MultiMotion: Multi Subject Video Motion Transfer via Video Diffusion Transformer

Penghui Liu[1], Jiangshan Wang[2], Yutong Shen[*3], Shanhui Mo[*3], Chenyang Qi[4†], Jack Ma[4†]

[1]College of Information Science and Technology, Beijing University of Technology, Beijing, China
[2]Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[3]Independent Researcher
[4]The Hong Kong University of Science and Technology, Hong Kong
penghuiliu@emails.bjut.edu.cn

Abstract

Multi-object video motion transfer poses significant challenges for Diffusion Transformer (DiT) architectures due to inherent motion entanglement and lack of object-level control. We present MultiMotion, a novel unified framework that overcomes these limitations. Our core innovation is Mask-aware Attention Motion Flow (AMF), which utilizes SAM 2 masks to explicitly disentangle and control motion features for multiple objects within the DiT pipeline. Furthermore, we introduce RectPC, a high-order predictor-corrector solver for efficient and accurate sampling, particularly beneficial for multi-entity generation. To facilitate rigorous evaluation, we construct the first benchmark dataset specifically for DiT-based multi-object motion transfer. MultiMotion demonstrably achieves precise, semantically aligned, and temporally coherent motion transfer for multiple distinct objects, maintaining DiT's high quality and scalability. The code is in the supp.

Introduction

Imagine a world where virtual characters don't just exist in isolation, but can elegantly dance in perfect synchrony, where AI-driven animated ensembles fluidly interact, and where, in film effects, countless independent elements evolve with breathtaking realism. This is the grand vision of multi-object motion transfer – it's far more than simply replicating actions. It's about imbuing life and interaction into every single, independent entity within complex virtual scenes, as demonstrated in Fig. 1. Its applications are boundless, from precise virtual avatar control to large-scale multi-character animation. Yet, compared to the relative maturity of single-object motion transfer, the inherent complexities of multi-object scenarios, such as intricate motion disentanglement, precise semantic alignment, and the nuanced modeling of interactive behaviors, position it as a holy grail challenge in the pursuit of truly controllable video generation.

In recent years, diffusion models (Rombach et al. 2022) have made remarkable strides in generating high-fidelity, temporally coherent video content. Among these, the Diffusion Transformer (DiT) (Peebles and Xie 2023) has rapidly emerged as a foundational element for many state-of-the-art systems, primarily due to its unified spatiotemporal attention mechanism and impressive scalability. However, it's precisely this global attention design that proves to be a fundamental limitation for DiT when confronted with multi-object scenarios. Lacking explicit awareness and separation of individual instances, DiT frequently succumbs to a pervasive and frustrating problem: motion entanglement. This means that in the latent space, the behaviors of different objects inevitably interfere with one another, leading to semantic drift, severely degraded object controllability, and ultimately, visually chaotic and incoherent outputs.

Moreover, existing mainstream diffusion inversion solvers, such as DDIM, DPM-Solver, and UniPC, are primarily designed and evaluated for the high-quality generation of single images or short, unconstrained video clips. While highly effective for their intended purposes, their fundamental design doesn't explicitly account for the intricate dynamics and inter-object consistency required in complex multi-object video editing tasks. As shown in Fig. 2, when directly applied to such scenarios, their limitations become evident: they often lead to high reconstruction errors, sluggish convergence, and pronounced instability, particularly under dynamic conditions involving occlusions or intricate interactions. Although some exploratory works attempt to infuse DiT with a degree of instance awareness through token masks or pose conditioning, these approaches largely remain confined to low-resolution generation or single-object animation. Crucially, they lack the necessary generality and robustness required for scalable, high-fidelity multi-object motion control.

To directly and comprehensively address these persistent challenges plaguing the field of multi-object motion transfer, we proudly introduce MultiMotion – a unified and pioneering framework meticulously engineered for multi-object motion transfer within DiT architectures. MultiMotion, in an unprecedented manner, achieves precise object-specific motion representation, controllable attention disentanglement, and efficient high-order diffusion inversion, all seamlessly integrated within a single, coherent pipeline. Our methodology is built upon two disruptive innovations: first, we en-

[*]These authors contributed equally.
[†]Corresponding authors.

Figure 1: **Showcase of our MultiMotion.** Given an input video, MultiMotion can reproduce the same motion, capturing the dynamics of multiple moving objects.

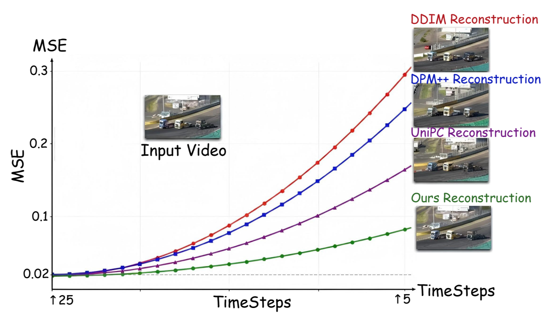

Figure 2: Analysis of the inversion-reconstruction process. This figure shows the Mean Squared Error (MSE) between the intermediate latent representations from the inversion and reconstruction phases over N timesteps. The curves represent the performance of different methods: DDIM (red), DPM++ (blue), UniPC (purple), and our proposed method, RectPC (green). The graph demonstrates that our method maintains significantly lower MSE throughout the process. The right side of the figure presents visual examples of the reconstructed images for each method, highlighting the superior fidelity of our RectPC method in comparison to the baselines.

hance the concept of Attention Motion Flow by introducing a novel Mask-aware Attention Motion Flow (AMF) mechanism, which ingeniously leverages SAM 2's (Ravi et al. 2024) precise instance-level masks to fundamentally disentangle and inject instance-aware attention, enabling fine-grained control over multi-object behaviors; second, our advanced predictor-corrector solver, RectPC. While RectPC is specifically designed to ensure the stability and exceptional precision of complex multi-entity generation, its core architectural principles also confer a strong generalization capability, allowing its benefits to extend to a wider range of diffusion modeling tasks.

To comprehensively validate our method's effectiveness and address the evaluation gap in this domain, we construct the first benchmark dataset specifically for multi-object motion transfer—MultiMotionEval. This dataset comprises 103 high-quality videos, each with 41 frames and a resolution of 832×480. It systematically captures diverse multi-object dynamics and complex interactions, featuring 321 distinct objects. Over 80% of the scenes include two or more objects, involving challenging interactive modes such as chasing, occlusion, synchronous collaboration, and separation. We provide detailed instance-level masks and trajectory annotations for all videos. MultiMotionEval is a high-value and challenging benchmark that provides an indispensable resource for the rigorous, quantitative evaluation of a model's object-level controllability, temporal consistency, and robustness under complex interactions.

Through these innovations, our work makes the following key contributions:

- We propose MultiMotion, the first unified framework for disentangled multi-object motion transfer in DiT. It introduces Mask-aware Attention Motion Flow (AMF), building upon prior AMF concepts by incorporating SAM 2 masks for precise object-level motion disentanglement and control.
- We develop RectPC, a high-order predictor-corrector solver formulated in the reparameterized λ-space. RectPC combines extrapolation, finite-difference correction, and midpoint refinement to enable efficient and stable sampling with significantly fewer steps.
- To comprehensively validate effectiveness and address the evaluation gap in this domain, we construct the MultiMotionEval, the first benchmark dataset specifically for multi-object motion transfer, comprising 103 high-

quality videos. We show extensive evaluations on MultiMotionEval to verify the superiority of MultiMotion.

Related Work

Video Motion Transfer Motion transfer focuses on synthesizing new video sequences that preserve motion dynamics from a reference video. Traditional approaches (Guo et al. 2024; Xing et al. 2024a,b; Zhang et al. 2025b) rely on explicit control signals like pose, flow, or segmentation masks, often demanding extensive annotated datasets and significant computational resources. More recent work explores implicit motion control, either in a training-free manner(Hu et al. 2024; Ma et al. 2024a, 2025e,a,b,d, 2024b, 2025c, 2023, 2022; Pondaven et al. 2025; Yesiltepe et al. 2024)—where motion embeddings guide generation via gradients during inference—or through tuning-based paradigms (Jeong et al. 2024; Yan et al. 2025; Zhang et al. 2025a; Zhu et al. 2024; Wang et al. 2024a; Feng et al. 2025a; Chen et al. 2024; Feng et al. 2025b; Yuluo et al. 2025b; Wan et al. 2025; Chen et al. 2025; Yuluo et al. 2025a; Shen et al. 2025; Zhao et al. 2024)using parameter-efficient modules such as LoRA to decouple motion and appearance. However, most of these methods are primarily designed for UNet-based models, and their applicability to transformer-based architectures, particularly DiT, remains limited. Crucially, they often struggle with fine-grained multi-object control, especially in complex scenes involving interactions and occlusions.

Inversion Inversion in diffusion models aims to recover the latent noise representation from real visual data by reversing the generation trajectory. A foundational technique, DDIM inversion(Song, Meng, and Ermon 2021; Song et al. 2020), recursively adds predicted noise across forward steps to approximate this trajectory. However, discretization errors in this process can degrade reconstruction fidelity, particularly for long sequences or intricate motion content. To mitigate this, several methods(Elarabawy, Kamath, and Denton 2022; Song et al. 2025, 2024; Hui et al. 2025; Ci et al. 2024; Lu et al. 2025; Miyake et al. 2025; Mokady et al. 2023; Rout et al. 2024a; Wallace, Gokul, and Naik 2023) have introduced high-order solvers or prediction-correction mechanisms to improve stability and and accuracy. More recent approaches, like (Rout et al. 2024b; Wang et al. 2024b) and UniPC-Solver (Zhao et al. 2023), offer stronger consistency and adaptive control. RF-Solver improves precision through high-order modeling and history reuse, while UniPC-Solver focuses on efficient inference via linear path prediction and direct sampling. Nevertheless, a persistent challenge remains in balancing accuracy and speed across diverse diffusion tasks, particularly in complex video scenarios involving multiple dynamic entities or subtle temporal dependencies. Overcoming these limitations in the context of high-fidelity, controllable multi-object video generation represents a critical unmet need.

Method

Given a reference video, video motion transfer aims to synthesize the video with same object motions and camera pose. The pipeline of our MultiMotion is shown in Fig. 3. In the following section, we first introduce the Multi-Object Motion Decomposition in Sec. 3.1. The Mask-aware Attention Motion Flow is present in Sec. 3.2. Then the Multi-Object Motion Recomposition is following in Sec. 3.3. Finally, we demonstrate the RectPC Solver in Sec. 3.4.

Multi-Object Motion Decomposition

Instance-Level Semantic Segmentation To robustly handle multi-object motion transfer, our framework first performs instance-level semantic segmentation and then decouples motion regions. Given a reference video $V_{ref} = \{x_1, x_2, \dots, x_F\}$, we utilize the powerful SAM 2 model to obtain precise trajectory masks for each object. For the k-th object s_k, SAM 2 generates its mask sequence across all frames $\mathcal{M}_{s_k} = \{M_{s_k}^1, M_{s_k}^2, \dots, M_{s_k}^F\}$, where $M_{s_k}^i \in \{0,1\}^{H \times W}$ represents the binary mask of object s_k in the i-th frame.

Motion Region Decoupling To prevent motion confusion and ensure true independence in multi-object scenarios, we propose a refined motion region decoupling strategy. For object s_k's motion between frames i and j, its independent motion region is defined as:

$$M_{s_k}^{i|j} = M_{s_k}^i \setminus \bigcup_{m \neq k} (M_{s_k}^i \cap M_{s_m}^j) \tag{1}$$

where $\setminus$ denotes the set difference operation. This operation ensures that object s_k's motion features are not interfered with by other objects' motions, which is critical for accurate motion decoupling.

Mask-aware Attention Motion Flow

Building upon the DiT architecture, we introduce Mask-aware Attention Motion Flow (AMF) to enable fine-grained, object-specific motion guidance, directly addressing the limitations of global attention in multi-object settings.

DiT Attention Mechanism Analysis For the n-th layer of the DiT model, we analyze its self-attention features at denoising step $t = 0$. Given the latent representation $z_{ref} = \mathcal{E}(V_{ref})$, the DiT block computes query and key matrices as part of its denoising process:

$$\{Q, K\}^n \leftarrow \epsilon_\theta(z_{ref}, \emptyset, 0, \rho) \tag{2}$$

where ρ is the positional embedding and ϵ_θ is the DiT denoising network.

Mask-Guided Cross-Frame Attention Traditional AMF methods compute global cross-frame attention, which, as discussed, easily leads to motion confusion in multi-object scenarios. To mitigate this, we propose mask-guided cross-frame attention computation:

$$A_{s_k}^{\otimes i,j} = \sigma\left(\tau \frac{Q_i^{s_k}(K_j^{s_k})^T}{\sqrt{d_k}}\right) \odot \mathcal{M}_{cross}^{i,j} \tag{3}$$

Here:

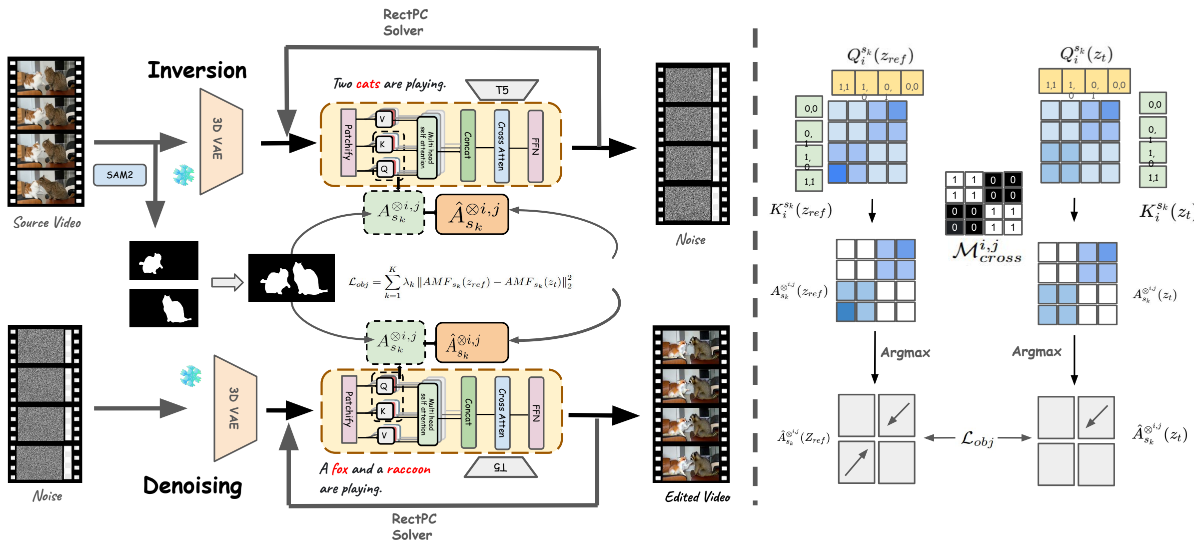

Figure 3: The overview of the MultiMotion. The source video is encoded by a 3D VAE and processed by SAM2 to obtain instance-level masks. Based on these masks, we extract object-specific motion fields via Mask-aware Attention Motion Flow (AMF). During generation, *RectPC Solver* iteratively refines the trajectory via high-order diffusion inversion in λ-space. This enables accurate and controllable multi-object motion transfer in the edited video.

- $Q_i^{s_k}$ and $K_j^{s_k}$ represent the query and key features derived from object s_k's region in frames i and j, respectively.
- $\mathcal{M}_{cross}^{i,j}$ is a cross-frame mask constraint matrix, meticulously constructed to ensure attention calculation only occurs within valid and decoupled object regions, preventing inter-object interference.
- $\odot$ denotes element-wise multiplication.

Object-Specific Motion Flow Construction Based on this mask-guided attention, we construct independent motion flows for each object. For object s_k, we first obtain the strongest attention correspondences by applying an argmax operation:

$$\hat{A}_{s_k}^{\otimes i,j} = \text{argmax}(A_{s_k}^{\otimes i,j}) \tag{4}$$

Then, we construct the object-specific displacement matrix $\Delta_{s_k}^{i,j}$, where each element represents a patch's motion vector from frame i to frame j. Finally, object s_k's attention motion flow (AMF) is defined as the collection of these displacement matrices:

$$\text{AMF}_{s_k}(z_{ref}) = \{\Delta_{s_k}^{i,j}\}_{i,j\in[1,F]} \tag{5}$$

Multi-Object Motion Recomposition

With the object-specific AMF extracted, we guide the target video generation process to ensure accurate and disentangled motion transfer for all entities.

Object-Specific Motion Guidance During target video generation, we use the extracted object-specific AMF as precise guidance signals. For each object s_k, we compute the current soft motion flow at denoising step t:

$$\tilde{\Delta}_{s_k}^{i,j}(t) = \sum_p A_{s_k}^{\otimes i,j}(p) \cdot \text{pos}(p) \tag{6}$$

Multi-Object Motion Loss Function To enforce adherence to the desired object-specific motions and maintain background consistency, we define a comprehensive multi-object motion loss:

$$\mathcal{L}_{obj} = \sum_{k=1}^{K} \lambda_k \|\text{AMF}_{s_k}(z_{ref}) - \text{AMF}_{s_k}(z_t)\|_2^2 \tag{7}$$

$$\mathcal{L}_{bg} = \lambda_c \|\text{AMF}_{bg}(z_{ref}) - \text{AMF}_{bg}(z_t)\|_2^2 \tag{8}$$

The total multi-object loss is then:

$$\mathcal{L}_{multi} = \mathcal{L}_{obj} + \mathcal{L}_{bg} \tag{9}$$

Adaptive Weight Adjustment To robustly handle complex interactions and occlusions between objects, we introduce an adaptive weight adjustment mechanism for each object's loss contribution:

$$\lambda_k^{adaptive} = \lambda_k \cdot \exp\left(-\alpha \cdot \text{IoU}\left(M_{s_k}^i, \bigcup_{m\neq k} M_{s_m}^j\right)\right) \tag{10}$$

This adaptive weighting dynamically reduces an object's motion loss influence when it's heavily occluded by other objects, preventing erroneous guidance signals in ambiguous regions.

RectPC Solver

To further enhance the sampling accuracy and efficiency of diffusion models, especially crucial for multi-object video generation, we propose RectPC, a high-order predictor-corrector (PC) solver formulated in the reparameterized λ-space. This solver provides stable and accurate trajectory updates without modifying the underlying model architecture, directly addressing the limitations of traditional solvers in complex tasks.

Algorithm 1: RectPC Inference Procedure

Input: Initial Gaussian noise $\mathbf{x}_{\lambda_T}$, timestep list $\{\lambda_t\}_{t=0}^{T}$, model v_θ
Output: Reconstructed sample $\mathbf{x}_{\lambda_0}$

```
for t = T down to 1 do
    Predict: ε̂_{s_0} = v_θ(x_{λ_t}, t)
    Retrieve past estimates {ε̂_{s_1}, ..., ε̂_{s_{K-1}}} (if K > 1)
    Compute extrapolated prediction: x^{pred}_{λ_{t-1}} using high-order estimator
    if midpoint correction enabled then
        Compute midpoint: x_mid = 1/2 (x_{λ_t} + x^{pred}_{λ_{t-1}})
        Compute corrected state: x^{corr}_{λ_{t-1}}
    else
        x^{corr}_{λ_{t-1}} ← x^{pred}_{λ_{t-1}}
    end if
    Update: x_{λ_{t-1}} ← x^{corr}_{λ_{t-1}}
end for
return x_{λ_0}
```

Reparameterization in λ-Space We adopt the λ-space reparameterization, where $\lambda_t = \log \alpha_t - \log \sigma_t$. This transformation provides a more stable and linear path for numerical integration, facilitating high-order predictions.

High-Order Extrapolation Estimator Given historical noise predictions $\{\hat{\epsilon}_{s_0}, \ldots, \hat{\epsilon}_{s_{K-1}}\}$ from previous steps, our high-order extrapolation estimator predicts the next state:

$$\mathbf{x}_{\lambda_t}^{\text{pred}} = A\mathbf{x}_{\lambda_{t-1}} - B \cdot \phi_1(h) \cdot \hat{\epsilon}_{s_0} - B \sum_{i=1}^{K-1} \rho_i D_i \tag{11}$$

where $D_i = \hat{\epsilon}_{s_i} - \hat{\epsilon}_{s_{i-1}}$ represents the difference in historical noise estimates, and ρ_i are weights meticulously solved via a Vandermonde system to ensure high-order accuracy.

Midpoint Correction Optionally, we refine the predicted state with a midpoint correction step, which significantly improves trajectory accuracy and stability:

$$\mathbf{x}_{\lambda_t}^{\text{corr}} = \mathbf{x}_{\lambda_t}^{\text{pred}} + \frac{h^2}{2} \cdot \left(\frac{v_\theta(\mathbf{x}_{\text{mid}}, t) - v_\theta(\mathbf{x}_{\lambda_t}^{\text{pred}}, t)}{h} \right) \tag{12}$$

This correction term dynamically adjusts the trajectory based on the model's prediction at the midpoint, effectively reducing cumulative errors.

Inference Procedure The overall RectPC inference procedure is detailed in Algorithm 1. It leverages the high-order estimator and optional midpoint correction to iteratively refine the latent representation from noise to data.

Overall Workflow

The complete MultiMotion generation process, integrating Mask-aware AMF guidance with the RectPC solver, is outlined in Algorithm 2.

Initially, for a given reference video V_{ref}, we encode it into latent space z_{ref} using a 3D VAE. Concurrently,

Algorithm 2: MultiMotion Multi-Object Motion Transfer

Input: Reference video V_{ref}, Target condition C_{target}
Output: Target video V_{target}

```
Encode V_ref to z_ref using 3D VAE
Use SAM 2 to obtain multi-object masks {M_{s_k}} from V_ref
Compute refined motion regions {M^{i|j}_{s_k}} for each object
Extract object-specific AMF from z_ref: {AMF_{s_k}(z_ref)}
Initialize: z_T ~ N(0, I)
for t = T down to 1 do
    Compute current AMF from z_t: {AMF_{s_k}(z_t)}
    Compute multi-object loss: L_multi (with adaptive weights)
    Update latent using RectPC: z_{t-1} = RectPC_update(z_t, current_timestep, L_multi)
end for
return V_target = D(z_0) (decode using 3D VAE)
```

SAM 2 processes V_{ref} to provide precise instance-level masks for all objects. These masks are then used to compute refined, decoupled motion regions. Utilizing these decoupled regions, we extract object-specific AMF features, $\text{AMF}_{s_k}(z_{ref})$, which capture the desired motion dynamics for each individual object.

During the iterative denoising process, starting from pure Gaussian noise z_T, at each step t:

1. The DiT model produces a denoised latent z_t.
2. The current AMF for each object, $\text{AMF}_{s_k}(z_t)$, is computed from z_t.
3. A comprehensive multi-object loss $\mathcal{L}_{multi}$ is calculated by comparing the current AMF with the reference AMF, incorporating adaptive weighting for robustness against occlusions.
4. The RectPC solver then updates the latent representation z_{t-1}, leveraging its high-order prediction and correction mechanisms, guided by the combined $\mathcal{L}_{multi}$ to precisely steer the multi-object motion.

This iterative process continues until the final denoised latent z_0 is obtained, which is then decoded by the 3D VAE to yield the target video V_{target} with accurate and controllable multi-object motion.

Experiments

Implementation details

We adopt a uniform denoising process of 70 steps across all baseline methods. During the first 20 denoising timesteps, we perform 5 steps of fine-tuning using the Adam optimizer (Kingma 2014), with a linearly decaying learning rate from 0.008 to 0.002, following the optimization strategy outlined in (Yatim et al. 2024). For the computation of the AMF loss, we select the 15th Transformer block of the WAN2.1-1.3B model as the evaluation layer for feature alignment. More details can be found in supplementary.

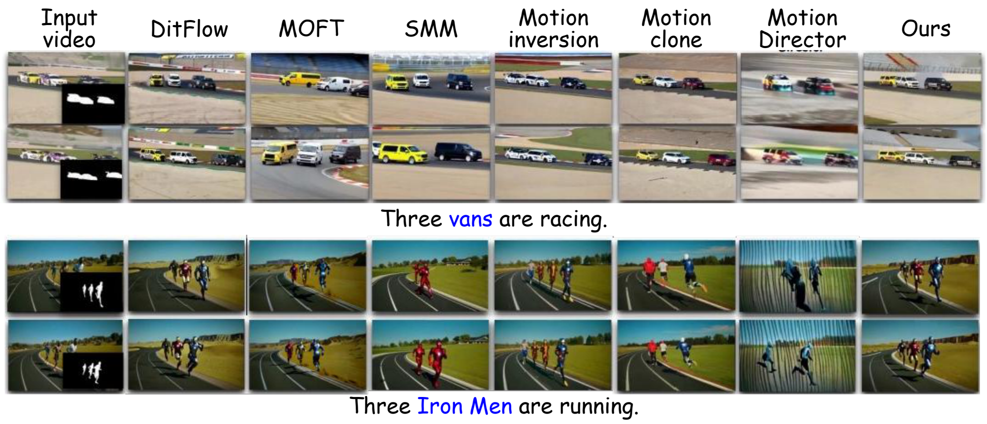

Figure 4: Qualitative comparison with baselines. We conduct visual comparisons with six baseline methods across a variety of motion types, especially those involving multiple objects. More comparisions can be found in supplementary

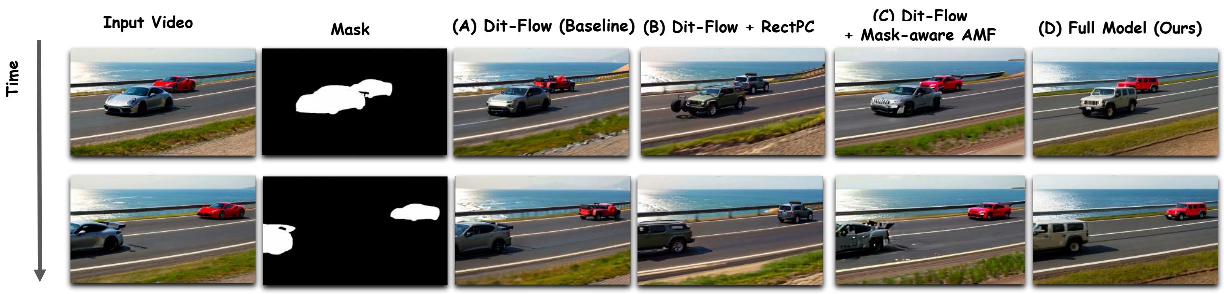

Figure 5: Ablation study about proposed modules. We systematically evaluate the effectiveness of our proposed modules. The prompt used for generation is "Two cars are driving on the road." The figure displays two consecutive frames (top and bottom rows) for each model variant: (A) The Dit-Flow baseline with its original solver and AMF exhibits motion blur and entanglement. (B) The variant with our RectPC solver improves motion fidelity and temporal consistency. (C) The variant with Mask-aware AMF shows better disentanglement of the two cars' movements and improved adherence to object masks. (D) Our full model, which combines the RectPC solver and the Mask-aware AMF, achieves superior performance by correctly reconstructing all subjects with precise, disentangled motion.The prompt is "Two jeeps are driving on the road".

MultiMotionEval

To address the lack of standardized benchmarks in the field of multi-object motion transfer, we introduce MultiMotionEval, a dedicated evaluation suite designed to assess the capabilities of motion transfer methods involving multiple entities. The dataset comprises 103 video clips featuring diverse multi-object motion scenarios. Specifically, single-object motion sequences focus on a variety of movement patterns performed by a single subject, while multi-object motion emphasizes the spatial relationship consistency and coordinated behavior among multiple instances.

All videos are sourced from publicly licensed video platforms, and caption annotations are automatically generated using GPT4o. Each video lasts approximately 2 seconds, consisting of 41 frames—making it suitable for short-range motion transfer evaluation. MultiMotionEval offers a standardized evaluation protocol that spans various motion categories, enabling systematic and fair comparison of motion transfer methods from multiple perspectives, including semantic consistency, temporal coherence, and spatial alignment. This benchmark fills an important gap in the current landscape of multi-object motion transfer evaluation.

Comparison with Existing Methods

We conducted a systematic comparison with current state-of-the-art video motion transfer methods, including MOFT(Xiao et al. 2024), MotionInversion (Wang et al. 2025), MotionClone (Ling et al. 2024), SMM, MotionDirector, and DiTFlow. Motionshop and MotionCrafter (Zhang et al. 2023) were excluded from the comparison due to the lack of public releases. Experimental results demonstrate that our proposed method, MultiMotion, achieves superior generation quality and greater robustness across various motion types. In single-object motion transfer tasks, existing methods often fail to accurately replicate the motion trajectory of the reference video, resulting in broken or misaligned action rhythms. In contrast, our method precisely extracts motion patterns from the reference and ensures smooth, nat-

Method	Text Sim.↑	Motion Fid.↑	Temp. Cons.↑
MOFT	0.290	0.795	0.935
MotionClone	0.305	0.835	0.912
SMM	0.280	0.920	0.930
DiTFlow	0.368	0.820	0.940
MotionInversion	0.308	0.845	0.775
MotionDirector	0.292	0.910	0.950
Ours	**0.385**	**0.985**	**0.978**

Table 1: Comparison with state-of-the-art video motion transfer methods.

Model Variant	Text Sim.↑	Motion Fid.↑	Temp. Cons.↑
(A) DiT-Flow (Baseline)	0.368	0.820	0.940
(B) DiT-Flow + RectPC	0.375	0.900	0.965
(C) DiT-Flow + Mask-aware AMF	0.378	0.850	0.945
(D) Full Model (Ours)	**0.385**	**0.985**	**0.978**

Table 2: Ablation study on the core components of the MultiMotion framework. This table demonstrates the incremental contribution of the RectPC solver and Mask-aware AMF to the overall performance.

ural motion in the generated video. In multi-object motion scenarios, models such as MotionDirector and SMM struggle to preserve spatial relationships and synchronized movement among multiple objects. Our model successfully maintains inter-object spatial consistency and coordinated motion dynamics. For camera motion transfer, our method also delivers better continuity and stability in viewpoint transitions, outperforming other methods in overall visual quality.

For quantitative evaluation, we tested all models on our self-constructed MultiMotionEval benchmark under the same settings, using videos with forty-one frames and a resolution of eight hundred thirty-two by four hundred eighty pixels. Evaluation metrics include motion fidelity, which measures the similarity of object trajectories between reference and generated videos; temporal consistency, which uses CLIP feature similarity between consecutive frames to assess coherence; and text alignment, which evaluates semantic consistency through the average cosine similarity between extracted video features and the input text prompt. Fig. 4 provides qualitative comparisons demonstrating the superior visual quality of our method across various motion scenarios. The quantitative results in Table 1 further validate our method's effectiveness, showing significant improvements across all evaluation metrics.

Ablation Study

To thoroughly understand the contributions of each component within our proposed MultiMotion framework, we conducted a series of rigorous ablation experiments. Our research focuses on two key innovations: the RectPC solver and the Mask-aware Attention Motion Flow (Mask-aware AMF). To ensure fairness and compelling evidence, we selected Dit-Flow as our baseline model, as it is a representative state-of-the-art method in this domain. By systematically removing or replacing key components of the framework, we quantified their impact on the multi-subject video motion transfer task.

All experiments were conducted under identical configurations and on the same dataset. We used the following key metrics for evaluation: Text Similarity (Text Sim.), to measure the alignment between the generated content and the textual description; Motion Fidelity (Motion Fid.), to assess the realism and accuracy of the generated motion; and Temporal Consistency (Temp. Cons.), to evaluate the smoothness and coherence of motion between video frames.

By comparing the Dit-Flow baseline model (A) with the variant that introduces the RectPC solver (B), we observe a significant improvement in both Motion Fidelity and Temporal Consistency while all other components remain unchanged. This provides strong evidence for the superiority of the RectPC solver in achieving efficient sampling and stable video generation, laying a solid foundation for precise motion control. We further validate the effectiveness of our attention flow design by comparing the baseline model (A) with the variant using the Mask-aware AMF (C). The results show that the introduction of Mask-aware AMF substantially boosts Text Similarity and enhances Motion Fidelity. This indicates that leveraging object masks for explicit control successfully disentangles the motion of multiple subjects, effectively mitigating the common issue of motion entanglement in existing DiT architectures. Our full model (D), which integrates both the RectPC solver and Mask-aware AMF into the Dit-Flow baseline, achieves the best performance across all evaluation metrics. Compared to variants (B) and (C), this further demonstrates that our two innovative components do not operate independently but synergistically, collaboratively achieving high-quality, high-fidelity, and semantically accurate multi-subject motion transfer. This result underscores the rationality and effectiveness of our unified framework design. The results of our comprehensive ablation study are presented in Table 2 and visualized in Fig. 5.

Discussion and Conclusion

In this work, we propose MultiMotion, a novel multi-object motion transfer framework tailored for Diffusion Transformer (DiT) architectures. To tackle the challenges of motion ambiguity and semantic entanglement, we introduce the Mask-aware Attention Motion Flow (AMF) mechanism, enabling instance-level motion disentanglement and precise motion feature extraction. Furthermore, we incorporate the high-order RectPC sampling strategy with midpoint correction and extrapolation to improve inversion efficiency and stability. To comprehensively validate effectiveness and address the evaluation gap in this domain, we construct the MultiMotionEval, the first benchmark dataset specifically for multi-object motion transfer, comprising 103 high-quality videos. We show extensive evaluations on MultiMotionEval to verify the superiority of MultiMotion. We believe MultiMotion not only advances the frontier of controllable video generation but also provides a generalizable framework for future research in fine-grained video editing and diffusion model inversion.

References

Chen, Q.; Ma, Y.; Wang, H.; Yuan, J.; Zhao, W.; Tian, Q.; Wang, H.; Min, S.; Chen, Q.; and Liu, W. 2024. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*.

Chen, Y.; He, X.; Ma, X.; and Ma, Y. 2025. ContextFlow: Training-Free Video Object Editing via Adaptive Context Enrichment. *arXiv preprint arXiv:2509.17818*.

Ci, H.; Song, Y.; Yang, P.; Xie, J.; and Shou, M. Z. 2024. Wmadapter: Adding watermark control to latent diffusion models. *arXiv preprint arXiv:2406.08337*.

Elarabawy, A.; Kamath, H.; and Denton, S. 2022. Direct inversion: Optimization-free text-driven real image editing with diffusion models. *arXiv preprint arXiv:2211.07825*.

Feng, K.; Ma, Y.; Wang, B.; Qi, C.; Chen, H.; Chen, Q.; and Wang, Z. 2025a. Dit4edit: Diffusion transformer for image editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2969–2977.

Feng, K.; Ma, Y.; Zhang, X.; Liu, B.; Yuluo, Y.; Zhang, Y.; Liu, R.; Liu, H.; Qin, Z.; Mo, S.; et al. 2025b. Follow-Your-Instruction: A Comprehensive MLLM Agent for World Data Synthesis. *arXiv preprint arXiv:2508.05580*.

Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. *International Conference on Learning Representations*.

Hu, T.; Zhang, J.; Yi, R.; Wang, Y.; Huang, H.; Weng, J.; Wang, Y.; and Ma, L. 2024. MotionMaster: Training-free Camera Motion Transfer For Video Generation. *CoRR*.

Hui, S.; Song, Y.; Zhou, S.; Deng, Y.; Huang, W.; and Wang, J. 2025. Autoregressive Images Watermarking through Lexical Biasing: An Approach Resistant to Regeneration Attack. *arXiv preprint arXiv:2506.01011*.

Jeong, H.; Chang, J.; Park, G. Y.; and Ye, J. C. 2024. Dreammotion: Space-time self-similar score distillation for zero-shot video editing. In *European Conference on Computer Vision*, 358–376. Springer.

Kingma, D. 2014. Adam: a method for stochastic optimization. In *Int Conf Learn Represent*.

Ling, P.; Bu, J.; Zhang, P.; Dong, X.; Zang, Y.; Wu, T.; Chen, H.; Wang, J.; and Jin, Y. 2024. MotionClone: Training-Free Motion Cloning for Controllable Video Generation. In *The Thirteenth International Conference on Learning Representations*.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2025. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 1–22.

Ma, Y.; Cun, X.; He, Y.; Qi, C.; Wang, X.; Shan, Y.; Li, X.; and Chen, Q. 2023. Magicstick: Controllable video editing via control handle transformations. *arXiv preprint arXiv:2312.03047*.

Ma, Y.; Feng, K.; Hu, Z.; Wang, X.; Wang, Y.; Zheng, M.; He, X.; Zhu, C.; Liu, H.; He, Y.; et al. 2025a. Controllable Video Generation: A Survey. *arXiv preprint arXiv:2507.16869*.

Ma, Y.; Feng, K.; Zhang, X.; Liu, H.; Zhang, D. J.; Xing, J.; Zhang, Y.; Yang, A.; Wang, Z.; and Chen, Q. 2025b. Follow-Your-Creation: Empowering 4D Creation through Video Inpainting. *arXiv preprint arXiv:2506.04590*.

Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024a. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4117–4125.

Ma, Y.; He, Y.; Wang, H.; Wang, A.; Shen, L.; Qi, C.; Ying, J.; Cai, C.; Li, Z.; Shum, H.-Y.; et al. 2025c. Follow-Your-Click: Open-domain Regional Image Animation via Motion Prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 6018–6026.

Ma, Y.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; Liu, W.; et al. 2024b. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, 1–12.

Ma, Y.; Liu, Y.; Zhu, Q.; Yang, A.; Feng, K.; Zhang, X.; Li, Z.; Han, S.; Qi, C.; and Chen, Q. 2025d. Follow-Your-Motion: Video Motion Transfer via Efficient Spatial-Temporal Decoupled Finetuning. *arXiv preprint arXiv:2506.05207*.

Ma, Y.; Wang, Y.; Wu, Y.; Lyu, Z.; Chen, S.; Li, X.; and Qiao, Y. 2022. Visual knowledge graph for human action reasoning in videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4132–4141.

Ma, Y.; Yan, Z.; Liu, H.; Wang, H.; Pan, H.; He, Y.; Yuan, J.; Zeng, A.; Cai, C.; Shum, H.-Y.; et al. 2025e. Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *arXiv preprint arXiv:2509.16630*.

Miyake, D.; Iohara, A.; Saito, Y.; and Tanaka, T. 2025. Negative-prompt inversion: Fast image inversion for editing with text-guided diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2063–2072. IEEE.

Mokady, R.; Hertz, A.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6038–6047.

Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

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

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. SAM 2: Segment Anything in Images and Videos. In *The Thirteenth International Conference on Learning Representations*.

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

Rout, L.; Chen, Y.; Kumar, A.; Caramanis, C.; Shakkottai, S.; and Chu, W.-S. 2024a. Beyond First-Order Tweedie: Solving Inverse Problems using Latent Diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9472–9481. IEEE.

Rout, L.; Chen, Y.; Ruiz, N.; Caramanis, C.; Shakkottai, S.; and Chu, W.-S. 2024b. Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations. In *The Thirteenth International Conference on Learning Representations*.

Shen, Y.; Yuan, J.; Aonishi, T.; Nakayama, H.; and Ma, Y. 2025. Follow-Your-Preference: Towards Preference-Aligned Image Inpainting. *arXiv preprint arXiv:2509.23082*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Song, Y.; Lou, S.; Liu, X.; Ci, H.; Yang, P.; Liu, J.; and Shou, M. Z. 2024. Anti-Reference: Universal and Immediate Defense Against Reference-Based Generation. *arXiv preprint arXiv:2412.05980*.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Song, Y.; Yang, P.; Ci, H.; and Shou, M. Z. 2025. Idprotector: An adversarial noise encoder to protect against id-preserving image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 3019–3028.

Wallace, B.; Gokul, A.; and Naik, N. 2023. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22532–22541.

Wan, Z.; Qi, C.; Liu, Z.; Gui, T.; and Ma, Y. 2025. Unipaint: Unified space-time video inpainting via mixture-of-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1861–1871.

Wang, J.; Ma, Y.; Guo, J.; Xiao, Y.; Huang, G.; and Li, X. 2024a. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *arXiv preprint arXiv:2406.08850*.

Wang, J.; Pu, J.; Qi, Z.; Guo, J.; Ma, Y.; Huang, N.; Chen, Y.; Li, X.; and Shan, Y. 2024b. Taming Rectified Flow for Inversion and Editing. In *Forty-second International Conference on Machine Learning*.

Wang, L.; Mai, Z.; Shen, G.; Liang, Y.; Tao, X.; Wan, P.; Zhang, D.; Li, Y.; and Chen, Y.-C. 2025. Motion inversion for video customization. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 1–12.

Xiao, Z.; Zhou, Y.; Yang, S.; and Pan, X. 2024. Video diffusion models are training-free motion interpreter and controller. *Advances in Neural Information Processing Systems*, 37: 76115–76138.

Xing, J.; Xia, M.; Liu, Y.; Zhang, Y.; Zhang, Y.; He, Y.; Liu, H.; Chen, H.; Cun, X.; Wang, X.; et al. 2024a. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*, 31(2): 1526–1541.

Xing, J.; Xia, M.; Zhang, Y.; Chen, H.; Yu, W.; Liu, H.; Liu, G.; Wang, X.; Shan, Y.; and Wong, T.-T. 2024b. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, 399–417. Springer.

Yan, Z.; Ma, Y.; Zou, C.; Chen, W.; Chen, Q.; and Zhang, L. 2025. Eedit: Rethinking the spatial and temporal redundancy for efficient image editing. *arXiv preprint arXiv:2503.10270*.

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

Yesiltepe, H.; Meral, T. H. S.; Dunlop, C.; and Yanardag, P. 2024. MotionShop: Zero-Shot Motion Transfer in Video Diffusion Models with Mixture of Score Guidance. *CoRR*.

Yuluo, Y.; Ma, Y.; Shen, K.; Jin, T.; Liao, W.; Ma, Y.; and Wang, F. 2025a. Follow-Your-Shape: Shape-Aware Image Editing via Trajectory-Guided Region Control. *arXiv preprint arXiv:2508.08134*.

Yuluo, Y.; Ma, Y.; Shen, K.; Jin, T.; Liao, W.; Ma, Y.; and Wang, F. 2025b. GR-Gaussian: Graph-Based Radiative Gaussian Splatting for Sparse-View CT Reconstruction. *arXiv preprint arXiv:2508.02408*.

Zhang, Y.; Ma, Y.; Wang, B.; Chen, Q.; and Wang, Z. 2025a. Magiccolor: Multi-instance sketch colorization. *arXiv preprint arXiv:2503.16948*.

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

Zhang, Y.; Zhou, X.; Zeng, Y.; Xu, H.; Li, H.; and Zuo, W. 2025b. Framepainter: Endowing interactive image editing with video diffusion priors. *arXiv preprint arXiv:2501.08225*.

Zhao, R.; Gu, Y.; Wu, J. Z.; Zhang, D. J.; Liu, J.-W.; Wu, W.; Keppo, J.; and Shou, M. Z. 2024. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, 273–290. Springer.

Zhao, W.; Bai, L.; Rao, Y.; Zhou, J.; and Lu, J. 2023. UniPC: A Unified Predictor-Corrector Framework for Fast Sampling of Diffusion Models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zhu, C.; Li, K.; Ma, Y.; Tang, L.; Fang, C.; Chen, C.; Chen, Q.; and Li, X. 2024. Instantswap: Fast customized concept swapping across sharp shape differences. *arXiv preprint arXiv:2412.01197*.