DRA-MTransfer: Physically Realistic Video Motion Transfer with Dual-Grained Re-Adaptation

Guoli Jia[1*] Zhiyuan Ma[2*] Junyao Hu[3] Xinwei Long[1] Kai Tian[1] Kaikai Zhao[1,4]
Zhaoxiang Liu[4] Kai Wang[4] Shiguang Lian[4] Bowen Zhou[1,5†]

[1] Tsinghua University [2] Huazhong University of Science and Technology
[3] The Hong Kong Polytechnic University [4] China Unicom
[5] Shanghai Artificial Intelligence Laboratory

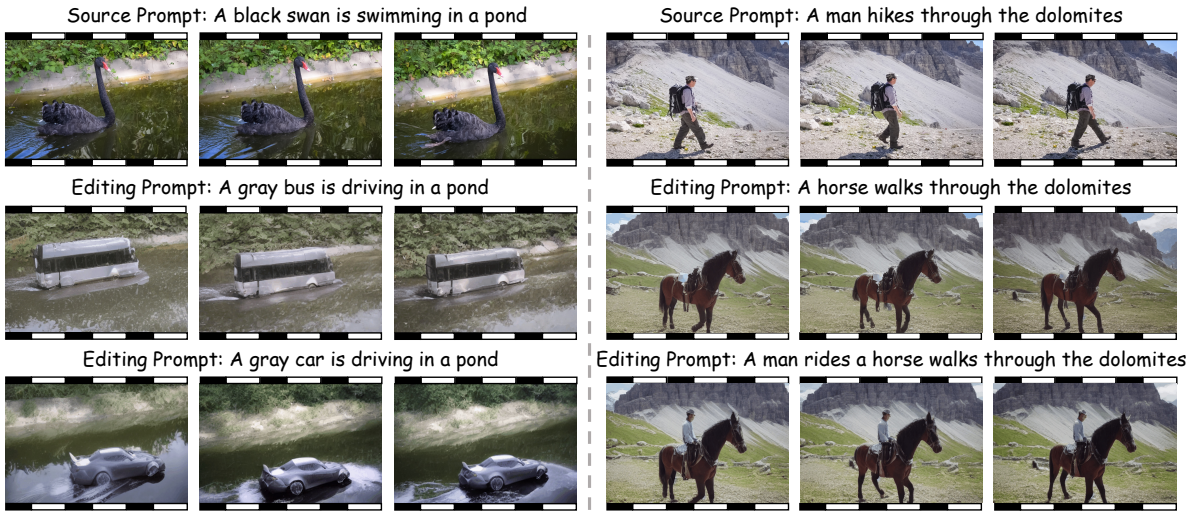


Figure 1. Transferring motion from the source video to the edited video. Our proposed DRA-MTransfer is able to edit the subjects with significant shape difference (left, swan to bus and car), and the subjects with ***distinct intrinsic motion patterns*** (right, two legs to four legs).

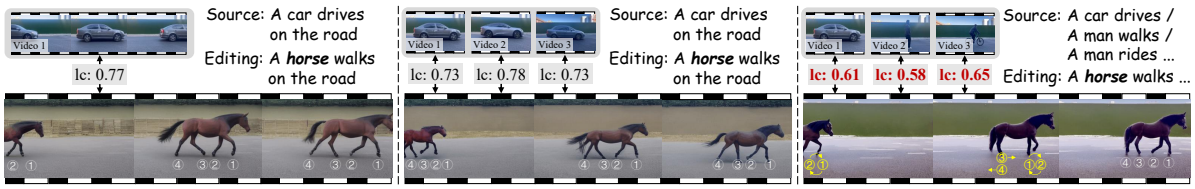
Abstract

Video motion transfer is a challenging video editing task that requires preserving the source motion while generating an edited subject faithful to the target instruction. Recent leading methods adapt video diffusion models (VDM) to a single source video via lightweight fine-tuning, e.g., LoRA, but often overlook an important issue. The edited subject is usually trajectory-consistent, yet physically unrealistic, especially when the subjects exhibit substantially different shapes. For example, when editing a car into a horse, the model tends to preserve the rotational tires, resulting in stiff legs and unrealistic postures. The intrinsic motion pattern of the edited subject is limited by the source video. To this end, we propose DRA-MTransfer, a dual-grained re-adaptation framework that reactivates motion priors already encoded in the pre-trained VDM. Specifically, we introduce Spatial Consistency Guided Re-Weighting (SCR) to alleviate overly strong source constraints, which improves global posture realism. Further, we design Synergistic Temporal Re-Attention (STR) to inject subject compatible motion cues into temporal attention for fine-grained motion refinement. Extensive experiments on V2VBench and MT-Bench show that DRA-MTransfer consistently improves intrinsic motion realism and physical plausibility, while maintaining strong motion transfer fidelity, text alignment, and temporal coherence.

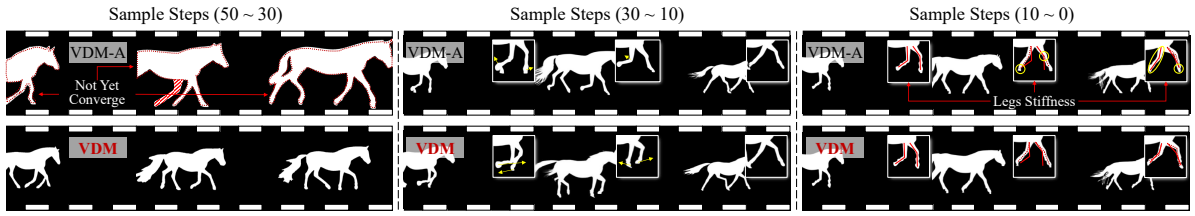
1. Introduction

With the rapid development of video diffusion models (VDM) [21, 54, 66], video editing [12, 13, 44] has achieved impressive progress in subject replacement, background manipulation, and style transformation. Among these tasks,

*Equal contribution.
†Corresponding author.



(a) Fine-tuning on one source video (left), three source videos contain a car as subject (middle), three source videos contain different subjects (right).



(b) Masks from VDM-A (up) and VDM (bottom), merging from 50 ~ 30 denoising steps (left), 30 ~ 10 steps (left), and 10 ~ 0 steps (right).

Figure 2. (a) Evaluating the IMP of motion transfer method (MotionDirector) under three training settings. All source videos are captured from the same position and exhibit the same trajectories. Local consistency (LC) is measured by computing the cosine similarity of the attention maps between the source and target videos within the foreground region. In the left and middle cases, the leg indices are unchanged, showing almost no movement of horse's legs. (b) Analyzing the masks from VDM-A and VDM in three stages of the denoising process, total of 50 steps. For clearer visualization, the masks are extracted from UpBlock.2, and refined using DenseCRF [31].

motion transfer is particularly challenging [19, 32, 93], as the edited video should simultaneously preserve the motion trajectory of the source video [39, 80], remain temporally coherent [91], and satisfy the target editing instruction. Recent methods achieve this goal by adapting a pre-trained VDM to a single source video (VDM-A), typically through lightweight fine-tuning modules such as LoRA [45, 86, 88]. This paradigm significantly improves robustness over zero-shot transfer, especially when the edited subject undergoes large appearance or shape changes.

However, we observe that current motion transfer methods still suffer from an crucial limitation [46, 98]. Although the generated motion often follows the source trajectory [86, 88], it may become **physically unrealistic** for the edited subject itself. As shown in Fig. 2, when transferring the motion of a driving car to a horse, the model tends to inherit the rotational motion patterns of the tires in the source video, causing the horse's legs to become stiff and the overall posture to remain unnatural. Similar failures also appear when editing between subjects with distinct structures, such as two-legged humans and four-legged animals. These results indicate that current methods are often insensitive to the intrinsic motion pattern (IMP) of the edited subject.

This issue is closely related to a broader goal in world modeling [60, 76], the edited subject should not only follow the intended motion pattern, but also move in a manner compatible with its own physical structure and dynamics. Although our goal is not to build an explicit world model, our insight that preserving subject-consistent dynamics is a basic requirement for realistic video editing [27, 73]. The challenge, therefore, is how to preserve source motion while allowing the edited subject to recover its own plausible IMP. ***Our key observation is that the pre-trained VDM already contains rich motion priors [20, 68] learned from large-scale videos [3, 51], including subject dependent IMP***. The single video adaptation, *i.e.*, VDM-A, excessively amplifies source specific motion, suppressing these general priors. This is supported by the mask visualizations during denoising in Fig. 2. As the trajectory gradually converges, the frozen VDM exhibits clearer leg-swinging patterns, whereas the VDM-A still tends to generate stiff legs.

Based on the insight, we propose DRA-MTransfer, a Dual-Grained Re-Adaptation framework that reactivates useful motion priors in the pre-trained model to improve intrinsic-motion realism. Specifically, we design Spatial Consistency Guided Re-Weighting (SCR) to generate plausible motion at a coarse-grained level. During training, SCR employs Gaussian re-weighting on spatial consistency to relax the source video constraint in foreground regions. During inference, it further adaptively adjusts the LoRA weight according to spatial consistency, enabling the model to better accommodate subjects with realistic motion patterns. To refine fine-grained motion, we further propose Synergistic Temporal Re-Attention (STR), which injects motion related features from the frozen pre-trained VDM into the temporal attention of the adapted model within foreground regions. In addition, a sharpening strategy is utilized to suppress artifacts caused by redundant or inaccurate motion cues. Our main contributions are summarized as follows:

- We explore the intrinsic motion pattern in video editing, which severely impacts the physical realism of video motion transfer.
- We propose a new insight to utilize the potential motion prior within the pre-trained VDM to tackle this issue, and meticulously design DRA-MTransfer, which consists of SCR and STR built upon this insight.
- Extensive qualitative and quantitative experiments con-

sistently demonstrate the advantages (at least + 4%) of our method in physical plausible video motion transfer.

2. Related Work

2.1. Video Editing

Benefiting from superior controllability, video editing has attracted increasing attention [28, 37, 49]. Video editing faces three main challenges [41, 77]: (1) *Generating high quality videos*, *e.g.*, spatial resolution and temporal coherence, (2) *Ensuring high consistency with editing instructions*, (3) *Preserving high fidelity with the source videos*. Previous methods usually use adversarial objective [87], decoupling [1], and discrete representation [17, 72] to cope with these challenges. Recently, with the rising of diffusion models [16, 47], notable improvements have been witnessed in video editing [18, 18, 52].

In terms of high quality, key-frame attention [8, 57, 65, 77] and neural layered atlas (NLA) [9, 15, 23, 33] are utilized to improve temporal coherence [8, 9, 33, 97]. In terms of high consistency, subject [34, 35], motion [70, 71, 100], and grounding [24] information are injected into the denoising network, providing additional reference information beyond prompt. In terms of high fidelity, optimizing inversion latent [41, 42, 61, 79] and score-based guidance [25, 71] are effective strategies. Currently, appearance editing, such as color and texture, has demonstrated excellent performance [36, 43, 81]. As a challenging task in video editing, ***motion transfer*** [6, 78] has attracted increasing attention.

2.2. Motion Transfer

Motion transfer is a typical research topic in computer vision, which has been investigated in tasks such as video *prediction* [63], *generation* [22], and *editing* [93]. There are four typical approaches for motion transfer [69].

As a leading approach, *optical flow* [64, 89] is utilized [14, 63] to guide the motion transfer in pixel level. Moreover, *trajectory* is another commonly used approach for video generation [50, 80, 95], which is usually extracted by pre-trained point tracking models [59, 99]. Recently, *attention mechanism* is improved for motion transfer [32, 86, 93]. Motion-related information in temporal attention maps [55, 93, 94] is extracted to guide the target video generation. The advantage of these three approaches is the zero-shot motion transfer capability. However, the lack of tailored training on source video leads to overly dense or sparse motion information [39, 86, 93], reducing robustness of motion transfer when editing subjects with significant shape variations, as discussed in Sec. 4. Therefore, current mainstream methods *fine-tune* a small scale of parameters to memorize the motion in the source video [7, 40, 62, 90], which exhibits remarkable robustness [11, 83, 96]. To enhance the upper limit of current methods, we contribute along the line of the fine-tuning approach. In this paper, we focus on the IMP stiffness issue, allowing for physically realistic motion transfer.

2.3. Physically Plausible and World-Consistent Video Generation

Recent video generation research has increasingly emphasized realism beyond appearance quality [73], including consistency of subject appearance [38], temporal evolution [48], and physically plausible motion [73]. This trend is also closely related to the broader theme of world modeling [48], which concerns whether a model can represent and generate temporally coherent, causally plausible, and structurally consistent visual worlds [67, 74].

In video editing, the physical plausibility of motion transfer remains underexplored. Most methods are evaluated based on text alignment, source fidelity, and temporal coherence, while the realism of the edited subject's own motion is rarely considered explicitly [46, 69]. Our work focuses on a concrete but important aspect of world-modeling, *i.e.*, world-consistent video editing. When the edited subject changes, how to control the generated motion remains compatible with the target subject's IMP. We address this challenge by reactivating the subject's IMP from VDM's pre-trained motion priors.

3. Methodology

3.1. Preliminary and Problem Formulation

Given a source video V^S and an editing prompt P^T, motion transfer aims to generate a target video V^T that preserves the motion of V^S while satisfying P^T. Recent methods adapt a pre-trained VDM to the source video V^S via lightweight modules such as LoRA, using an appearance loss $\mathcal{L}_{\mathcal{S}}$ for random-frame denoising, a temporal loss $\mathcal{L}_T$, and a debiased temporal loss $\mathcal{L}_{\mathcal{D}}$ for learning appearance-disentangled temporal representations from frame differences. However, such adaptation may overemphasize source motion and suppress the general motion priors of the pre-trained VDM. As a result, the generated subject may remain trajectory-consistent yet fail to exhibit subject compatible IMP, we term *intrinsic motion insensitivity*.

To tackle this issue, we propose DRA-MTransfer, which reactivates motion priors in the pre-trained VDM at two complementary granularities. The pipeline is shown in Fig. 3, which consists of Spatial Consistency Guided Re-Weighting (SCR) and Synergistic Temporal Re-Attention (STR). At the coarse-grained level, SCR aims to improve the global posture and motion pattern of the edited subject by reducing excessive source constraints. At the fine-grained level, STR aims to recover realistic motion details for local body parts or moving components.

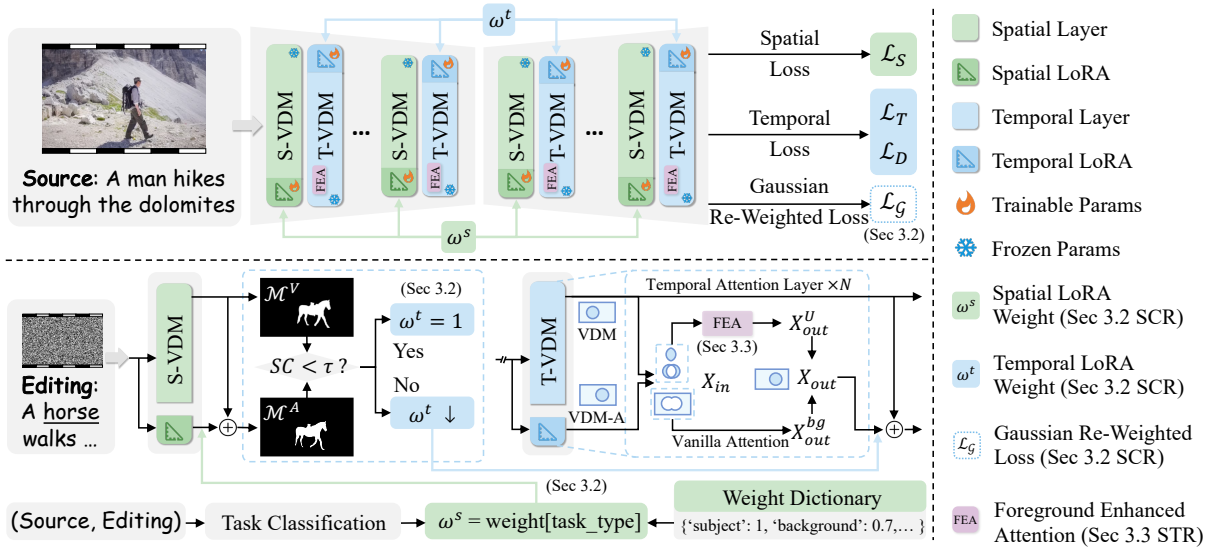


Figure 3. Framework of DRA-MTransfer. (1) Spatial Consistency Guided Re-Weighting (SCR): During training, $\mathcal{L}_\mathcal{G}$ uses Gaussian distribution to re-weight the loss on SC. During inference, temporal LoRA weight (ω^t) is adaptively adjusted guided by SC, and spatial LoRA weight (ω^s) is determined by a classifier model based on the prompt pair and weighting dictionary. (2) Synergistic Temporal Re-Attention (STR): During inference, foreground enhanced attention (FEA) integrates the IMP knowledge of edited subject within VDM into VDM-A by concatenating their keys and values.

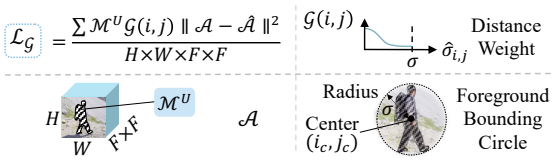


Figure 4. Illustration of $\mathcal{L}_\mathcal{G}$. The smaller the distance to the center of the foreground, the greater the loss weight.

3.2. Spatial Consistency Guided Re-Weighting

SCR aims to alleviate the overly strong motion constraint of VDM-A imposed by the source-adapted LoRA, and re-activate the coarse-grained motion prior in the pre-trained model. We use spatial consistency (SC) between source and generated foreground regions as the key cue, since SC naturally reflects whether the transferred motion is overly constrained by the source. During training, SCR employs the Gaussian distribution to re-weight the temporal loss based on SC. Further, to support adaptation to varying edited subjects during inference, SCR adaptively decreases the weight of the LoRA based on SC. Through coarse-grained re-weighting during both training and inference stages, SCR enables the model to generate physically realistic global postures for diverse edited subjects.

Gaussian distribution re-weighted loss $\mathcal{L}_\mathcal{G}$ aims to smooth the strict motion constraint on subject. We first calculate subject's mask $\mathcal{M} \in \mathbb{R}^{F\times H\times W}$ from cross-attention maps following standard strategy [57, 84], where high attention values are regarded as foreground regions. The masks from VDM and VDM-A are defined as $\mathcal{M}^V$ and $\mathcal{M}^A$.

Next, uniformly decreasing the foreground weight inevitably impacts motion trajectory learning. Therefore, we apply Gaussian re-weighting, assigning larger weights to the central region for trajectory learning, and smaller weights to marginal regions to alleviate the constraints imposed by the source video. This design is motivated by the fact that the subject's motion can be decomposed [2] into translation (*trajectory*) and rotation (*intrinsic motion*), where rotation causes larger relative displacement (intrinsic motion) for pixels farther from the center. The procedure is depicted in Fig. 4. After calculating the union mask $\mathcal{M}^U = \bigcup_{i=1}^{F} \mathcal{M}_i^V \cup \mathcal{M}_i^A$, where $\mathcal{M}^U \in \mathbb{R}^{H\times W}$, the $\mathcal{L}_\mathcal{G}$ is formulated as:

$$\mathcal{L}_\mathcal{G} = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{F\times F} \mathcal{G}(i,j)||\mathcal{A}_{i,j,k} - \hat{\mathcal{A}}_{i,j,k}||^2}{H\times W\times F\times F},$$

where $\mathcal{A} \in \mathbb{R}^{N\times F\times F}$ is the predicted temporal attention map, $N = H \times W$, $\mathcal{M}^U$ is broadcast along the temporal dimension to all frames. $\hat{\mathcal{A}}$ is the attention map from the source video. The weight $\mathcal{G}(i,j)$ of location i,j is calculated as:

$$\mathcal{G}(i,j) = \begin{cases} e^{-\frac{\hat{\sigma}_{i,j}^2}{2\sigma^2}}, & \text{if } (i,j) \in \mathcal{M}^U, \\ 1, & \text{else,} \end{cases} \quad (1)$$

where $\hat{\sigma}_{i,j}$ indicates the distance between (i,j) and (i_c, j_c), (i_c, j_c) is the center of bounding box enclosing the foreground region. The radius σ uses the half of diagonal length. Note that we find the maximum contours to filter the noisy pixels. Finally, the total temporal loss $\mathcal{L}_{total}$ is

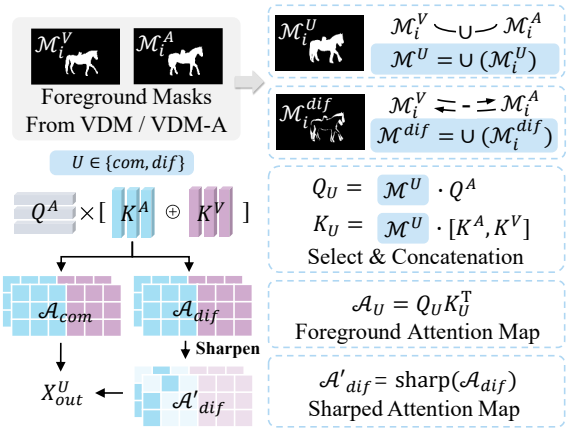


Figure 5. Synergistic Temporal Re-Attention (STR) processes masks in three parts: (1) For the common region in the foreground ($\mathcal{M}^{com}$), the keys and values from both VDM and VDM-A are merged for attention. (2) For the differential region in the foreground ($\mathcal{M}^{dif}$), on this basis, sharpen strategy is applied to the attention map. (3) For the background ($\mathcal{M}^{bg}$), using vanilla attention, which is not shown here. Details can be found in Sec. 3.3.

formalized as:

$$\mathcal{L}_{total} = (\mathcal{L}_T + \mathcal{L}_{\mathcal{G}})/2 + \mathcal{L}_D. \tag{2}$$

Temporal LoRA weight ω^t re-weighting aims to support DRA-MTransfer to handle video motion transfer across diverse edited subjects during the inference stage. First, the SC is computed as the *IoU* of masks between VDM and VDM-A:

$$SC = \frac{1}{F}\sum_{i=1}^{F}\frac{|\mathcal{M}_i^V \cap \mathcal{M}_i^A|}{|\mathcal{M}_i^V \cup \mathcal{M}_i^A|}. \tag{3}$$

If SC is small, the motion trajectory is inconsistent with $\mathcal{V}^S$, while an excessively large SC indicates high consistency of the IMP with the source video, leading to motion stiffness. To keep SC near a desirable range, we introduce a hyperparameter τ as its target value and design the following function:

$$\omega^t = \begin{cases} 1, & \text{if } SC < \tau, \\ \frac{e^{-SC}+e^{-(T-t)}}{2}, & \text{else.} \end{cases} \tag{4}$$

where t is the current denoising step. When SC $< \tau$, w^t is set to 1 to speed up SC convergence. Otherwise, SC is utilized for negative feedback adjustment. Besides, SC tends to increase as t decreases ($T \to 0$), hence we incorporate t as an additional empirical term to jointly adjust w^t.

In addition, existing methods [57, 98] rely on manually tuning the weights of spatial modules to handle diverse editing tasks, *e.g., subject, background*. To address this issue, we develop a simple yet effective strategy to automatically adjust the spatial LoRA weight ω^s.

3.3. Synergistic Temporal Re-Attention

STR aims to enhance fine-grained, physically plausible motion details for local body or moving components that may not be fully captured by coarse-grained adaptation. Although SCR reduces source over-constraint and improves global posture, local motions such as legs, and limbs may still appear temporally inconsistent. To address this, STR injects motion knowledge from the frozen pre-trained VDM into the temporal attention of the adapted model VDM-A, enhancing foreground dynamics and suppressing noisy or conflicting signals through a sharpening mechanism.

Foreground Enhanced Attention (FEA), shown in Fig. 5, integrates motion prior within VDM to improve the local motion details of the edited subject. First, we split the attention map into three parts: *common region in the foreground* ($\mathcal{M}^{com}$), *differential region in the foreground* ($\mathcal{M}^{dif}$), and *background* ($\mathcal{M}^{bg}$). The regions are calculated as:

$$\begin{cases} \mathcal{M}^{com} = \mathcal{M}^U - \mathcal{M}^{dif}, \\ \mathcal{M}^{dif} = \bigcup_{i=1}^{F}(\mathcal{M}_i^V - \mathcal{M}_i^A) \cup (\mathcal{M}_i^A - \mathcal{M}_i^V), \\ \mathcal{M}^{bg} = \mathcal{M}^{all} - \mathcal{M}^U, \ \mathcal{M}^{all} = 1^{H\times W}. \end{cases} \tag{5}$$

We calculate query, key, value of corresponding regions:

$$\begin{cases} Q_U = \mathcal{M}^U \cdot Q^A \in \mathbb{R}^{N_U \times F \times d}, \\ K_U = \mathcal{M}^U \cdot [K^A, K^V] \in \mathbb{R}^{N_U \times 2F \times d}, \\ V_U = \mathcal{M}^U \cdot [V^A, V^V] \in \mathbb{R}^{N_U \times 2F \times d}, \\ Q_{bg} = \mathcal{M}^{bg} \cdot Q^A \in \mathbb{R}^{N_{bg} \times F \times d}, \\ K_{bg} = \mathcal{M}^{bg} \cdot K^A \in \mathbb{R}^{N_{bg} \times F \times d}, \\ V_{bg} = \mathcal{M}^{bg} \cdot V^A \in \mathbb{R}^{N_{bg} \times F \times d}, \end{cases} \tag{6}$$

where N_U and N_{bg} are the number of pixels selected by masks $\mathcal{M}^U$ and $\mathcal{M}^{bg}$, $N_U + N_{bg} = N$. $Q^A, K^A, V^A \in \mathbb{R}^{N\times F\times d}$ are the query, key and value from VDM-A. $K^V, V^V \in \mathbb{R}^{N\times F\times d}$ are the key and value from VDM. d is the dimension. In foreground $\mathcal{M}^U$, key and value from VDM are concatenated with those from VDM-A. In background $\mathcal{M}^{bg}$, we use vanilla temporal attention.

Next, to mitigate the artifacts caused by inaccurate IMP in the VDM-A, we sharpen the differential attention maps $\mathcal{A}_{dif} = Q_{dif}K_{dif}^{\mathrm{T}} \in \mathbb{R}^{N_{dif}\times F\times 2F}$ in $\mathcal{M}^{dif}$ as follows:

$$\mathcal{A}'_{dif,i} = exp(\mathcal{A}_{dif,i}/\alpha) \Big/ \sum_{j=1}^{2F} exp(\mathcal{A}_{dif,j}/\alpha), \tag{7}$$

where $i \in \{1, 2, \ldots, 2F\}$ and α is temperature.

4. Experiments

4.1. Implementation Details

Following [98], we utilize Adam [30] as optimizer with 400 iterations to train LoRAs. Betas and epsilon are set to 0.9,

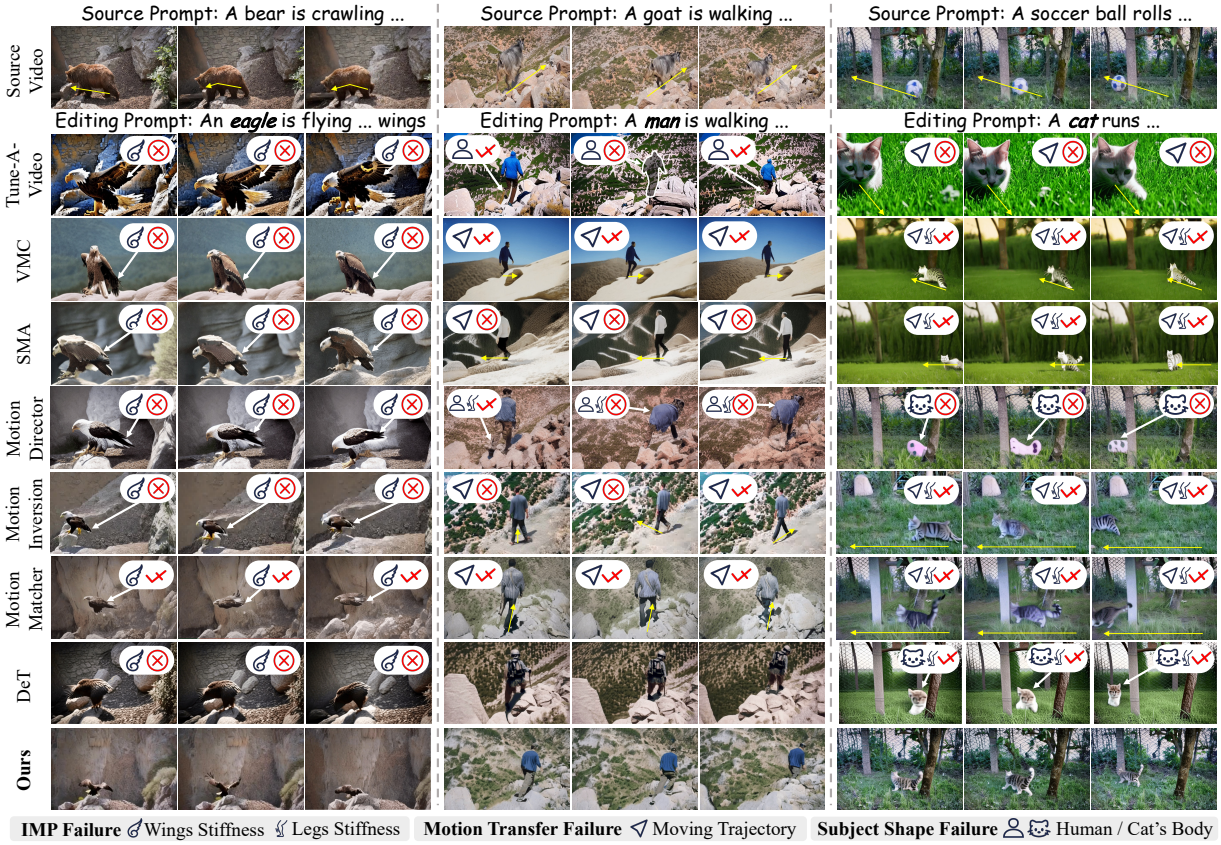


Figure 6. Qualitative comparisons with sota fine-tuning methods. All three cases evaluate motion transfer with different IMP. Specifically, the left case evaluates the subject with significant shape differences. The middle case evaluates motion transfer in camouflaged scenario. The right case evaluates small subject. Note that baseline methods use the same ω^s as our DRA-MTransfer.

0.999, and $1e$-8. The learning rate is set to $1e$-4 with a weight decay of $5e$-4. For videos in V2VBench, 16 frames are sampled with 240 $\times$ 384 resolution. For videos in MT-Bench, following [62], the generated videos are 49 $\times$ 480 $\times$ 720.

4.2. Evaluation Settings

Benchmark. To evaluate the motion transfer ability of DRA-Transfer, we conduct experiments on the commonly used V2VBench [69] and MTBench [62]. V2VBench consists of 50 videos, each paired with three editing instructions, extending the videos from DAVIS [56] dataset. MT-Bench consists of 100 videos, paired with 500 editing instructions. However, most existing editing instructions contain the same IMP as their source videos, which makes them unsuitable for IMP evaluation. Therefore, we use GPT-4o to construct 50 and 100 new editing instructions to evaluate the IMP metrics on V2VBench and MTBench, respectively.
Metrics. Following [46, 62], we utilize text alignment [58], temporal coherence [58], and motion fidelity [93] to measure the motion transfer capability of video editing. In terms of IMP, typical metrics in motion transfer mainly focus on motion consistency between source and target videos [69, 93], while the physical realism of the subject's IMP is not effectively evaluated. Fortunately, we find that VideoPhy [4, 5] focuses on the physical commonsense of motion, the proposed metric is abbreviated as PC. Moreover, inspired by [10, 48], we utilize Qwen2.5-VL-72B (IMP-Q) and GPT-4o (IMP-G) to evaluate the IMP. Therefore, we employ PC [4, 5], IMP-Q, and IMP-G to evaluate IMP.

4.3. Comparison with SOTA Methods

We conduct both qualitative and quantitative evaluations to comprehensively probe the effectiveness of our method.
Baselines. First, we compare DRA-MTransfer with seven fine-tuning methods, *i.e.*, ***Tune-A-Video*** [82], ***VMC*** [26], ***SMA*** [53], ***MotionInversion*** [75], ***MotionDirector*** [98], ***MotionMatcher*** [85], and ***DeT*** [62]. Besides, we perform qualitative comparison with six typical zero-shot methods, *i.e.*, ***FLATTEN*** [14], ***FateZero*** [57], ***MOFT*** [86], ***MotionClone*** [39], ***DMT*** [93], and ***DiTFlow*** [55].
Qualitative comparison. According to the results shown in Fig 6, it can be observed that: 1) ***Most fine-tuning methods are capable of editing subjects with shape variation***. As shown in the left case, all the methods are capable to edit the bear to the eagle. Moreover, even though DeT employs a more powerful VDM, *i.e., CogVideoX* [92], the lim-

Methods	*Objective Evaluations*						*Subjective Evaluations*		
	Text Align.	Temporal Coh.	Motion Fid.	PC	IMP-Q	IMP-G	Temporal Coh.	Motion Fid.	IMP Real.
Tune-A-Video	0.269	0.90	0.83	0.47	0.78	0.77	0.86	0.90	0.84
VMC	0.265	**0.95**	0.81	0.34	0.75	0.74	**0.92**	0.85	0.82
SMA	0.266	**0.95**	0.82	0.36	0.76	0.74	**0.92**	0.84	0.82
MotionInversion	0.271	0.91	0.84	0.53	0.81	0.78	0.89	0.84	0.84
MotionDirector	0.274	0.93	0.89	0.53	0.80	0.79	0.89	0.92	0.85
MotionMatcher	0.278	0.93	0.86	0.61	0.83	0.81	0.87	0.90	0.88
DeT*	0.281	**0.95**	0.90	0.55	0.79	0.79	0.90	**0.94**	0.84
DeT* + **ours** ω^t	**0.282**	**0.95**	0.89	0.58	0.82	0.81	0.91	0.91	0.86
DRA-MTransfer	0.278	0.94	**0.91**	**0.68**	**0.88**	**0.87**	0.90	**0.94**	**0.95**

Table 1. Quantitative comparison with sota fine-tuning methods on V2VBench. Align., Coh. Fid. and Real. denote alignment, coherence, fidelity, and realism. PC is normalized to 0-1. * means the VDM using CogVideoX [92], others using ZeroScope [68]. For subjective evaluation, we employ 26 workers for rating, and report the average score. For DeT, the trajectory of videos in V2VBench is obtained follows its original strategy [29, 62]. ***Note that most original editing instructions contain the same IMP as the source instructions, making them unsuitable for evaluating IMP. Therefore, PC, IMP-Q, IMP-G, and IMP-R are evaluated using newly constructed editing instructions, other metrics are evaluated using the original instructions***. **Bold** and underline denote best, 2nd, respectively.

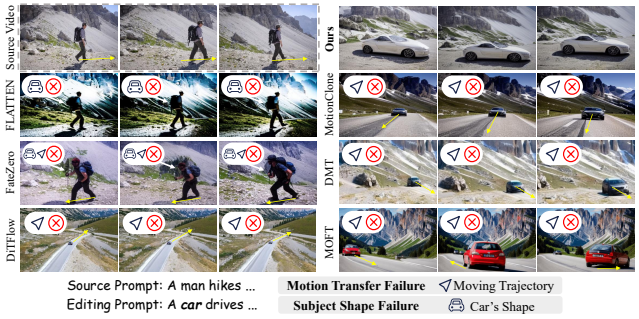


Figure 7. Qualitative comparisons with six state-of-the-art zero-shot motion transfer methods.

Method	FLATTEN	FateZero	MotionClone	DMT	MOFT	Ours
Inference time (m) ↓	3.3	9.2	5.1	7.2	6.1	0.7
Inference storage (GB) ↓	10.8	13.8	14.2	16.1	13.4	8.1

Table 2. Inference time (in minutes) and GPU memory consumption (in GB) for training-free methods. To ensure fairness, we report the results of all methods using ZeroScope as the VDM, except for DiTFlow, which uses DiT. The setting is the same as the evaluation on V2VBench. All the experiments are conducted on a single NVIDIA A6000 GPU.

Method	Tune	VMC	SMA	M-D	MM	MI	Ours
Training time (m) ↓	16.4	5.5	5.5	15.1	21.1	15.5	16.7
Inference time (m) ↓	0.7	16.8	18.2	0.4	1.1	0.6	0.7
Training storage (GB) ↓	7.7	27.5	27.5	11.3	11.4	9.9	11.9
Inference storage (GB) ↓	6.9	26.7	26.8	5.6	5.8	9.0	8.1

Table 3. Training / inference time (in minutes) and GPU memory consumption (in GB) for fine-tuning methods. Tune, M-D, MM, MI denote Tune-A-Video, MotionDirector, MotionMatcher, and MotionInversion respectively.

itation of containing only a single source video still leads to IMP stiffness (eagle's wings). 2) ***DRA-MTransfer exhibits a remarkable advantage in motion transfer***. As shown in the middle and right cases, the trajectories generated by MotionInversion and MotionMatcher are not fully consistent with the source video, since they enhance motion consistency only at the feature level. MotionDirector and DeT retain the trajectory, but the shapes of the subjects are distorted. DRA-MTransfer adaptively balances trajectory transfer and IMP realism in SCR, thereby demonstrating remarkable trajectory transfer capability. 3) ***DRA-MTransfer shows significant improvements in the physical realism of IMP***. As shown in the left and right cases, the wings and legs in target videos generated by most methods are stiff. DRA-MTransfer utilizes motion priors within pre-trained VDM in a dual grained manner, enabling the model to generate physically realistic global IMP posture (*obvious wings, shape of man/cat*) and local motion details (*wings flipping, legs swinging*).

Besides, we compare DRA-MTransfer with zero-shot methods. As shown in Fig. 7, FLATTEN and FateZero are not robust to edit objects with significant shape differences. More importantly, ***most methods struggle to transfer the motion trajectory to the target video***. Therefore, we improve physical realism based on the current mainstream fine-tuning approaches to push the boundaries of video motion transfer capability.

Quantitative comparison. For comprehensive evaluation, we conduct both objective and subjective analysis to quantitatively evaluate DRA-MTransfer, the results are shown in Tab. 1 and Tab. 4. The subjective evaluation is annotated by 26 workers. For DeT, we transfer our adaptive weighting strategy of ω^t to its temporal kernel, denoted as DeT* + ours ω^t.

According to the results shown in Tab. 1, we can observe that: 1) ***DRA-MTransfer exhibits noticeable advantages in IMP***. On both V2VBench and MTBench, DRA-MTransfer achieves the best performance across all IMP metrics, including PC, IMP-Q, and IMP-G, IMP Real., and ***outperforms other methods by at least 4%***. Note that, unlike typical motion-fidelity metric [93] that decouple trajectory and shape, motion-fidelity score in MTBench is influenced by subject's shape, and DiT-based methods exhibit a significant advantage. With the same VDM to ensure fairness, our method achieves superior motion fidelity. 2) ***Our insight of utilizing motion priors within pre-trained VDM can be generalized to more VDMs to enhance the physical realism of IMP***. Incorporating our ω^t adaptive adjustment from

Methods	*Objective Evaluations*						*Subjective Evaluations*		
	Text Align.	Temporal Coh.	Motion Fid.	PC	IMP-Q	IMP-G	Temporal Coh.	Motion Fid.	IMP Real.
MotionDirector	**0.319**	**0.92**	0.68	0.50	0.80	0.78	0.88	0.91	0.84
MotionMatcher	0.311	0.89	0.64	0.56	0.83	0.82	0.88	0.88	0.87
MotionInversion*	0.266	0.85	0.85	0.47	0.82	0.81	0.86	0.85	0.84
DeT*	0.312	0.90	**0.86**	0.55	0.80	0.79	**0.91**	0.92	0.83
DeT* + ours ω^t	0.313	0.91	0.85	0.57	0.83	0.83	**0.91**	**0.93**	0.85
DRA-MTransfer	0.317	0.91	0.71	**0.63**	**0.89**	**0.87**	0.89	0.90	**0.93**

Table 4. Quantitative comparison with sota fine-tuning methods on MTBench. On this benchmark, we calculate temporal coherence and motion fidelity following its original evaluation methods [62], which differ from those on V2VBench.

Methods	Text Sim.	Temporal Coh.	Motion Fid.	PC	IMP-Q	IMP-G
w/o SCR	0.265	0.92	0.87	0.36	0.77	0.77
w/o STR	0.274	0.90	0.90	0.59	0.84	0.82
w/o $\mathcal{L}_\mathcal{G}$	0.272	0.91	0.89	0.44	0.79	0.79
w/o $\mathcal{G}$	0.269	0.91	0.85	0.57	0.82	0.81
w/o ω^t	0.271	0.92	0.88	0.53	0.81	0.79
w/o ω^s	0.263	0.93	0.86	0.66	**0.88**	**0.88**
w/o FEA	0.274	0.89	0.89	0.63	0.85	0.83
w/o sharpen	0.275	0.90	0.90	0.61	0.83	0.82
Ours	**0.278**	**0.94**	**0.91**	**0.68**	**0.88**	0.87

Table 5. Quantitative ablation of DRA-MTransfer on V2VBench.

SCR into DeT (*using CogVideoX as VDM*) yields clear improvements in IMP metrics, and the full set of modules has the potential to transfer to recent spatial–temporal decoupled DiT architectures [46].

Computation and memory costs. To validate the practicality of DRA-MTransfer, we conduct experiments on computation and memory costs. To ensure fairness, we compare methods using the same VDM [68]. DiTFlow and DeT utilize CogVideoX [92] as VDM, which significantly increases computational and storage overhead. According to the results shown in Tab. 2 and Tab. 3, we observe that: 1) Although fine-tuning methods introduce an additional one-shot training process, they generally show lower computational and memory consumption than training-free methods in inference stage. 2) DRA-MTransfer shows comparable computation and memory costs to previous fine-tuning methods, and significantly improves the physical realism of IMP, demonstrating its practicality.

4.4. Ablation Study

Moreover, we conduct detailed ablation experiments for in-depth evaluation. The results are shown in Fig. 8 and Tab. 5. For the module-level ablation, we observe that: 1) ***SCR enables the model to generate natural global posture, STR further refines local motion details***. As shown in Fig. 8, when SCR is removed, the horse's shape is severely distorted. Without STR, although the horse exhibits the correct IMP (i.e., walking), the number of legs is often incorrect. 2) ***With both SCR and STR, our DRA-MTransfer generates physically realistic edited videos***. As shown in Tab. 5, the dual-grained optimization achieves the best performance on IMP-related metrics.

For the detailed component-level ablation, we observe that: 1) $\mathcal{L}_\mathcal{G}$ ***significantly improves global posture realism,***

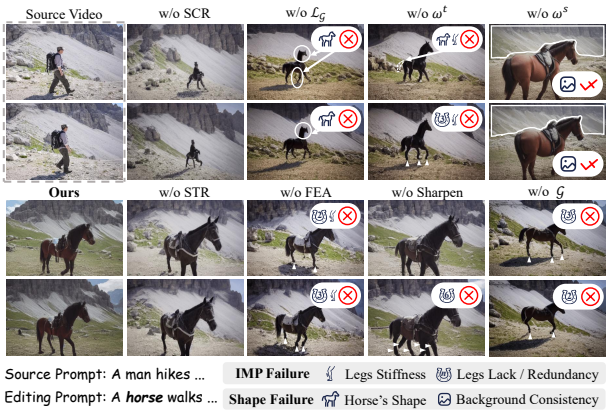


Figure 8. Ablation study. Taking four long legs of a horse as an example to provide a clear visualization.

and together with the adaptive adjustment of ω^t during inference makes the edited subjects more realistic. Moreover, replacing the Gaussian weighting $\mathcal{G}$ with a fixed value of 0.90 (see w/o $\mathcal{G}$) leads to an incorrect number of legs, and a smaller weight even affects the trajectory consistency of motion transfer. 2) According to the visualization of w/o ω^s, identifying the type of editing task and simply ***adjusting ω^s helps to maintain high background consistency with the source video***. 3) ***Without FEA, the two legs of man in the source video affect the horse's IMP, leading to the insufficient leg count***. Without synergistic sharpening, inaccurate IMP knowledge from VDM-A leads to artifacts (excessive number of legs).

5. Conclusion

This paper explores the physical implausibility of IMP in current video motion transfer methods. To this end, we introduce DRA-MTransfer to utilize the potential motion prior within pre-trained VDM at dual granularity. SCR adaptively activates the effect of VDM in a coarse grained manner, and STR utilizes IMP of edited subject in a fine grained attention manner. With the collaboration of SCR and STR, DRA-MTransfer achieves intrinsic-aware motion transfer, showing a remarkable improvement in the physical realism of IMP compared to previous works. We hope this work can encourage future research on subject-faithful dynamics and contribute to world-consistent video editing.

6. Acknowledgements

This work is supported by the National Science and Technology Major Project (2023ZD0121403), and the National Natural Science Foundation of China (No. 62406161).

References

[1] Moayed Haji Ali, Andrew Bond, Tolga Birdal, Duygu Ceylan, Levent Karacan, Erkut Erdem, and Aykut Erdem. Vidstyleode: Disentangled video editing via stylegan and neuralodes. In *ICCV*, pages 7523–7534, 2023. 3

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. *ACM Transactions on Graphics (TOG)*, 24(3):408–416, 2005. 4

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 2

[4] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. In *ICLR*, 2025. 6

[5] Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. Videophy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025. 6

[6] Xiuli Bi, Jian Lu, Bo Liu, Xiaodong Cun, Yong Zhang, Weisheng Li, and Bin Xiao. Customttt: Motion and appearance customized video generation via test-time training. In *AAAI*, 2025. 3

[7] Yufei Cai, Hu Han, Yuxiang Wei, Shiguang Shan, and Xilin Chen. Efficientmt: Efficient temporal adaptation for motion transfer in text-to-video diffusion models. In *ICCV*, 2025. 3

[8] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *ICCV*, pages 23206–23217, 2023. 3

[9] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *ICCV*, pages 23040–23050, 2023. 3

[10] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *ICML*, 2024. 6

[11] Fangda Chen, Shanshan Zhao, Chuanfu Xu, and Long Lan. Jointtuner: Appearance-motion adaptive joint training for customized video generation. *arXiv preprint arXiv:2503.23951*, 2025. 3

[12] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, pages 7310–7320, 2024. 1

[13] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*, 2024. 1

[14] Yuren Cong, Mengmeng Xu, Shoufa Chen, Jiawei Ren, Yanping Xie, Juan-Manuel Perez-Rua, Bodo Rosenhahn, Tao Xiang, Sen He, et al. Flatten: optical flow-guided attention for consistent text-to-video editing. In *ICLR*, 2024. 3, 6

[15] Paul Couairon, Clément Rambour, Jean-Emmanuel Haugeard, and Nicolas Thome. Videdit: Zero-shot and spatially aware text-driven video editing. *Transactions on Machine Learning Research*, 2023. 3

[16] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 3

[17] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883, 2021. 3

[18] Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccedit: Creative and controllable video editing via diffusion models. In *CVPR*, pages 6712–6722, 2024. 3

[19] Yuchao Gu, Yipin Zhou, Bichen Wu, Licheng Yu, Jia-Wei Liu, Rui Zhao, Jay Zhangjie Wu, David Junhao Zhang, Mike Zheng Shou, and Kevin Tang. Videoswap: Customized video subject swapping with interactive semantic point correspondence. In *CVPR*, pages 7621–7630, 2024. 2

[20] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 2

[21] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020. 1

[22] Yaosi Hu, Chong Luo, and Zhenzhong Chen. Make it move: controllable image-to-video generation with text descriptions. In *CVPR*, pages 18219–18228, 2022. 3

[23] Jiahui Huang, Leonid Sigal, Kwang Moo Yi, Oliver Wang, and Joon-Young Lee. Inve: Interactive neural video editing. *arXiv preprint arXiv:2307.07663*, 2023. 3

[24] Hyeonho Jeong and Jong Chul Ye. Ground-a-video: Zero-shot grounded video editing using text-to-image diffusion models. In *ICLR*, 2024. 3

[25] Hyeonho Jeong, Jinho Chang, Geon Yeong Park, and Jong Chul Ye. Dreammotion: Space-time self-similar score distillation for zero-shot video editing. In *ECCV*, pages 358–376, 2024. 3

[26] Hyeonho Jeong, Geon Yeong Park, and Jong Chul Ye. Vmc: Video motion customization using temporal attention adaption for text-to-video diffusion models. In *CVPR*, pages 9212–9221, 2024. 6

[27] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *ICCV*, pages 17191–17202, 2025. 2

[28] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *CVPR*, pages 6507–6516, 2024. 3

[29] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *ECCV*, pages 18–35. Springer, 2024. 7

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[31] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, pages 109–117, 2011. 2

[32] Max Ku, Cong Wei, Weiming Ren, Huan Yang, and Wenhu Chen. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 2, 3

[33] Yao-Chih Lee, Ji-Ze Genevieve Jang, Yi-Ting Chen, Elizabeth Qiu, and Jia-Bin Huang. Shape-aware text-driven layered video editing. In *CVPR*, pages 14317–14326, 2023. 3

[34] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: pre-trained subject representation for controllable text-to-image generation and editing. In *NeurIPS*, pages 30146–30166, 2023. 3

[35] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023. 3

[36] Maomao Li, Yu Li, Tianyu Yang, Yunfei Liu, Dongxu Yue, Zhihui Lin, and Dong Xu. A video is worth 256 bases: Spatial-temporal expectation-maximization inversion for zero-shot video editing. In *CVPR*, pages 7528–7537, 2024. 3

[37] Jun Hao Liew, Hanshu Yan, Jianfeng Zhang, Zhongcong Xu, and Jiashi Feng. Magicedit: High-fidelity and temporally coherent video editing. *arXiv preprint arXiv:2308.14749*, 2023. 3

[38] Minghui Lin, Xiang Wang, Yishan Wang, Shu Wang, Fengqi Dai, Pengxiang Ding, Cunxiang Wang, Zhengrong Zuo, Nong Sang, Siteng Huang, et al. Exploring the evolution of physics cognition in video generation: A survey. *arXiv preprint arXiv:2503.21765*, 2025. 3

[39] Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024. 2, 3, 6

[40] Huijie Liu, Jingyun Wang, Shuai Ma, Jie Hu, Xiaoming Wei, and Guoliang Kang. Separate motion from appearance: Customizing motion via customizing text-to-video diffusion models. *arXiv preprint arXiv:2501.16714*, 2025. 3

[41] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *CVPR*, pages 8599–8608, 2024. 3

[42] Tianyi Lu, Xing Zhang, Jiaxi Gu, Hang Xu, Renjing Pei, Songcen Xu, and Zuxuan Wu. Fuse your latents: Video editing with multi-source latent diffusion models. *arXiv preprint arXiv:2310.16400*, 2023. 3

[43] Haoyu Ma, Shahin Mahdizadehaghdam, Bichen Wu, Zhipeng Fan, Yuchao Gu, Wenliang Zhao, Lior Shapira, and Xiaohui Xie. Maskint: Video editing via interpolative non-autoregressive masked transformers. In *CVPR*, pages 7403–7412, 2024. 3

[44] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *AAAI*, 2024. 1

[45] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH*, pages 1–12, 2024. 2

[46] Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2025. 2, 3, 6, 8

[47] Zhiyuan Ma, Guoli Jia, and Bowen Zhou. Adapedit: Spatio-temporal guided adaptive editing algorithm for text-based continuity-sensitive image editing. In *AAAI*, pages 4154–4161, 2024. 3

[48] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Quanfeng Lu, Wenqi Shao, Kaipeng Zhang, Yu Cheng, Dianqi Li, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. In *ICML*, 2025. 3, 6

[49] Eyal Molad, Eliahu Horwitz, Dani Valevski, Alex Rav Acha, Yossi Matias, Yael Pritch, Yaniv Leviathan, and Yedid Hoshen. Dreamix: Video diffusion models are general video editors. *arXiv preprint arXiv:2302.01329*, 2023. 3

[50] Chong Mou, Mingdeng Cao, Xintao Wang, Zhaoyang Zhang, Ying Shan, and Jian Zhang. Revideo: Remake a video with motion and content control. *arXiv preprint arXiv:2405.13865*, 2024. 3

[51] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024. 2

[52] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *CVPR*, pages 8089–8099, 2024. 3

[53] Geon Yeong Park, Hyeonho Jeong, Sang Wan Lee, and Jong Chul Ye. Spectral motion alignment for video motion transfer using diffusion models. In *AAAI*, 2025. 6

[54] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 1

[55] Alexander Pondaven, Aliaksandr Siarohin, Sergey Tulyakov, Philip Torr, and Fabio Pizzati. Video motion transfer with diffusion transformers. In *CVPR*, pages 22911–22921, 2025. 3, 6

[56] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 6

[57] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *ICCV*, pages 15932–15942, 2023. 3, 4, 5, 6

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 6

[59] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[60] X Ren, T Shen, J Huang, and et al. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, pages 6121–6132, 2025. 2

[61] Fengyuan Shi, Jiaxi Gu, Hang Xu, Songcen Xu, Wei Zhang, and Limin Wang. Bivdiff: A training-free framework for general-purpose video synthesis via bridging image and video diffusion models. In *CVPR*, pages 7393–7402, 2024. 3

[62] Qingyu Shi, Jianzong Wu, Jinbin Bai, Jiangning Zhang, Lu Qi, Yunhai Tong, and Xiangtai Li. Decouple and track: Benchmarking and improving video diffusion transformers for motion transfer. In *ICCV*, pages 10995–11005, 2025. 3, 6, 7, 8

[63] Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Videoflow: Exploiting temporal cues for multi-frame optical flow estimation. In *ICCV*, pages 12469–12480, 2023. 3

[64] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *SIGGRAPH*, pages 1–11, 2024. 3

[65] Chaehun Shin, Heeseung Kim, Che Hyun Lee, Sang-gil Lee, and Sungroh Yoon. Edit-a-video: Single video editing with object-aware consistency. In *ACML*, pages 1215–1230, 2024. 3

[66] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1

[67] Zijian Song, Sihan Qin, Tianshui Chen, Liang Lin, and Guangrun Wang. Physical autoregressive model for robotic manipulation without action pretraining. *arXiv preprint arXiv:2508.09822*, 2025. 3

[68] Spencer Sterling. Zeroscope. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023. 2, 7, 8

[69] Wenhao Sun, Rong-Cheng Tu, Jingyi Liao, and Dacheng Tao. Diffusion model-based video editing: A survey. *arXiv preprint arXiv:2407.07111*, 2024. 3, 6

[70] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *CVPR*, pages 7882–7891, 2024. 3

[71] Shuyuan Tu, Qi Dai, Zihao Zhang, Sicheng Xie, Zhi-Qi Cheng, Chong Luo, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motionfollower: Editing video motion via lightweight score-guided diffusion. *arXiv preprint arXiv:2405.20325*, 2024. 3

[72] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NIPS*, 30, 2017. 3

[73] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3

[74] Dingrui Wang, Zhexiao Sun, Zhouheng Li, Cheng Wang, Youlun Peng, Hongyuan Ye, Baha Zarrouki, Wei Li, Mattia Piccinini, Lei Xie, et al. Enhancing physical consistency in lightweight world models. *arXiv preprint arXiv:2509.12437*, 2025. 3

[75] Luozhou Wang, Guibao Shen, Yixun Liang, Xin Tao, Pengfei Wan, Di Zhang, Yijun Li, and Yingcong Chen. Motion inversion for video customization. *arXiv preprint arXiv:2403.20193*, 2024. 6

[76] L Wang, W Zheng, D Du, and et al. Authentic 4d driving simulation with a video generation model. In *ICCV*, pages 28892–28902, 2025. 2

[77] Wen Wang, Yan Jiang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. *arXiv preprint arXiv:2303.17599*, 2023. 3

[78] Wenchuan Wang, Mengqi Huang, Yijing Tu, and Zhendong Mao. Dualreal: Adaptive joint training for lossless identity-motion fusion in video customization. *arXiv preprint arXiv:2505.02192*, 2025. 3

[79] Yukun Wang, Longguang Wang, Zhiyuan Ma, Qibin Hu, Kai Xu, and Yulan Guo. Videodirector: Precise video editing via text-to-video models. In *CVPR*, pages 2589–2598, 2025. 3

[80] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *SIGGRAPH*, pages 1–11, 2024. 2, 3

[81] Bichen Wu, Ching-Yao Chuang, Xiaoyan Wang, Yichen Jia, Kapil Krishnakumar, Tong Xiao, Feng Liang, Licheng Yu, and Peter Vajda. Fairy: Fast parallelized instruction-guided video-to-video synthesis. In *CVPR*, pages 8261–8270, 2024. 3

[82] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu

Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, pages 7623–7633, 2023. 6

[83] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. In *AAAI*, 2025. 3

[84] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *ICCV*, pages 1206–1217, 2023. 4

[85] Yen-Siang Wu, Chi-Pin Huang, Fu-En Yang, and Yu-Chiang Frank Wang. Motionmatcher: Motion customization of text-to-video diffusion models via motion feature matching. *arXiv preprint arXiv:2502.13234*, 2025. 6

[86] Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *NeurIPS*, 37:76115–76138, 2024. 2, 3, 6

[87] Yiran Xu, Badour AlBahar, and Jia-Bin Huang. Temporally consistent semantic video editing. In *ECCV*, pages 357–374, 2022. 3

[88] Jingyun Xue, Hongfa Wang, Qi Tian, Yue Ma, Andong Wang, Zhiyuan Zhao, Shaobo Min, Wenzhe Zhao, Kaihao Zhang, Heung-Yeung Shum, et al. Follow-your-pose v2: Multiple-condition guided character image animation for stable pose control. *arXiv e-prints*, pages arXiv–2406, 2024. 2

[89] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH*, pages 1–11, 2023. 3

[90] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *SIGGRAPH*, pages 1–12, 2024. 3

[91] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

[92] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 6, 7, 8

[93] Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *CVPR*, pages 8466–8476, 2024. 2, 3, 6, 7

[94] Hidir Yesiltepe, Tuna Han Salih Meral, Connor Dunlop, and Pinar Yanardag. Motionshop: Zero-shot motion transfer in video diffusion models with mixture of score guidance. *arXiv preprint arXiv:2412.05355*, 2024. 3

[95] Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 3

[96] Xinyu Zhang, Zicheng Duan, Dong Gong, and Lingqiao Liu. Training-free motion-guided video generation with enhanced temporal consistency using motion consistency loss. *arXiv preprint arXiv:2501.07563*, 2025. 3

[97] Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *arXiv preprint arXiv:2305.17098*, 2023. 3

[98] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *ECCV*, 2024. 2, 5, 6

[99] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *ECCV*, pages 523–542, 2022. 3

[100] Yi Zuo, Lingling Li, Licheng Jiao, Fang Liu, Xu Liu, Wenping Ma, Shuyuan Yang, and Yuwei Guo. Edit-your-motion: Space-time diffusion decoupling learning for video motion editing. *arXiv preprint arXiv:2405.04496*, 2024. 3