

Regional Anchor Routing for Multi-Object Motion Transfer

Anonymous submission

Abstract

Compositional motion transfer requires a model to extract motions from multiple reference videos, and apply each motion to a user-specified object in one target scene. Existing methods normally struggle to reliably assign different reference motions to the correct target objects, thus a motion intended for one object may affect another. Meanwhile, some motion-specific tuning approaches are still facing another issue: they must run a new optimization whenever the reference video changes, making repeated transfers inefficient. To address these limitations, we propose RART, a framework that unifies object-wise motion extraction, target-region anchoring, and frame-aligned injection into a frozen video Diffusion Transformer for multi-object motion transfer. For each reference motion, RART extracts temporally ordered features from the masked reference video, and then conditions them on the target first-frame mask to form mask-anchored motion tokens. Frame-aligned motion cross-attention allows each generated frame to read only the motion tokens at the corresponding time slots, composing all object motions in one denoising process. During training, a motion-attention routing loss encourages each target region to attend to its corresponding motion tokens and reduces interference between objects. Therefore, RART can handle novel reference-target combinations without per-video optimization. Experiments show that RART achieves accurate motion preservation, efficient inference, and strong compositional generalization.

Introduction

Diffusion Transformers now generate high-quality videos from text and images (Ho, Jain, and Abbeel 2020; Rombach et al. 2022; Peebles and Xie 2023; Blattmann et al. 2023; Kong et al. 2024; Wan et al. 2025), yet fine-grained object-wise motion control still remains challenging. Such a control paradigm is crucial for character animation, video editing, and multi-subject video creation, where objects usually need to follow different existing motion patterns.

Here, we study the problem of compositional motion transfer: given a target image, multiple reference video-mask pairs, and target object masks, the model transfers each reference motion to its assigned target object in the image while preserving the target image’s original appearance and background. Compared with single-object animation (Karras et al. 2023; Xu et al. 2024; Hu 2024), this setting introduces three coupled requirements. The model must iso-

late the desired motion from each masked reference without carrying over its appearance or source layout, assign each extracted motion to the correct target region, and ensure that each object follows its reference motion at the correct time step. Existing approaches address only part of these requirements. Explicit-control methods typically rely on poses, optical flow, depth, trajectories, or camera motion (Wang et al. 2023b; Zhang et al. 2024; Wang et al. 2024b; Yin et al. 2023; Ma et al. 2024; Chu et al. 2026) to control motion. The signals provide effective structural guidance when reliable priors are available, but obtaining and aligning them across objects often requires task-specific extraction or annotation. Implicit approaches transfer motion directly from reference videos without converting it into explicit pose or optical-flow conditions, offering greater flexibility. Among them, training-free methods reuse diffusion features or attention for zero-shot transfer (Yatim et al. 2024; Ku et al. 2024; Ling et al. 2025), but mainly handle a single motion source and lack object-wise motion reassignment. Training-based methods learn motion representations, but mostly focus on single-object transfer (Zhao et al. 2024; Jeong, Park, and Ye 2024; Ouyang et al. 2024; Zhang et al. 2025) and require retraining for each new motion (Zhang et al. 2025), limiting their generalization. Recent studies have explored feed-forward implicit motion transfer (Ressler-Antal et al. 2025; Yan et al. 2026). DisMo focuses on single-object transfer, and SCAIL-2 does not support motion transfer from multiple independent reference videos.

To address these limitations, we propose RART, a newly-minted framework that factorizes compositional motion transfer into object-wise motion extraction, target-region anchoring, and frame-aligned injection. For each reference video, the reference mask isolates the desired motion source, and a pretrained motion encoder extracts a temporally ordered feature sequence. RART projects these features into the DiT token space and conditions them on the corresponding target mask, producing motion tokens that encode both the reference motion and its target-region anchor. Motion tokens are then grouped by latent-frame index, and concatenated across objects within each group. Tokens associated with different video frames remain separate. At each denoising step, the video latent tokens attend to the motion tokens belonging to corresponding frame index, not only preserving alignment between each reference motion and its target ob-

This is an anonymized submission for review purposes only. Distribution, citation, or public sharing of this manuscript is strictly prohibited. Copyright and publication details will appear in the final version if accepted.



Figure 1: Compositional motion transfer with RART. Masked reference videos provide object-wise motions, which are assigned to colored target regions and jointly rendered in the generated video. The examples show (a) and (b) motion reassignment, (c) cross-species transfer, and (d) multi-object composition.

ject’s motion but also allowing multiple reference motions to be injected jointly. During training, a motion-attention routing loss further discourages each target region from attending to motion tokens of other objects, while preserving attention to its own tokens. After optimization, RART can transfer motions from new reference videos without video-specific optimization. Representative results are shown in Figure 1.

Our contributions are summarized as follows:

- We propose RART, a framework for compositional multi-object motion transfer on a frozen video Diffusion Transformer, supporting flexible reference-target combinations without per-video optimization.
- We introduce mask-anchored motion encoding and frame-aligned motion cross-attention, separating motion content from its target location and preserving temporal correspondence during control injection, respectively.
- We propose motion-attention routing loss, which encourages each target region to attend to its assigned motion tokens and suppresses cross-object information leakage.

Related Work

Video Diffusion Models. Diffusion models have become the dominant backbone for visual generation (Song and Ermon 2019; Ho, Jain, and Abbeel 2020; Song et al. 2021; Song, Meng, and Ermon 2021; Dhariwal and Nichol 2021; Ho and Salimans 2021; Rombach et al. 2022; Peebles and Xie 2023; Esser et al. 2024; Lipman et al. 2023). Recent video generation systems extend these priors with temporal modules, latent video diffusion, or transformer-based video modeling (Ho et al. 2022b; Singer et al. 2023; Ho et al. 2022a; Villegas et al. 2023; Hong et al. 2023; He et al. 2022; Wang et al. 2023a; Chen et al. 2023; Blattmann et al. 2023; Brooks et al. 2024; Kong et al. 2024; Wan et al. 2025). These models provide strong priors for realistic video synthesis, but text or image conditions alone are still insufficient for fine-grained object-wise motion control, which should be emphasized in research field.

Explicit Motion Transfer. One line of controllable video generation relies on explicit motion conditions such as poses, trajectories, depth, optical flow, or camera motion. VideoComposer (Wang et al. 2023b) uses compositional spatial and temporal conditions for motion controllability. ControlVideo (Zhang et al. 2024) adapts structural controls to text-to-video generation. MotionCtrl (Wang et al. 2024b) separates camera and object motion control. DragNUWA (Yin et al. 2023) integrates text, image, and trajectory conditions for fine-grained control. Wan-Move (Chu et al. 2026) uses latent trajectory guidance for motion-controllable generation. Pose-guided animation methods drive a target subject according to a frame-wise pose sequence (Ma et al. 2024; Karras et al. 2023; Xu et al. 2024; Hu 2024). In summary, these methods are effective when reliable explicit structures are available, but they introduce extra extraction and alignment costs and often specialize to particular object categories. RART, in contrast, extracts motion directly from reference videos, and can be applied to cross-object motion transfer.

Implicit Motion Transfer. Another line of work transfers motion through implicit features, which can be divided into two categories. Training-free approaches keep model parameters fixed but exploit pretrained representations in different ways. Space-Time Diffusion Features (Yatim et al. 2024) guides sampling with an internal feature loss. AnyV2V (Ku et al. 2024) injects source-video features into an image-to-video process, and MotionClone (Ling et al. 2025) reuses sparse temporal-attention weights. These methods avoid reference-specific training but generally process one motion source at a time. In comparison, training-based methods (Zhao et al. 2024; Ouyang et al. 2024) fit motion representations to individual inputs. For instance, MotionDirector (Zhao et al. 2024) learns spatial and temporal LoRAs. VMC (Jeong, Park, and Ye 2024) adapts temporal layers through motion distillation. FlexiAct optimizes frequency-aware embeddings for each reference video (Zhang et al. 2025). FlexiMMT further extends FlexiAct to multiple objects by learning object-wise motion tokens and constraining

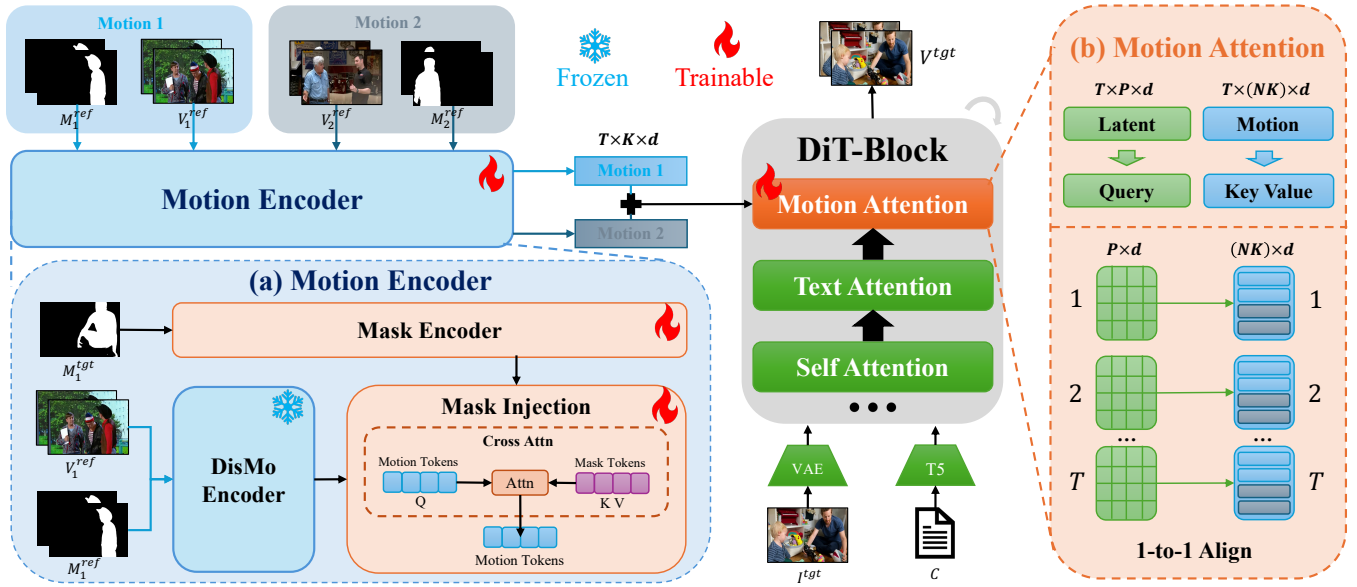


Figure 2: Overview of RART. (a) The motion encoder extracts temporally ordered features from each masked reference video and injects the corresponding target-mask anchor to produce object-wise motion tokens. (b) Frame-aligned motion cross-attention injects the motion tokens of each latent frame into the frozen DiT.

them with masks. However, it still needs to optimize motion tokens for each reference video (Li et al. 2026). Recent feed-forward approaches avoid such repeated optimization. DisMo (Ressler-Antal et al. 2025) learns appearance-agnostic motion representations for open-world transfer. SCAIL-2 (Yan et al. 2026) directly conditions generation on a driving video, and uses an in-context mask to bind driving and target characters across single- and multi-character animation. Nevertheless, DisMo does not provide target-region routing for multiple motion sources, while SCAIL-2 obtains multi-character motions jointly from a driving video rather than composing motions from multiple independent reference videos. In our work, RART extracts each motion, anchors it with its target region, and inject it into the temporally corresponding latent frames, without per-video optimization.

Methodology

Task Formulation

Given a target first-frame image I^{tgt} , a text condition C , and N object-wise control signals, the i -th control signal contains a reference video V_i^{ref} , its per-frame mask M_i^{ref} , and a target mask M_i^{tgt} on I^{tgt} . RART aims to generate a target video V^{tgt} in which the region specified by M_i^{tgt} follows the motion of (V_i^{ref}, M_i^{ref}) while preserving the target appearance and background, which can be formulated as:

$$V^{tgt} = G_{\theta}(I^{tgt}, C, \{(V_i^{ref}, M_i^{ref}, M_i^{tgt})\}_{i=1}^N), \quad (1)$$

where $G_{\theta}(\cdot)$ denotes RART, (V_i^{ref}, M_i^{ref}) specifies which object motion to extract, and M_i^{tgt} specifies where it should be applied. Superscripts ref and tgt distinguish reference and target quantities. Uppercase M denotes pixel-space masks,

while lowercase m denotes target masks on the latent grid. Note that the masks can be obtained from segmentation models (Cheng et al. 2022; Kirillov et al. 2023; Ravi et al. 2025; Carion et al. 2026).

Overview of RART

As illustrated in Figure 2, RART builds upon a video Diffusion Transformer (Peebles and Xie 2023; Wan et al. 2025). For i -th control signal, RART extracts its reference motion from V_i^{ref} with M_i^{ref} , anchors it with the target object mask M_i^{tgt} , and finally produces corresponding motion tokens Z_i :

$$\begin{aligned} F_i &= E_{\text{mot}}(V_i^{ref}, M_i^{ref}), \\ Z_i &= E_{\text{anchor}}(F_i, M_i^{tgt}), \end{aligned} \quad (2)$$

where $E_{\text{mot}}(\cdot)$ denotes motion extraction and feature projection into the DiT feature space, and F_i is the projected reference-motion feature sequence. E_{anchor} denotes the function that encodes the reference-motion feature and the target object mask into the motion token Z_i . For all the attained motion tokens of N objects, we concatenate them along token dimension as $Z = \text{Concat}(Z_1, \dots, Z_N)$. We denote the DiT hidden dimension by d , and all token features below use this dimension. The frozen DiT reads Z through frame-aligned motion cross-attention, while a motion-attention routing loss directly supervises object-wise assignment during training.

The design separates three roles that are often entangled in implicit motion transfer: the masked reference video provides *what* motion should be transferred, the target object mask specifies *where* the motion should act, and the frozen video generator determines *how* the target scene should be rendered. This separation is especially important for multi-object composition.

Mask-Anchored Motion Conditioning

Motion Tokenization. As shown in Figure 2, for the i -th control signal, the reference mask M_i^{ref} suppresses unrelated regions in V_i^{ref} before motion extraction. RART then uses the DisMo encoder (Ressler-Antal et al. 2025) as a pretrained motion feature extractor to obtain the temporally ordered sequence F_i . Since these features describe reference motion but do not specify where it should act, we first encode the target first-frame mask using mask encoder $E_{\text{mask}}(\cdot)$ as:

$$A_i = E_{\text{mask}}(M_i^{\text{tgt}}). \quad (3)$$

The mask encoder consists of convolutional blocks, adaptive pooling and a linear projection to the DiT feature space. Its output A_i is a compact set of spatial tokens that encodes the designated target region. We then inject this spatial anchor into the temporally ordered motion features and group the resulting tokens according to the latent video frames:

$$Z_i = \text{Chunk}(\text{MaskCA}((F_i + P_e), A_i)). \quad (4)$$

Here, P_e denotes a sinusoidal temporal encoding. $\text{MaskCA}(\cdot)$ denotes a residual cross-attention operation, with the projected motion features as queries, and mask tokens A_i as keys and values. This associates the motion features with the designated target region while preserving their temporal order. $\text{Chunk}(\cdot)$ groups consecutive features according to video latent temporal length, producing $Z_i \in R^{T \times K \times d}$, where T is the number of latent frames and K is the number of motion tokens per object at each latent frame. The control signals are encoded independently, and then concatenated for frame-aligned motion cross-attention.

Frame-Aligned Motion Cross-Attention. Given the video latent $H \in R^{T \times P \times d}$ in DiT, and the concatenated motion tokens $Z \in R^{T \times NK \times d}$ of N objects, we conduct the cross-attention between H and Z of each corresponding time slot for motion condition injection. Such frame-aligned motion cross-attention is formulated as:

$$H'_f = \text{CA}(H_f, Z_f), f \in [1, T] \quad (5)$$

where $\text{CA}(\cdot)$ is the cross-attention operation with a zero-initialized output projection to preserve the pretrained generator at the beginning of training. H_f provides the queries, while Z_f provides the keys and values for the f -th time slot. Each sample-frame pair is processed independently, so the P latent tokens of a frame can access only the corresponding NK motion tokens, for the K tokens from each object. Object-wise tokens are concatenated only within the same frame, allowing multiple motions to be injected jointly without mixing motion tokens across frames; the updated features are then restored to the original layout.

Compositional Training

The preceding modules determine how each motion enters the generator; training must further enforce correct composition across multiple active objects. For each multi-object training clip, we sample an active object subset $\mathcal{S}$ and concatenate its motion tokens as $Z_{\mathcal{S}}$. Let m_i denote target mask sequence of object i on the latent grid. These sequences

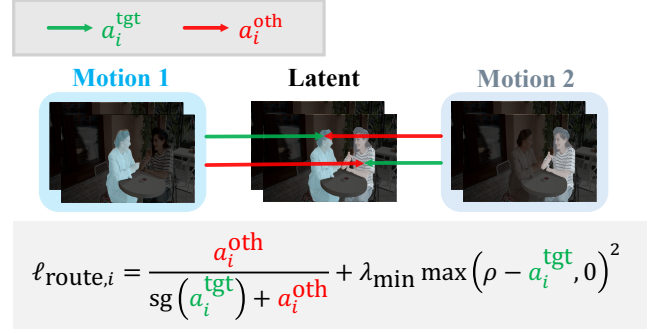


Figure 3: Motion-Attention Routing Loss. For each object, RART encourages its target region to attend to the corresponding motion tokens while discouraging other object regions from attending to the same token block.

provide training-only supervision; no target mask sequence is needed at inference beyond M_i^{tgt} . We also apply frame-consistent spatial augmentation to the reference video-mask pairs during training.

Foreground-Weighted Flow Matching. The active target masks define the foreground weights for the flow-matching objective (Lipman et al. 2023). Let x be the latent encoding of V^{tgt} , ϵ be Gaussian noise, τ be a sampled flow time, and $x_\tau = (1 - \tau)x + \tau\epsilon$. The target velocity is $v = \epsilon - x$, and the predicted velocity is $\hat{v}_\theta = v_\theta(x_\tau; I^{\text{tgt}}, C, Z_{\mathcal{S}})$. For the active object set $\mathcal{S}$, we define

$$w_{\mathcal{S}} = \lambda_b + (1 - \lambda_b) \max_{i \in \mathcal{S}} m_i, \quad (6)$$

$$\mathcal{L}_{\text{fm}} = E_{x, \tau, \epsilon} \left[\frac{\sum w_{\mathcal{S}} \odot (\hat{v}_\theta - v)^2}{\sum w_{\mathcal{S}}} \right].$$

In Eq. (6), $w_{\mathcal{S}}$ is broadcast over latent channels, and the sums cover all latent elements. Foreground positions receive unit weight, while background positions receive λ_b , with $0 < \lambda_b < 1$. We also apply motion dropout by replacing $Z_{\mathcal{S}}$ with zeros with a small probability, enabling classifier-free-style motion guidance at inference.

Motion-Attention Routing Loss. Reconstruction alone does not specify which region should read which motion tokens. For multi-object samples, we regularize the motion-attention maps so that each object region routes to its corresponding motion tokens. Figure 3 illustrates this operation. For object i , let a_i^{tgt} denote the attention mass from latent positions within its target-mask sequence m_i to the corresponding motion-token block, and let a_i^{oth} denote the attention mass from the other active object regions to the same block. Both quantities are obtained directly from the motion-attention maps and averaged over heads, frames, and valid spatial positions. The per-object routing loss is

$$\ell_{\text{route},i} = \frac{a_i^{\text{oth}}}{\text{sg}(a_i^{\text{tgt}}) + a_i^{\text{oth}}} + \lambda_{\min} \max(\rho - a_i^{\text{tgt}}, 0)^2. \quad (7)$$

Here $\text{sg}(\cdot)$ stops gradients through the target-region term in the denominator. The first term penalizes attention routed

from other object regions, while the second is activated when $a_i^{\text{tgt}} < \rho$ to prevent all motion-token attention from collapsing. The weight $\lambda_{\min}$ controls this minimum-attention constraint, and ρ is the target attention threshold. Averaging Eq. (7) over active objects and motion-attention modules gives $\mathcal{L}_{\text{route}}$. The final objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{fm}} + \lambda_{\text{route}} \mathcal{L}_{\text{route}}, \quad (8)$$

where λ_{route} balances the routing loss. The routing loss in Eq. (8) is applied only to non-dropped multi-object training batches and is obtained from the same forward-pass attention maps, without an additional denoiser evaluation. It therefore directly supervises object-to-motion assignment rather than relying on reconstruction alone.

Experiments

In this section, we report quantitative and human evaluation results in Table 1, qualitative comparisons in Figure 4, and component analyses in Table 2, Figure 5, and Figure 6.

Experimental Setup

Datasets. We train RART on a mixture of OpenVidHD (Nan et al. 2025) and HumanVid (Wang et al. 2024a) videos. After filtering invalid and low-quality samples, the training set contains 286,573 video entries. We extract per-frame object masks with Grounding-SAM2 (Liu et al. 2024; Ravi et al. 2025). For evaluation, we extend the FlexiMMT test set (Li et al. 2026) to a multi-object benchmark with 266 test cases and 25 reference videos. Each test sample consists of a target first frame, a text prompt, target first-frame object masks, and multiple reference video-mask pairs. Each target mask specifies where a motion should be applied in the target image, while the corresponding reference-mask sequence marks the object providing that motion in the reference video.

Baselines. We compare RART with four representative implicit methods closely related to motion transfer. AnyV2V (Ku et al. 2024) is a training-free method that injects source-video features into an image-to-video process. DisMo (Ressler-Antal et al. 2025) is a feed-forward method that learns appearance-agnostic motion representations for open-world motion transfer. SCAIL-2 (Yan et al. 2026) is a character-animation method that directly conditions generation on driving videos and uses in-context masks to bind driving and target characters. FlexiMMT (Li et al. 2026) is the closest multi-object multi-motion baseline, which learns object-wise motion tokens under mask constraints but requires reference-specific optimization.

To evaluate all baselines in our setting, we implement the necessary task-specific adaptations while retaining their released backbone weights and core generation mechanisms. For AnyV2V, we independently extract intermediate features from each reference video, fuse them while preserving their mean and variance, and use the fused features for image-to-video inference conditioned on the target image. For DisMo, we independently extract the motion embedding of each reference and fuse the embeddings before its CogVideoX generator. For SCAIL-2, we assemble the independently masked

reference clips into a unified multi-character driving sequence and use color-coded masks to preserve object correspondence. As for FlexiMMT, we use an object-specific text description of each designated object during training to distinguish the motion sources in multi-object reference videos. All methods receive the same target image and reference videos. Object assignments are provided directly when supported; otherwise, they are encoded through the task-specific adaptations described above. Apart from these adaptations, all other configurations follow the respective original papers.

Evaluation metrics. We mainly follow the evaluation protocol of FlexiMMT (Li et al. 2026). Specifically, we report CLIP-based text similarity, temporal consistency, and appearance consistency, together with two object-wise motion metrics: flow fidelity and motion fidelity. Flow fidelity compares mask-restricted optical-flow statistics between the reference object and the generated target object, while motion fidelity compares tracking trajectories inside the corresponding masks. For multi-object samples, we compute object-level scores and average them over all controlled objects. We additionally conduct human evaluation along three dimensions: target-appearance consistency, temporal consistency, and reference-motion consistency. For each of the 266 test cases, 5 annotators are shown the target first frame, all reference videos, and the outputs of five methods in randomized order. They independently select the best result for each criterion. We aggregate the selections over all test cases and report per-method selection rates, which sum to 100% for each criterion. We also report runtime to assess efficiency.

Implementation details. RART is built on a frozen Wan2.2 TI2V-5B backbone (Wan et al. 2025). We use the DisMo encoder as the pretrained motion feature extractor (Ressler-Antal et al. 2025), and train only the motion encoding and motion-attention injection modules. The final model is trained for 200K steps with a learning rate of 1×10^{-5} under our mixed-object training configuration. We sample 49 frames from each training video. Training begins at 320×480 resolution and is increased to 480×736 at 120K steps for high-resolution finetuning. We use 4 motion tokens per latent frame and 2 residual target-mask cross-attention layers. During training, we sample 2 to 6 active objects for multi-object clips, use motion dropout with probability 0.1, caption dropout with probability 0.1, and apply mask perturbation with probability 0.6. As an auxiliary augmentation, each reference video-mask pair undergoes an independent spatial transform shared across frames, while the target condition remains unchanged. We set the background weight to $\lambda_b = 0.1$. The motion-attention routing loss is enabled from step 170K with $\lambda_{\text{route}} = 0.2$, $\lambda_{\min} = 0.05$, and $\rho = 0.3$. All experiments are conducted on eight NVIDIA H200 GPUs.

Comparisons with State-of-the-Art Methods

Quantitative Evaluation. Table 1 presents the quantitative results. RART achieves the best flow fidelity and motion fidelity, demonstrating more accurate transfer of reference motions to the designated target objects. It also obtains competitive text similarity, temporal consistency, and appearance

Method	Automatic Metrics					Human Evaluations (%)			
	TS $\uparrow$	TC $\uparrow$	AC $\uparrow$	FF $\uparrow$	MF $\uparrow$	App. $\uparrow$	Temp. $\uparrow$	Motion $\uparrow$	Time $\downarrow$
AnyV2V	0.261	0.879	0.742	0.571	0.338	0.0	0.0	0.0	176s/ref + 50s
DisMo	0.266	0.944	0.913	0.652	0.451	3.0	3.0	2.6	113s
SCAIL-2	0.266	0.922	0.906	0.648	0.428	15.8	9.0	24.1	233s
FlexiMMT	0.267	0.925	0.899	0.671	0.462	14.7	27.1	21.8	2h/ref + 432s
RART	0.265	0.928	0.903	0.713	0.480	66.5	60.9	51.5	41s

Table 1: Quantitative and human evaluations. TS, TC, AC, FF, and MF denote text similarity, temporal consistency, appearance consistency, flow fidelity, and motion fidelity, respectively. Human evaluation columns report five-way selection rates (%), while Time denotes runtime. For AnyV2V and FlexiMMT, the first time term denotes per-reference preparation.



Figure 4: Qualitative comparison. The top row shows the reference motions, target image, and motion-to-object assignments; the remaining rows show frames generated by the baselines and RART. RART transfers each reference motion to its designated target object while preserving target appearance.

consistency, showing that the improvement in motion accuracy does not compromise overall video quality.

The latter three metrics should nevertheless be interpreted together with motion fidelity. Temporal consistency and appearance consistency are computed from frame-to-frame similarities and may therefore favor videos with limited motion, while text prompts rarely describe fine-grained object dynamics and therefore make text similarity an incomplete measure of motion transfer. Consequently, strong scores on these generic video-quality metrics are less informative when the generated motion does not follow the reference.

RART also has the shortest runtime. AnyV2V performs additional per-video noising and caches intermediate features; FlexiMMT optimizes motion tokens for each reference motion and includes an additional mask-extraction stage at inference; and SCAIL-2 uses a larger generative backbone. DisMo remains efficient, but its original formulation does not support compositional multi-object motion transfer. Overall,

RART provides the strongest balance of motion-transfer accuracy, visual quality, and inference efficiency.

Human evaluation. The human evaluation results in Table 1 show that RART achieves the highest selection rate across all three criteria, receiving 66.5% for target-appearance consistency, 60.9% for temporal consistency, and 51.5% for reference-motion consistency. These results confirm that RART transfers reference motions more accurately while maintaining coherent dynamics and target appearance.

Qualitative results. Figure 4 compares the generated videos of different methods. AnyV2V fails to perform multi-motion transfer and produces videos with poor visual quality. DisMo better preserves video quality, but its architecture lacks explicit object-wise spatial assignment and therefore cannot transfer the desired motions precisely. FlexiMMT can transfer motions extracted from single-object reference videos, but struggles to isolate and transfer the motion of a designated region when the reference video contains mul-

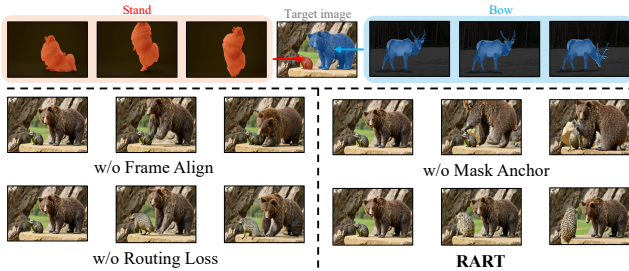


Figure 5: Qualitative ablation. The top row shows reference motions and their assignments to target regions; the bottom panels compare RART with variants without frame alignment, target-mask anchoring, or routing loss.

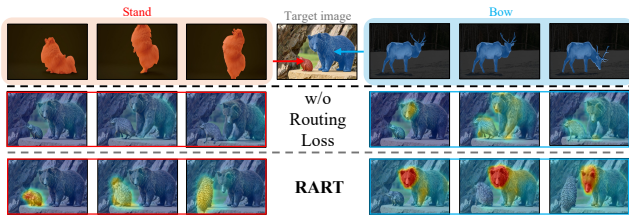


Figure 6: Motion-attention heat maps. Columns correspond to the two reference motions, while rows compare the model without the routing loss and RART. Warmer colors indicate stronger attention to each motion-token block.

multiple objects. SCAIL-2 does not natively support extracting motions from multiple reference videos; although our adaptation enables this input setting, its generated videos often drift from the target first frame and exhibit weak temporal consistency. In contrast, RART consistently extracts the specified motions, assigns them to the intended target objects, and preserves both target appearance and temporal coherence.

Ablation Study

w/o Frame Alignment. We train this variant by replacing frame-aligned motion cross-attention with global motion cross-attention while keeping all other settings unchanged. As shown in Figure 5, neither the bear nor the hedgehog follows the temporal motion pattern of its reference video, whereas the full model with frame-aligned motion cross-attention generates correctly aligned motions. Consistently, Table 2 shows that this variant obtains lower FF and MF scores than the full model. These results demonstrate that frame-aligned motion cross-attention is essential for establishing temporal correspondence.

w/o Target-Mask Anchoring. To evaluate the effect of target-mask anchoring, we replace every target mask M_i^{tgt} with an all-ones mask at inference. Consequently, all objects receive the same mask condition and no longer have object-specific target-region information. As shown in Figure 5, the bear incorrectly follows the motion intended for the hedgehog when the spatial anchors are removed. Table 2 also reports lower FF and MF scores for this variant than for the full

Variant	FF↑	MF↑
w/o Frame Alignment	0.678	0.400
w/o Mask Anchor	0.690	0.413
w/o Routing Loss	0.727	0.448
RART	0.713	0.480

Table 2: Quantitative ablations. We replace frame-aligned motion attention with global attention, use all-ones target masks, or remove the motion-attention routing loss.

model. Together, these results demonstrate that target-mask anchoring is crucial for assigning each reference motion to its intended target region.

w/o Motion-Attention Routing Loss. We continue training from step 170K without the motion-attention routing loss, while keeping all other settings unchanged. Figure 5 shows that, although the hedgehog partially follows its reference motion, the bear incorrectly performs the motion intended for the hedgehog, revealing cross-object motion leakage.

Table 2 further quantifies this effect. The routing loss raises MF from 0.448 to 0.480, a 7.1% relative improvement, with only a small FF change from 0.727 to 0.713. FF captures local flow magnitude and direction, whereas MF measures trajectory-level correspondence, making them complementary indicators of motion transfer. The full model therefore retains strong local flow fidelity while substantially improving trajectory-level correspondence.

To examine the routing behavior directly, we further visualize the attention maps in Figure 6. For each object, we aggregate the weights assigned to its motion-token block at every latent spatial position and upsample the resulting heat map to the original video resolution. With the routing loss, attention to each object’s motion tokens is concentrated in its designated target region. Without the loss, the attention maps of different motion sources become substantially mixed. Together with the qualitative and quantitative results, these heat maps show that the routing loss improves object-wise motion assignment and suppresses cross-object leakage.

Conclusions

We presented RART, a unified framework for compositional multi-object motion transfer. RART represents each reference motion with object-wise tokens anchored by the corresponding target mask, and injects temporally matched tokens into a frozen video Diffusion Transformer through frame-aligned motion cross-attention. The motion-attention routing loss further strengthens the correspondence between target regions and motion sources, reducing cross-object motion leakage. RART jointly composes multiple reference motions in one denoising pass, without separate optimization for each reference video. Quantitative, qualitative, and human evaluations show that RART delivers accurate motion transfer, strong visual quality, and efficient inference; ablations and attention visualizations support our key design choices. Future work will explore broader applications of RART in long-form generation and interactive multi-object video creation.

References

- Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv:2311.15127*.
- Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. 2024. Video generation models as world simulators. *OpenAI Blog*, 1(8): 1.
- Carion, N.; Gustafson, L.; Hu, Y.-T.; Debnath, S.; Hu, R.; Coll-Vinent, D. S.; Ryali, C.; Alwala, K. V.; Khedr, H.; Huang, A.; et al. 2026. SAM 3: Segment Anything with Concepts. In *ICLR*.
- Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; et al. 2023. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv:2310.19512*.
- Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 1290–1299.
- Chu, R.; He, Y.; Chen, Z.; Zhang, S.; Xu, X.; WANG, D.; Yi, H.; Liu, X.; Zhao, H.; Liu, Y.; et al. 2026. Wan-move: Motion-controllable video generation via latent trajectory guidance. *NeurIPS*, 38: 404–432.
- Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *NeurIPS*, 34: 8780–8794.
- Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*.
- He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv:2211.13221*.
- Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022a. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33: 6840–6851.
- Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS*.
- Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *NeurIPS*, 35: 8633–8646.
- Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2023. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. In *ICLR*.
- Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 8153–8163.
- Jeong, H.; Park, G. Y.; and Ye, J. C. 2024. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *CVPR*, 9212–9221.
- Karras, J.; Holynski, A.; Wang, T.-C.; and Kemelmacher-Shlizerman, I. 2023. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *ICCV*, 22623–22633. IEEE.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *CVPR*, 4015–4026.
- Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuan-video: A systematic framework for large video generative models. *arXiv:2412.03603*.
- Ku, M.; Wei, C.; Ren, W.; Yang, H.; and Chen, W. 2024. AnyV2V: A Tuning-Free Framework For Any Video-to-Video Editing Tasks. *TMLR*. Reproducibility Certification.
- Li, Y.; Gong, D.; Cao, X.; Yuan, J.; Li, D.; Zhou, L.; Koh, Y. S.; Yan, C.; and Zhang, X. 2026. Let Your Image Move with Your Motion!—Implicit Multi-Object Multi-Motion Transfer. In *CVPR*, 11207–11217.
- Ling, P.; Bu, J.; Zhang, P.; Dong, X.; Zang, Y.; Wu, T.; Chen, H.; Wang, J.; and Jin, Y. 2025. MotionClone: Training-Free Motion Cloning for Controllable Video Generation. In *ICLR*.
- Lipman, Y.; Chen, R. T. Q.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. In *ICLR*.
- Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 38–55. Springer.
- Ma, Y.; He, Y.; Cun, X.; Wang, X.; Chen, S.; Li, X.; and Chen, Q. 2024. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *AAAI*, volume 38, 4117–4125.
- Nan, K.; Xie, R.; Zhou, P.; Fan, T.; Yang, Z.; Chen, Z.; Li, X.; Yang, J.; and Tai, Y. 2025. OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-video Generation. In *ICLR*.
- Ouyang, W.; Dong, Y.; Yang, L.; Si, J.; and Pan, X. 2024. I2vedit: First-frame-guided video editing via image-to-video diffusion models. In *SIGGRAPH Asia*, 1–11.
- Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *CVPR*, 4195–4205.
- Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollar, P.; and Feichtenhofer, C. 2025. SAM 2: Segment Anything in Images and Videos. In *ICLR*.
- Ressler-Antal, T.; Fundel, F.; Alaya, M. B.; Baumann, S. A.; Krause, F.; Gui, M.; and Ommer, B. 2025. DisMo: Disentangled Motion Representations for Open-World Motion Transfer. In *NeurIPS*.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, 10684–10695.
- Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; Parikh, D.; Gupta, S.; and Taigman, Y. 2023. Make-A-Video: Text-to-Video Generation without Text-Video Data. In *ICLR*.
- Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *ICLR*.
- Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *NeurIPS*, 32.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*.

Villegas, R.; Babaeizadeh, M.; Kindermans, P.-J.; Moraldo, H.; Zhang, H.; Saffar, M. T.; Castro, S.; Kunze, J.; and Erhan, D. 2023. Phenaki: Variable Length Video Generation from Open Domain Textual Descriptions. In *ICLR*.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv:2503.20314*.

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

Wang, X.; Yuan, H.; Zhang, S.; Chen, D.; Wang, J.; Zhang, Y.; Shen, Y.; Zhao, D.; and Zhou, J. 2023b. Videocomposer: Compositional video synthesis with motion controllability. *NeurIPS*, 36: 7594–7611.

Wang, Z.; Li, Y.; Zeng, Y.; Fang, Y.; Guo, Y.; Liu, W.; Tan, J.; Chen, K.; Xue, T.; Dai, B.; et al. 2024a. Humanvid: Demystifying training data for camera-controllable human image animation. *NeurIPS*, 37: 20111–20131.

Wang, Z.; Yuan, Z.; Wang, X.; Li, Y.; Chen, T.; Xia, M.; Luo, P.; and Shan, Y. 2024b. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*, 1–11.

Xu, Z.; Zhang, J.; Liew, J. H.; Yan, H.; Liu, J.-W.; Zhang, C.; Feng, J.; and Shou, M. Z. 2024. Magicanimate: Temporally consistent human image animation using diffusion model. In *CVPR*, 1481–1490.

Yan, W.; Guo, F.; Yang, Z.; and Tang, J. 2026. SCAIL-2: Unifying Controlled Character Animation with End-to-end In-Context Conditioning. *arXiv:2606.10804*.

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 *CVPR*, 8466–8476.

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

Zhang, S.; Zhuang, J.; Zhang, Z.; Shan, Y.; and Tang, Y. 2025. Flexiact: Towards flexible action control in heterogeneous scenarios. In *ACM SIGGRAPH*, 1–11.

Zhang, Y.; Wei, Y.; Jiang, D.; ZHANG, X.; Zuo, W.; and Tian, Q. 2024. ControlVideo: Training-free Controllable Text-to-video Generation. In *ICLR*.

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 *ECCV*, 273–290. Springer.