arXiv:2506.05207v2 [cs.CV] 13 Aug 2025

Follow-Your-Motion: Video Motion Transfer via Efficient Spatial-Temporal Decoupled Finetuning

Yue Ma[1]*, **Yulong Liu**[1]*, **Qiyuan Zhu**[1]*, **Ayden Yang**, **Kunyu Feng**[2], **Xinhua Zhang**[3], **Zhifeng Li**[4], **Sirui Han**[1]†, **Chenyang Qi**[1]†, **Qifeng Chen**[1]

[1]HKUST, [2]HKUST(GZ), [3]Tsinghua Univerisity, [4]Tencent

https://follow-your-motion.github.io

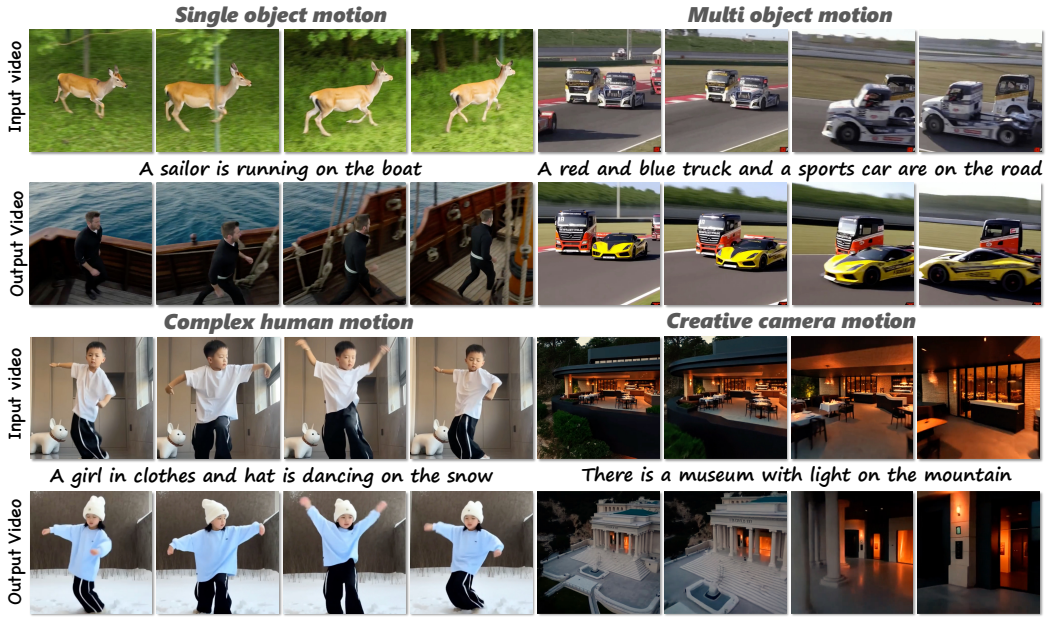


Figure 1: **Showcases of our Follow-Your-Motion**. Given an input video, Follow-Your-Motion enables generating the video with the same motion, including motion of single or multiple objects, complex poses of humans, and movements of the camera view.

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. Additionally, they require time-consuming fine-tuning processes in both stages. To tackle these issues, we propose Follow-Your-Motion, an efficient two-stage video motion transfer framework that finetunes a powerful video diffusion transformer to synthesize complex motion. Specifically, we propose a spatial-temporal decoupled LoRA to decouple the attention architecture for spatial appearance and temporal motion processing. During the second training stage, we design the sparse motion sampling and adap-

*Equal contribution. † Corresponding author.

Preprint. Under review.

tive 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 Follow-Your-Motion.

1 Introduction

Motion transfer aims to synthesize new videos with the same motion as the reference video. Unlike video-to-video translation [28, 39], motion transfer focuses on extracting and reproducing camera view changes and object motions, rather than preserving the low-level and 2D spatial structure. This task has potential applications in diverse fields, including the film industry, augmented reality, automated advertising systems, and social media content creation.

In recent years, the diffusion model [31] has demonstrated remarkable capabilities in generating high-quality visual content due to its stability when optimized for Gaussian noise removal objectives. In addition, Diffusion Transformer (DiT) has further shown its superiority in scalability of model parameters, computation, and large-scale video data. Utilizing pretrained video diffusion models, various motion transfer techniques are proposed using training-free or fine-tuning paradigms. Training-free motion transfer methods [7, 27, 28, 41, 47] replace or optimize the generated motion representation with the reference video in inference time, while freezing the parameters of neural networks. *e.g.*, . SMM [50] proposes a new descriptor by averaging video features along spatial dimensions. Motionshop [51] uses the latent updates in the denoising pipeline as Motion Score representation for DiT models. Although these training-free methods avoid the network training cost and generalize in both early UNet and recent DiT models, their generated motion is bottlenecked by the knowledge of the pretrained diffusion model.

To generate complex motions beyond the domain of the pretrained diffusion model, tuning-based motion transfer [57] has emerged to learn the motion in the reference video precisely by optimizing the model parameters. In the training stage, early UNet-based methods (MotionDirector) optimize the parameters of separated temporal layers to learn the input motion. In the inference stage, new videos are synthesized by combining the motion learned in the optimized parameters and prior knowledge in the frozen foundation model. Despite the improvements in motion transfer quality based on UNet models, applying tuning-based motion transfer to modern DiT is still challenging because of its high training cost and spatial-temporal mixed computation in its 3D self-attention blocks. In the context of tuning DiT for motion transfer, a naive baseline is to finetune the whole DiT model using Low-Rank Adaptation (LoRA) in the 3d self-attention layer directly.

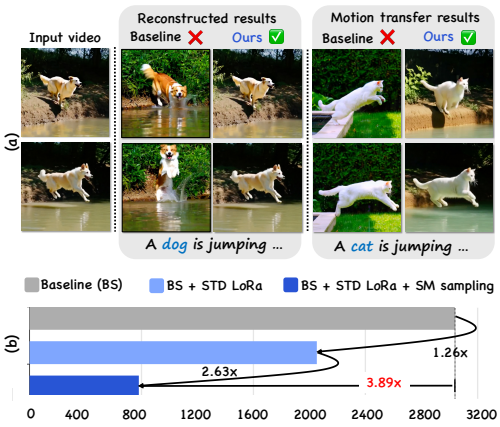


Figure 2: **Comparison between Follow-Your-Motion and baseline**. We finetune the baseline and our method 3,000 steps using Wan2.1 [34]. Our method gets better reconstruction and motion preservation.

Tuning-based motion transfer, including U-Net [57] and DiT [1] employ a spatial–temporal decoupled tuning paradigm: They first optimize a spatial LoRA on sampled frames to capture spatial structure, then freeze those weights and optimize a temporal LoRA over the full video frames to learn motion dynamics. However, we argue that this two-stage procedure is inherently inefficient. Specifically, the limitations are listed as follows:

(1) ***Motion inconsistency***: In the first stage, both spatial and temporal heads in self-attention are tuned on a fixed frame, which leads to spatial-temporal coupling. As shown in the top Fig. 2, the naive baseline tunes both spatial and temporal heads, but both the reconstructed results and motion transfer results fail to follow the dog's original motion. Therefore, tuning both spatial and temporal heads for appearance preservation is not reasonable.

(2) ***Tuning inefficiency***: Motivated by the observation in [40], the 3D full attention mechanism in DiT-based models contains attention heads that specialize in either spatial or temporal modeling. However, existing methods tune all attention heads in the spatial/temporal tuning stages, regardless of their roles, leading to inefficient parameter usage and redundancy. Moreover, since 3D VAEs compress temporal sequences into compact latent representations, using all reference frames for motion transfer ignores 3D VAEs' interpolation ability and introduces redundancy.

To address these challenges, we propose Follow-Your-Motion, an efficient video motion transfer framework. Specifically, to mitigate the motion inconsistency problem, we first apply robust head matching to categorize attention heads into spatial and temporal types. During tuning, only spatial heads are updated in the spatial stage, and only temporal heads in the temporal stage. As shown in Figure 2 (a), our method preserves the motion consistency in both reconstruction and transfer results compared to the naive baseline. To address tuning inefficiency, we introduce sparse motion sampling during the temporal tuning stage to accelerate. Furthermore, we propose adaptive RoPE to enhance the model's ability to learn motion interpolation, allowing it to effectively capture motion dynamics even from sparse frames. As evidence in Figure 2 (b), our decoupled tuning strategy reduces latency by 1.26×, and with sparse sampling, achieves a 3.89× speed-up compared to tuning with all frames.

With the aforementioned designs, our approach can effectively generate high-quality videos following the motion in the reference video, as illustrated in Figure 1. Additionally, to address the lack of a benchmark in video motion transfer, we introduce a comprehensive benchmark called MotionBench, which consists of 53 various video motions, including single object motion, complex human motion, multi-object motion, and camera motion that showcase a wide range of scenes and styles. Compared with the existing baseline methods, we achieve remarkable results on 8 various evaluation metrics on MotionBench. Our method can facilitate more accurate video motion transfer using powerful DiT. Overall, our contributions can be summarized as follows:

- We propose Follow-Your-Motion, a two-stage video motion transfer framework that tunes the powerful video diffusion transformer to synthesize video clips with complex motion.
- We analyze motion inconsistency and tuning inefficiency in DiT-based video motion transfer. To ensure motion consistency, we separately tune spatial and temporal LoRA heads for appearance and motion, respectively. To address tuning inefficiency, we introduce adaptive RoPE and sparse motion sampling to accelerate training.
- 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 [8, 21, 24, 52, 56, 54, 22, 58, 17, 46, 33, 3, 36] are proposed to synthesize consistent results using a pretrained image diffusion model. Previous works [4, 5, 9, 10, 37, 23, 44, 35, 6, 59] 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 [18], CogVideoX [49], EasyAnimate [45], HunyuanVideo [14], and Wan2.1 [34], 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 [20, 21, 42, 43, 53] 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 [11, 27, 51], 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 [12, 57] always fine-tune model parameters to utilize different attention for temporal

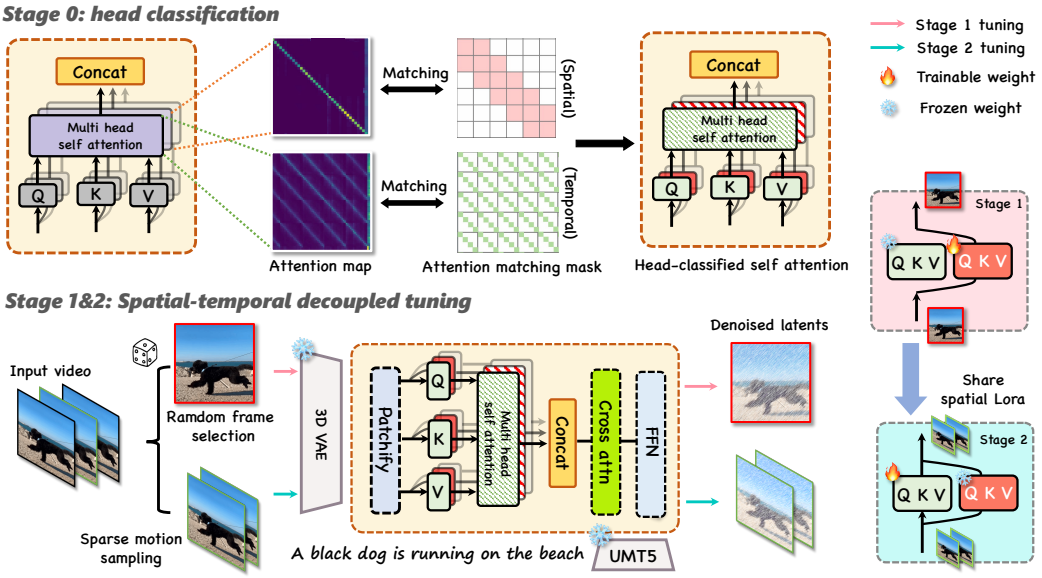


Figure 3: **Overview of our methods.** ***Stage 1***: We first classify the attention heads using a pseudo spatial attention map. ***Stage 2***: After attention classification, we first tune the spatial LoRA using a random frame in the video. ***Stage 3***: After finishing spatial LoRA tuning, we load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

and spatial information. Current works [12, 30, 57] employ the dual-path LoRA structure to separate motion and appearance. However, these methods are developed on the UNet-based pretrained model [2], 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

We target video motion transfer from the input source video to the new synthesized one by optimizing generative models to learn the object motions and camera pose changes in the source video. As proposed in previous work [1, 57], a naive baseline (Fig. 2) of motion transfer is to first optimize the LoRA weight for sampled video frames as a text-to-image task, where the LoRA obtained in the first stage is called spatial LoRA (ΔW_s). In the second stage, consecutive multiple frames are treated as ground truth to optimize the temporal LoRA (ΔW_t), while the spatial LoRA is loaded and kept frozen. During the inference stage, only spatial lora (ΔW_t) is loaded to transfer the video motion in the source video while presenting the appearance bias leakage. Although this strategy can generate results for the UNet-based diffusion model, the computation cost for the DiT-based video diffusion model is still *too expensive*. As shown in Fig. 2, such naive LoRA tuning cannot reproduce the motion in the input video after tuning 3000 steps (3042 seconds on a single H20 GPU). 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.

To address these challenges, in Sec 3.1, we first present an attention head classification strategy to decouple the spatial and temporal model parameters by analyzing the sparsity of attention in the pretrained video diffusion transformer. Then we introduce an efficient and economical tuning framework (Sec. 3.1) for learning spatial appearance and temporal motion in the source video. Note that we use WAN as our pretrained backbone in our experiments, but our method can generalize to other video DiT models.

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 [27], training efficiency, and storage cost [27]. Inspired by evidence in previous work [40], we leverage the inherent sparsity in 3D Full Attention of video DiT to decouple the parameters for temporal motion and spatial appearance.

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 $\alpha = 1.3$(empirically set to balance the number of spatial and temporal heads).

Algorithm 1: Dual attention decoupling

Input: $Q, K, \in \mathbb{R}^{H\times S\times D}$: query and key where sequence $S = F \times H \times W$
Output: Closest head type: t_{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, K)
▷ Calculate similarity metrics
$\text{Sim}_s \leftarrow \|M_{\text{input}} \odot M_s\|$ .mean(dim $= (1,2)$)
$\text{Sim}_t \leftarrow \|M_{\text{input}} \odot M_t\|$ .mean(dim $= (1,2)$)
▷ Classify head type: Boolean tensor [head]
$t_{\text{head}} \leftarrow (\text{Sim}_s < \alpha\text{Sim}_t)$

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 [1, 57], 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 1. We randomly sample a batch of frames 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:

$$\mathcal{L}_{spat} = E_{x_1,x_0\sim N(0,I),i\sim U(0,F)} \left\| v_{t,i} - v_{\theta_{spat}}(x_{t,i}, t, p) \right\|_2^2. \quad (1)$$

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_{spat} of temporal heads branch $(q_{temp}, k_{temp}, v_{temp}, o_{temp})$. Since the Wan [34] 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 spatial LoRA θ_{spat} 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) [32] 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 [14, 48], 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

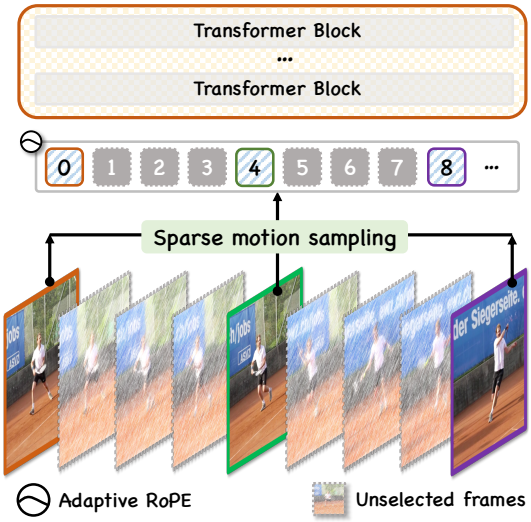


Figure 4: **Illustration of sparse motion sampling and adaptive RoPE**. The adaptive RoPE is utilized to represent frame position in the video.

Algorithm 2: Dual Attention Fusion

Input: $x \in \mathbb{R}^{H\times S\times D}$: input sequence
f_{regs}: positional frequencies
d_{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

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 [15] 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 the difference between the current frame latent v_i and the previous frame latent v_{i-1} as: $\hat{v}_i = v_{t,i} - v_{t,i-1}$.

Then, we define the motion loss following [57] as the negative cosine similarity between the ground truth motion latent and predicted motion latent:

$$\mathcal{L}_{Motion} = E_{x_1,x_0\sim N(0,I),i\sim U(0,F)}[1 - \text{CosineSim}(\hat{v}_{t,i}, \hat{v}_\theta\,(x_{t,i}, t, p))]. \tag{3}$$

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 [34] as the base text-to-video generation model. The LoRA ranks are 128 in both stages. We first randomly select a single frame and take about 3,000 steps for spatial appearance learning. The AdamW [19] optimizer is utilized, and the learning rate is 1×10^{-5}. The spatial weight decay is 0.1. During the second 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. 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 [16] with a sampling step of 30, and a text-guidance ratio of 7.0. The LoRA weights are set as 0.5. More details and evaluation metrics can be found in the supplementary materials.

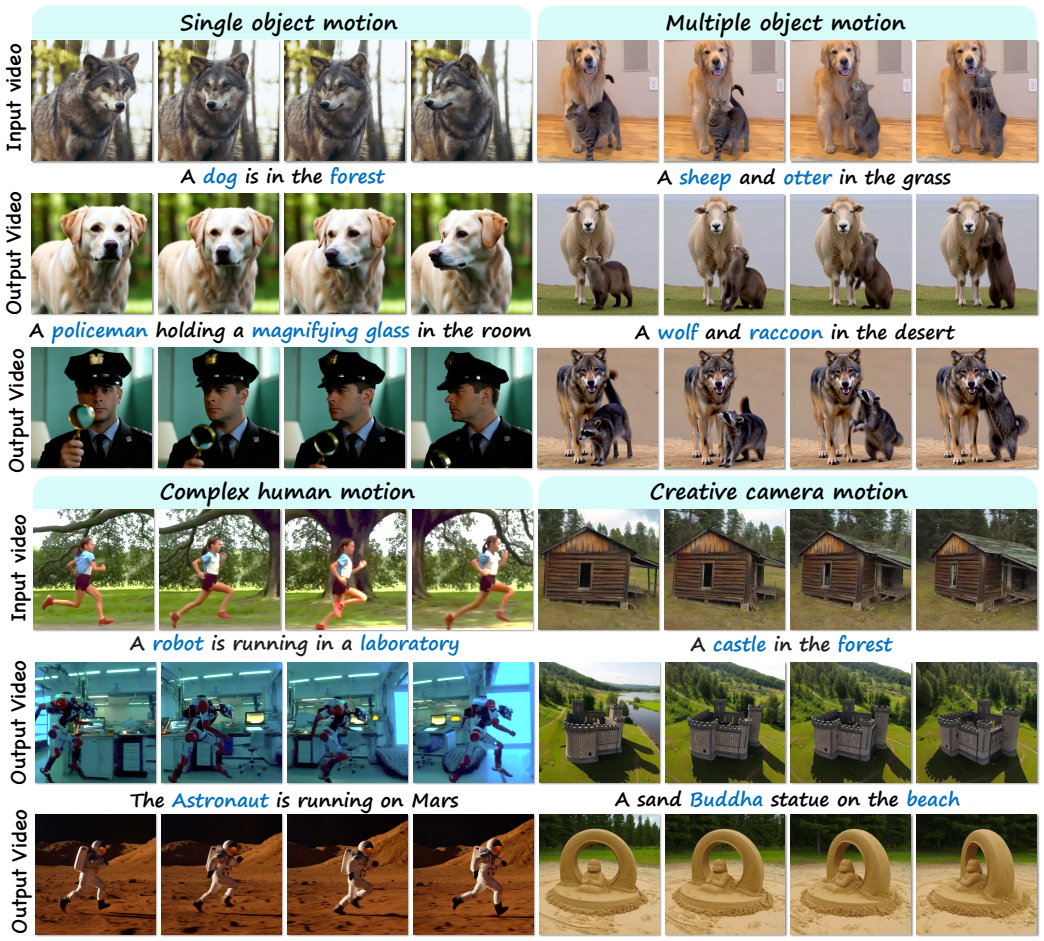


Figure 5: **Gallery of our proposed methods.** Given a reference video, our Follow-Your-Motion capability of generating a high-quality video clip with the same motion, including single object motion, multiple object motion, complex human motion, and camera motion.

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 53 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 [34], and other videos are obtained from publicly licensed video websites. We use the GPT4o [25] 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

Qualitative comparison. We compare our approach with previous video motion transfer methods visually, including state-of-the-art video motion transfer methods: MOFT [41], MotionInversion [38], MotionClone [15], SMM [50], MotionDirector [57], DiTFlow [26]. We exclude Motionshop [51] and MotionCrafter [55]. from our comparisons as no public release exists. Our experimental results exhibit Follow-Your-Motion better performance and versatility across diverse motion transfer scenarios. As

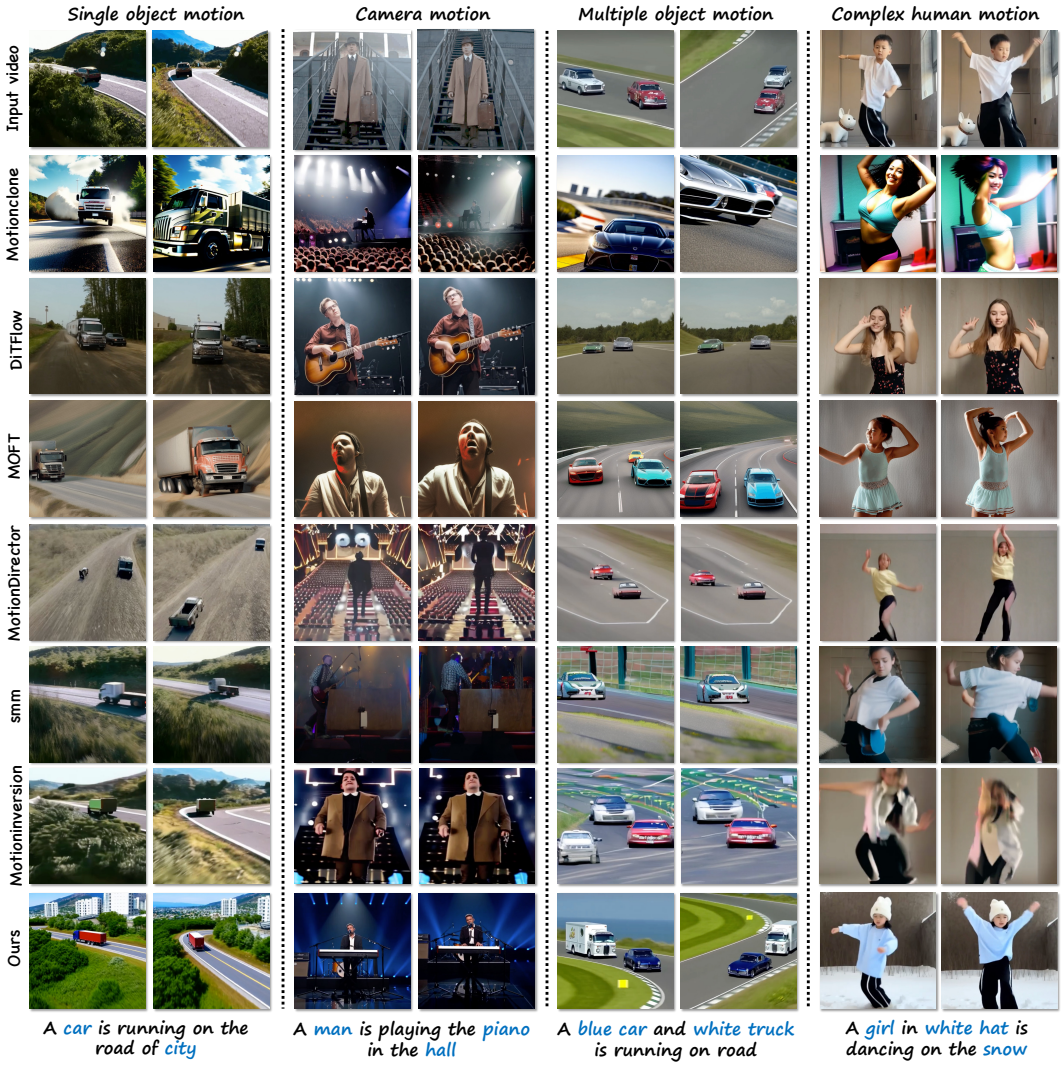


Figure 6: **Qualitative comparison with baselines.** We perform the visual comparison with six baselines using various kinds of motions. Our method obtains better performance in various motions.

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 [57] and SMM [50] 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, Follow-Your-Motion 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 [50], 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 [29] feature similarity among consecutive video frames. (d)**Text similarity**: We use CLIP to

Table 1: **Comparison with state-of-the-art video motion transfer methods**. **Red** and **Blue** denote the best and second best results, respectively. User study scores report the average rank (1=best, 7=worst) for all the methods; lower is better.

Method	Quantitative Metrics				User Study			
	Text Sim.↑	Motion Fid.↑	Temp. Cons.↑	Time (s)↓	Motion Pres.↓	App.↓	Text Align.↓	Overall↓
Training-Free Methods								
MOFT [41]	0.283	0.786	0.928	1241	6.883	6.631	5.894	6.639
MotionClone [15]	0.298	0.825	0.906	1028	6.283	5.874	6.642	4.192
SMM [50]	0.276	**0.927**	0.924	785	4.350	5.086	4.205	5.883
DiTFlow [27]	**0.371**	0.813	0.937	**714**	3.326	**2.417**	**2.215**	3.284
Tuning-Based Methods								
MotionInversion [13]	0.301	0.838	0.764	2359	5.417	3.295	5.117	5.074
MotionDirector [57]	0.285	0.903	**0.943**	3022	**2.217**	4.208	3.298	**2.216**
Ours	**0.375**	**0.976**	**0.971**	**781**	**1.123**	**1.335**	**1.174**	**1.132**

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—motion preservation, appearance diversity, text alignment, and overall quality—on a 1 (best) to 7 scale. The average rank per method (lower is better) is shown in Tab. 1. 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.

Table 2: Quantitative ablation. **Red** and **Blue** denote best, 2nd.

Method	Text Sim.↑	Motion Fid.↑	Temp. Cons.↑	Time(s)↓
w/o STD LoRa	0.365	0.542	0.842	983
w/o Adaptive RoPE	0.368	0.658	0.814	788
w/o Sparse Sampling	**0.369**	**0.979**	**0.968**	2053
Ours	**0.372**	**0.976**	**0.971**	**781**

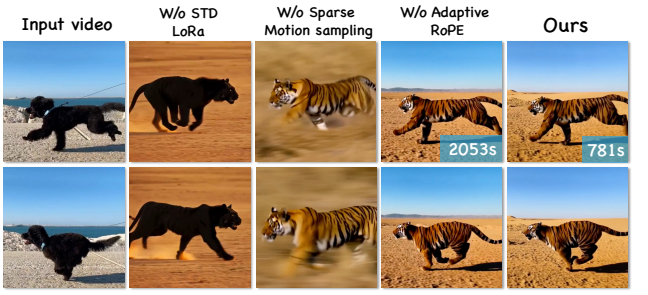


Figure 7: **Ablation study about proposed modules**. We remove the proposed modules to evaluate their effectiveness. "STD" means spatial–temporal decoupled LoRA.

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 from 2053s to 781s—an approximately 2.6× speed-up. 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 (0.979 *vs.* 0.976) while incurring the higher time cost.

5 Conclusion

In this paper, we propose Follow-Your-Motion, a two-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.

References

[1] 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.

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

[3] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*, 2024.

[4] Siran Chen, Yue Ma, Yu Qiao, and Yali Wang. M-bev: Masked bev perception for robust autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1183–1191, 2024.

[5] Siran Chen, Qinglin Xu, Yue Ma, Yu Qiao, and Yali Wang. Attentive snippet prompting for video retrieval. *IEEE Transactions on Multimedia*, 26:4348–4359, 2023.

[6] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4edit: Diffusion transformer for image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2969–2977, 2025.

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

[8] 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.

[9] 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.

[10] Yin-Yin He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. 2022.

[11] 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.

[12] 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*, 2024.

[13] 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)*, pages 9212–9221, June 2024.

[14] 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.

[15] 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.

[16] 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.

[17] 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*, 2025.

[18] 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.

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

[20] 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.

[21] 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*, pages 4117–4125, 2024.

[22] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Chenyang Qi, Chengfei Cai, Xiu Li, Zhifeng Li, Heung-Yeung Shum, Wei Liu, et al. Follow-your-click: Open-domain regional image animation via short prompts. *arXiv preprint arXiv:2403.08268*, 2024.

[23] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900*, 2024.

[24] Yue Ma, Yali Wang, Yue Wu, Ziyu Lyu, Siran Chen, Xiu Li, and Yu Qiao. Visual knowledge graph for human action reasoning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4132–4141, 2022.

[25] OpenAI. GPT-4o technical report. Technical report, OpenAI, 2024. Accessed: 2025-05-12.

[26] Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers. *arXiv preprint arXiv:2412.07776*, 2024.

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

[28] 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.

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

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

[31] 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*, pages 10684–10695, 2022.

[32] 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.

[33] Zhen Wan, Yue Ma, Chenyang Qi, Zhiheng Liu, and Tao Gui. Unipaint: Unified space-time video inpainting via mixture-of-experts. *arXiv preprint arXiv:2412.06340*, 2024.

[34] 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*, 2025.

[35] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *arXiv preprint arXiv:2406.08850*, 2024.

[36] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming rectified flow for inversion and editing. *arXiv preprint arXiv:2411.04746*, 2024.

[37] 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.

[38] 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*, 2024.

[39] 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.

[40] 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.

[41] 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)*, 2024.

[42] 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*, 2024.

[43] 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*, pages 399–417. Springer, 2024.

[44] 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.

[45] 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*, 2024.

[46] Zexuan Yan, Yue Ma, Chang Zou, Wenteng Chen, Qifeng Chen, and Linfeng Zhang. Eedit: Rethinking the spatial and temporal redundancy for efficient image editing. *arXiv preprint arXiv:2503.10270*, 2025.

[47] Xiangpeng Yang, Linchao Zhu, Hehe Fan, and Yi Yang. Videograin: Modulating space-time attention for multi-grained video editing. In *ICLR*, 2025.

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

[49] 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, 2024.

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

[51] 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.

[52] Beiyuan Zhang, Yue Ma, Chunlei Fu, Xinyang Song, Zhenan Sun, and Ziqiang Li. Follow-your-multipose: Tuning-free multi-character text-to-video generation via pose guidance. *arXiv preprint arXiv:2412.16495*, 2024.

[53] 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.

[54] Yinhan Zhang, Yue Ma, Bingyuan Wang, Qifeng Chen, and Zeyu Wang. Magiccolor: Multi-instance sketch colorization. *arXiv preprint arXiv:2503.16948*, 2025.

[55] 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*, 2023.

[56] 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*, 2023.

[57] 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*, 2023.

[58] 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, pages 10923–10931, 2025.

[59] Chenyang Zhu, Kai Li, Yue Ma, Longxiang Tang, Chengyu Fang, Chubin Chen, Qifeng Chen, and Xiu Li. Instantswap: Fast customized concept swapping across sharp shape differences. *arXiv preprint arXiv:2412.01197*, 2024.