Motion4Motion: Motion Transfer Across Subjects at Inference

LING-HAO CHEN, Tsinghua University, China and Stepfun, China
ZIXIN YIN, Hong Kong University of Science and Technology, China and Stepfun, China
DUOMIN WANG, Stepfun, China
XIANFANG ZENG, Stepfun, China
GANG YU*, Stepfun, China

Fig. 1. **Introducing Motion4Motion, a framework transferring motion from one subject to another.** Motion4Motion achieves cross-species (*e.g.*, human → panda, or human → goose) motion transfer without a uniform skeleton at inference.

*Corresponding author.

Authors' Contact Information: Ling-Hao Chen, Tsinghua University, Shenzhen, China and Stepfun, Shanghai, China, thu.lhchen@gmail.com; Zixin Yin, Hong Kong University of Science and Technology, Hong Kong, China and Stepfun, Shanghai, China, zixin.yin@connect.ust.hk; Duomin Wang, Stepfun, Shanghai, China, wangduomin@gmail.com; Xianfang Zeng, Stepfun, Shanghai, China, zzlongjuanfeng@zju.edu.cn; Gang Yu, Stepfun, Shanghai, China, skicy@outlook.com.

SIGGRAPH Conference Papers '26, Los Angeles, CA, USA

ACM ISBN 979-8-4007-2554-8/26/07
https://doi.org/10.1145/3799902.3811062

This work explores the motion transfer from one video to another, which is crucial in animation for diverse characters. Previously, video motion transfer has been largely explored between human and human-like characters, enabling a lot of applications in digital creation. However, these approaches encounter a main limitation. Specifically, related technical pipelines heavily rely on a predefined human skeleton structure and accordingly require skeleton-conditional model training. On the one hand, these methods are difficult to generalize to diverse characters, such as animals from different species, while preserving their unique motion styles. On the other hand, labeled data in diverse skeletons is limited, which additionally restricts the large-scale training for the task. In this paper, we jump out of the skeleton-based motion transfer framework and propose a training-free motion transfer framework, named Motion4Motion. Motion4Motion models the motion flow of the character in a video instead of skeletons, which makes motion

transfer across species easier. Extensive experimental results and novel applications show our methods outperform baselines impressively.

CCS Concepts: • **Computing methodologies** → **Computer vision**; **Animation**.

ACM Reference Format:
Ling-Hao Chen, Zixin Yin, Duomin Wang, Xianfang Zeng, and Gang Yu. 2026. Motion4Motion: Motion Transfer Across Subjects at Inference. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers (SIGGRAPH Conference Papers '26), July 19–23, 2026, Los Angeles, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3799902.3811062

1 Introduction

Motion transfer [Guo et al. 2024; Hu 2024; Zhang et al. 2025b] has a wide application in digital creation and animation workflows, such as character animation [Chen et al. 2025], virtual reality, and movie post-production. In recent years, the mainstream research has primarily focused on human-centric scenarios, where the core objective is to migrate movements from a source person to a target person. To achieve this, existing methodologies heavily rely on skeletal representations to bridge the geometric gap between different subjects. By extracting structural poses, these skeleton-based pipelines have achieved remarkable success in transfer quality.

Despite these progresses, existing frameworks face significant hurdles in real-world applications, particularly when extending beyond human-centric domains. Most current paradigms are intrinsically "hard-coded" for specific structural priors, which severely restricts their flexibility across different species. When tasked with transferring motion between characters with vastly different morphologies, such as among various animal species, the lack of a shared skeletal template makes spatial alignment ill-defined. This rigidity prevents current methodologies from generalizing to "in-the-wild" scenarios where characters may possess arbitrary shapes and motion styles that deviate significantly from standard human proportions. Notably, even the most relevant attempt, FlexiAct [Zhang et al. 2025b], remains unsatisfactory, as it relies on per-case optimization that leads to overfitting and consequent information leakage.

Two fundamental challenges impede the realization of robust, cross-species motion transfer in a more general space. The first challenge is the critical scarcity of high-quality, paired motion data across diverse topologies. Unlike human-centric research, which benefits from massive video datasets and mature pose estimation tools, obtaining synchronized motion sequences for diverse characters is both labor-intensive and often impractical. This data bottleneck forces data-driven models to rely on narrow distributions, leading to severe artifacts when encountering unseen species. The second challenge involves the ambiguity of defining semantic correspondences between source and target subjects without skeletons, such as ambiguity between the legs of a chair and a quadrupedal animal. Establishing consistent mappings becomes exceptionally difficult when the target character possesses completely different physical semantics, making it hard to maintain motion fidelity while ensuring the visual plausibility of the transferred results.

To overcome the inherent constraints of predefined skeletal templates in driving novel character animations, we depart from the conventional motion transfer pipeline and propose Motion4Motion, a novel framework designed for video-based motion synthesis. Unlike previous kinematics-based methods that rely on skeletal priors, Motion4Motion operates on dense pixel-level motion flows, treating them as the fundamental primitives for motion representation. By capturing the temporal dynamics of source pixels and mapping them onto the target subject via our proposed TransPE module, our method achieves high-fidelity transfer without structural limitations. Notably, Motion4Motion is a training-free approach implemented entirely during inference, which enables the good interpretability.

Before delving into details, our core contributions are as,

- We present Motion4Motion, the first training-free framework capable of transferring motion across general subjects without relying on predefined skeletal priors.
- We propose a simple yet effective module, namely TransPE, for injecting motion flows to the target subject.
- Extensive evaluations demonstrate that Motion4Motion not only delivers high-fidelity motion transfer results but also exhibits potential applications in novel concept composition and cross-morphology motion transfer (Sec. 5).

2 Related Work

2.1 Video Generation

While early diffusion models predicated on U-Net backbones [Guo et al. 2024; Ho et al. 2020; Rombach et al. 2022] significantly outperformed GAN-based systems [Reed et al. 2016; Wang et al. 2023; Yu et al. 2023] in image fidelity, their scalability constraints have precipitated a paradigm shift toward Diffusion Transformers (DiTs) [Esser et al. 2024; Peebles and Xie 2023; Wang et al. 2025d]. DiTs have since emerged as the foundational backbone for state-of-the-art video diffusion models, exemplified by CogVideo-X [Yang et al. 2025] and WAN [Wan et al. 2025]. In this work, we propose a novel attention control method designed to integrate with the WAN.

2.2 Attention Control

Parallel to these architectural advancements, the field of controllable generation has expanded significantly. Originating with Prompt-to-Prompt [Hertz et al. 2023], attention control methods have been extensively deployed to modulate pre-trained U-Nets for image and video editing tasks [Cao et al. 2023; Liu et al. 2024b; Tumanyan et al. 2023; Yin et al. 2025a,b,c]. In particular, key-value (KV) injection and concatenation in attention layers have been explored for various purposes, including style-consistent generation [Hertz et al. 2024], zero-shot style transfer [Deng et al. 2024], subject-driven consistent generation [Tewel et al. 2024], and appearance transfer [Alaluf et al. 2024]. Although recent endeavors have begun extending these controls to DiTs [Cai et al. 2025; Wang et al. 2025c; Yin et al. 2025a,c], they predominantly target MM-DiT architectures [Esser et al. 2024], which rely on a unified self-attention mechanism for fusing visual and textual modalities. Consequently, the efficacy of attention control within DiT architectures employing *decoupled* self-attention and cross-attention layers, such as WAN, remains underexplored. Furthermore, while existing literature addresses general editing and long-video synthesis, the potential of DiT-based attention control specifically for motion transfer has yet to be investigated.

2.3 Motion Transfer

3D-based motion transfer. The problem of motion transfer between different characters was first established and extensively explored in the 3D animation community as motion retargeting [Gleicher 1998]. Traditional methods relied on kinematic optimization to satisfy spatial constraints [Feng et al. 2012; Lee and Shin 1999], while modern neural-based approaches [Aberman et al. 2020; Lim et al. 2019] leverage deep architectures to decouple pose and structure. Although some recent iterations [Chen et al. 2025; Li et al. 2022] attempt to improve generalization across disparate morphologies, they remain bound to 3D skeletal topologies [Chen et al. 2024, 2023; Dai et al. 2024; Lu et al. 2023] and often require manual joint correspondences. Transitioning these concepts to the video domain presents unique challenges, as explicit structural information is often absent or noisy in raw pixels.

Video-based motion transfer. Specialized human animation frameworks [Cheng et al. 2025; Hu 2024; Zhang et al. 2025a] built upon pre-trained video diffusion models rely heavily on large-scale pre-training and explicit skeletal guidance, which limits their applicability to fixed topologies and demands massive computational resources. Conversely, general motion editing approaches like FlexiAct [Zhang et al. 2025b], and others [Burgert et al. 2025a; Gokmen et al. 2025; Ling et al. 2025; Zhao et al. 2024] avoid skeletal constraints but typically necessitate time-consuming per-video fine-tuning. Motion Prompting [Geng et al. 2025] proposed ControlNet-based training for general trajectory-guided motion control. Go-with-the-Flow [Burgert et al. 2025b] introduces warped noise for real-time motion-controllable generation. Diffusion-As-Shader [Gu et al. 2025] and WAN-Move [Chu et al. 2025] train dedicated trajectory-conditioned modules for video control. ATI [Wang et al. 2025b] further unifies trajectory instructions for controllable generation, while MotionStream [Shin et al. 2026] achieves real-time interactive motion control. Despite the impressive performance, these training-based methods are inherently limited by the coverage of their training data and may not generalize to unseen motion patterns. In contrast, our approach establishes a fully training-free method that eliminates the need for both large-scale model training and auxiliary driving signals, enabling cross-species motion transfer seamlessly.

3 Methodology

In this section, we will introduce the whole pipeline of our proposed system, Motion4Motion. Motion4Motion is a training-free framework via manipulating the attention calculation of the denoising process. To introduce our method, we begin with the introduction of foundational concepts of our base generative framework in Sec. 3.1. As our method is not based on the skeleton correspondence, we track the motion flow of subjects in the video playback and build correspondence between the source and target subjects in images (Sec. 3.2). To achieve the motion transfer from the source video to the subject, we introduce a novel module, TransPE, for attention manipulation in Sec. 3.3.

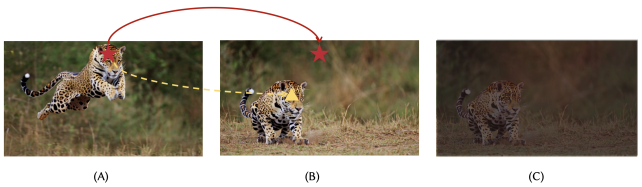

Fig. 2. **Position awareness in self-attention.** We sample two frames from a video. The source point in (A) is marked by a point ★. We found the attention weight (overlaid on (C)) of this point ★ is more aware of the spatial-temporal neighbors ★, but not semantically similar ones ▲.

3.1 Video Generation Framework

Diffusion transformer. Our framework is built upon WAN-T2V [Wan et al. 2025], which utilizes a DiT architecture integrated with Flow Matching [Lipman et al. [n. d.]]. For efficient processing, an input video is first compressed into a latent space $\mathbf{z} \in \mathbb{R}^{(1+\lfloor \frac{F}{4} \rfloor) \times \frac{H}{8} \times \frac{W}{8} \times C}$ via a 3D causal VAE, where F, H, and W denote the number of frames, height, and width of the video, respectively, and C represents the latent channel dimension. Unlike traditional Gaussian diffusion, this paradigm models the generative process as a continuous-time probability path where the DiT model predicts a velocity $\mathbf{v}_t$ that transforms initial noise $\mathbf{z}_0$ into the target latent $\mathbf{z}_1$ through linear interpolation $\mathbf{z}_t = t\mathbf{z}_1 + (1-t)\mathbf{z}_0$. Our Motion4Motion operates within this latent space, intervening in the denoising process by manipulating the spatial-temporal attention maps of the DiT blocks to achieve training-free motion transfer for general subjects. For convenience, we take the simplification of $F \leftarrow 1 + \lfloor \frac{F}{4} \rfloor$, $H \leftarrow \frac{H}{8}$, and $W \leftarrow \frac{W}{8}$ to denote the temporal and spatial dimensions of the latent space in the following sections.

Self-attention and positional encoding. The denoising backbone of WAN consists of L successive transformer blocks, each integrating multi-head self-attention (SA), cross-attention for text conditioning, and a feed-forward network (FFN). Within each SA layer, the input latent $\mathbf{X}$ is first projected into query $\mathbf{Q}$, key $\mathbf{K}$, and value $\mathbf{V}$ tensors, through linear transformations. To capture the complex spatio-temporal dependencies of video data, WAN employs 3D Rotary Positional Embedding (RoPE) [Su et al. 2024]. Unlike absolute positional encodings, RoPE injects relative position information by rotating pairs of dimensions in the $\mathbf{Q}$ and $\mathbf{K}$ tensors according to their temporal and spatial coordinates (f, h, w). The attention mechanism then computes the weighted sum of values based on the similarity between queries and keys embedded with positions, which can be formulated as,

$$\mathbf{X} \leftarrow \mathrm{Softmax}\left(\frac{\mathsf{RoPE}(\mathbf{Q})\ \mathsf{RoPE}(\mathbf{K})^\top}{\sqrt{d}}\right)\mathbf{V}, \tag{1}$$

where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{L \times d}$, $L = F \times H \times W$, and d is the latent dimension of query, key, and value. The resulting attention output $\mathbf{X}$ is subsequently processed by the FFN and fed into the next block. *Particularly*, by injecting positional priors into each latent token, RoPE enables the model to perceive the relative distance between pixels effectively. This mechanism ensures a heightened sensitivity to local structures [Wang et al. 2025a; Yin et al. 2025c], as tokens in

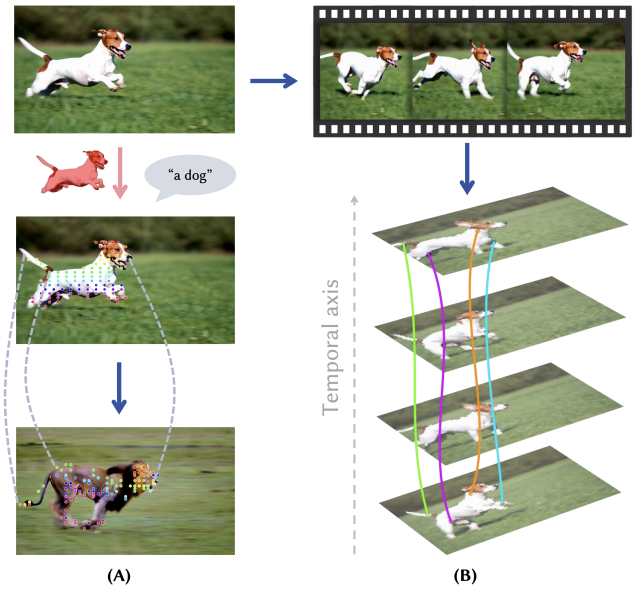

Fig. 3. **Building correspondences across images and video.** (A) Cross-image correspondence: Anchor points P^1_{src} sampled within the source mask are semantically matched to the target subject P_{tgt} via point matching. (B) Motion flow extraction: The motion flow $\mathcal{M}^{src}$ is constructed by tracking the trajectories of these points across the temporal axis of the source video.

close proximity exhibit stronger positional correlations during the attention calculation, shown in Fig. 2. Particularly, points with similar semantics might not have a higher attention weight, as shown in Fig. 2. This phenomenon reveals the crucial role that positional encoding plays in video generation. Such a relative distance-aware property is crucial for maintaining structural integrity and capturing the nuanced motion dynamics within the spatial-temporal grid.

3.2 Motion Flows in-and-cross Video Playback

An essential component in MOTION4MOTION is establishing the motion flow in the video and then transferring to the target. As shown in Fig. 3, we take an example of transferring the running motion of a dog in a video to a lion. In this stage, we target two essential correspondences, (1) cross-image correspondence between the source and target subjects, and (2) motion flow of the source subjects in the source video.

Cross-image correspondence. To enable motion transfer between subjects with potentially disparate topologies, we first establish a semantic bridge between the source and target subjects in a skeleton-free manner. As shown in Fig. 3-(A), given the first frame of the source video $\mathbf{I}^1_{src}$ and a target subject image $\mathbf{I}_{tgt}$, we employ a subject mask to sample a set of N representative anchor points $\mathrm{P}^1_{src} = \{\mathrm{p}^1_{src,i}\}_{i=1}^N$ on the source character using Grounded SAM-2 [Liu et al. 2024a; Ravi et al. 2024; Ren et al. 2024]. Technically, the target subject image $\mathbf{I}_{tgt}$ is the first frame of the video generated by unedited prompt. For real target image transfer, we generate its video by an I2V model and then use the inversion technique to obtain the noise.

We then leverage a semantic matching algorithm (*e.g.*, diffusion features [Tang et al. 2023]) to find their corresponding coordinates $\mathrm{P}_{tgt} = \{\mathrm{p}_{tgt,i}\}_{i=1}^N$ on the target subject. This point-to-point mapping $\mathcal{C} : \mathrm{P}^1_{src} \rightarrow \mathrm{P}_{tgt}$ serves as the foundation for cross-image transfer, ensuring that the movement of specific semantic parts (*e.g.*, the "legs" of a dog *vs.* the "legs" of a lion) is accurately migrated despite differences in their categories.

Initialization for real image transfer. When the target subject is provided as a real image $\mathbf{I}_{tgt}$ rather than generated by the T2V model, we first synthesize a video $\mathbf{V}_{gen}$ from $\mathbf{I}_{tgt}$ using a WAN-I2V model. We then apply inversion [Jiao et al. 2025] on $\mathbf{V}_{gen}$ to obtain the initial noise latent, which serves as the starting point for the subsequent attention manipulation.

Motion flow in source video. While cross-image correspondence handles semantic alignment across subjects, the temporal dynamics must be extracted from the source video playback to drive the animation. As illustrated in Fig. 3-(B), for the sampled anchor points P^1_{src}, we track their trajectories across the subsequent $F-1$ frames using a point tracking framework implemented with DIFT features. This mapping yields a sequence of coordinates for each point, denoted as the motion flow $\mathcal{M}^{src} = \{\mathrm{P}^f\}_{f=1}^F$, where $\mathbf{P}^f$ represents the positions of all anchor points at frame f. We mark this mapping as $\mathcal{I} : \mathrm{P}^1_{src} \rightarrow \mathrm{M}^{src}$ Specifically, the motion flow (*aka.* trajectory) of the i-th point is defined as a set of spatio-temporal coordinates $\mathbf{P}^f_i = \{(h^f_i, w^f_i)\}_{f=1}^F$, illustrated as the motion flow lines in Fig. 3-(B). By decoupling the motion from the source subject into these flow lines, M^{src} provides a topology-agnostic representation of the dynamics, which is subsequently injected into the generative process of the target subject through our attention manipulation in Sec. 3.3.

Note that the point matching and tracking are conducted within the downsampled coordinate system, downsampled by factors of 4, 8, and 8 along the temporal, height, and width axes, respectively.

3.3 Video Motion Transfer with TRANSPE

Here, we detail how the extracted motion flows are utilized to drive the target subject's generation process. Our strategy consists of two primary steps: mapping the source dynamics to the target subject coordinates and performing a training-free attention manipulation.

Target motion flow construction. In Sec. 3.2, we introduce the point correspondence across images and the video motion flow. These two mappings are presented as $\mathcal{C} : \mathrm{P}^1_{src} \rightarrow \mathrm{P}_{tgt}$ and $\mathcal{I} : \mathrm{P}^1_{src} \rightarrow \mathrm{M}^{src}$, respectively. The core objective of motion transfer is to reproject the points of the target subject onto the spatio-temporal coordinates of the extracted source motion flow, such that $\mathrm{M}^{tgt} = \mathrm{M}^{src}$. Consequently, the trajectory mapping for the target subject points P_{tgt} can be formulated as a composite function: $\mathcal{M} = \mathcal{I} \circ \mathcal{C}^{-1}$. By leveraging this derived mapping $\mathcal{M} : \mathrm{P}_{tgt} \rightarrow \mathrm{M}^{tgt}$, we establish a relationship between the target subject's semantic points and their intended spatio-temporal destinations in the synthesized video, for inheriting the dynamics of the source.

Attention manipulation with TRANSPE.. In practical applications, the target image used for cross-image matching is the first frame of a target subject video. Our goal, therefore, is to reposition the

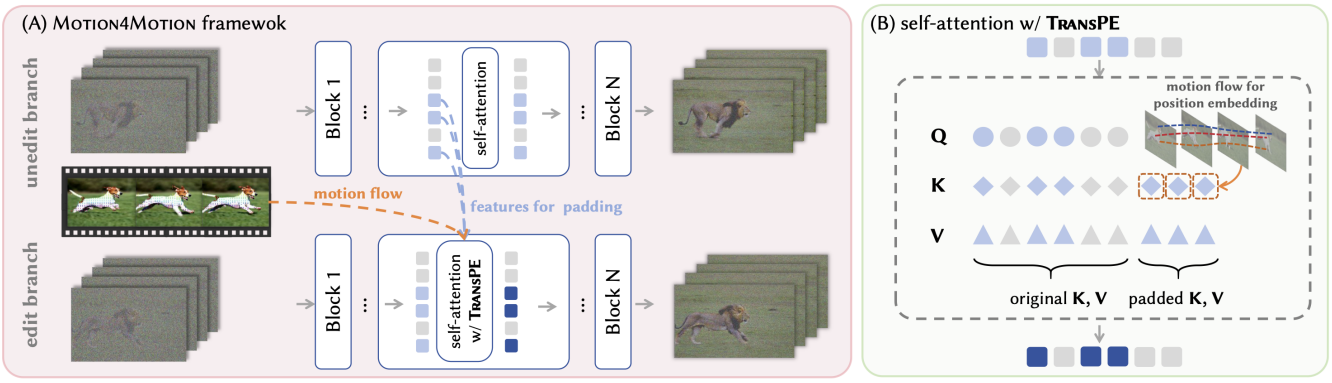

Fig. 4. **System overview of Motion4Motion.** (A) **Overall Framework.** Our system adds standard self-attention with the TransPE module within DiT blocks to achieve training-free motion transfer. (B) **Mechanism of TransPE.** The original Query $\mathbf{Q}$ (●) keeps unchanged. Key $\mathbf{K}$ (◆) is padded with the appearance key features of the target subject, marked as light blue key tokens ◆. Additionally, padded keys are also applied to positional encoding using the motion flow trajectories $\mathbf{M}^{tgt}$ from the source video, as marked in the orange dashed boxes. Value $\mathbf{V}$ (▲) is augmented by concatenating appearance value features of the target subject, marked as light blue value tokens ▲. With the proposed pipeline, Motion4Motion faithfully transfers the motion of a source video to a new subject in a another video that follows the source motion. All blue notations means tokens or features of the "lion" in this example. The darker blue tokens denote the *edited* tokens after TransPE manipulation, in contrast with the lighter blue tokens representing the original (unedited) target features.

target subject's features to their corresponding locations within the transferred video. This editing process is performed in the latent space of the video via diffusion inversion. To faithfully migrate the target subject's appearance along the constructed trajectories, we intervene in the self-attention calculation of the DiT blocks. As introduced in the previous text, self-attention calculates the similarity between points with positional encoding. Motivated by this, we introduce a Transferring Positonal Encoding method (TransPE) to rearrange points of the target subject in editing the attention.

During the denoising inference, let $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ be the query, key, and value tensors projected from the original noisy latent, *i.e.*, unedited video branch. We introduce the TransPE module to inject the target subject's anchor features into the attention mechanism. Given the latent $\mathbf{K}$ feature of the target subject's first frame $\mathbf{K}^1_{tgt}$ (cached by diffusion inversion), we extract the matched point as $\mathbf{K}^\star_{tgt} = \mathbf{K}^1_{tgt}[P_{tgt}]$ via the slice operation, which provides the appearance of the target subject. We replicate it F times to initialize the sequence of these appearance features $\hat{\mathbf{K}} = [\mathbf{K}^\star_{tgt}, \ldots, \mathbf{K}^\star_{tgt}] \in \mathbb{R}^{F\times H\times W\times d}$. Similarly, we can slice the feature of the subject from the original $\mathbf{V}$ and repeat F times, as $\hat{\mathbf{V}}$, with the same dimension of $\hat{\mathbf{K}}$. After that, we re-embed $\hat{\mathbf{K}}$ with positional information using RoPE based on the target motion flow $\mathbf{M}^{tgt}$. This allows the model to "look for" the target subject's features at the newly transferred coordinates. We then augment the original key and value tensors via concatenation,

$$\mathbf{K}_{new} = [\mathrm{RoPE}(\mathbf{K}),\ \mathrm{RoPE}(\hat{\mathbf{K}}, \mathbf{M}^{tgt})], \quad \mathbf{V}_{new} = [\mathbf{V},\ \hat{\mathbf{V}}], \tag{2}$$

while keeping the query $\mathbf{Q}$ unchanged. The updated self-attention operation is then performed as,

$$\mathbf{X} \leftarrow \mathrm{Softmax}\left(\frac{\mathrm{RoPE}(\mathbf{Q})\ \mathbf{K}^\top_{new}}{\sqrt{d}}\right)\mathbf{V}_{new}. \tag{3}$$

By padding position-aware features into the attention space, Motion4Motion effectively forces the denoising process to synthesize the target subject at the specific coordinates dictated by the motion flow, achieving high-fidelity motion transfer without tuning.

The complete pipeline of Motion4Motion is summarized in Alg. 1. Given a source video and target subject, we first extract the motion flow and cross-image correspondence, then perform the denoising process with TransPE attention manipulation within specified step and layer ranges.

4 Experiments

4.1 Setting

4.1.1 Implementation details. Our method is implemented on top of the WAN-14B-T2V [Wan et al. 2025] (480p resolution) model. By default, attention manipulation is applied across layers [0, 40] until step 35 out of 50 denoising steps. TransPE is applied in all self-attention layers within this range. Point matching is performed via diffusion feature matching, while subject segmentation is obtained using SAM-2 [Ravi et al. 2024]. To align the segmentation masks and point coordinates with the latent space resolution, both are downsampled by a factor of 1/8. All experiments are conducted on one NVIDIA H-800 GPU.

4.1.2 Baselines. Most existing methods in this field focus on human or human-like motion transfer and typically rely on predefined global motion signals or explicit pose representations. For example, approaches such as WAN-animate [Cheng et al. 2025] heavily depend on human pose detectors, whose performance may be unstable in non-human-like character scenarios. In this paper, to evaluate general fine-grained character motion control, we first adopt two state-of-the-art methods, FlexAct [Zhang et al. 2025b] and MotionClone [Ling et al. 2025], as primary baselines. We further compare our approach with global motion transfer methods, including MotionDirector [Zhao et al. 2024] and RoPECraft [Gokmen et al. 2025]. Additionally, we compare with recent training-required

Algorithm 1: Motion Transfer with TransPE

Input: Source video $\mathbf{V}_{src}$, target subject $\mathbf{I}_{tgt}$, text prompt $\mathbf{c}$, total steps T, total layers L, begin/end step s_b/s_e, begin/end layer l_b/l_e
Output: Transferred video $\mathbf{V}_{tgt}$

// Stage 1: Motion Flow Extraction
$\mathrm{P}^1_{src} \leftarrow \texttt{SampleAnchors}(\mathbf{V}_{src}[0])$; // anchor points on source first frame
$\mathrm{P}_{tgt} \leftarrow \texttt{SemanticMatch}(\mathrm{P}^1_{src}, \mathbf{I}_{tgt})$; // cross-image correspondence $\mathcal{C}$
$\mathrm{M}^{src} \leftarrow \texttt{PointTrack}(\mathrm{P}^1_{src}, \mathbf{V}_{src})$; // motion flow in source video $\mathcal{I}$
$\mathrm{M}^{tgt} \leftarrow \mathrm{M}^{src}$; // target flow via $\mathcal{M} = \mathcal{I} \circ \mathcal{C}^{-1}$
// Stage 2: Latent Initialization
$\mathbf{V}_{gen} \leftarrow \texttt{Generate}(\mathbf{I}_{tgt}, \mathbf{c})$; // T2V/I2V generation
$\mathbf{z}_0 \leftarrow \texttt{Inversion}(\mathbf{V}_{gen})$; // deterministic inversion
Cache $\mathbf{K}^1_{tgt}, \mathbf{V}^1_{tgt}$ from inversion at each layer;
// Stage 3: Denoising with TransPE
for $t = 0$ **to** $T - 1$ **do**
  for $l = 0$ **to** $L - 1$ **do**
    Compute $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ from $\mathbf{z}_t$ at layer l;
    if $t < s_e$ ***and*** $l_b \le l < l_e$ **then**
      // Apply TransPE manipulation
      $\hat{\mathbf{K}} \leftarrow \texttt{Repeat}(\mathbf{K}^1_{tgt}[\mathrm{P}_{tgt}], F)$;
      $\hat{\mathbf{V}} \leftarrow \texttt{Repeat}(\mathbf{V}^1_{tgt}[\mathrm{P}_{tgt}], F)$;
      $\mathbf{K}_{new} \leftarrow [\texttt{RoPE}(\mathbf{K}), \texttt{RoPE}(\hat{\mathbf{K}}, \mathrm{M}^{tgt})]$;
      $\mathbf{V}_{new} \leftarrow [\mathbf{V}, \hat{\mathbf{V}}]$;
      $\mathbf{X} \leftarrow \texttt{Softmax}\left(\frac{\texttt{RoPE}(\mathbf{Q}) \cdot \mathbf{K}^\top_{new}}{\sqrt{d}}\right)\mathbf{V}_{new}$;
    else
      $\mathbf{X} \leftarrow$ standard self-attention;
  $\mathbf{z}_{t+1} \leftarrow$ update with velocity prediction;
$\mathbf{V}_{tgt} \leftarrow \texttt{Decode}(\mathbf{z}_T)$;

trajectory-based methods, Diffusion-As-Shader [Gu et al. 2025] and WAN-Move [Chu et al. 2025], which train dedicated motion control modules on large-scale trajectory data. Since the implementation of the recent method MotionV2V [Burgert et al. 2025a] and MotionShot [Liu et al. 2025] was not publicly available before submission, we discuss them in related work.

4.1.3 Benchmark and evaluation protocol. We evaluate Motion4Motion on the established animal (33 pairs) and human motion (123 pairs) transfer benchmarks introduced by Zhang et al. [2025b]. Since our framework is built upon the WAN-T2V-14B foundation model, we adapt the evaluation protocol by first performing a deterministic inversion [Jiao et al. 2025] of the source video to obtain the base latent path, followed by our attention manipulation to synthesize the transferred results. To quantify the quality of the transferred results, we employ four standard metrics: (1) *Textual Similarity* (TS): calculated via CLIP [Radford et al. 2021] to measure the semantic consistency between the generated frames and the prompt. (2) *Motion Fidelity* (MF): which utilizes tracklets computed by Co-tracker [Karaev et al. 2025] to measure the similarity between motion trajectories in unaligned videos. (3) *Temporal Consistency* (TC): quantified by the average CLIP image feature similarity between all frame pairs to ensure smoothness and coherence. (4) *Appearance Consistency* (AC): which reflects the identity preservation by calculating the average CLIP [Radford et al. 2021] similarity between the target image and the generated video frames. Additionally, to evaluate the fine-grained pose alignment between the source and target motion, we introduce a new benchmark with 50 image-video pairs with different animals. Specifically, the category of these examples can be covered by a cross-category pose detector. Consequently, we evaluate the *pose similarity (PS)* between the source and target motions with a precise detector [Yang et al. 2024].

4.2 Evaluation

4.2.1 Quantitative evaluation. We compare our method with several state-of-the-art baselines in Tab. 1, where Motion4Motion consistently achieves the best performance across all metrics. The superiority of our approach can be attributed to its fine-grained control mechanism. Compared to earlier frameworks like MotionDirector and RoPECraft, which primarily rely on global representations or model-level tuning, Motion4Motion operates at a point-to-pixel level. This enables our model to capture local motion nuances that are typically lost in holistic control paradigms, leading to significant gains in Motion Fidelity (MF). Furthermore, Motion4Motion outperforms recent mask-based methods such as MotionClone and FlexiAct. While these approaches improve motion quality by constraining attention within the subject's silhouette, they lack explicit semantic guidance between the source and target. In contrast, our point-to-point mapping establishes a direct semantic bridge, ensuring that each part of the target subject follows the intended trajectory with high precision. This is reflected in the substantial improvement in Pose Similarity (PS), demonstrating our method's robustness in handling complex character transfers.

4.2.2 Qualitative evaluation. We present a visual comparison between our method and several state-of-the-art baselines in Fig. 5, focusing on challenging cross-species motion transfer tasks (*e.g.*, transferring a fox's gait to a giraffe, and a lion's pounce to a zebra). As illustrated, existing baselines struggle to maintain a balance between motion fidelity and visual quality. Specifically, RoPECraft [Gokmen et al. 2025] often produces stiff movements and fails to capture the fluid leg dynamics of the reference. FlexiAct [Zhang et al. 2025b] suffers from significant identity distortion and texture blurring, particularly around the limbs (as seen in the giraffe's shadow and the zebra's legs), failing to preserve the structural integrity of the target animal. While MotionClone [Ling et al. 2025] achieves better motion magnitude, it introduces noticeable *ghosting artifacts* and *temporal appearance drifting*. A key limitation of these baselines is their over-reliance on global motion features or their lack of explicit structural guidance, which makes precise **fine-grained pose alignment** extremely difficult. In contrast, our approach achieves superior semantic pose alignment even across disparate morphologies. Despite the vast differences in limb proportions between a fox and a giraffe, our method precisely retargets the reference gait while maintaining high-fidelity appearance and sharp textures without the need for skeletal priors or fine-tuning. More results are in Fig. 9.

Table 1. **Main quantitative comparison with baselines.**

Method	Human				Animal				Ours				
	TS ↑	MF ↑	TC ↑	AC ↑	TS ↑	MF ↑	TC ↑	AC ↑	TS ↑	MF ↑	TC ↑	AC ↑	PS ↑
MotionDirector [2024]	0.255	0.312	0.915	0.887	0.248	0.298	0.902	0.875	0.251	0.305	0.908	0.881	0.342
RoPECraft [2025]	0.241	0.330	0.907	0.894	0.235	0.315	0.895	0.882	0.238	0.322	0.901	0.888	0.355
MotionClone [2025]	0.258	0.381	0.937	0.900	0.252	0.365	0.924	0.891	0.255	0.373	0.931	0.896	0.408
FlexiAct [2025b]	0.269	0.391	0.928	0.945	0.261	0.378	0.915	0.932	0.265	0.384	0.922	0.938	0.415
Motion4Motion	**0.288**	**0.452**	**0.955**	**0.971**	**0.282**	**0.441**	**0.948**	**0.962**	**0.285**	**0.448**	**0.952**	**0.967**	**0.543**

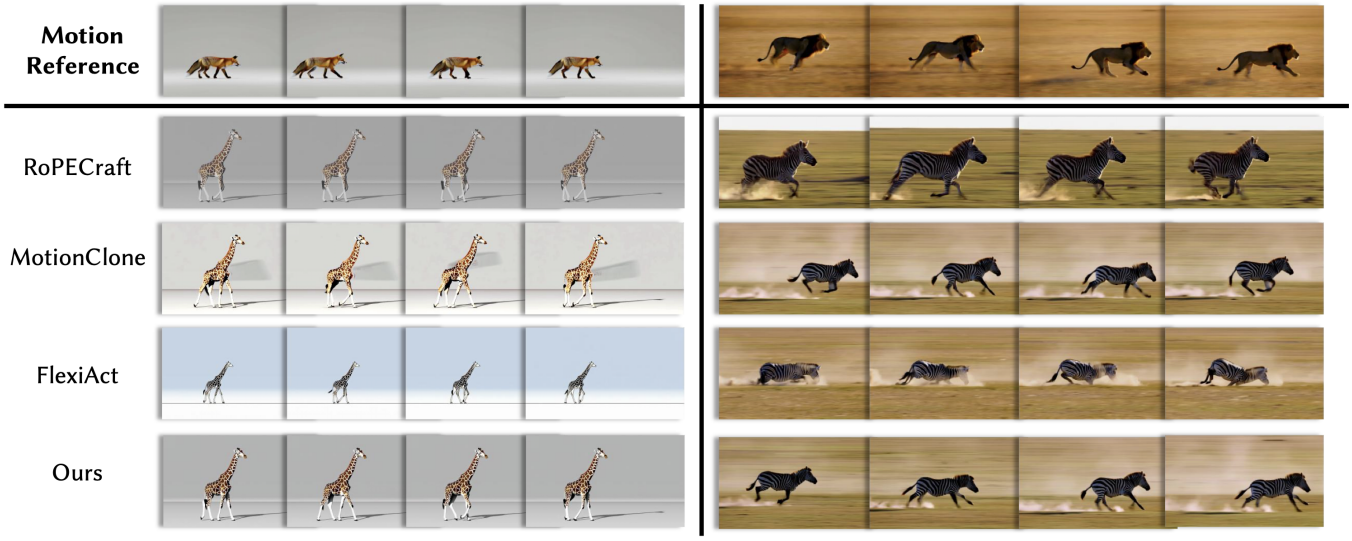

Fig. 5. **Comparison of Motion4Motion with baselines.** We focus on the preservation of pose details across individual frames. Most baseline methods exhibit varying degrees of visual artifacts or appearance drifting. In contrast, our method maintains high-fidelity pose alignment and appearance consistency.

4.2.3 Real-image/video evaluation. To further evaluate the robustness and generalization capabilities of Motion4Motion, we perform motion transfer using "in-the-wild" images and videos sourced from the Internet. As shown in Fig. 6, we animate static portraits of diverse subjects, including public figures with distinct stylistic features, using a complex dance sequence as the source motion. Despite our method being without any training, Motion4Motion successfully retargets the dynamic motion while faithfully preserving the subjects' identity, clothing textures, and structural integrity. This zero-shot performance on diverse real-world data demonstrates that our approach does not overfit to specific datasets and can handle diverse in-the-wild scenarios without additional fine-tuning.

4.3 User Study

We conducted a subjective evaluation to assess the perceptual quality of the generated videos. The study followed a blind pairwise comparison protocol [Zhang et al. 2025b; Zhao et al. 2024], where participants were presented with two videos side-by-side: one generated by the Base Model (WAN-I2V-14B) and the other by a competing method (or ours). For each of the 50 randomly selected test cases, all 10 raters were asked to select the preferred result based on two specific criteria: *Motion Consistency* (how accurately the transferred motion matches the source motion) and *Appearance Consistency* (how well the visual identity of the subject is preserved). As reported in Tab. 2, our method demonstrates a significant advantage over the baseline, on both motion and appearance consistency. These results confirm that our approach not only ensures faithful motion transfer but also maintains higher visual fidelity compared to baselines.

5 Application: Teaching a Table Walking

Although a large video generation model shows impressive generation ability in various scenarios, it still struggles with some novel concept composition [Shi et al. 2025], such as "`A desk coming to life, running rapidly along a muddy riverside.`". As the **novel concept composition** of "`running`" and "`desk`" differs substantially, the T2V model is quite hard to compose a reasonable result for the model. As shown in Fig. 8(E-F), the model always synthesizes the static desk motions or generates a sliding motion, due to the limited imagination ability of the base model.

To address this, we introduce an external source to control the motion. We explore the boundaries of our algorithm by performing **cross-morphology transfer**: driving a static table (Fig. 8B) using a

Table 2. **Human evaluation results compared to others.** Our method outperforms baselines across all dimensions. The "Base Model" refers to WAN-I2V-14B with TransPE (K-V concatenation) applied, which serves as the anchor for pairwise comparison.

v.s. Base Model	**Motion Consistency**	**Appearance Consistency**
MotionDirector	44.8 *v.s.* 55.2	49.7 *v.s.* 50.3
RoPECraft	67.1 *v.s.* 32.9	62.6 *v.s.* 37.4
MotionClone	71.7 *v.s.* 28.3	68.6 *v.s.* 31.4
FlexAct	78.4 *v.s.* 21.6	62.1 *v.s.* 37.9
Motion4Motion	**92.5** *v.s.* **7.5**	**87.8** *v.s.* **12.2**

video of a walking human (Fig. 8A). This presents a significant challenge due to the distinct structural differences. To render this problem tractable, inspired by "bone binding" in 3D animation [Artell 2025; Rokoko 2021], we explicitly bind the human legs to the table's legs using SAM2 masks, as shown in Fig. 8(C-D). Consequently, Motion4Motion constrains the diffusion feature matching within these masked regions, mitigating the structural ambiguity caused by morphological disparities and successfully transferring the gait to the table (Fig. 8H), synced with the motion in source video (Fig. 8G).

6 Conclusion and Discussion

In this work, we presented Motion4Motion, a novel training-free framework that achieves high-fidelity motion transfer across diverse subjects without relying on skeletal priors or specific fine-tuning. By leveraging the proposed TransPE module, our approach effectively injects motion flows from a source video into the self-attention mechanism of a pre-trained Diffusion Transformer, enabling precise semantic alignment even across disparate morphologies. Extensive evaluations demonstrate that Motion4Motion significantly outperforms state-of-the-art baselines in both motion fidelity and appearance preservation. Furthermore, the ability to animate inanimate objects (*e.g.*, a walking table) highlights the generalization capability of our method. We believe this work offers a flexible and efficient solution for video generation, opening new avenues for creative animation workflows.

Acknowledgments

The author team of Motion4Motion would like to convey sincere appreciation to all reviewers and committee members for their significant efforts in enhancing the quality of this work. We sincerely acknowledge Bohong Chen, Yukai Shi, and Shiyi Zhang for their thoughtful suggestions and discussions on this paper.

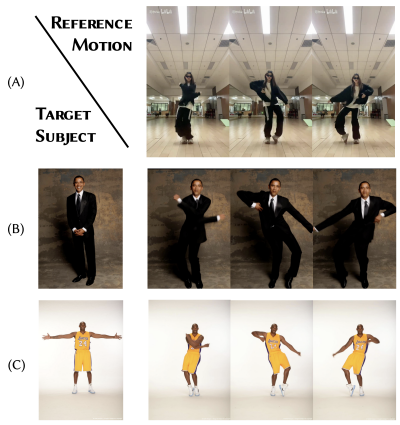

Fig. 6. **Real image motion motion transfer Motion4Motion.** The image of the driven subject and the motion driving video are from the Internet. The transfer results show the good performance of Motion4Motion.

Fig. 7. **Ablation on manipulation steps.** (A) Without manipulation (Step 0), the model fails to relocate target features along the intended trajectories. (B) Our default strategy (Step 35) achieves an optimal balance between motion fidelity and visual realism. (C) Full-step manipulation (Step 50) causes semantic confusion, resulting in texture stretching and artifacts.

Fig. 8. **Application of Motion4Motion in Cross-domain Motion Synthesis.** We empower a vanilla T2V model with the capability to animate inanimate objects using biological motion trajectories. Given a reference video of a walking person (A) and a target static image of a table (B), our method extracts the motion semantics (G) and applies them to the target. While baseline approaches (WAN-I2V-14B, WAN-T2V-14B) without our control (E, F) produce either static motions or simple sliding motions. With the binding legs from a man to two specified legs in (C) and (D), Motion4Motion (H) successfully retargets the walking gait onto the table's legs while maintaining structural integrity and environmental consistency.

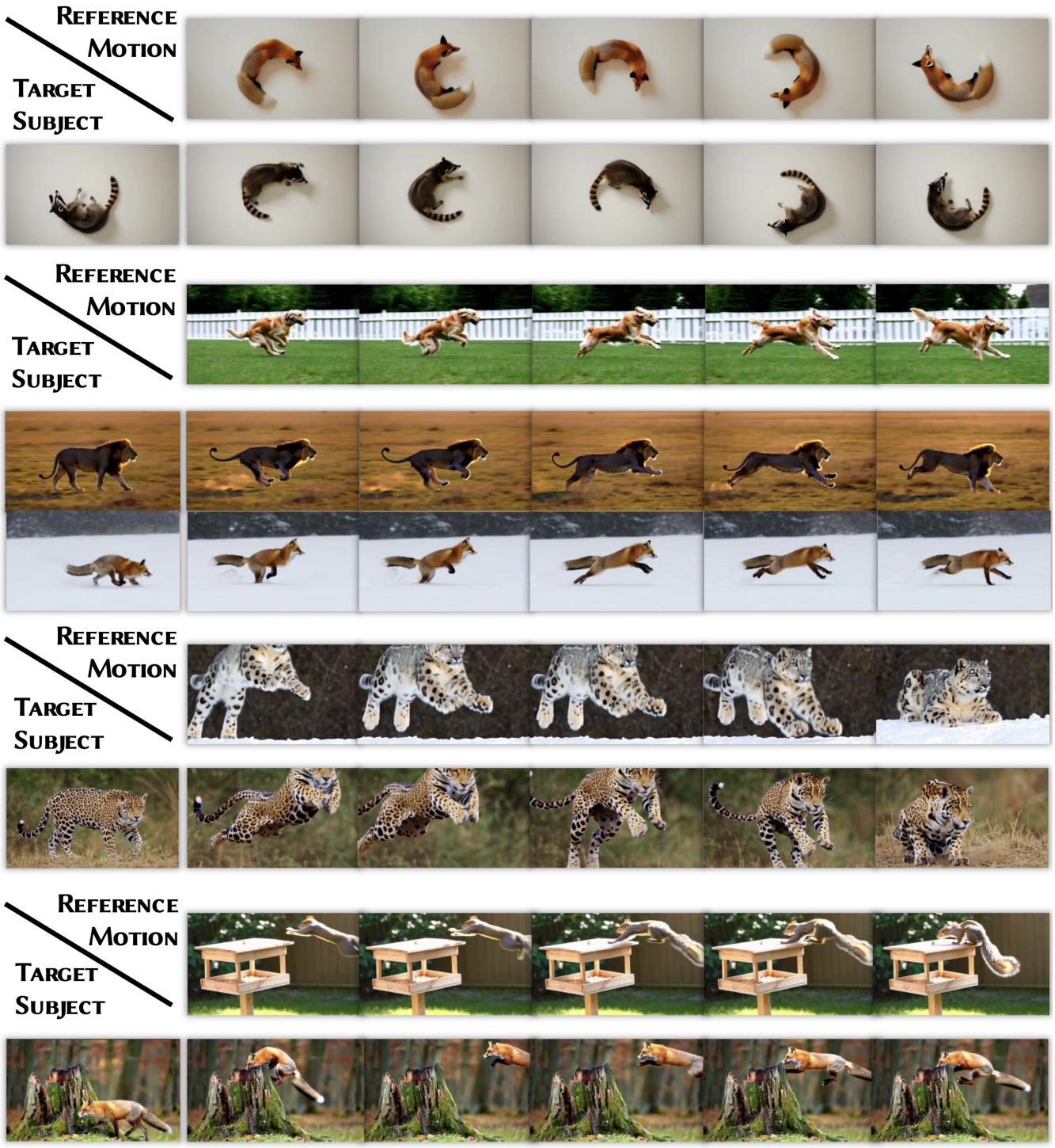

Fig. 9. **Examples of the motion transfer results bt Motion4Motion.**

References

Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. 2020. Skeleton-aware networks for deep motion retargeting. *ACM TOG* 39, 4 (2020), 62–1.

Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2024. Cross-image attention for zero-shot appearance transfer. In *ACM SIGGRAPH 2024 Conference Papers*.

Artell. 2025. Auto-Rig Pro. https://superhivemarket.com/products/auto-rig-pro. Accessed: 2025-05-21.

Ryan Burgert, Charles Herrmann, Forrester Cole, Michael S Ryoo, Neal Wadhwa, Andrey Voynov, and Nataniel Ruiz. 2025a. MotionV2V: Editing Motion in a Video. arXiv:2511.20640 [cs.CV] https://arxiv.org/abs/2511.20640

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. 2025b. Go-with-the-Flow: Motion-Controllable Video Diffusion Models Using Real-Time Warped Noise. In *CVPR*.

Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. 2025. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 7763–7772.

Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*. 22560–22570.

Ling-Hao Chen, Shunlin Lu, Wenxun Dai, Zhiyang Dou, Xuan Ju, Jingbo Wang, Taku Komura, and Lei Zhang. 2024. Pay attention and move better: Harnessing attention for interactive motion generation and training-free editing. *arXiv preprint arXiv:2410.18977* (2024).

Ling-Hao Chen, Jiawei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, and Tongliang Liu. 2023. Humanmac: Masked motion completion for human motion prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9544–9555.

Ling-Hao Chen, Yuhong Zhang, Zixin Yin, Zhiyang Dou, Xin Chen, Jingbo Wang, Taku Komura, and Lei Zhang. 2025. Motion2Motion: Cross-topology Motion Transfer with Sparse Correspondence. In *SIGGRAPH Asia 2025*. ACM, Hong Kong, China. doi:10.1145/3757377.3763811

Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. 2025. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055* (2025).

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, Yu Liu, Yingya Zhang, and Yujiu Yang. 2025. Wan-Move: Motion-controllable Video Generation via Latent Trajectory Guidance. *arXiv preprint arXiv:2512.08765* (2025).

Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. 2024. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*. Springer, 390–408.

Yingying Deng, Xiangyu He, Fan Tang, and Weiming Dong. 2024. Z*: Zero-shot Style Transfer via Attention Rearrangement. In *CVPR*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first ICML*.

Andrew Feng, Yazhou Huang, Yuyu Xu, and Ari Shapiro. 2012. Automating the transfer of a generic set of behaviors onto a virtual character. In *Motion in Games: 5th International Conference, MIG 2012, Rennes, France, November 15-17, 2012. Proceedings 5*. Springer, 134–145.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Carl Doersch, Yusuf Aytar, Michael Rubinstein, Chen Sun, Oliver Wang, Andrew Owens, and Deqing Sun. 2025. Motion Prompting: Controlling Video Generation with Motion Trajectories. In *CVPR*.

Michael Gleicher. 1998. Retargetting motion to new characters. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*. 33–42.

Ahmet Berke Gokmen, Yigit Ekin, Bahri Batuhan Bilecen, and Aysegul Dundar. 2025. RoPECraft: Training-Free Motion Transfer with Trajectory-Guided RoPE Optimization on Diffusion Transformers. *arXiv preprint arXiv:2505.13344* (2025).

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, Wenping Wang, and Yuan Liu. 2025. Diffusion as Shader: 3D-aware Video Diffusion for Versatile Video Generation Control. *arXiv preprint arXiv:2501.03847* (2025).

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *ICLR*.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations*.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. 2024. Style aligned image generation via shared attention. In *CVPR*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *NeurIPS* 33 (2020), 6840–6851.

Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*. 8153–8163.

Guanlong Jiao, Biqing Huang, Kuan-Chieh Wang, and Renjie Liao. 2025. UniEdit-Flow: Unleashing Inversion and Editing in the Era of Flow Models. arXiv:2504.13109 [cs.CV] https://arxiv.org/abs/2504.13109

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. 2025. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *CVPR*. 6013–6022.

Jehee Lee and Sung Yong Shin. 1999. A hierarchical approach to interactive motion editing for human-like figures. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 39–48.

Shujie Li, Lei Wang, Wei Jia, Yang Zhao, and Liping Zheng. 2022. An iterative solution for improving the generalization ability of unsupervised skeleton motion retargeting. *Computers & Graphics* 104 (2022), 129–139.

Jongin Lim, Hyung Jin Chang, and Jin Young Choi. 2019. Pmnet: Learning of disentangled pose and movement for unsupervised motion retargeting. In *30th British Machine Vision Conference (BMVC 2019)*. British Machine Vision Association, BMVA.

Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. 2025. MotionClone: Training-Free Motion Cloning for Controllable Video Generation. In *ICLR*.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. [n. d.]. Flow Matching for Generative Modeling. In *ICLR*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. 2024a. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*. Springer, 38–55.

Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. 2024b. Video-p2p: Video editing with cross-attention control. In *CVPR*. 8599–8608.

Yanchen Liu, Yanan Sun, Zhening Xing, Junyao Gao, Kai Chen, and Wenjie Pei. 2025. MotionShot: Adaptive Motion Transfer across Arbitrary Objects for Text-to-Video Generation. In *ICCV*. 11861–11871.

Shunlin Lu, Ling-Hao Chen, Ailing Zeng, Jing Lin, Ruimao Zhang, Lei Zhang, and Heung-Yeung Shum. 2023. Humantomato: Text-aligned whole-body motion generation. *arXiv preprint arXiv:2310.12978* (2023).

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *ICCV*. 4195–4205.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PmLR, 8748–8763.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *ICML*. PMLR, 1060–1069.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024).

Rokoko. 2021. Rokoko Motion Capture Solutions. https://www.rokoko.com/.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*. 10684–10695.

Yukai Shi, Jiarong Ou, Rui Chen, Haotian Yang, Jiahao Wang, Xin Tao, Pengfei Wan, Di Zhang, and Kun Gai. 2025. Imbalance in Balance: Online Concept Balancing in Generation Models. In *ICCV*. 17432–17442.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. 2026. MotionStream: Real-Time Video Generation with Interactive Motion Controls. In *ICLR*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. 2023. Emergent Correspondence from Image Diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=ypOiXjdfnU

Yoad Tewel, Omri Kaduri, Rinon Gal, Yoni Kasten, Lior Wolf, Gal Chechik, and Yuval Atzmon. 2024. Training-free consistent text-to-image generation. *ACM Transactions on Graphics (TOG)* 43, 4 (2024), 1–18.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*. 1921–1930.

Team Wan, 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. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).

Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. 2025b. ATI: Any Trajectory Instruction for Controllable Video Generation. *arXiv preprint arXiv:2505.22944* (2025).

Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. 2023. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *CVPR*. 17979–17989.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang Yu. 2025d. UniVerse-1: Unified Audio-Video Generation via Stitching of Experts. *arXiv preprint arXiv:2509.06155* (2025).

Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. 2025c. Taming Rectified Flow for Inversion and Editing. In *Forty-second ICML*.

Mengyu Wang, Henghui Ding, Jianing Peng, Yao Zhao, Yunpeng Chen, and Yunchao Wei. 2025a. Characonsist: Fine-grained consistent character generation. In *ICCV*. 16058–16067.

Jie Yang, Ailing Zeng, Ruimao Zhang, and Lei Zhang. 2024. X-pose: Detecting any keypoints. In *European Conference on Computer Vision*. Springer, 249–268.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2025. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *ICLR*.

Zixin Yin, Ling-Hao Chen, Lionel Ni, and Xili Dai. 2025a. ConsistEdit: Highly Consistent and Precise Training-free Visual Editing. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. 1–11.

Zixin Yin, Xili Dai, Ling-Hao Chen, Deyu Zhou, Jianan Wang, Duomin Wang, Gang Yu, Lionel M Ni, Lei Zhang, and Heung-Yeung Shum. 2025b. Training-Free Text-Guided Color Editing with Multi-Modal Diffusion Transformer. *arXiv preprint arXiv:2508.09131* (2025).

Zixin Yin, Xili Dai, Duomin Wang, Xianfang Zeng, Lionel M Ni, Gang Yu, and Heung-Yeung Shum. 2025c. LazyDrag: Enabling Stable Drag-Based Editing on Multi-Modal Diffusion Transformers via Explicit Correspondence. *arXiv preprint arXiv:2509.12203* (2025).

Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, and Baoyuan Wang. 2023. Talking head generation with probabilistic audio-to-visual diffusion priors. In *ICCV*. 7645–7655.

Shiyi Zhang, Junhao Zhuang, Zhaoyang Zhang, Ying Shan, and Yansong Tang. 2025b. Flexiact: Towards flexible action control in heterogeneous scenarios. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–11.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Yuefeng Zhu, FangYuan Zou, et al. 2025a. MimicMotion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance. In *Forty-second ICML*.

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