EffiVMT: Video Motion Transfer via Efficient Spatial-Temporal Decoupled Finetuning

Anonymous authors
Paper under double-blind review

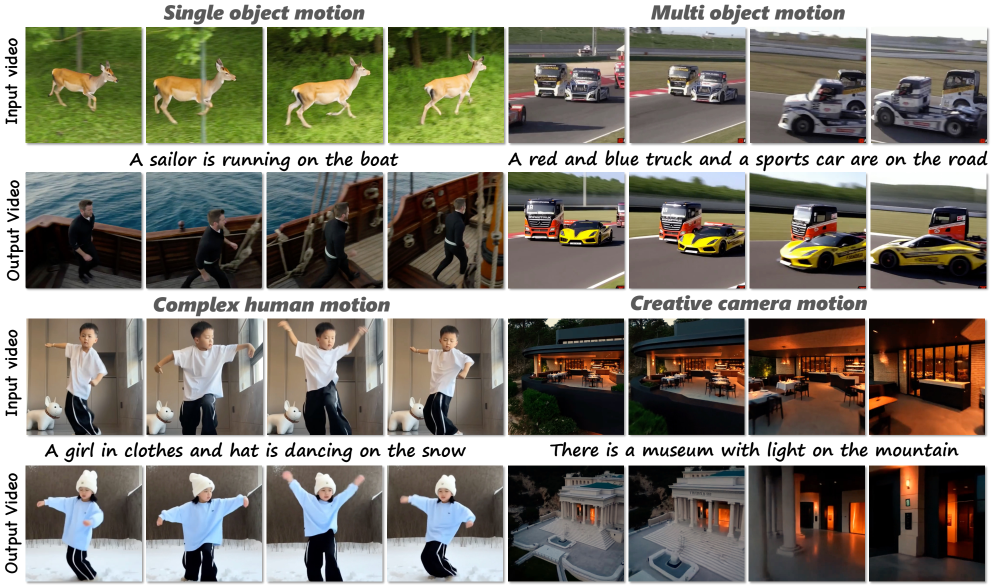

Single object motion
Multi object motion
Input video
Output Video
A sailor is running on the boat
A red and blue truck and a sports car are on the road
Complex human motion
Creative camera motion
Input video
Output Video
A girl in clothes and hat is dancing on the snow
There is a museum with light on the mountain

Figure 1: **Showcases of our EffiVMT**. Given an input video, EffiVMT enables generating the video with the same motion, including motion of single or multiple objects, complex poses of humans, and movements of the camera view.

Abstract

Recently, breakthroughs in the video diffusion transformer have shown remarkable capabilities in diverse motion generations. As for the motion-transfer task, current methods mainly use two-stage Low-Rank Adaptations (LoRAs) finetuning to obtain better performance. However, existing adaptation-based motion transfer still suffers from ***motion inconsistency*** and ***tuning inefficiency*** when applied to large video diffusion transformers. Naive two-stage LoRA tuning struggles to maintain motion consistency between generated and input videos due to the inherent spatial-temporal coupling in the 3D attention operator. In addition, they require time-consuming fine-tuning processes in both stages. To tackle these issues, we propose EffiVMT, an efficient ***three-stage*** video motion transfer framework that finetunes a powerful video diffusion transformer to synthesize complex motion. In ***stage 1***, we propose a spatial-temporal head classification technique to decouple the heads of 3D attention to distinct groups for spatial-appearance and temporal motion processing. We then finetune the spatial heads in the ***stage 2***. In the ***stage 3*** of temporal head tuning, we design the sparse motion sampling and adaptive RoPE to accelerate the tuning speed. To address the lack of a benchmark for this field, we introduce MotionBench, a comprehensive benchmark comprising diverse motion, including creative camera motion, single object motion, multiple object motion, and complex human motion. We show extensive evaluations on MotionBench to verify the superiority of EffiVMT.

1 Introduction

Motion transfer aims to synthesize novel videos that faithfully replicate the motion dynamics,including camera movements and object trajectories from a given reference video. Unlike video-to-video translation methods (Qi et al., 2023; Wu et al., 2022), which prioritize preserving low-level appearance and 2D spatial structure, motion transfer focuses exclusively on disentangling and reapplying motion patterns. This capability holds significant promise across diverse domains such as cinematic production, augmented reality, automated advertising, and social media content generation.

Recent advances in generative models have been dominated by diffusion models (Rombach et al., 2022), which excel in producing high-fidelity visual content through stable optimization over Gaussian noise trajectories. The emergence of Diffusion Transformers (DiTs) has further elevated scalability in terms of model size, computational efficiency, and compatibility with large-scale video datasets. Leveraging pretrained video diffusion models, researchers have developed a spectrum of motion transfer techniques, broadly categorized into *training-free* and *tuning-based* paradigms.

Training-free approaches (Geyer et al., 2023; Pondaven et al., 2025; Qi et al., 2023; Xiao et al., 2024b; Yang et al., 2025) operate entirely during inference by manipulating intermediate motion representations, such as attention maps or latent trajectories, without modifying model parameters. For instance, SMM (Yatim et al., 2024) introduces a spatially averaged feature descriptor to guide motion consistency, while MotionShop (Yesiltepe et al., 2024) repurposes latent-space updates in the denoising process as a “Motion Score” for DiT models. Although these methods offer zero-training-cost generalization across both UNet and DiT architectures, their fidelity is inherently constrained by the motion priors embedded in the pretrained model.

To overcome this limitation and capture complex, out-of-distribution motions, tuning-based methods (Zhao et al., 2023b) optimize model parameters to explicitly encode reference motion. In early UNet-based frameworks like MotionDirector (first row in the Figure 2(c).), temporal layers are fine-tuned independently to learn motion dynamics, while spatial layers remain frozen or jointly optimized. During inference, the learned motion is composited with the frozen model’s prior knowledge to generate novel videos. While effective, extending this paradigm to modern DiT architectures remains challenging due to their high computational cost and the entangled nature of spatial-temporal modeling in 3D self-attention blocks.

A naive baseline for DiT-based motion transfer involves applying Low-Rank Adaptation (LoRA) directly to all parameters within the 3D self-attention layers, as shown in the second row of Figure 2(c). More sophisticated methods, such as the approach proposed by Abdal et al. (2025), employ a two-stage spatial–temporal decoupled tuning strategy: first, spatial LoRAs are optimized on a subset of key frames to preserve appearance consistency; these are then frozen, and temporal LoRAs are tuned over the full video sequence to capture and transfer motion dynamics. However, we argue that this two-stage procedure is inherently inefficient. Specifically, the limitations are listed as follows:

(1) ***Motion inconsistency***: During the spatial tuning stage, both spatial and temporal attention heads are updated using static frames, inadvertently coupling spatial appearance with temporal dynamics. As shown in the top Fig. 2(a), for the naive baseline, both the reconstructed results and motion transfer results fail to follow the reference video. Therefore, tuning both spatial and temporal heads for appearance preservation is not reasonable.

(2) ***Tuning inefficiency***: Recent analysis (Xi et al., 2025) reveals that 3D self-attention heads in DiTs naturally specialize, some focus on spatial relations, others on temporal coherence. Yet current methods indiscriminately tune all heads in each stage, resulting in parameter redundancy and suboptimal adaptation. Furthermore, since 3D VAEs inherently compress and interpolate temporal sequences, processing all reference frames during tuning ignores this latent interpolation capacity and introduces unnecessary computational overhead.

To tackle these challenges, we propose EffiVMT, an efficient video motion transfer framework. First, to resolve motion inconsistency, we employ robust head matching to classify attention heads into spatial and temporal types. During tuning, spatial heads are updated only in the spatial stage, and temporal heads only in the temporal stage for preserving motion consistency in both reconstruction and transfer, as shown in Figure 2(a). To improve tuning efficiency, we introduce sparse motion sampling during temporal tuning, significantly accelerating training. We further propose adaptive RoPE to enhance motion interpolation learning, enabling accurate motion capture even from sparse

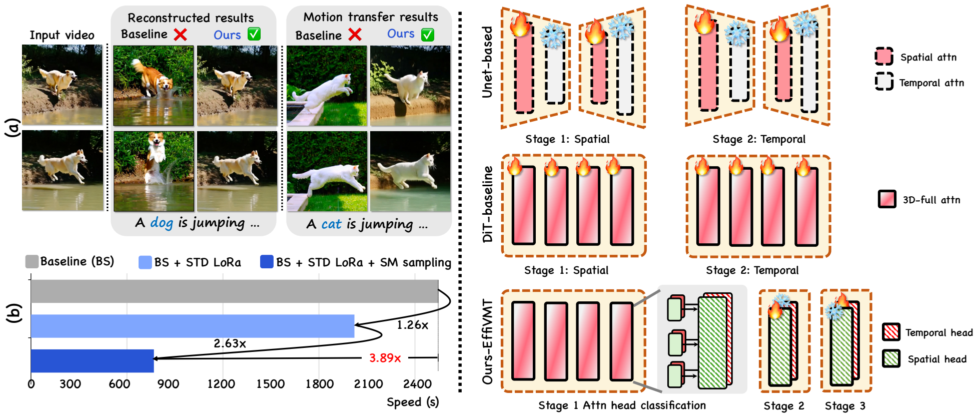


Figure 2: **Comparison between EffiVMT and baseline and Motivation**. ***(a)&(b):*** We finetune the baseline and our method 3,000 steps using Wan2.1 (Wang et al., 2025a). Our method gets better reconstruction and motion preservation. ***(c):*** Despite the decoupling of temporal and spatial in UNet is common, applying it to modern DiT is still challenging because of its *spatial-temporal mixed 3D full self-attention* blocks. To address it, we propose the spatial-temporal decoupled tuning for DiT, sparse motion sampling, and adaptive RoPE to synthesize video with complex motion efficiently.

frames. As demonstrated in Figure 2(b), our decoupled strategy reduces latency by 1.26×, and with sparse sampling, achieves a 3.89× speed-up over full-frame tuning.

Together, these designs enable EffiVMT to generate high-fidelity videos that faithfully follow reference motion (See Figure 1). Additionally, to address the lack of benchmark in video motion transfer, we introduce MotionBench, which is a comprehensive benchmark covering single-object, complex human, multi-object, and camera motions across diverse scenes and styles. Our method outperforms existing baselines across various evaluation metrics, demonstrating its effectiveness in leveraging powerful DiTs for accurate motion transfer. Overall, our key contributions are summarized as follows:

- We propose EffiVMT, a three-stage motion transfer framework that efficiently adapts powerful video Diffusion Transformers (DiTs) to synthesize videos with complex, high-fidelity motion.
- We identify and address two core challenges in DiT-based motion transfer: motion inconsistency and tuning inefficiency. To preserve motion coherence, we decouple spatial and temporal adaptation via specialized LoRA heads. To accelerate training, we introduce sparse motion sampling and adaptive RoPE for efficient yet accurate motion interpolation.
- To validate the effectiveness of our methods, we construct a benchmark MotionBench. We perform extensive experiments and user studies to evaluate our approach, which shows our method achieves state-of-the-art performance.

2 Related Work

Text-to-video generation. Text-to-video generation aims to produce realistic videos that precisely match the spatial visuals and temporal dynamics described in the input prompt. To generate the complicated motion in the videos, diffusion-based video generation models (Guo et al., 2024; Zhao et al., 2023a; Zhu et al., 2025; Liu et al., 2025a) are proposed to synthesize consistent results using a pretrained image diffusion model. Previous works (Guo et al., 2023; He et al., 2022; Wang et al., 2023; Xiong et al., 2025; Yang et al., 2024c) design the temporal module of UNet to generate consistent results. Recently, the emergence of Diffusion Transformer-based methods for text-to-video generation has exhibited superior performance in quality and consistency. These powerful scaling transformers, including Sora (Liu et al., 2024), CogVideoX (Yang et al., 2024d), EasyAnimate (Xu et al., 2024a), HunyuanVideo (Kong et al., 2024), and Wan2.1 (Wang et al., 2025a), enable generating more realistic video clips from given detailed prompts, paving the way for various downstream video generation tasks.

Video Motion transfer. Motion transfer involves an important demand: creating a novel video and maintaining the motion from the reference one. Some methods leverage the explicit control signal (Ma et al., 2023; 2024; Xing et al., 2024a;b; Zhang et al., 2025; Yang et al., 2024b) to achieve motion transfer from the reference video. However, these methods rely on a huge control signal dataset and cost large computational resources. Thanks for the powerful pretrained text-to-video generation model, the researchers pay attention to motion transfer using implicit control, including training-free or tuning-based paradigm. For training-free methods (Hu et al., 2024; Pondaven et al., 2025; Yesiltepe et al., 2024), they extract a motion embedding in the inference stage and use the gradient to guide optimization. However, these methods fail to transfer the complex motion. For tuning-based methods, they (Jeong et al., 2024a; Zhao et al., 2023b) always fine-tune model parameters to utilize different attention for temporal and spatial information. Current works (Jeong et al., 2024a; Ren et al., 2024; Zhao et al., 2023b) employ the dual-path LoRA structure to separate motion and appearance. However, these methods are developed on the UNet-based pretrained model (Chai et al., 2023), making them unsuitable for DiTs. In contrast, our proposed method is the first one-shot DiT-based motion transfer framework. Using the video diffusion transfer as the foundation model, our method extends the boundary of motion transfer performance.

3 Method

Following prior work (Abdal et al., 2025; Zhao et al., 2023b), a naive baseline first optimizes spatial LoRA weights (ΔW_s) by treating sampled frames as independent text-to-image instances. Subsequently, temporal LoRA (ΔW_t) is learned by fine-tuning on consecutive frame sequences while freezing ΔW_s. At inference, only ΔW_t is applied to transfer motion. However, this leads to appearance leakage and remains computationally expensive for DiT-based video diffusion models. As shown in Fig. 2, even after 3,000 optimization steps (3,042s on a single H20 GPU), motion fidelity is unsatisfactory.

Previous naive LoRA tuning faces two main challenges. (1) Recent Video DiT models leverage 3D attention block without explicit temporal blocks, which makes it difficult to disentangle temporal parameters, and fine-tuning LoRA on whole attention parameters results in larger parameter number (*e.g.*, 29.5 M for naive LoRA) (2) Finetuning on multiframe videos increases token sequence length (*e.g.*, 24276 tokens for 81 frames) and computation cost.

Algorithm 1 Dual attention decoupling

```
Input: $Q, K \in \mathbb{R}^{H \times S \times D}$: query and key
        where $S = F \times H \times W$
Output: Closest head type: $t_{\text{head}}$
▷ Target spatial & temporal attention maps:
 [head, $S$, $S$]
$M_{\text{spatial}} \leftarrow$ gen_spatial_maps($F, H, W$)
  $M_{\text{temporal}} \leftarrow$ gen_temporal_maps($F, H, W$)
▷ Get attention maps of input data: [head, $S$, $S$]
$M_{\text{input}} \leftarrow$ Softmax($Q \cdot K^{\top}/\sqrt{D}$)
▷ Calculate similarity metrics
Sim$_s \leftarrow \|M_{\text{input}} \odot M_{\text{spatial}}\|_{\text{mean}}$        // mean
 over (1,2)
Sim$_t \leftarrow \|M_{\text{input}} \odot M_{\text{temporal}}\|_{\text{mean}}$
▷ Classify head type: Boolean tensor [head]
$t_{\text{head}} \leftarrow (\text{Sim}_s < \alpha \cdot \text{Sim}_t)$
```

To address these challenges, we first propose an attention head classification strategy (Sec. 3.1) that decouples spatial and temporal parameters by analyzing attention sparsity in the pretrained Video DiT. Building on this, we introduce an efficient tuning framework (Sec. 3.2) to separately learn spatial appearance and temporal motion from the source video. While we use WAN as the pretrained backbone in our experiments, our method is model-agnostic and readily generalizes to other Video DiT architectures(See Appendix E.2).

3.1 Stage 1: Spatial-temporal attention classification

The pretrained video DiT model Wan utilizes unified 3D attention instead of separated spatial and temporal attention, which brings challenges to motion information decoupling (Pondaven et al., 2025), training efficiency, and storage cost (Pondaven et al., 2025). Inspired by evidence in previous work (Xi et al., 2025), we leverage the inherent sparsity in 3D Full Attention of video DiT to decouple the parameters for temporal motion and spatial appearance.

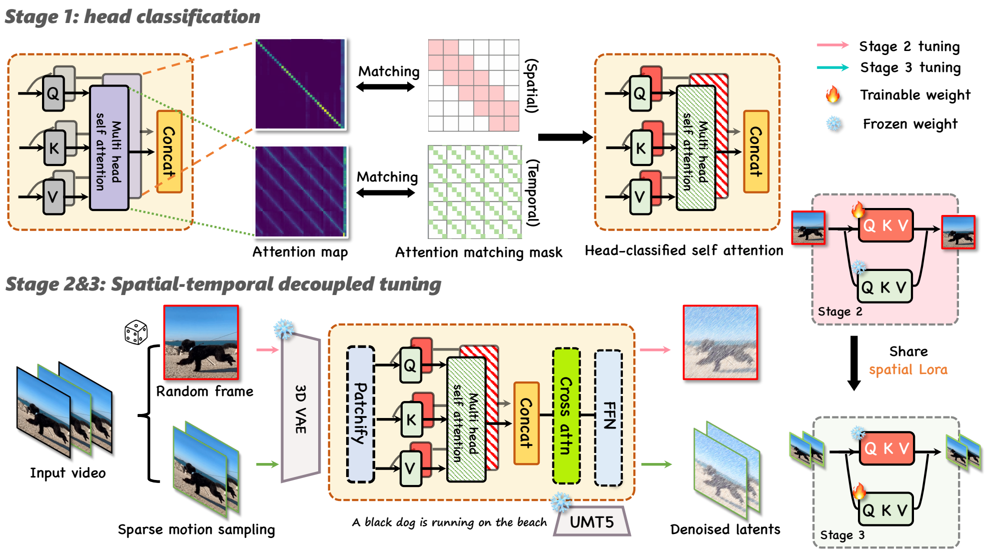


Figure 3: **Overview of our methods.** ***Stage 1***: We first classify the attention heads using a pseudo spatial attention map. ***Stage 2***: After attention classification, we first tune the spatial LoRA using a random frame in the video. ***Stage 3***: After finishing spatial LoRA tuning, we load the spatial LoRA weight and conduct temporal tuning using sparse motion sampling and adaptive RoPE.

Dual attention decoupling. As shown in Alg. 1, our method classifies attention heads in Wan into temporal or spatial types. We take query and key tokens $Q, K \in \mathbb{R}^{H \times S \times D}$ as input, where H is the number of heads, S is the sequence length, and D is the feature dimension.

We prepare pseudo ground truths: for spatial attention map, $M_{\text{spatial}}[i, j] = 1$ if points (i, j) are near the main diagonal (within a predefined range), otherwise 0; for temporal attention map, $M_{\text{temporal}}[i, j] = 1$ if points (i, j) are near diagonals parallel to the main diagonal (identical spatial positions in different frames), otherwise 0.

We compute the cosine similarity Sim_s between the input attention map M_{input} and M_{spatial}, and Sim_t between M_{input} and M_{temporal}. A head is classified as temporal if $\text{Sim}_s < \alpha \cdot \text{Sim}_t$, where $\alpha = 1.25$(empirically set to balance the number of spatial and temporal heads).

Dual attention fusion. Then, we rearrange the channels of the linear layers in full 3D attention q, k, v, o to two parallel branches for temporal attention and spatial attention. The forward algorithm of a single rearranged block is shown in Alg. 2. Given input sequence x, we concatenate the features from the temporal and spatial branches along the channel dimensions to get the tokens of query Q, key K, and value V. After applying rotary position embedding and scaled dot product attention, feature x is split along the channel dimension, and fed to o_{temp} and o_{spat}. Finally, the summed feature is returned at the end of the attention block.

3.2 Stages 2&3: Spatial-temporal decoupled tuning

Spatial LoRA tuning. As the parameters of the attention block are decoupled in the previous stage, we can use the spatial and temporal branches in two stages to learn the appearance and motion in the reference videos, respectively. Following previous work (Abdal et al., 2025; Zhao et al., 2023b), we first inject LoRAs θ_{spat} into the spatial heads branch $(q_{spat}, k_{spat}, v_{spat}, o_{spat})$ to learn the spatial appearance in stage 2. In each iteration, We randomly sample a single frame x_i from index $\{0, 1, 2, ..., F-1\}$, and optimize the spatial LoRA θ_{spat} as a text-to-image model using the training loss:

$$\mathcal{L}_{spat} = E_{x_{i,1} \sim P_{data}, x_{i,0} \sim N(0,I), i \sim U(0,F)} \left\| v_{i,t} - v_{\theta_{spat}} \left(x_{i,t}, t, p \right) \right\|_2^2, \tag{1}$$

where t is time step, p is positional embedding and v represents the velocity in the diffusion model(See Appendix C.

Temporal LoRA tuning. Once the spatial LoRA θ_{spat} gets converged, we freeze θ_{spat} in the model, and continue to finetune temporal LoRA parameters θ_{temp} of temporal heads branch

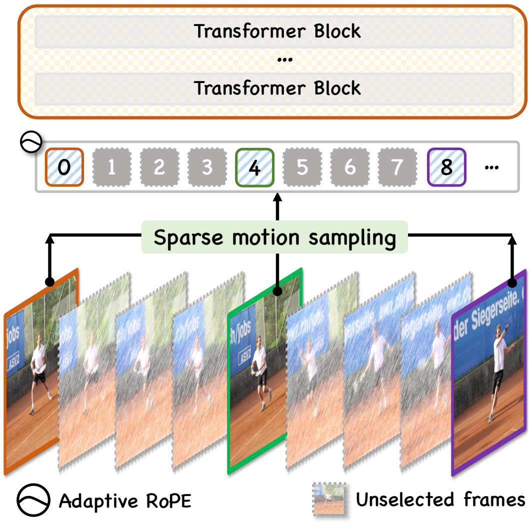


Figure 4: **Illustration of sparse motion sampling and adaptive RoPE**. The adaptive RoPE is utilized to represent frame position in the video.

Algorithm 2 Dual Attention Fusion

```
Input:  $x \in \mathbb{R}^{H \times S \times D}$: input sequence
        $f_{\text{regs}}$: positional frequencies
        $d_{\text{temp}}$: temporal dimension size
Output: Fused output $y \in \mathbb{R}^{H \times S \times D}$

▷ Channel concatenate, and normalize
$Q \leftarrow \text{Norm}([q_{\text{temp}}(x) \| q_{\text{spat}}(x)])$
$K \leftarrow \text{Norm}([k_{\text{temp}}(x) \| k_{\text{spat}}(x)])$
$V \leftarrow [v_{\text{temp}}(x) \| v_{\text{spat}}(x)]$

▷ Rotary Position Embeddings for Q, K
$\widetilde{Q} \leftarrow \text{RoPE}(Q, f_{\text{regs}}, H)$
$\widetilde{K} \leftarrow \text{RoPE}(K, f_{\text{regs}}, H)$

▷ Multi-Head Attention
$x \leftarrow \text{Attention}(\widetilde{Q}, \widetilde{K}, V; H)$

▷ Dual Output Projection Fusion
$y \leftarrow o_{\text{temp}}(x[: d_{\text{temp}}]) + o_{\text{spat}}(x[d_{\text{temp}} :])$
return $y$
```

$(q_{temp}, k_{temp}, v_{temp}, o_{temp})$. Since the Wan (Wang et al., 2025a) is pretrained on a large frame number $F = 81$, fine-tuning on the original number $F = 81$ costs too expensive computation (Fig. 2). To alleviate the high computation requirement for videos, we propose the ***sparse motion sampling***, which finetune our temporal LoRA θ_{temp} on a sampled video with fewer frame number $F_{samp} = 17$ and then infer with the original frame number. While recent transformer models apply Rotary Positional Embedding (RoPE) (Vaswani et al., 2017) to encode the relative position dependency according to the frame index, sampling frames from F to F_{samp} breaks the original dependency and thus deteriorates the motion quality. Motivated by previous text-to-image DiT models (Kong et al., 2024; Yang et al., 2024e), we propose the ***adaptive RoPE***, a centralized scaling positional encoding along the frame index to align the position range with different total frame numbers. For each frame with temporal index $i \in [0, 1, ..., F_{samp} - 1]$, its temporal positional embedding is assigned as:

$$\text{PE}_{x_i} = f(\frac{F}{2} + \frac{F}{F_{samp}}(i - \frac{F_{samp}}{2})), \tag{2}$$

which ensures that videos with less frame number F_{samp} have the same input range $[0, F]$ for the embedding function f, as the pertaining stage of video DiT.

To further decouple the temporal motion from spatial appearance, we further introduce a motion loss (Ling et al., 2024) by eliminating the appearance and focusing on the changes in the temporal dimensions. We first define the motion latent $\hat{v}$ for each frame i as: $\hat{v}_{i,t} = v_{i,t} - v_{i-1,t}$.

Then, we define the motion loss following (Zhao et al., 2023b) as the negative cosine similarity between the ground truth motion latent and predicted motion latent:

$$\mathcal{L}_{Motion} = E_{x_{i,1} \sim P_{data}, x_{i,0} \sim N(0,I), i \sim U(0,F)}[1 - \text{CosineSim}(\hat{v}_{i,t}, \hat{v}_{\theta temp}(x_{i,t}, t, p))]. \tag{3}$$

Finally, the total loss for temporal LoRAs is the combination of general video denoising loss and motion loss as $\mathcal{L}_{temp} = \mathcal{L}_{\text{video\_denoise}} + \mathcal{L}_{Motion}$.

4 Experiments

4.1 Implementation details

In our experiment, we employ the open-sourced video generation model WAN-2.1 (Wang et al., 2025a) as the base text-to-video generation model. The LoRA ranks are 128 in both stages. We

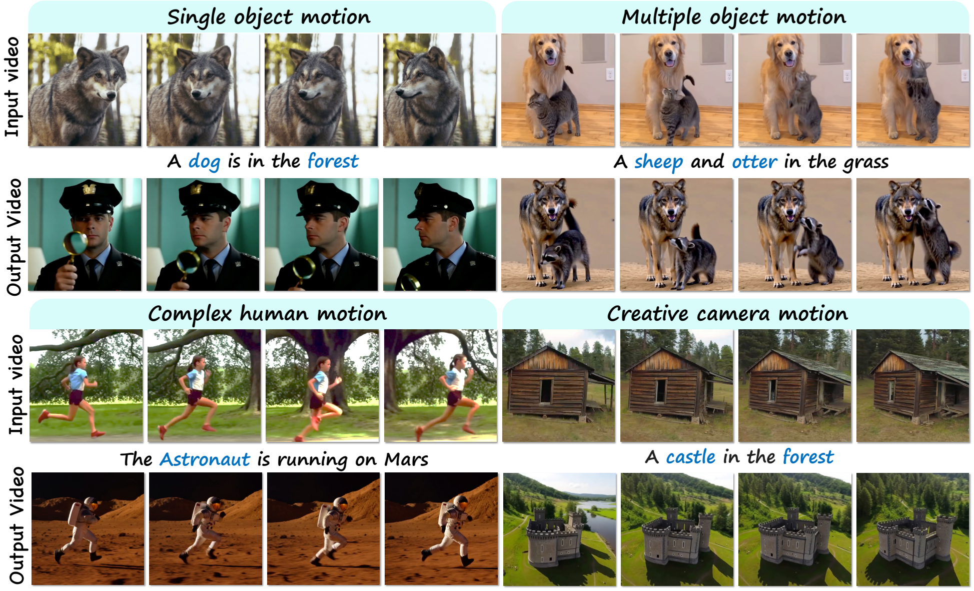


Figure 5: **Gallery of our proposed methods.** Given a reference video, our EffiVMT capability of generating a high-quality video clip with the same motion, including single object motion, multiple object motion, complex human motion, and camera motion.

first randomly select a single frame and take about 3,000 steps for spatial appearance learning. The AdamW (Loshchilov & Hutter, 2017) optimizer is utilized, and the learning rate is 1×10^{-5}. The spatial weight decay is 0.1. During the third tuning stage, we freeze the spatial head LoRA and only train the temporal head LoRA for 2000 steps with learning rate 1×10^{-5} and weight decay 0.99. More details and evaluation metrics can be found in Appendix A.

4.2 MotionBench

In order to address the lack of a benchmark in video motion transfer, we introduce MotionBench, a comprehensive benchmark to evaluate the ability of current motion transfer approaches. In detail, we collect 200 videos from four aspects, including 1). camera motion, 2). single object motion, 3). multiple object motion, and 4). complex human motion. Single object motion sequences focus on diverse motion patterns from a single subject. Multiple object motion involves the consistency of spatial relationships between different instances. Camera motion evaluates viewpoint changes through both simple camera trajectories (zoom, tilt, pan) and complex camera operations. Here, single/multi-object refers to general objects and animals, while human motion contains more non-rigid deformations, so we treat it separately. The 30% videos in our benchmark are generated by text-to-video generation models (Wang et al., 2025a), and other videos are obtained from publicly licensed video websites. We use the GPT4o (OpenAI, 2024) to get the video captions. In MotionBench, each video is approximately 5 seconds long with 150 frames. MotionBench provides a standardized evaluation protocol across diverse motion categories, enabling systematic assessment and comparison of motion transfer methods.

4.3 Comparison with baselines

In the following paragraphs, we qualitatively and quantitatively compare our method with previous state-of-the-art methods. We also apply their methods to Wan-2.1 (Wang et al., 2025a) and CogVideo (Yang et al., 2024e) for fair comparison (See Appendix E.2).

Qualitative comparison. We compare our approach with previous video motion transfer methods visually, including state-of-the-art video motion transfer methods: MOFT (Xiao et al., 2024b), MotionInversion (Wang et al., 2024a), MotionClone (Ling et al., 2024), SMM (Yatim et al., 2024), MotionDirector (Zhao et al., 2023b), DiTFlow (Pondaven et al., 2024). We exclude Motionshop (Yesiltepe et al., 2024) and MotionCrafter (Zhang et al., 2023b) from our comparisons as no public release

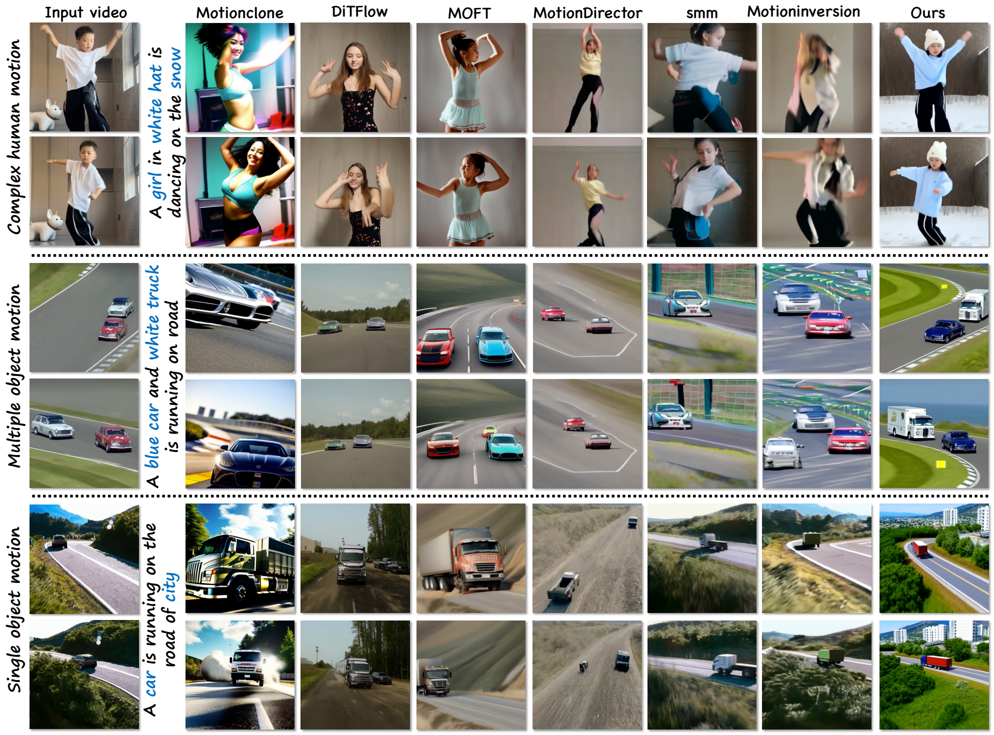


Figure 6: **Qualitative comparison with baselines.** We perform the visual comparison with various baselines using various kinds of motions. Our method obtains better performance in various motions.

Table 1: **Comparison with state-of-the-art video motion transfer methods**. **Red** and **Blue** denote the best and second best results, respectively.

Method	Quantitative Metrics				User Study			
	Text Sim.↑	Motion Fid.↑	Temp. Cons.↑	Time (s)↓	Motion Pres.↓	App.↓	Text Align.↓	Overall↓
Training-free methods								
MOFT (Xiao et al., 2024b)	0.286	0.792	0.922	1230	6.883	6.631	5.894	6.639
MotionClone (Ling et al., 2024)	0.302	0.831	0.901	1015	6.283	5.874	6.642	4.192
SMM (Yatim et al., 2024)	0.279	**0.932**	0.918	775	4.350	5.086	4.205	5.883
DiTFlow (Pondaven et al., 2025)	**0.375**	0.807	**0.941**	**712**	3.326	**2.417**	**2.215**	3.284
Tuning-based methods								
MotionInversion (Jeong et al., 2024b)	0.295	0.831	0.771	2315	5.417	3.295	5.117	5.074
MotionDirector (Zhao et al., 2023b)	0.292	0.896	0.939	3008	**2.217**	4.208	3.298	**2.216**
Ours	**0.380**	**0.971**	**0.976**	**727**	**1.123**	**1.335**	**1.174**	**1.132**

exists. Our experimental results exhibit EffiVMT better performance and versatility across diverse motion transfer scenarios. As illustrated in Fig. 6, in single object motion cases (first column), we find that the previous works fail to follow source motion. In contrast, our approaches effectively transform the motion from the source video into the target object, maintaining a consistent motion pattern. For multi-object cases, MotionDirector (Zhao et al., 2023b) and SMM (Yatim et al., 2024) have the challenge of handling multi-object interaction motion. Our method enables generating videos with aligned movement patterns, preserving the spatial relationships between moving subjects. Additionally, we provide a visual comparison of complex camera motion. The visual results demonstrate the superiority of our methods in camera motion transfer capabilities.

Quantitative comparison. We compare our method with state-of-the-art video motion transfer on our MotionBench, and the results are shown in Tab. 1. Due to the limited video length of previous works, all evaluations are performed in 32 frames at a resolution of 512 × 512. Here, we classify the SOTA methods as two classes, training-free or tuning-based, according to whether they use spatial/temporal LoRA to optimize complex motion patterns. (a) **Time**: Thanks to sparse motion sampling, EffiVMT is the fastest tuning-based method. Moreover, our running time is on par with training-free approaches while delivering superior performance. (b) **Motion Fidelity**: Following (Yatim et al., 2024), motion fidelity is applied to evaluate tracklet similarity between reference and output videos. (c) **Temporal Consistency**: We evaluate the average frame-to-frame coherence using CLIP (Radford et al., 2021) feature similarity among consecutive video frames. (d)**Text similarity**: We use CLIP to extract target video features and compute the average cosine similarity between the input prompt and all video frames. (f) **User study**: Since automatic metrics often fail to reflect real preferences, we invited 20 volunteers to rank methods on MotionBench across four aspects including motion preservation, appearance diversity, text alignment, and overall quality from 1 (best) to 7(worst). The average rank per method (lower is better) is shown in Tab. 1 (1=best, 7=worst). Our method achieves the top result in both automatic metrics and human preference.

4.4 Ablation Study

In this section, we conduct a systematic ablation study to isolate and quantify the contribution of each key component in our framework. The qualitative and quantitative ablation study results are shown in Fig. 7 and Tab. 2, respectively. More ablation studies can be found in the Appendix F.

Table 2: **Quantitative ablation**. **Red** and **Blue** denote best, 2nd. Baseline means we disables all three proposed components simultaneously.

Method	Text Sim.↑	Motion Fid.↑	Temp. Cons.↑	Time(s)↓
Baseline	0.362	0.658	0.824	2493
w/o STD LoRa	0.364	0.546	0.845	971
w/o Adaptive RoPE	0.371	0.655	0.817	**792**
w/o Sparse Sampling	**0.369**	**0.975**	**0.967**	2068
Ours	**0.380**	**0.971**	**0.976**	**727**

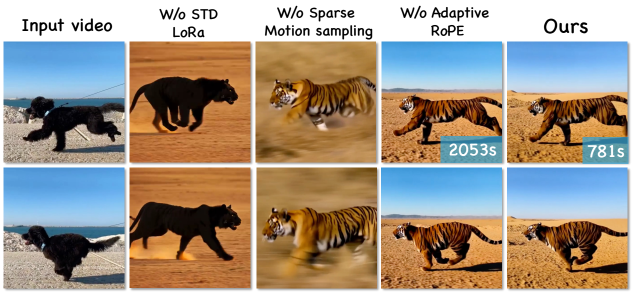


Figure 7: **Ablation study about proposed modules**. We remove the proposed modules to evaluate their effectiveness. "STD" means spatial–temporal decoupled LoRA.

Effectiveness of spatial–temporal decoupled LoRA. As shown in the second row of Fig. 7 and the "w/o STD LoRA" ablation in Tab. 2, the naive baseline jointly tunes without separating spatial and temporal attention heads, failing to decouple the dog's appearance and causing the edited tiger to look unnaturally black. In contrast, our decoupled LoRA preserves motion while effectively modifying appearance, as evidenced by the improved text similarity in Tab. 2.

Effectiveness of adaptive RoPE. Thanks to our adaptive RoPE design, the model can precisely infer each sampled frame's original index under sparse motion sampling, ensuring the edited motion remains aligned with the source. Without adaptive RoPE, the tiger's motion becomes disordered and fails to match the original video dynamics. In Tab. 2, an improvement of about 48.3% over motion fidelity, quantitatively confirms the benefit of our adaptive RoPE.

Effectiveness of sparse motion sampling. By employing sparse motion sampling in the temporal tuning phase, we reduce the tuning time to 727s. Note that in the "w/o sparse sampling" setting, we still apply adaptive RoPE but tune on all video frames, resulting in identical motion fidelity (0.975 *vs.* 0.971) while incurring the higher time cost.

5 Conclusion

In this paper, we propose EffiVMT, a three-stage video motion transfer framework that tunes the video diffusion transformer to synthesize video clips with complex motion. In detail, we analyze the motion inefficiency and tuning inefficiency in DiT-based video motion transfer. Through the proposed efficient spatial-temporal decoupled LoRA, we achieve better motion consistency. To address the tuning inefficiency, we introduce adaptive RoPE and sparse motion sampling to accelerate training. Extensive experimental results demonstrate the effectiveness of our method, which outperforms a wide range of previous methods, achieving state-of-the-art video motion transfer quality.

Reproducibility Statement

All quantitative tables, qualitative images, and video results in this work are reproducible and correspond to raw model outputs without manual editing or post-hoc alteration, except for minimal format conversion and compression. After the review process, we will release a partial public repository to support reproduction, including inference scripts, example data, and example videos under **CC-BY-NC-4.0**. The datasets, configurations, and procedures used for training and evaluation are documented in Section 4.1 and Appendix D. User study participants were compensated, gave informed consent, and could withdraw at any time. All visual results in the paper and demo are unedited.

Ethics Statement

Our work studies motion-transfer video editing and has social potential impact as shown in Appendix. L. The proposed dataset contains videos of people, vehicles, and landscape camera motions. To mitigate representational bias in demonstrations, we curated and display examples spanning different races, genders, and styles in the main text and appendix. All illustrative videos shown in this paper are sourced from publicly available web content; we respect the original licenses and terms of service and use the content solely for research purposes. We will not publicly release the dataset prior to completing the insertion of AI-generated watermarks and an ethics/content-safety audit. We explicitly prohibit harmful or deceptive uses of our methods and data, including deepfake attacks and other malicious generative behaviors. When any portion of our code is made public, we will enforce visible and/or machine-detectable watermarking during inference to help deter misuse. Any future releases will be accompanied by usage terms that forbid impersonation, harassment, or other malicious applications, and we will remove or restrict content that raises privacy, legal, or safety concerns.

References

Rameen Abdal, Or Patashnik, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Dynamic concepts personalization from single videos. *arXiv preprint arXiv:2502.14844*, 2025.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, 2023.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, 2023.

Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *ICCV*, 2023.

Mengting Chen, Xi Chen, Zhonghua Zhai, Chen Ju, Xuewen Hong, Jinsong Lan, and Shuai Xiao. Wear-any-way: Manipulable virtual try-on via sparse correspondence alignment. In *ECCV*, 2024a.

Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, 2024b.

Xi Chen, Zhifei Zhang, He Zhang, Yuqian Zhou, Soo Ye Kim, Qing Liu, Yijun Li, Jianming Zhang, Nanxuan Zhao, Yilin Wang, et al. Unireal: Universal image generation and editing via learning real-world dynamics. In *CVPR*, 2025.

Yuren Cong, Mengmeng Xu, Christian Simon, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, and Sen He. Flatten: optical flow-guided attention for consistent text-to-video editing. *arXiv preprint arXiv:2310.05922*, 2023.

Yutao Cui, Xiaotong Zhao, Guozhen Zhang, Shengming Cao, Kai Ma, and Limin Wang. Stabledrag: Stable dragging for point-based image editing. In *ECCV*, 2024.

Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

Yuchao Gu, Yipin Zhou, and Mike Zheng et al. Videoswap: Customized video subject swapping with interactive semantic point correspondence. *arXiv preprint arXiv:2312.02087*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

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. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhen Han, Zeyinzi Jiang, Yulin Pan, Jingfeng Zhang, Chaojie Mao, Chenwei Xie, Yu Liu, and Jingren Zhou. Ace: All-round creator and editor following instructions via diffusion transformer. *arXiv preprint arXiv:2410.00086*, 2024.

Yin-Yin He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022. URL https://api.semanticscholar.org/CorpusID:257631986.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. In *ICLR*, 2023.

Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789*, 2024.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv:2404.09990*, 2024.

Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. *arXiv preprint arXiv:2310.01107*, 2023.

Hyeonho Jeong, Jinho Chang, Geon Yeong Park, and Jong Chul Ye. Dreammotion: Space-time self-similarity score distillation for zero-shot video editing. *arXiv preprint arXiv:2403.12002*, 2024a.

Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. 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 (CVPR)*, pp. 9212–9221, June 2024b.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *ECCV*, 2024.

Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. *ACM TOG*, 40(6):1–12, 2021.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *CVPR*, 2023.

Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *CVPR*, 2024.

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.

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 e-prints*, pp. arXiv–2403, 2024.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, 2024a.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024b.

Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024. URL https://arxiv.org/abs/2406.05338.

Yotam Lipman, Belinda Tzen, Yihan Zhang, Yang Song, and Stefano Ermon. Rectified flow: Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2209.15571.

Hongyu Liu, Xuan Wang, Ziyu Wan, Yue Ma, Jingye Chen, Yanbo Fan, Yujun Shen, Yibing Song, and Qifeng Chen. Avatarartist: Open-domain 4d avatarization. In *CVPR*, 2025a.

Shaoteng Liu, Tianyu Wang, Jui-Hsien Wang, Qing Liu, Zhifei Zhang, Joon-Young Lee, Yijun Li, Bei Yu, Zhe Lin, Soo Ye Kim, et al. Generative video propagation. In *CVPR*, 2025b.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, Lifang He, and Lichao Sun. Sora: A review on background, technology, limitations, and opportunities of large vision models. *ArXiv*, abs/2402.17177, 2024. URL https://api.semanticscholar.org/CorpusID:268032569.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *ICML*, 2023a.

Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. In *NeurIPS*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yue Ma, Xiaodong Cun, Yingqing He, Chenyang Qi, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Magicstick: Controllable video editing via control handle transformations. *arXiv preprint arXiv:2312.03047*, 2023.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 4117–4125, 2024.

Yue Ma, Xiaodong Cun, Sen Liang, Jinbo Xing, Yingqing He, Chenyang Qi, Siran Chen, and Qifeng Chen. Magicstick: Controllable video editing via control handle transformations. In *WACV*, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *CVPR*, 2023.

Chong Mou, Xintao Wang, Jiechong Song, Ying Shan, and Jian Zhang. Dragondiffusion: Enabling drag-style manipulation on diffusion models. In *ICLR*, 2024.

Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *Advances in Neural Information Processing Systems*, 37:18481–18505, 2025.

OpenAI. GPT-4o technical report. Technical report, OpenAI, 2024. URL https://openai.com/research/gpt-4o. Accessed: 2025-05-12.

Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. *arXiv preprint arXiv:2308.07926*, 2023.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, and Van Gool et al. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 724–732, 2016.

Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers. *arXiv preprint arXiv:2412.07776*, 2024. URL https://arxiv.org/abs/2412.07776.

Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers. In *CVPR*, 2025.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10318–10327, 2021.

Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. *arXiv preprint arXiv:2303.09535*, 2023.

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*, pp. 8748–8763. PMLR, 2021.

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*, pp. 332–349. Springer, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

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 *CVPR*, 2023.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *CVPR*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025a.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningying Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Rui Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wen-Chao Zhou, Wente Wang, Wen Shen, Wenyuan Yu, Xianzhong Shi, Xiaomin Huang, Xin Xu, Yan Kou, Yan-Mei Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhengbin Han, Zhigang Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *ArXiv*, abs/2503.20314, 2025b. URL https://api.semanticscholar.org/CorpusID:277321639.

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.

Luozhou Wang, Ziyang Mai, Guibao Shen, Yixun Liang, Xin Tao, Pengfei Wan, Di Zhang, Yijun Li, and Yingcong Chen. Motion inversion for video customization. *arXiv preprint arXiv:2403.20193*, 2024a. URL https://arxiv.org/abs/2403.20193.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv:2401.07519*, 2024b.

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Weixian Lei, Yuchao Gu, 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. *arXiv preprint arXiv:2212.11565*, 2022.

Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse videogen: Accelerating video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340*, 2024a.

Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b. URL https://arxiv.org/abs/2405.14864.

Chenxi Xie, Minghan Li, Shuai Li, Yuhui Wu, Qiaosi Yi, and Lei Zhang. Dnaedit: Direct noise alignment for text-guided rectified flow editing. *arXiv preprint arXiv:2506.01430*, 2025.

Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 2024a.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024b.

Zhen Xiong, Yuqi Li, Chuanguang Yang, Tiao Tan, Zhihong Zhu, Siyuan Li, and Yue Ma. Enhancing image generation fidelity via progressive prompts. *arXiv preprint arXiv:2501.07070*, 2025.

Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024a.

Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv:2403.01779*, 2024b.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, 2023a.

Ling Yang, Bohan Zeng, Jiaming Liu, Hong Li, Minghao Xu, Wentao Zhang, and Shuicheng Yan. Editworld: Simulating world dynamics for instruction-following image editing. *arXiv:2405.14785*, 2024a.

Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *ACM SIGGRAPH Asia Conference Proceedings*, 2023b.

Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Eva: Zero-shot accurate attributes and multi-object video editing. *arXiv e-prints*, pp. arXiv–2403, 2024b.

Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. Dgl: Dynamic global-local prompt tuning for text-video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6540–6548, 2024c.

Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Videograin: Modulating space-time attention for multi-grained video editing. *arXiv preprint arXiv:2502.17258*, 2025.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *ArXiv*, abs/2408.06072, 2024d. URL https://api.semanticscholar.org/CorpusID:271855655.

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*, 2024e.

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*, pp. 8466–8476, 2024.

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. URL https://arxiv.org/abs/2412.05355.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10459–10469, 2023.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NeurIPS*, 2024.

Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023a.

Yabo Zhang, Xinpeng Zhou, Yihan Zeng, Hang Xu, Hui Li, and Wangmeng Zuo. Framepainter: Endowing interactive image editing with video diffusion priors. *arXiv preprint arXiv:2501.08225*, 2025.

Yuxin Zhang, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Motioncrafter: One-shot motion customization of diffusion models. *arXiv preprint arXiv:2312.05288*, 2023b. URL https://arxiv.org/abs/2312.05288.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. In *NeurIPS*, 2024.

Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023a.

Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. *arXiv preprint arXiv:2310.08465*, 2023b.

Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Xiu Li. Multibooth: Towards generating all your concepts in an image from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10923–10931, 2025.

Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *ECCV*, 2024.

APPENDIX

A Implementation Details

For sparse motion sampling, we set the sampling stride to 5. The input videos are fed into the model as 512×512. The sampled frame number is 16 in the second stage. In the inference stage, we leverage the flow matching scheduler (Lipman et al., 2023) with a sampling step of 30, and a text-guidance ratio of 7.0. The LoRA weights are set as 0.5. For the user study, to achieve a more comprehensive evaluation of human preferences in video quality, we perform the user study with four aspects. *Motion preservation* assesses the motion's adherence between reference videos and generated ones. *Appearance diversity* measures the diversity according to the reference video. *Text alignment* means the semantic alignment between generated videos and prompts. *Overall* assesses the subjective quality of the generated videos. We invite 20 volunteers to provide human feedback. The questionnaire includes 30 cases about our method and other baselines. The volunteers are asked to rank the video clips in terms of the performance of various motion transfer results. (The smaller the score, the better; 1 point is the best.). Then, we calculate the average result for each baseline.

B Related Work

Diffusion-based video editing. Image editing is a broad and impactful field with diverse applications. Early works (Mokady et al., 2023; Meng et al., 2022; Kawar et al., 2023; Cao et al., 2023; Hertz et al., 2023) explore training-free or fine-tuning-based methods to modify image attributes via text prompts. Subsequent approaches (Zhao et al., 2024; Hui et al., 2024; Zhang et al., 2024; Yang et al., 2024a) advance instruction-based editing by training on curated datasets. A line of research explores additional control signals, such as masked regions (Zhuang et al., 2024; Ju et al., 2024), compositing content (Chen et al., 2024b; Yang et al., 2023a; Song et al., 2023), customized ID with reference images (Li et al., 2024a; Ruiz et al., 2023; Liu et al., 2023a;b; Kumari et al., 2023; Kim et al., 2024; Chen et al., 2024a; Xu et al., 2024b; Li et al., 2024b; Wang et al., 2024b), drag points (Cui et al., 2024; Mou et al., 2024). However, most of these works are limited to single editing tasks, making them inadequate for diverse real-world application scenarios. To address these limitations, unified frameworks (Chen et al., 2025; Xiao et al., 2024a; Han et al., 2024) are introduced to support various image editing and generation tasks. Recent advances in video editing can be categorized into two main approaches based on their underlying architectures. **Image-based methods** typically extend pretrained text-to-image models to the video domain. Tune-A-Video (TAV) (Wu et al., 2022) pioneeRed this direction by adapting latent diffusion models for spatial-temporal generation through one-shot tuning. Subsequent works (Qi et al., 2023; Ceylan et al., 2023; Ma et al., 2025) improved temporal consistency through attention map fusion during inversion. Alternative approaches relying on Neural Atlas (Kasten et al., 2021), dynamic NeRF deformation fields (Pumarola et al., 2021; Chai et al., 2023; Ouyang et al., 2023), optical flow, (Yang et al., 2023b; Cong et al., 2023; Zhang et al., 2023a), feature aggregation (Geyer et al., 2023; Jeong & Ye, 2023) significantly mitigate the temporal inconsistency issue. At the same time, they still suffer from artifacts when handling videos with large motions. **Video-based methods** leverage emerging video foundation models (Yu et al., 2023; Guo et al., 2024; Yang et al., 2024e) to overcome some limitations of image-based approaches. Prior research efforts (Gu et al., 2023; Mou et al., 2025; Liu et al., 2025b; Ku et al., 2024) demonstrate improved capabilities in motion transfer and editing by exploiting rich motion priors on single tasks. Recent works also investigate the merit of unified video generation and editing frameworks (Jiang et al., 2025).

C Preliminaries: Low-Rank Adaptation for Video Diffusion Transformer Model

Video diffusion models. Following pioneering Latent Diffusion Model (Chai et al., 2023), video diffusion models first compress the input video V in pixel space into a latent space $x = \mathcal{E}(V)$ utilize a pretrained encoder $\mathcal{E}$, where the latent space x can be reconstructed back to pixel space video by a decoder $\mathcal{D}$. The encoder $\mathcal{E}$ and decoder $\mathcal{D}$ are built with causal 3D convolution blocks, which can encode single-frame images and multi-frame videos into the same latent space. The size of a video latent x is $F \times C \times W \times H$, where F, C, W, H stand for the video length, latent channels, width, and height, respectively.

Recent video diffusion models (Wang et al., 2025a) leverage flow matching to formulate the diffusion and denoising process in the latent space. During straining, a timestep $t \in [0, 1]$ is sampled from a logit-normal distribution, and the intermediate latent x_t is defined as the linear interpolation between image or video latent x_1 and a random noise $x_0 \in \mathcal{N}(0, I)$ as

$$x_t = tx_1 + (1 - t)x_0. \tag{4}$$

The velocity v_t is further defined as

$$v_t = \frac{dx_t}{dt} = x_1 - x_0 \tag{5}$$

The diffusion models (Wang et al., 2025a; Kong et al., 2024) take intermediate latents x_t as input and are trained to estimate the velocity v_t using mean squaRed error loss.

$$\min_{\theta} E_{x_1, x_0 \sim N(0,I)} \left\| v_t - v_\theta \left(x_t, t, p \right) \right\|_2^2, \tag{6}$$

where p is embedding the text description for the input clean video.

The inference stage starts from a Gaussian noise x_0, then the pretraiend diffusion model gradually removes the noise in N discrete timesteps $t = t_N, ..., t_0$ as $x_{t_{i-1}} = x_{t_i} + (t_{i-1} - t_i)v_\theta(x_{t_i}, t_i)$ Finally, the pRedicted latent x_1 is decoded to pixel space by the pretrained decoder $\mathcal{D}$.

Diffusion Transformers and Low-Rank Adaptation. Recently, Diffusion Transformer (DiT) (Wang et al., 2025b) demonstrated better motion consistency and visual quality over the previous UNet (Chai et al., 2023) backbone in text-to-video generation. In the DiT model architecture v_θ, the noisy latent x_t is first divided into patches of size $P \times P$, and then rearranged into the token sequence of shape $(F \cdot \frac{H}{P} \cdot \frac{W}{P}) \times D$ with token dimensionality D.

The patchified latent token sequence is fed into a stack of N DiT blocks (Wang et al., 2025a). In each block, latent tokens are processed by feedforward layers and multi-head self-attention layers, while text embedding p is injected through the cross-attention block (Wang et al., 2025a) or the multimodal self-attention block.

To preserve spatial relationships between patches during attention computation, a positional embedding $PE = f(i)$ is introduced. This embedding captures the positions i of patches within the sequence and conditions the denoising process $v_\theta(x_t, p, t, PE)$. Different positional encoding methods (Vaswani et al., 2017) can be applied, including adding PE to the input patches at the initial stage of v_θ directly, or incorporating it into the attention mechanism by rotating query and key vectors (Kong et al., 2024).

To alleviate the high computation cost of video DiT (e.g, WAN (Wang et al., 2025a) has 14B parameters), low-rank adaptation (LoRA) has been applied in downstream fine-tuning and appearance customization (Ma et al., 2024; Jeong et al., 2024a). Specifically, LoRA proposes to optimize a Low-Rank factorized residual ΔW of the parameters as

$$W = W_0 + \Delta W = W_0 + BA, \tag{7}$$

where $W_0 \in \mathcal{R}^{d \times k}$ is the weights of the attention block in the pretrained model, $B \in \mathcal{R}^{d \times r}$ and $A \in \mathcal{R}^{r \times d}$ are factors where r is much smaller than d and k so the updated parameters are Reduced compaRed with optimizing the whole model.

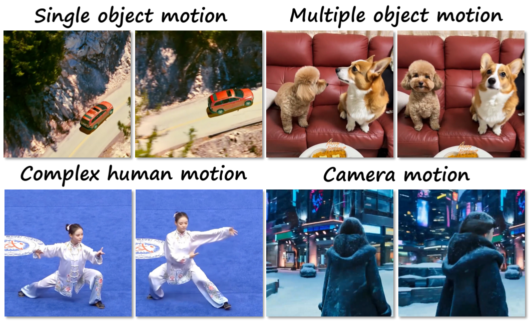


Figure 1: **MotionBench**. We collect four categories of motion, including single object motion, multiple object motion, complex human motion, and camera motion.

D MotionBench

D.1 Benchmark Construction

In Fig. 1, the 30% videos in our benchmark are generated by text-to-video generation models (Kong et al., 2024; Wang et al., 2025b). The other videos are obtained from publicly licensed video websites. We also use the GPT4o (OpenAI, 2024) to get the video captions. Note that the video clips with excessive motion or overly large subjects are manually filteRed out, which often exhibit noticeable blur. Each caption is about 20 words, and each video is approximately 5 seconds long with 150 frames.

MotionBench contains **200 videos**, categorized as:

- Camera motion: 49 videos
- Single object motion: 52 videos
- Multiple object motion: 51 videos
- Complex human motion: 48 videos

All selections were **random** after automated filtering without cherry-picking. The benchmark includes both real and synthetic videos, each paiRed with GPT-4o and human-written prompts for semantic diversity. Such a benchmark with various motions would be beneficial for the development of the community.

D.2 Comparison with Established Benchmarks

MotionBench improves upon DAVIS and FIVE (Xie et al., 2025) by:

- Covering more motion types (including camera and separated multi-object motion)
- Larger scale (200 vs. FIVE's 100)
- Inclusion of synthetic data and diverse prompts
- Randomized, automated selection pipeline

E Additional Experimental Results

E.1 Results on DAVIS Dataset

For comprehensive comparison, We report the results of different baselines and ours on the DAVIS dataset in Tab. 1. These results demonstrate that our method generalizes well not only on our MotionBench but also DAVIS.

Table 1: **Comparison on DAVIS dataset (following DiTFlow's protocol)**. We randomly select 50 high-quality videos from the DAVIS dataset (Perazzi et al., 2016). **Red** and **Blue** denote the best and second best results, respectively.

Methods	Text Sim. ↑	Motion FID ↑	Temp. Cons. ↑	Time (s) ↓
Training-free methods				
MOFT	0.244	0.659	0.884	1267
MotionClone	0.249	0.680	0.859	1049
SMM	0.333	**0.765**	0.883	795
DiTFlow	0.318	0.680	**0.914**	**734**
Training-based methods				
MotionInversion	0.239	0.697	0.726	2429
MotionDirector	**0.349**	0.727	0.901	3104
Ours (EffiVMT)	**0.424**	**0.883**	**0.936**	**762**

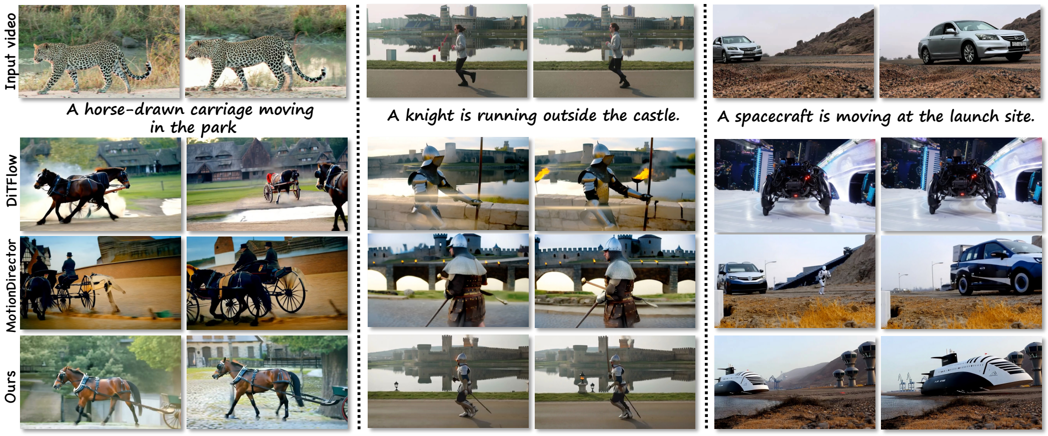

Input video
A horse-drawn carriage moving in the park
A knight is running outside the castle.
A spacecraft is moving at the launch site.
DiTFlow
MotionDirector
Ours

Figure 2: **More visual comparisons with WAN-2.1.** For fair comparison, we present the qualitative comparison with baselines using WAN-2.1 (Wang et al., 2025a).

E.2 Fair Comparison Using WAN-2.1 and CogVideo Backbone

We re-implement DiTFlow and MotionDirector with WAN-2.1 and CogVideo backbone for fair comparison. The results are shown in Tab. 9 and Tab. 4, respectively. The performance gain of EffiVMT is influenced by the backbone's motion modeling capability:

- On WAN-2.1 (strong motion prior): Motion FID gain = 0.971 - 0.931 = **0.04**
- On CogVideoX (weaker motion prior): Motion FID gain = 0.944 - 0.928 = **0.016**

This validates that our sparse motion sampling and adaptive RoPE benefit more from stronger base models. Future work will explore adaptation to multimodal DiTs (e.g., CogVideoX's expert attention).

Table 2: **Comparison with ReVideo**.

Methods	Text Sim. ↑	Motion FID ↑	Temp. Cons. ↑	Time (s) ↓
ReVideo	0.247	0.793	0.882	1013
Ours (EffiVMT)	**0.380**	**0.971**	**0.976**	**727**

Table 3: **Comparison using WAN-2.1 backbone** (re-implemented baselines). We select two SOTA training-free/training-based approaches, DiTFlow (Pondaven et al., 2025) and MotionDirector (Zhao et al., 2023b) for fair comparison. **Red** and **Blue** denote the best and second best results, respectively.

Methods	Text Sim. ↑	Motion FID ↑	Temp. Cons. ↑	Time (s) ↓
DiTFlow	**0.369**	0.872	0.947	**713**
MotionDirector	0.352	**0.931**	**0.963**	4641
Ours (EffiVMT)	**0.380**	**0.971**	**0.976**	**727**

Across both backbones, our method consistently achieves the top results. On WAN-2.1, it yields a Motion FID Reduction of 0.040 relative to the strongest baseline (from 0.971 to 0.931), while on CogVideo2 it Reduces Motion FID by 0.022 (from 0.944 to 0.922). These gains, alongside improvements in Text Similarity and Temporal Consistency and competitive or faster generation time, indicate superior intent adherence, motion stability, and efficiency. The larger improvement under the stronger motion prior (WAN-2.1) further suggests that our approach better exploits backbone motion priors, validating the effectiveness of our design across diverse generative settings.

Table 4: **Comparison using CogVideoX backbone**. Performance gap is smaller due to CogVideoX's weaker motion modeling capability. **Red** and **Blue** denote the best and second best results, respectively.

Methods	Text Sim. ↑	Motion FID ↑	Temp. Cons. ↑	Time (s) ↓
DiTFlow	**0.371**	0.813	0.937	**716**
MotionDirector	0.343	**0.928**	**0.952**	4287
Ours (EffiVMT)	**0.373**	**0.944**	**0.963**	**732**

Table 5: **Extended evaluation using VBench metrics and Warp Error**. **Red** and **Blue** denote the best and second best results, respectively.

Methods	Subj. Consis. ↑	Temp. Flicker ↑	Motion Smooth ↑	Overall Consis. ↑	Warp Err. ↓
Training-based methods					
MOFT	0.7527	0.7438	0.7041	0.1932	4.62
MotionClone	0.7619	0.7821	0.7628	0.2315	3.22
SMM	0.7845	0.7764	0.7543	0.2087	2.89
DiTFlow	0.8128	0.8236	0.8017	0.2213	**2.26**
Training-free methods					
TokenFlow	0.8314	0.8217	0.8114	0.2106	2.86
StreamV2V	0.8125	0.8169	0.8251	0.1987	3.13
MotionInversion	0.8425	**0.8673**	**0.8515**	0.2326	2.31
MotionDirector	**0.8763**	0.8432	0.8423	**0.2418**	2.57
Ours	**0.9113**	**0.8931**	**0.8842**	**0.2915**	**1.74**

E.3 Additional Metrics: VBench and Warp Error

We follow SMM (Yatim et al., 2024) in not using warp error as a primary metric, as it cannot evaluate structural deviations in motion transfer tasks. However, for completeness, we report it in Tab. 5 and our method achieves the lowest error. Across all VBench metrics, our method consistently achieves the top performance, indicating superior subject consistency and Reduced temporal flicker relative to prior work. Training-based methods exhibit weaker overall VBench scores and higher warp error.

F More Ablation Studies

F.1 Attention Head Assignment Strategy

Table 6: **Ablation on attention head assignment** (random vs. pseudo-label based).

Methods	Text Sim. ↑	Motion FID ↑	Temp. Cons. ↑
Random Assigning	0.317	0.922	0.864
Ours (Pseudo-label)	**0.380**	**0.971**	**0.976**

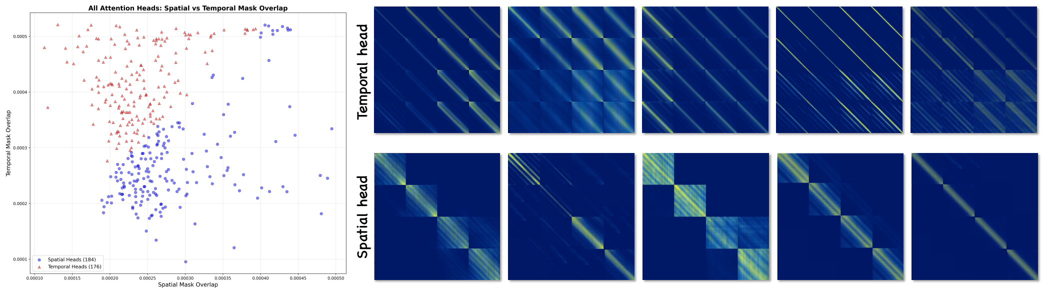
Rationale for pseudo-labels: InspiRed by U-Net's decoupled attention, we derive pseudo-labels $M_{spatial}$ and $M_{temporal}$ from DiT's attention maps:

- $M_{spatial}$: High activation within same/adjacent frames → focuses on spatial structure.
- $M_{temporal}$: High activation at same positions across frames → focuses motion trajectory.

This guides head classification without manual annotation. We also compare our head classification method with random assigning, the results are shown in Tab. 6.

Table 7: **Ablation on alpha** (weighting factor for head classification).

Alpha	Text Sim. ↑	Motion FID ↑	Temp. Cons. ↑
0.75	0.347	0.925	0.923
1.00	0.362	0.948	0.957
1.25	**0.380**	**0.971**	**0.976**
1.50	0.374	0.962	0.958
1.75	0.368	0.957	0.954



(a) Overlap between temporal head and spatial head.

(a) Visualization between temporal head and spatial head.

Figure 3: **Visualization and overlap illustration**. We note that most of the attention heads are easy to separate effectively.

F.2 Hyperparameter Sensitivity of Alpha

Alpha,the weighting factor for head classification, is dataset-dependent; we found 1.25 optimal for MotionBench. The detailed results for different Alpha value are shown in Tab. 7. The ablation study on the hyperparameter α reveals a clear unimodal trend across evaluation metrics: performance improves as α increases from small to moderate values, peaks at $\alpha = 1.25$, and then declines as α grows further. At $\alpha = 1.25$, the model attains its best overall results. These findings indicate that α materially affects generation quality, including intent adherence, motion fidelity, and temporal coherence.

G Additional Metrics: VBench and Warp Error

H Comparison with Additional Baselines

In Tab. 8, we compare with two more baselines(VideoComposer and SMA) on MotionBench. As shown in the Tab. 8, our method surpasses competing approaches and achieves the best performance on Text Similarity, Motion FID, Temporal Consistency, and Time. These results demonstrate that, relative to other baselines, our approach attains state-of-the-art intent adherence, motion consistency, and generation speed.

I Visualized Comparison with Baseline

In Fig. 8, we provide qualitative comparisons across four representative single-object motion (a swan is swimming), camera motion (a blockhouse captuRed with an arc shot), multiple-object motion (a camel is turning and a panda is sniffing), and complex human motion (a spiderman is climbing on a wall) to assess motion fidelity, temporal coherence, and intent adherence. Prior approaches (e.g., DiTFlow, MotionDirector, MOST, VMD, MotionDriver) frequently exhibit inconsistent motion patterns, including foreground trajectory jitter and identity drift, inaccurate or drifting camera control with unstable parallax, asynchronous or conflicting dynamics between multiple objects, and discontinuous human pose transitions with temporal flicker and limb distortions. In contrast, our method produces: (i) stable single-object trajectories with smooth velocities and preserved appearance, (ii) precise and temporally consistent camera motion (e.g., a smooth arc shot) with

Table 8: Comparison with VideoComposer and SMA.

Methods	Text Sim. ↑	Motion FID ↑	Temp. Cons. ↑	Time (s) ↓
VideoComposer	0.354	0.942	0.963	1103
SMA	0.358	0.935	0.956	3216
Ours	**0.380**	**0.971**	**0.976**	**727**

Table 9: **Ablation with sparse sampling ratio**.The input video is 81 frame. We select 1/2,1/4,1/8,1/10 of the video frames for ablation. **Red** and **Blue** denote best, 2nd.

Methods	Text Sim. ↑	Motion FID ↑	Temp. Cons. ↑	Time (s) ↓
1/2	**0.379**	**0.970**	**0.975**	912
1/4	**0.380**	**0.971**	**0.976**	727
1/8	0.345	0.943	0.951	**583**
1/10	0.328	0.915	0.938	**374**

coherent background alignment and realistic parallax, (iii) coordinated, temporally aligned behaviors for multiple subjects without cross-object interference, and (iv) plausible complex human motion with consistent body geometry, articulated limb kinematics, and realistic contact dynamics. Overall, these results indicate that our approach effectively resolves motion inconsistency observed in prior work, yielding coherent dynamics and faithful adherence to the textual intent across diverse motion regimes.

J More Results

We show more video motion transfer results produced by our method in an MP4 file, which can be found in the file: `demo.mp4`. The accompanying video further demonstrates our method's motion transfer capabilities across a broad spectrum of scenarios, including single-object motion, camera motion, multiple-object motion, and complex human motion. The demonstrations span natural landscapes, animals, vehicles, close-up facial footage, and architectural cinematography, highlighting robust camera control and high temporal coherence. The video also presents ablation analyses and side-by-side comparisons with prior methods. Together, these comprehensive examples substantiate the effectiveness of our approach and underscore its advantages over competing solutions, clearly conveying the quality, consistency, and intent adherence of the generated results.

K Limitation

- As a tuning-based method, our method optimizes LoRA for each input video. CompaRed to tuning-free methods, LoRA tuning is more time-consuming but can generalize to more complex motion.
- Since our base model WAN has more learnable parameters than previous video diffusion models, optimizing WAN with LoRA requires more training cost. In the future, this issue will be addressed with better base models and more acceleration strategies.

L Social Potential Impact

The development of EffiVMT, an advanced framework for video motion transfer using a spatial-temporal decoupled LoRA, holds significant social potential across various domains. By facilitating the generation of complex motions in videos, this technology can greatly enhance creative industries such as film, animation, and gaming, allowing artists and creators to easily produce high-quality, dynamic content that was previously time-consuming and resource-intensive to achieve.

Moreover, the introduction of MotionBench as a benchmark will promote standardization and collaboration within the research community, driving further advancements in video diffusion technologies.

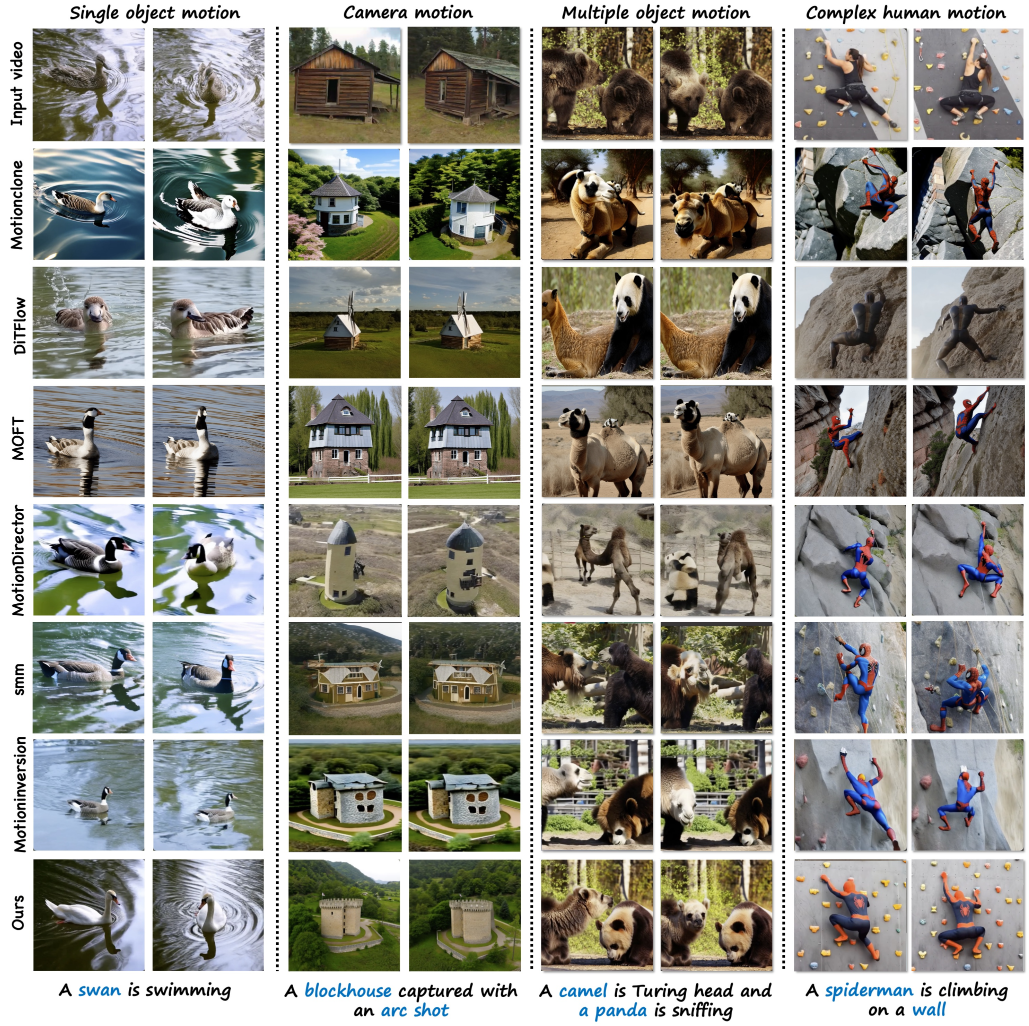


Figure 4: More comparisons of our methods against baselines on motion of objects and cameras.

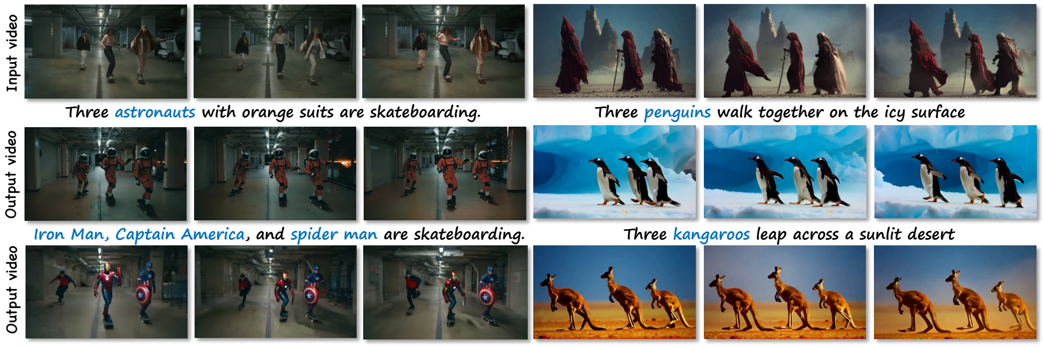

Input video
Three astronauts with orange suits are skateboarding.
Three penguins walk together on the icy surface
Output video
Iron Man, Captain America, and spider man are skateboarding.
Three kangaroos leap across a sunlit desert
Output video

Figure 5: **More object cases.** We are able to animate multiple objects with consistent motion.

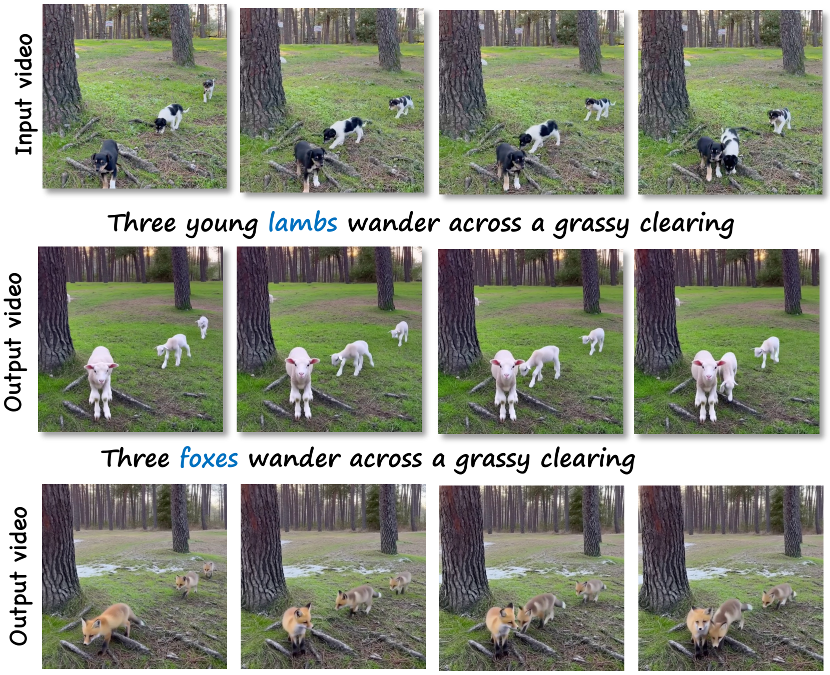

Input video
Three young lambs wander across a grassy clearing
Output video
Three foxes wander across a grassy clearing
Output video

Figure 6: **More object cases**. We are able to animate multiple objects with consistent motion.

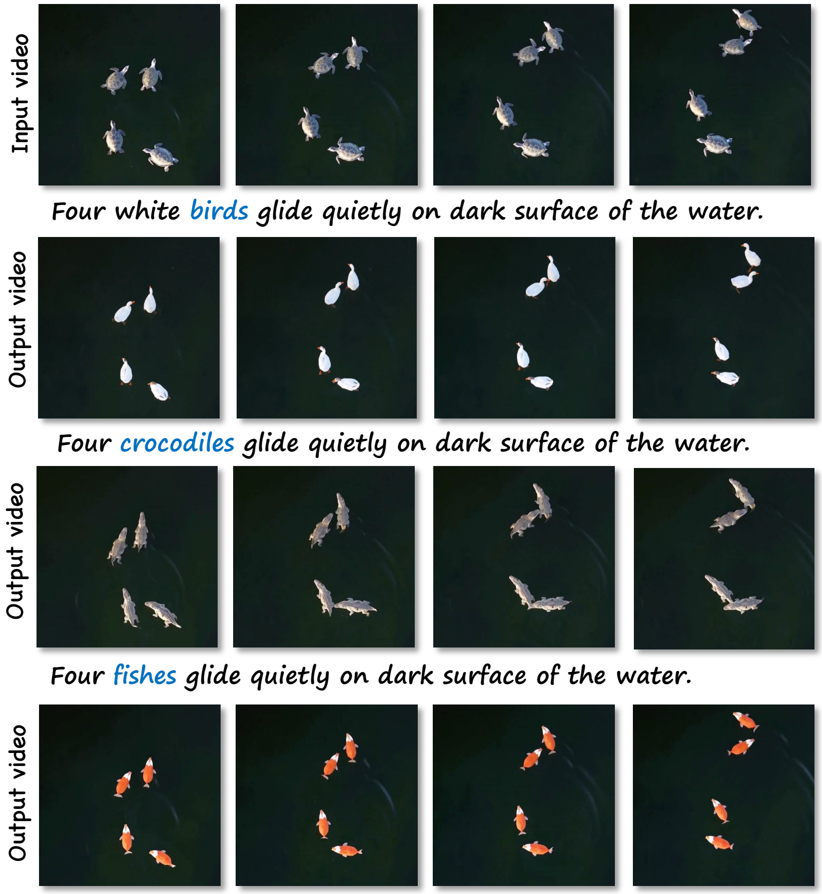

Input video
Four white birds glide quietly on dark surface of the water.
Output video
Four crocodiles glide quietly on dark surface of the water.
Output video
Four fishes glide quietly on dark surface of the water.
Output video

Figure 7: **More object cases**. We are able to animate multiple objects with consistent motion.

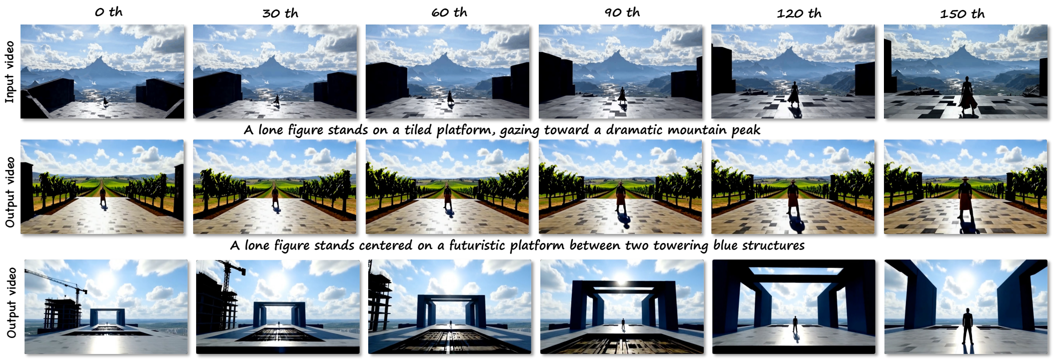

0 th
30 th
60 th
90 th
120 th
150 th
Input video
A lone figure stands on a tiled platform, gazing toward a dramatic mountain peak
Output video
A lone figure stands centered on a futuristic platform between two towering blue structures
Output video

Figure 8: **Longer video**. We are able to achieve longer video motion transfer.

This could lead to improved applications in fields like education, where engaging video content can enhance learning experiences, or in virtual reality, enabling more immersive environments.

Additionally, the efficiency of EffiVMTcan democratize access to high-quality video production, making it more accessible to individuals and small businesses, fostering innovation and creativity in the digital landscape. However, as with any advanced technology, it is vital to consider ethical implications and ensure responsible use to prevent potential misuse in areas such as misinformation or deepfakes. Overall, EffiVMThas the potential to significantly impact not only artistic fields but also education, virtual experiences, and the broader digital economy.

M The Usage of Large Language Models

In this paper, the usage of the LLM mainly falls into the following aspects:

- **Grammar checking and format optimization**: In the paragraphs of the paper, LLMs are used for grammar error checking and format checking of charts and graphs.
- **Language polishing**: The text description part of the paper uses LLMs to polish and optimize the language expression.
- All authors are responsible for the content generated by the LLMs.