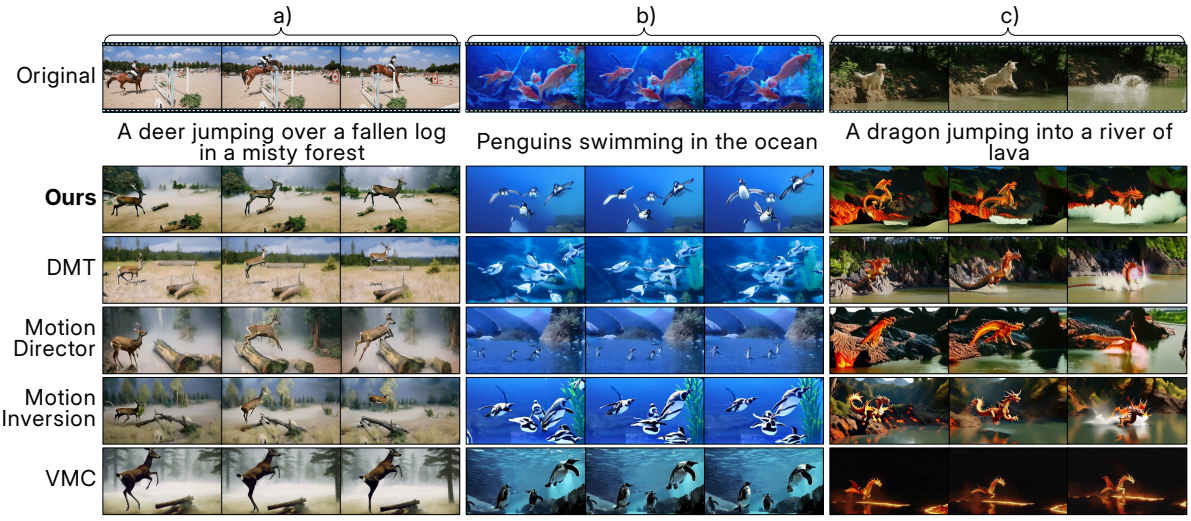


Figure 5: **Comparison.** Results of MotionFlow compared with DMT (Yatim et al. 2024), MotionDirector (Zhao et al. 2025), Motion Inversion (Wang et al. 2024b), and VMC (Jeong, Park, and Ye 2024) on various video editing tasks. Please see Supplementary Material for full videos.

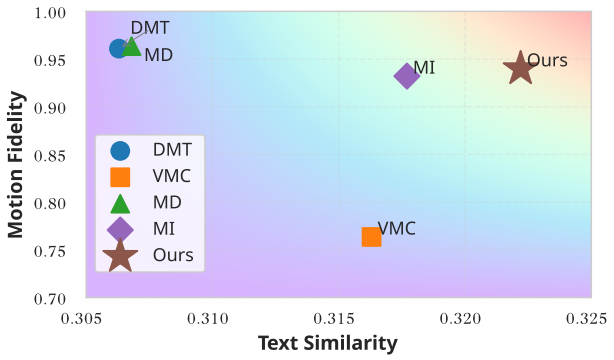


Figure 6: **Evaluation.** CLIP text similarity vs Motion Fidelity scores for each baseline. Our method exhibits a better balance between these metrics.

port Motion Fidelity Score (Yatim et al. 2024), Temporal Consistency (average cosine similarity between CLIP (Radford et al. 2021) features of consecutive frames), Text Similarity (CLIP similarity between frames and prompt), and FID. Higher Motion Fidelity, Temporal Consistency, and Text Similarity and lower FID are better.

## Qualitative Experiments

Fig. 4 demonstrates the versatility of MotionFlow in transferring motion across diverse scenarios while maintaining control over scene composition according to the target prompt. MotionFlow can transfer motion between semantically different subjects, a challenge for many existing methods. For example, in Fig. 4(a), the motion of a bear is transferred to both a robot and a fox, demonstrating novel scene layouts with different subjects. Fig. 4(f) shows the mapping of a train's motion to a motorbike, and Fig. 4(b) maps a car's motion to a camel, preserving characteristic movement patterns. See the supplementary material for additional results and full videos.

Fig. 5 provides a qualitative comparison of MotionFlow against several state-of-the-art methods: DMT (Yatim et al. 2024), MotionDirector (Zhao et al. 2025), Motion Inversion (Wang et al. 2024b), and VMC (Jeong, Park, and Ye 2024). In terms of motion fidelity, MotionFlow, in Fig. 5(a), accurately transfers the horse's motion to the deer, preserving the nuances of the gait and overall movement, while some competing methods struggle to capture these details. Concerning object count and motion consistency, as shown in Fig. 5(b), MotionFlow best captures both the number of fish and their characteristic swimming motion, demonstrating superior handling of multiple objects and complex movements. Finally, in Fig. 5(c), while both MotionFlow and VMC (Jeong, Park, and Ye 2024) succeed in generating content that aligns with the target prompt, MotionFlow exhibits superior motion alignment with the original video. VMC, in this instance, fails to capture the intended movement of the target. Overall, MotionFlow demonstrates a superior ability to balance motion fidelity, target prompt adherence, and the generation of visually coherent results. Additional qualitative comparisons, including results for Video-P2P (Liu et al. 2024), FateZero (Qi et al. 2023), and Pix2Video (Ceylan, Huang, and Mitra 2023), are provided in the supplementary material.

Table 1 and Fig. 6 summarize our approach's performance across key metrics, highlighting its advantages over existing methods. MotionFlow achieves the highest text similarity while preserving strong motion fidelity and temporal consistency, resulting in a better balance between these metrics than competing methods. This behavior can be attributed to three key factors: (1) By leveraging cross-attention maps, our framework achieves precise motion transfer, preserving

| Methods | Quantitative Results | | | | User Study | | | Processing Times (sec) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Text ↑ Similarity | Motion ↑ Fidelity | Temporal ↑ Consistency | FID ↓ | Text Alignment | Motion Fidelity | Visual Quality | Initial Setup | Video Generation | Total |
| DMT | 0.306 | 0.960 | 0.934 | **218.99** | 0.20 | 0.17 | 0.19 | 258 | 332 | 590 |
| MD | 0.307 | **0.963** | 0.928 | 240.80 | 0.10 | 0.12 | 0.11 | 410 | 67 | 477 |
| VMC | 0.306 | 0.763 | **0.961** | 273.19 | 0.15 | 0.15 | 0.16 | 227 | 503 | 730 |
| MI | 0.317 | 0.930 | 0.941 | 239.68 | 0.13 | 0.13 | 0.15 | 195 | 30 | 225 |
| **Ours** | **0.322** | 0.940 | 0.941 | 230.89 | **0.42** | **0.43** | **0.39** | 49 | 376 | 425 |

Table 1: **Quantitative Comparisons, User Study, and Processing Times.** We compare MotionFlow with several baselines across Text Similarity, Motion Fidelity, Temporal Consistency, FID, user preferences, and processing times. User study scores report the fraction of trials ($[0, 1]$) in which each method was preferred under the corresponding criterion.


Figure 7: Ablation study on latent updates. Without latent updates guided by cross-attention, inverted latents may fail to preserve motion or generate the intended subject. Please see Supplementary Material for full videos.

the integrity of original motion patterns without requiring fine-tuning. This sets it apart from training-intensive methods like DMT and MotionDirector. (2) As shown in Fig. 6, our method trades off text similarity and motion fidelity in a controlled way, enabling accurate motion adaptation to prompt specifications. High motion fidelity alone can sometimes penalize necessary edits, but our method maintains fidelity to the original while allowing flexibility for editing tasks. (3) Our approach better aligns with the target prompt, resulting in higher text similarity scores and improved semantic coherence. This alignment surpasses methods such as VMC, which struggles with motion fidelity, and Motion Inversion, although it uses an alternative diffusion model backbone (MotionCraftV2 (Zhang et al. 2023a)). In summary, MotionFlow achieves a strong balance across metrics, combining competitive motion fidelity and temporal consistency with superior adherence to target prompts, as evidenced by higher text similarity scores.

**User Study.** To evaluate the perceptual quality of our motion transfer results, we conducted a user study on Amazon Mechanical Turk with 50 participants. Each participant viewed 30 sets of videos, each set containing five generated videos: one from our method and four from baseline methods (Motion Inversion, DMT, VMC, and MotionDirector), along with the original video and the edit prompt. Participants were asked to select the best video based on three criteria: Prompt Alignment (how accurately the video matched the edit prompt), Motion Fidelity (how well the original motion was preserved), and Visual Quality (overall appearance and coherence). Table 1 shows our method is preferred most often across all three criteria. Additional details about the user study are provided in the supplementary material.

**Processing Times.** Runtime comparison on a single NVIDIA L40 GPU is reported in Table 1.

**Ablation Study.** To evaluate the importance of latent updates, we conduct a qualitative ablation study (Fig. 7) by using only the initial noise latent from DDIM inversion and omitting the latent update step. The results show that while DDIM inversion provides a high-level structure, it often fails to capture detailed subject motion and even the intended subject. For example, for the prompt "A robot walking across ancient stone ruins", the model without latent updates struggles to generate a walking motion or a recognizable robot. In contrast, our method incorporating attention-guided latent updates, accurately captures both the subject and its motion, ensuring correct spatial placement. This ablation demonstrates the critical role of latent updates in achieving precise motion transfer and subject generation. Additional ablations over the number of latent update steps and over the cross-, self-, and temporal-attention loss terms are provided in the supplementary material.

## Conclusion

In this paper, we introduced MotionFlow, a training-free framework for high-fidelity motion transfer in video diffusion models. By optimizing latent representations at test time to align with attention maps from a source video, our method successfully transfers complex dynamics across diverse subjects and scenes without model retraining. Our experiments show that MotionFlow provides a favorable trade-off between motion fidelity, temporal consistency, and prompt adherence, and is consistently preferred in user studies compared to prior motion-transfer methods. By making our code public, we aim to enable new creative applications.

## References


Bai, J.; He, T.; Wang, Y.; Guo, J.; Hu, H.; Liu, Z.; and Bian, J. 2024. UniEdit: A unified tuning-free framework for video motion and appearance editing. *arXiv preprint arXiv:2402.13185*.

cerspense. 2023. zeroscope_v2. https://huggingface.co/cerspense/zeroscope_v2_576w. Accessed: 2024-11-14.

Ceylan, D.; Huang, C.-H. P.; and Mitra, N. J. 2023. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23206–23217.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4): 1–10.

Chen, T.-S.; Lin, C. H.; Tseng, H.-Y.; Lin, T.-Y.; and Yang, M.-H. 2023a. Motion-conditioned diffusion model for controllable video synthesis. *arXiv preprint arXiv:2304.14404*.

Chen, W.; Ji, Y.; Wu, J.; Wu, H.; Xie, P.; Li, J.; Xia, X.; Xiao, X.; and Lin, L. 2023b. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*.

Dahary, O.; Patashnik, O.; Aberman, K.; and Cohen-Or, D. 2024. Be yourself: Bounded attention for multi-subject text-to-image generation. *arXiv preprint arXiv:2403.16990*, 2(5).

Guo, Y.; Yang, C.; Rao, A.; Wang, Y.; Qiao, Y.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

Helbling, A.; Meral, T. H. S.; Hoover, B.; Yanardag, P.; and Chau, D. H. 2025. ConceptAttention: Diffusion Transformers Learn Highly Interpretable Features. *arXiv preprint arXiv:2502.04320*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868*.

Jain, Y.; Nasery, A.; Vineet, V.; and Behl, H. 2024. Peekaboo: Interactive video generation via masked-diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8079–8088.

Jeong, H.; Park, G. Y.; and Ye, J. C. 2024. 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*, 9212–9221.

Kara, O.; Kurtkaya, B.; Yesiltepe, H.; Rehg, J. M.; and Yanardag, P. 2024. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6507–6516.

Li, M.; Wan, B.; Moens, M.-F.; and Tuytelaars, T. 2024. Animate Your Motion: Turning Still Images into Dynamic Videos. *arXiv preprint arXiv:2403.10179*.

Lian, L.; Shi, B.; Yala, A.; Darrell, T.; and Li, B. 2023. Llm-grounded video diffusion models. *arXiv preprint arXiv:2309.17444*.

Lin, H.; Zala, A.; Cho, J.; and Bansal, M. 2023. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*.

Ling, P.; Bu, J.; Zhang, P.; Dong, X.; Zang, Y.; Wu, T.; Chen, H.; Wang, J.; and Jin, Y. 2024. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*.

Liu, S.; Zhang, Y.; Li, W.; Lin, Z.; and Jia, J. 2024. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8599–8608.

Ma, W.-D. K.; Lewis, J. P.; and Kleijn, W. B. 2023. Trailblazer: Trajectory control for diffusion-based video generation. *arXiv preprint arXiv:2401.00896*.

Meral, T. H. S.; Simsar, E.; Tombari, F.; and Yanardag, P. 2023. CONFORM: Contrast is All You Need For High-Fidelity Text-to-Image Diffusion Models. *arXiv preprint arXiv:2312.06059*.

Meral, T. H. S.; Simsar, E.; Tombari, F.; and Yanardag, P. 2024. CLoRA: A Contrastive Approach to Compose Multiple LoRA Models. *arXiv preprint arXiv:2403.19776*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

OpenAI. 2024. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/. [Accessed 11-11-2024].

Pondaven, A.; Siarohin, A.; Tulyakov, S.; Torr, P.; and Pizzati, F. 2025. Video motion transfer with diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22911–22921.

Pont-Tuset, J.; Perazzi, F.; Caelles, S.; Arbeláez, P.; Sorkine-Hornung, A.; and Gool, L. V. 2017. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv: Computer Vision and Pattern Recognition*.

Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ren, Y.; Zhou, Y.; Yang, J.; Shi, J.; Liu, D.; Liu, F.; Kwon, M.; and Shrivastava, A. 2024. Customize-a-video: One-shot motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2402.14780*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S. K. S.; Ayan, B. K.; Mahdavi, S. S.; Lopes, R. G.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tang, R.; Pandey, A.; Jiang, Z.; Yang, G.; Kumar, K.; Lin, J.; and Ture, F. 2022. What the daam: Interpreting stable diffusion using cross attention. *arXiv preprint arXiv:2210.04885*.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Wang, J.; Yuan, H.; Chen, D.; Zhang, Y.; Wang, X.; and Zhang, S. 2023. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*.

Wang, J.; Zhang, Y.; Zou, J.; Zeng, Y.; Wei, G.; Yuan, L.; and Li, H. 2024a. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*.

Wang, L.; Mai, Z.; Shen, G.; Liang, Y.; Tao, X.; Wan, P.; Zhang, D.; Li, Y.; and Chen, Y. 2024b. Motion inversion for video customization. *arXiv preprint arXiv:2403.20193*.

Wei, Y.; Zhang, S.; Qing, Z.; Yuan, H.; Liu, Z.; Liu, Y.; Zhang, Y.; Zhou, J.; and Shan, H. 2024. Dreamvideo: Composing your dream videos with customized subject and motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6537–6549.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2022. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. *arXiv preprint arXiv:2212.11565*.

Yang, S.; Zhou, Y.; Liu, Z.; and Loy, C. C. 2023. Rerender A Video: Zero-Shot Text-Guided Video-to-Video Translation. *arXiv preprint arXiv:2306.07954*.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*.

Yatim, D.; Fridman, R.; Bar-Tal, O.; Kasten, Y.; and Dekel, T. 2024. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8466–8476.

Yin, S.; Wu, C.; Liang, J.; Shi, J.; Li, H.; Ming, G.; and Duan, N. 2023. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*.

Yu, S.; Nie, W.; Huang, D.-A.; Li, B.; Shin, J.; and Anandkumar, A. 2024. Efficient Video Diffusion Models via Content-Frame Motion-Latent Decomposition. *arXiv preprint arXiv:2403.14148*.

Yuan, H.; Zhang, S.; Wang, X.; Wei, Y.; Feng, T.; Pan, Y.; Zhang, Y.; Liu, Z.; Albanie, S.; and Ni, D. 2024. InstructVideo: instructing video diffusion models with human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6463–6474.

Zhang, Y.; Tang, F.; Huang, N.; Huang, H.; Ma, C.; Dong, W.; and Xu, C. 2023a. Motioncrafter: One-shot motion customization of diffusion models. *arXiv preprint arXiv:2312.05288*.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023b. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*.

Zhao, R.; Gu, Y.; Wu, J. Z.; Zhang, D. J.; Liu, J.-W.; Wu, W.; Keppo, J.; and Shou, M. Z. 2025. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, 273–290. Springer.

Zhou, D.; Wang, W.; Yan, H.; Lv, W.; Zhu, Y.; and Feng, J. 2022. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*.