Collaboratively "Copy & Paste" 2D–3D Features for Complex Video-to-Video Motion Editing

Jia-Xing Zhong[1*†], **Shijie Zhao**[1✉], **Junlin Li**[1], **Li Zhang**[1]

[1]ByteDance Inc.

Abstract

Video-to-video human motion editing aims to transfer motion from a driving video to a reference video while preserving the background dynamics and the protagonist's original appearance. We identify critical limitations in existing methods that fail to capture the full complexity of human motions, particularly regarding: 1) location changes, 2) orientation variations, and 3) complicated non-upright poses. To address these challenges, we propose a framework that collaboratively "copies and pastes" 2D and 3D features across spatio-temporal dimensions into a shared representation space for motion guidance. Our approach achieves this through: 1) a mutual distillation mechanism that enhances the robustness and capability of individual encoders, and 2) a selective fusion module that adaptively weights and combines complementary information from spatio-temporal representations. To evaluate motion editing algorithms under challenging scenarios, we introduce a comprehensive benchmark dataset comprising real-world video clips from artistic gymnastics and figure skating competitions. These sports disciplines naturally encompass the three aforementioned aspects of motion complexity. Extensive experiments demonstrate that our approach significantly outperforms existing methods, particularly in handling intricate human motions.

Project Page —
https://jx-zhong-for-academic-purpose.github.io/Copy-Paste-2D-3D-Video-Motion-Editing

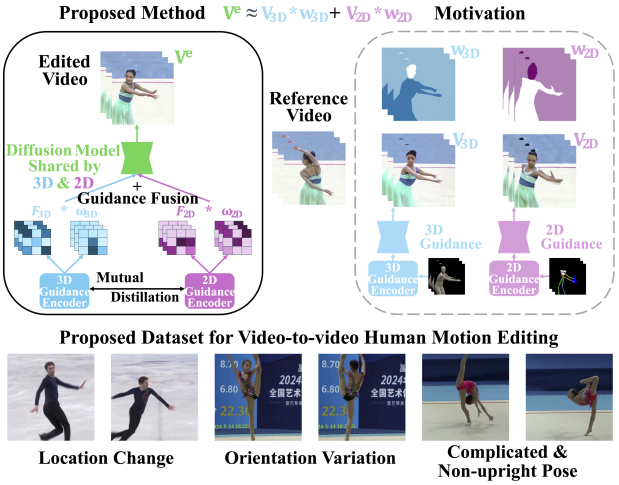

Figure 1: Illustration of our proposed dataset and method motivation. We present a new benchmark for evaluating video motion editing under challenging conditions (zoom in for best viewing). Our approach collaboratively "copies" 2D and 3D features across spatio-temporal dimensions & "pastes" them together by employing a *mutual distillation* strategy to enhance feature learning and a *selective fusion* mechanism to adaptively weight and combine complementary information into a shared representation space.

1 Introduction

Recent advances in diffusion models have demonstrated remarkable success in various generative vision tasks, including image synthesis (Rombach et al. 2022; Mou et al. 2024b; Zhang, Rao, and Agrawala 2023), and video generation (Wu et al. 2023; Guo et al. 2023; Zhang et al. 2023). While significant progress has been made in video editing through diffusion-based approaches (Wu et al. 2023; Bar-Tal et al. 2022; Qi et al. 2023), existing methods predominantly focus on attribute-level manipulation, such as style transfer and appearance editing. Motion information, which stands out as one of the most distinctive and sophisticated features in videos compared to static images, remains largely unexplored.

Video-to-video human motion editing (Tu et al. 2024a), which aims to transfer motion from a specific video to a reference video while preserving the original protagonist's appearance and background in the reference clip, presents several unique challenges. Unlike conventional video editing tasks that modify low-level attributes, motion editing requires precise control over complex spatio-temporal dynamics while maintaining visual consistency. This becomes particularly challenging when handling videos featuring significant camera movements, intricate human poses, and dynamic backgrounds. Furthermore, the inherent entanglement between motion and appearance information in video representations makes it difficult to modify one aspect without affecting the other.

*Email: jxzhong@pku.edu.cn.

†Work done while the author was an intern at ByteDance Inc.

Previous approaches to related tasks, such as image-to-video human motion transfer (Siarohin et al. 2019; Zhai et al. 2024) and pose-guided video generation (Ma et al. 2024; Zhang et al. 2023), have shown limitations in addressing this challenge. Human motion transfer methods typically focus on animating static images based on reference motions, while pose-guided video generation aims to create videos based on coarse-grained motion cues (*e.g.*, trajectories (Mou et al. 2024a; Wang et al. 2024b; Shi et al. 2024; Wu et al. 2025), texts (Blattmann et al. 2023; Ho et al. 2022; Chen et al. 2023), or box-based guidance (Wang et al. 2024a; Ma, Lewis, and Kleijn 2023)). Neither approach adequately addresses the complex requirements of video motion editing, where maintaining the reference video's dynamic elements (e.g., camera movements, background variations, and temporal consistency) is crucial.

Despite the progress made by prior methods of video-to-video motion editing, existing models are evaluated on the benchmarks (detailed in Section 4.1) that primarily rely on self-collected dance performances (Tu et al. 2024b), Tai-Chi demonstrations (Tu et al. 2024a; Zuo et al. 2024). While these benchmarks have facilitated initial progress in the field, they fail to capture the full complexity of human motions in real-world scenarios. We identify three critical aspects of human motion complexity that are under-explored by previous models: 1) large-scale *location changes*, particularly in terms of various depths, where performers move significantly in the anterior-posterior direction; 2) substantial *orientation variations*, where people frequently rotate and present multiple viewing angles; 3) *complicated and non-upright poses* that deviate significantly from standard standing or dancing postures.

To address these limitations and establish a more comprehensive evaluation standard, we introduce a carefully curated test set comprising 130 real-world video clips from "artistic gymnastics" and "figure skating" competitions. As shown in Figure 1, this dataset is specifically designed to encompass the three aforementioned aspects of motion complexity. The videos feature athletes performing intricate movements with dramatic changes in location, orientation, and pose configuration, providing a more challenging and realistic benchmark for evaluating motion editing algorithms. Our dataset is sourced from publicly available competition footage, ensuring accessibility while maintaining high production quality. To facilitate reproducibility and foster further research, we will release our data download scripts and preprocessing pipeline.

Existing approaches (Tu et al. 2024a,b; Zuo et al. 2024) for video-to-video motion editing primarily rely on 2D skeletons to guide video generation. However, our experiments (detailed in Sections 4.3 & 4.2) demonstrate that *2D guidance alone is insufficient to handle complex orientations, occlusions, and position changes*. While incorporating informative 3D representations appears to be a natural solution for complex motion, *estimating 3D information from 2D videos often leads to inaccurate predictions* in challenging conditions. This can compromise the quality of video generation. Moreover, unlike 2D skeletons that typically include fine-grained face keypoints, 3D representations often lack detailed facial and expression information.

Given the complementary nature of 2D and 3D guidance, we propose to leverage their respective strengths through a spatially and temporally aware "copy & paste" approach. As illustrated in Figure 1, when processing a reference video (where the person has her back turned to the camera), the 2D feature-guided video V^{2D} excels at head detail synthesis due to its fine-grained facial landmarks, particularly beneficial when frontal facial references are lacking. Conversely, the 3D feature-guided generation V^{3D} demonstrates superiority in modeling target motion poses and orientations that differ from the reference video. Ideally, if we have the videos generated from 2D and 3D guidances respectively, of which we know exactly "when and where" to select them (perfect spatio-time fusion weights w_{2D} and w_{3D} are obtainable), we could directly "copy" facial details from V^{2D} and "copy" body regions from V^{3D}, and then "paste" them together into the generated video. However, explicit image-space composition seems almost impossible in practice, so we instead perform a similar selective "copy & paste" integration in the feature space.

Specifically, as depicted in Figure 1, our method comprises two key components to encode 2D-3D motion guidance: 1) **Mutual Distillation**: To enhance separate representation learning, we implement a mutual distillation process between 2D and 3D information through an alternating optimization scheme. This iterative process allows each encoder to assist in the other's learning better individual representations and converge to a consistent feature space for subsequent fusion. 2) **Selective Fusion**: To achieve the optimal combination of features, we develop a spatial-temporal fusion module. The porposed method selectively "copies & pastes" 2D and 3D feature maps in a way that allows them to complement each other across spatio-temporal dimensions into a shared representation space.

Our main contributions can be summarized as follows:

- We introduce a comprehensive evaluation dataset for video-to-video motion editing, which features complex human motions from gymnastics and figure skating, addressing crucial gaps in existing benchmarks. Our data download scripts and preprocessing pipeline will be released publicly.
- We propose a framework that effectively combines 2D and 3D motion guidance through collaborative "copy & paste" operations, utilizing mutual distillation and selective fusion, improving individual robustness and enabling complementary integration for specific spatial-temporal representations. To the best of our knowledge, it is the first work to utilize 3D information in video-to-video human motion editing.
- Experiments demonstrate superior performance in handling complex human motions, particularly in scenarios involving significant spatial movement, orientation changes, and intricate poses.

2 Related Works

Due to page limitations, please refer to the **Supplementary Materials** for a comprehensive review of related works on

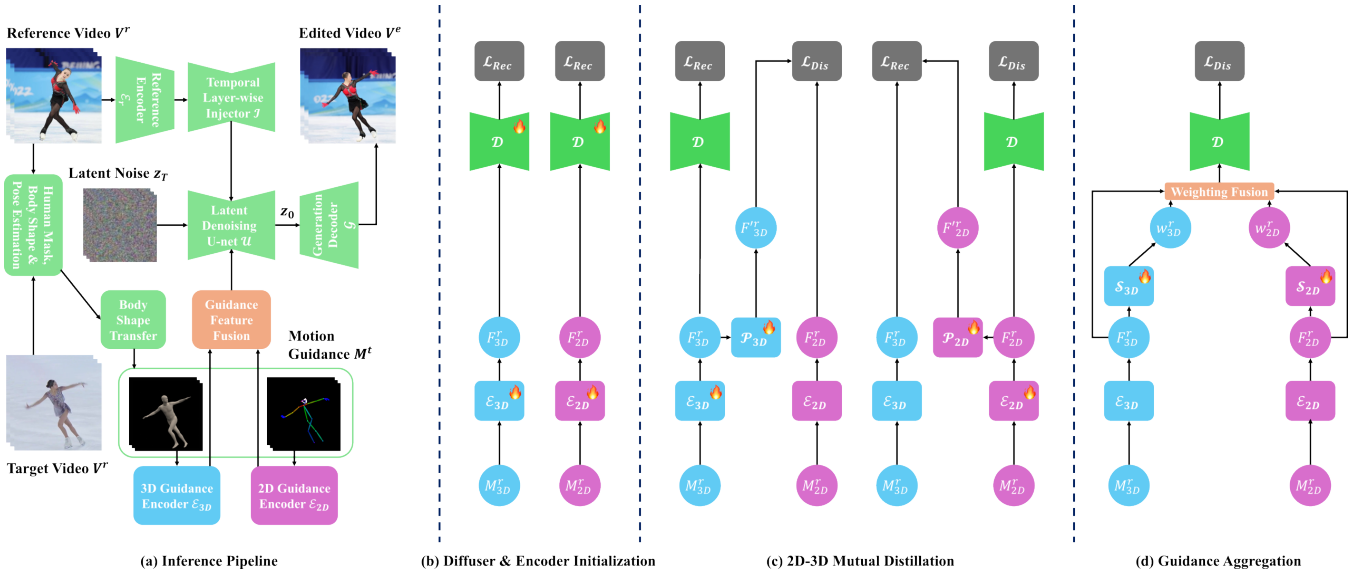

Figure 2: Inference pipeline and the three training stages. To make full use of 2D and 3D guidance during inference (a), we (b) initially train the diffusion model and guidance encoders, and then (c) alternately optimize the encoders with distillation loss, followed by (d) learning spatial-temporal fusion weights. 🔥 represents the modules under optimization; ● denotes intermediate variables.

human motion transfer, editing, and diffusion-based video editing.

3 Methodology

3.1 Problem Statement

Human motion editing in videos is a challenging task that aims to modify the motion of a human subject in a reference video while preserving the original appearance, background, and their movement. Formally, given a reference video of N frames $V^r = \{I_r^i\}_{i=1}^N$ and a target motion sequence $M^t = \{P_t^j\}_{j=1}^L$ obtained from an L-frame video V^t, where I_r^i represents the i-th frame of the source video and P_j^t denotes the corresponding target human pose, the goal is to generate a new video $V^e = \{I_e^t\}_{t=1}^L$ that satisfies the following conditions: 1) V^e follows the motion defined by M^t; 2) Human appearance in V^e matches V^r; 3) Background and its movement from V^r are preserved in V^e; 4) V^e maintains temporal consistency and visual quality. This task can be formulated as finding a function φ that maps the reference video V^r and target motion M^t to the edited video $V^e = \varphi(V^r, M^t)$.

3.2 Framework Overview & Inference Pipeline

Based on the problem formulation above, we now present our proposed framework by introducing the preliminaries, our framework's key components and inference pipeline, followed by our three-stage training strategy.

Preliminaries: Latent Diffusion Model Similarly to recent state-of-the-art approaches in motion editing (Tu et al. 2024a,b; Zuo et al. 2024), our framework is built upon the Latent Diffusion Model (LDM) (Rombach et al. 2022). The core idea of LDM is to learn a T-step *backward diffusion* process $z_T \to z_0$ that can progressively denoise a sequence of low-dimensional latent variables. Specifically, starting from Gaussian noise z_T, the model recovers latent features z_0 that can be synthesized into the edited video V^e using a generation decoder $\mathcal{G}$: $V^e = \mathcal{G}(z_0)$. This backward process reverses a *forward diffusion* process $z_0 \to z_T$, which gradually applies stochastic degradation to the initial latent representation z_0. Formally, at each timestep τ in the forward diffusion process, noise is added from its last step $\tau - 1$ according to a predefined variance schedule β_τ:

$$q(z_\tau | z_{\tau-1}) = \mathcal{N}(z_\tau; \sqrt{1-\beta_\tau} z_{\tau-1}, \beta_\tau \mathbf{I}), \quad (1)$$

where $\tau \in [1, T]$ denotes the diffusion timesteps, $\mathbf{I}$ is an identity matrix, and $\mathcal{N}$ represents a Gaussian distribution.

Structure Components & Inference Pipeline Given a reference video V^r and a target video V^t, our model aims to generate an edited video V^e that maintains the appearance of V^r while following the motion M^t from V^t. As illustrated in Figure 2 (a), our framework comprises several learnable components:

- **3D Guidance Encoder** $\mathcal{E}_{3D}$: Processes 3D motion representations including depth maps, normal maps, and dense pose information from M^t.
- **2D Guidance Encoder** $\mathcal{E}_{2D}$: Handles 2D skeleton keypoints information from M^t.
- **Feature Fusion Module**: Combines 2D and 3D features in a spatio-temporal aware manner. It consists of two key submodules: Mutual Distillation and Weighting Aggregation, as described in Sections 3.3 & 3.3.
- **Diffusion Model Shared by 2D and 3D Inputs** $\mathcal{D}$: A composite system comprising:

- U-Net Denoiser $\mathcal{U}$: Recovers latent noise z_T into z_0
- Reference Encoder $\mathcal{E}_r$: Extracts appearance and background features from V^r
- Temporal Layer-wise Injector $\mathcal{I}$: Injects sequential reference features into the denoiser $\mathcal{U}$ layer by layer
- Generation Decoder $\mathcal{G}$: Generates the video V^e from the latent representations z_0

For conciseness, we denote $\{\mathcal{U}, \mathcal{E}_r, \mathcal{I}, \mathcal{G}\}$ collectively as a diffuser $\mathcal{D}$, which is shared between 2D and 3D guidance paths. The detailed architecture of $\mathcal{D}$ is provided in the **Supplementary Materials**. The following description focuses on motion guidance extraction, encoding procedure, and the feature fusion mechanism.

Motion Guidance M^t Extraction from Target Video V^t. For 2D motion features, we follow common practice by utilizing DWPose (Yang et al. 2023) for skeleton estimation. For 3D motion features, we extract depth maps, normal maps, and dense pose maps through two alternative approaches:

1. **SMPL-based approach**: Inspired by Champ (Zhu et al. 2024), we first estimate the pose and body shape parameters using the SMPL model (Loper et al. 2015). Next, the body shape parameters are transferred from the target video to the reference video, followed by the rendering procedure of the guidance maps.
2. **SAPIENS-based approach**: We utilize the recent SAPIENS model (Khirodkar et al. 2024) to directly render guidance maps containing 3D information. Subsequently, affine transformation is applied to these guidance maps based on the reference human masks.

Our preliminary experiments showed superior performance with the SMPL-based approach, which we adopt for subsequent experiments (details in **Supplementary Materials**). Notably, our approach differs from Champ (Zhu et al. 2024) in handling SMPL-based parameters. While Champ aligns all context parameters (such as body shape, location, and camera pose) to a single reference image for image-to-video transfer, such a strategy proves inadequate for video-to-video editing, particularly when target motion involves significant spatial displacement and dynamic camera movements. Instead, we only transfer body shape parameters and implicitly operate spatial alignment in the feature space through our Feature Fusion Module.

Guidance Feature Encoding. Our framework processes target motion $M^t = \{M^t_{3D}, M^t_{2D}\}$ through two parallel paths:

$$F^t_{3D} = \mathcal{E}_{3D}(M^t_{3D}) \quad M^t_{3D} = \{D^t, N^t, S^t\}\,, \tag{2}$$

where D^t, N^t, and S^t represent depth maps, normal maps, and semantic maps respectively.

$$F^t_{2D} = \mathcal{E}_{2D}(M^t_{2D}) \quad M^t_{2D} = \{K^t\}\,, \tag{3}$$

where K_t denotes 2D human keypoint information. Both encoders share a similar architecture based on inflated 3D convolutions and transformer blocks. The implementation details of $\mathcal{E}_{2D}$ and $\mathcal{E}_{3D}$ (*e.g.*, network depth, and layer width) are moved to **Supplementary Materials** for brevity.

3.3 Training Strategy

As shown in Figure 2 (b)(c)(d), our training process consists of three stages:

Optimization for Diffuser & Initialized Encoders As illustrated in Figure 2 (b), this stage focuses on training two primary components: 1) the diffusion model $\mathcal{D}$ that is shared between 2D and 3D pathways, and (2) the initial guidance encoders $\mathcal{E}_{2D}$ and $\mathcal{E}_{3D}$. Leveraging the rich parametric representation of $\mathcal{D}$, which can be initialized from pre-trained backbones (*e.g.*, Stable Diffusion model (Rombach et al. 2022)), we can effectively optimize these modules using a reconstruction objective on a relatively small dataset consisting of reference videos V^r and their corresponding motion information M^r.

Taking the 3D guidance path as an example (the 2D path follows analogously), given the 3D motion M^r_{3D}, we obtain its feature $F^r_{3D} = \mathcal{E}_{3D}(M^r_{3D})$. The reconstruction loss is formulated as the expected mean squared error (MSE) between the real noise ϵ and its estimate $\epsilon_{\mathcal{D}}$ at timestep τ:

$$\mathcal{L}_{Rec} = \mathbb{E}_{(\epsilon,\tau)}[\eta_\tau \|\epsilon - \epsilon_{\mathcal{D}}(z_\tau, \tau, F^r_{3D})\|^2_2]\,, \tag{4}$$

where η_τ represents a timestep-dependent weighting factor that modulates the contribution of the reconstruction loss.

Feature Fusion: 2D-3D Mutual Distillation After obtaining a well-trained diffuser and initial guidance encoders, we employ knowledge distillation (Hinton 2015) to further enhance the encoding capabilities and constrain the feature space discrepancy between 2D and 3D guidance representations, facilitating their subsequent fusion.

During this phase, we alternate optimization between 2D and 3D encoders. As shown in Figure 2 (c), when training the 3D encoder $\mathcal{E}_{3D}$, the total loss comprises a reconstruction term and a distillation term $\mathcal{L} = \mathcal{L}_{Rec} + \lambda\mathcal{L}_{Dis}$, where λ is a hyperparameter defining the fusion weight. The distillation loss $\mathcal{L}_{Dis}$ is computed based on the projected 3D-guidance feature ${F'}^r_{3D}$ and the 2D feature F^r_{2D}:

$$\mathcal{L}_{Dis} = 1 - \textit{cosine}\langle {F'}^r_{3D}, F^r_{2D}\rangle\,, \tag{5}$$

where *cosine* denotes the cosine similarity. The projected feature ${F'}^r_{3D}$ is obtained through a 3D-to-2D feature projector $\mathcal{P}_{3D}$, which is implemented as a lightweight spatial self-attention mechanism: $\mathbf{Q}^{\mathcal{P}_{3D}} = \mathbf{W}_Q^{\mathcal{P}_{3D}} F^r_{3D}$, $\mathbf{K}^{\mathcal{P}_{3D}} = \mathbf{W}_K^{\mathcal{P}_{3D}} F^r_{3D}$, $\mathbf{V}^{\mathcal{P}_{3D}} = \mathbf{W}_V^{\mathcal{P}_{3D}} F^r_{3D}$ where $\mathbf{W}_Q^{\mathcal{P}_{3D}}$, $\mathbf{W}_K^{\mathcal{P}_{3D}}$, and $\mathbf{W}_V^{\mathcal{P}_{3D}}$ are learnable parameter matrices. The final projected feature is computed through the attention mechanism:${F'}^r_{3D} = \textit{attention}(\mathbf{Q}^{\mathcal{P}_{3D}}, \mathbf{K}^{\mathcal{P}_{3D}}, \mathbf{V}^{\mathcal{P}_{3D}})$. The optimization for $\mathcal{E}_{2D}$ proceeds analogously. Through this mutual distillation process, both encoders learn to generate more robust and compatible feature representations while maintaining their individual strengths in capturing different aspects of motion information.

Feature Fusion: Weighting Guidance Aggregation To effectively combine the information from 2D and 3D guidance, we propose a spatial-temporal fusion module that dynamically calculates fusion weights as depicted in Figure 2 (d). The fusion weights are computed as:

$$w^r_{3D} = \sigma[\mathcal{S}_{3D}(F^r_{3D})] \quad w^r_{2D} = \sigma[\mathcal{S}_{2D}(F^r_{2D})]\,, \tag{6}$$

Method	L1 ↓	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
LWG (Liu et al. 2019)	1.35E-04	0.69	28.45	0.47	55.42	42.33	445.12
MRAA (Siarohin et al. 2021)	1.28E-04	0.70	28.67	0.45	53.67	41.56	442.34
DisCo (Wang et al. 2023)	2.22E-04	0.68	27.82	0.54	68.45	51.23	534.56
MagicAnimate (Xu et al. 2023)	1.47E-04	0.71	28.50	0.47	55.67	41.34	467.89
AnimateAnyone (Hu et al. 2023)	1.33E-04	0.70	28.21	0.46	54.78	42.89	452.45
MagicPose (Chang et al. 2023)	1.42E-04	0.69	28.15	0.48	56.34	43.56	463.67
MusePose (Tong et al. 2024)	1.38E-04	0.70	28.74	0.46	54.89	42.67	455.78
Champ (Zhu et al. 2024)	1.01E-04	0.73	29.22	0.41	48.12	35.45	401.23
MotionFollower (Tu et al. 2024b)	1.94E-04	0.56	28.52	0.52	66.58	49.83	512.44
MotionEditor (Tu et al. 2024a) ‡	1.12E-04	0.73	28.80	0.45	51.37	39.12	428.77
Ours	**6.72E-05**	**0.78**	**30.69**	**0.31**	**37.25**	**27.14**	**358.67**

Table 1: Objective comparison on our dataset, including both image-to-video motion transfer and video-to-video motion editing methods. ‡ indicates the method needs case-specific fine-tuning on each input reference video. Models highlighted in light gray are video-to-video motion editing methods, while the rest are image-to-video motion transfer methods. AnimateAnyone is implemented upon Moore-AnimateAnyone (MooreThreads 2025).

Dataset	#Videos	Motion Type
MotionEditor (Tu et al. 2024a)	20	Dance & Tai Chi
Edit-Your-Motion (Zuo et al. 2024)	25	Dance & Tai Chi
MotionFollower (Tu et al. 2024b)	100	Dance
Ours	**130**	Rhythmic Gymnastics & Figure Skating

Table 2: Comparison on Video-to-video Motion Editing Dataset.

where σ denotes the Sigmoid activation function, and $\mathcal{S}_{3D}$ and $\mathcal{S}_{2D}$ are learnable weighting modules that capture both spatial and temporal context information. For instance, $\mathcal{S}_{3D}$ consists of a spatial cross-attention mechanism $\mathcal{S}_{3D}^{Spat.}$ followed by a temporal attention layer $\mathcal{S}_{3D}^{Temp.}$ (detailed architecture provided in **Supplementary Materials**). The spatial cross-attention operation processes the 3D guidance representations F_{3D}^r in conjunction with the injected U-net features $F_{(\mathcal{I},\mathcal{U},\tau)}^r$ at the timestep τ, computing query, key, and value matrices as follows:

$$\mathbf{Q}^{\mathcal{S}_{3D}^{Spat.}} = \mathbf{W}_Q^{\mathcal{S}_{3D}^{Spat.}} F_{(\mathcal{I},\mathcal{U},\tau)}^r;\ \mathbf{K}^{\mathcal{S}_{3D}^{Spat.}} = \mathbf{W}_K^{\mathcal{S}_{3D}^{Spat.}} F_{3D}^r;\ \mathbf{V}^{\mathcal{S}_{3D}^{Spat.}} = \mathbf{W}_V^{\mathcal{S}_{3D}^{Spat.}} F_{3D}^r, \quad (7)$$

where $\mathbf{W}_Q^{\mathcal{S}_{3D}^{Spat.}}$, $\mathbf{W}_K^{\mathcal{S}_{3D}^{Spat.}}$, and $\mathbf{W}_V^{\mathcal{S}_{3D}^{Spat.}}$ are learnable projection matrices. Based on this spital cross-attention, the temporal attention $\mathcal{S}_{3D}^{Temp.}$ is subsequently applied for modeling dynamics in the video. A parallel training procedure is implemented for $\mathcal{S}_{2D}$. The final fused features are computed through a weighted element-wise combination:

$$F^r = w_{3D}^r * F_{3D}^r + w_{2D}^r * F_{2D}^r\,, \quad (8)$$

where $*$ denotes the Hadamard product. The resulting fused feature F^r is then fed into the diffuser $\mathcal{D}$. The weighting modules $\mathcal{S}_{3D}$ and $\mathcal{S}_{2D}$ are training with the reconstruction loss, enabling them to learn optimal fusion strategies for different spatial regions and temporal contexts.

4 Experiments

4.1 Proposed Dataset

Dataset. We introduce an evaluation dataset for video-to-video motion editing. As detailed in Table 2, current testing data predominantly focus on dance or Tai Chi. In contrast, our dataset introduces more diverse and challenging motion types, specifically including rhythmic gymnastics and figure skating performances. With 130 video clips, our dataset not only offers greater scale but also encompasses more complex motion patterns, *i.e.*, significant spatial movement, orientation changes, and intricate non-upright poses. The data were sourced from publicly available competition videos, and we plan to release download links along with pre-processing scripts to facilitate future research. To ensure compatibility with existing works, we cropped 512×512 regions for evaluation. However, the released dataset maintains the original video quality. Each clip ranges from 25 to 569 frames (predominantly exceeding 100 frames). Visual examples from our dataset are illustrated in Figures 1 & 5, with additional specimens, implementation and evaluation details provided in the **Supplementary Materials**.

4.2 Performance & Comparisons

Baselines. we report comprehensive evaluations against state-of-the-art methods, including both video-to-video motion editing approaches and image-to-video human motion transfer models. For fair comparisons, we adapt image-to-video models to handle video inputs by feeding all reference frames to their reference encoders, rather than the conventional single source image input. Following the original implementation, for MotionEditor, we perform case-specific fine-tuning on each input video before testing. For MotionFollower, we directly employ their trained model on test cases.

Objective Evaluation. As shown in Table 1, our method consistently outperforms all baselines across various metrics, achieving superior performance in terms of both reconstruction accuracy (L1: 6.72E-05, SSIM: 0.78, PSNR:

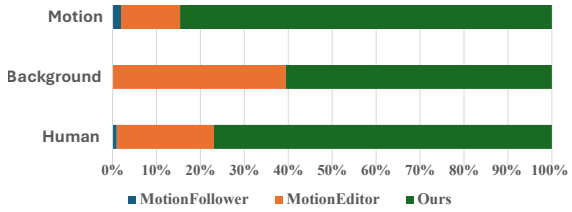

Figure 3: User preference ratio on our subjective settings. *Motion*: Target motion reproduction accuracy; *Appearance*: Reference subject visual preservation; *Background*: Reference scene element maintenance.

30.69) and perceptual quality (LPIPS: 0.31, FID: 37.25). Notably, our approach demonstrates significant advantages in video-specific metrics (FID-VID: 27.14, FVD: 358.67), indicating better temporal consistency and motion fidelity. The performance gap is particularly evident when compared to image-to-video methods like LWG and DisCo, which struggle to maintain temporal coherence across frames despite their adaptation for video inputs. Detailed visualizations of the objective results are provided in the **Supplementary Materials**.

Subjective Evaluation. Figure 5 illustrates both the challenges presented by our dataset and the superiority of our approach. Overall, MotionFollower, without case-specific fine-tuning, shows the poorest performance, particularly in color fidelity (with only a few exceptions, such as the last frame in Figure 5(a)). Remarkably, our method, despite not utilizing case-specific fine-tuning, significantly outperforms the MotionEditor fine-tuned on an input instance. The comparative results highlight several key aspects: (1) *Orientation Variation:* Figure 5(a) demonstrates the challenge of handling gymnastic rotations, where the reference video is profile-view while the target motion includes frontal and back views. MotionEditor, relying solely on 2D information, fails to differentiate orientations (maintaining consistent facial orientations across all poses), whereas our method successfully adapts the editing results based on orientation changes. (2) *Location Changes:* Figure 5(b) showcases rapid position variations, where MotionEditor distorts the athlete's leg in the second frame, while our results remain robust. (3)*Complex Poses:* Figure 5(c) presents non-upright, complex poses as reference videos, in which the rarely seen case (*e.g.*, folded legs) could compromise MotionEditor's generation quality.

Failure Case & Limitations. In Figure 5(d), we observe limitations when handling extreme cases involving distant subjects and unconventional overhead-view squatting poses. While our model demonstrates better performance than other methods in pose capture, it still exhibits notable distortions in generating human appearances.

User Studies. User Preference Ratio is calculated through blind testing, where users compare different approaches in a side-by-side manner with randomized video order. The subjective evaluation focuses on the following three quality aspects: (1) *Appearance Fidelity*: Preservation of reference

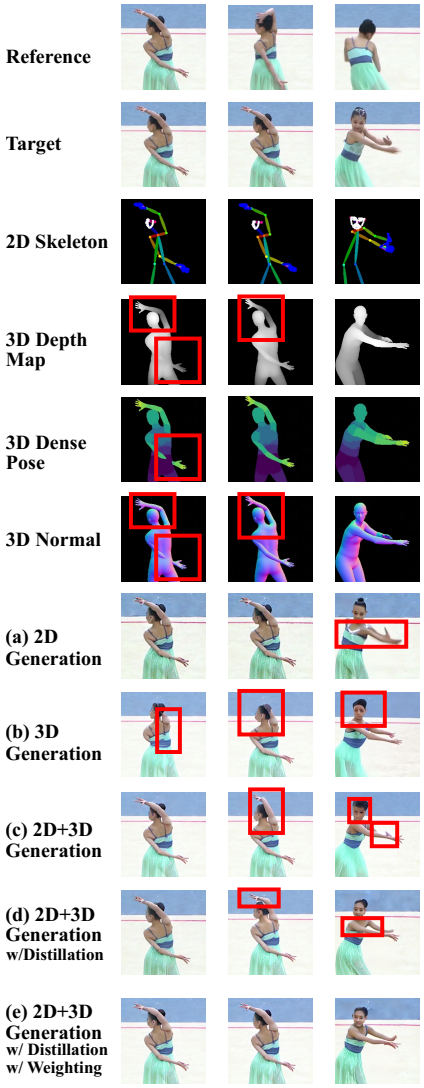

Figure 4: *Ablation studies.* □ highlights areas with significant generation errors in each setting.

subject's visual characteristics. (2) *Background Consistency*: Maintenance of reference scene elements. (3) *Human Motion Similarity*: Accuracy of target motion reproduction.

As illustrated in Figure 3, our method achieves the highest scores across all metrics among the three comparative models. Particularly noteworthy is our substantial lead in motion alignment (84.6% versus 13.5% for the second-best method). Even in background consistency, where case-specific fine-tuning typically excels, our approach maintains a significant advantage.

4.3 Ablation Studies

To thoroughly evaluate the effectiveness of our proposed components, we conduct comprehensive ablation studies on a validation set comprising 10 diverse videos. Quantitative results are reported in **Supplementary Materials**, and we provide intuitive visual examples in Figure 4 in the main

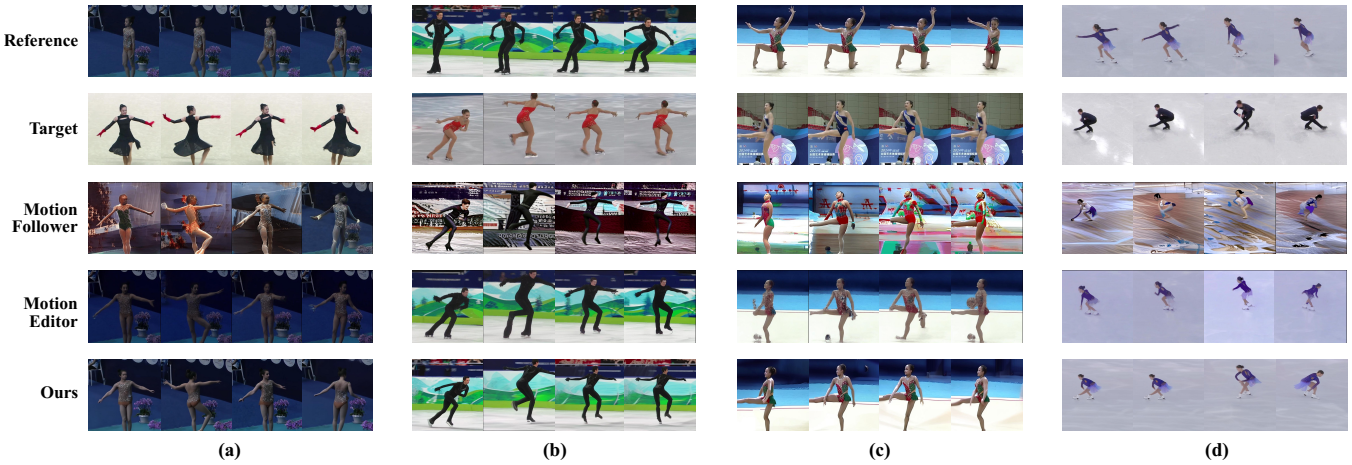

Figure 5: *Qualitative comparison on challenging scenarios.* (a) Orientation Variation; (b) Location Changes; (c) Complex Poses; (d) Failure Case. Zoom in for better view.

body. Our experiments are designed to investigate four key aspects of our method through five controlled settings.

What are the strengths and limitations of 2D and 3D guidance individually? As illustrated in Figure 4(a), 2D guidance exhibits limitations in modeling spatial transformations, particularly during rotational movements, due to its inherent lack of depth information. Conversely, while 3D guidance demonstrates advantages in handling spatial variations (Figure 4(b)), it presents two notable limitations: 1) unlike 2D guidance with fine-grained facial landmarks, it struggles to generate natural facial expressions; 2) the 3D estimation process itself introduces artifacts, such as incorrectly positioning hands placed behind the head or generating anatomically impossible poses where limbs intersect with the body, which adversely affects generation quality.

Does direct combination of 2D and 3D guidance help? Figure 4(c) reveals that while directly combining 2D and 3D guidance yields some complementary benefits, this naive approach often results in compromised outputs that represent a suboptimal middle ground between the two guidance sources (*e.g.*, $2^n d$ frame). The misalignment between different feature spaces can lead to mutual interference, sometimes producing results that do not fully capitalize on the strengths of either guidance type. For instance, facial generation quality may not necessarily surpass that achieved using 2D guidance alone.

How does Mutual Distillation contribute to the performance? The introduction of Mutual Distillation demonstrates improved feature fusion outcomes, as shown in Figure 4(d). This improvement can be attributed to the enhanced alignment between feature spaces. However, certain limitations persist, as evidenced by cases where 2D guidance dominates the generation process, potentially leading to suboptimal results (*e.g.*, in the rightmost example where clothing details are not ideally rendered).

What is the effect of Weighting Fusion? Our Weighting Fusion mechanism enables more selective feature utilization, resulting in further improvements in generation quality. This selective approach allows the model to leverage the strengths of each guidance type more effectively while mitigating their respective weaknesses, as Figure 4(e) illustrates.

5 Conclusion

In this paper, we aim to address the task of video-to-video human motion editing, with a particular focus on complex real-world scenarios. We introduce an evaluation dataset that captures three critical aspects of motion complexity: large-scale location changes, substantial orientation variations, and complicated non-upright poses. This dataset establishes a new benchmark for evaluating motion editing algorithms under challenging conditions. In addition, we propose a novel framework that collaboratively "copies & pastes" 2D and 3D features across spatio-temporal dimensions through mutual distillation and selective fusion, significantly improving the performance where traditional 2D-only methods struggle. Our approach selectively integrates complementary features from both guidance sources into a shared representation space, effectively overcoming the limitations of using either 2D or 3D source alone. While our method demonstrates promising results, limitations remain in handling extreme cases with distant subjects and unconventional viewing angles. Future work could focus on improving 3D estimation robustness and developing more sophisticated adaptive feature selection strategies for extreme poses. We believe our work provides valuable resources and insights for advancing video-to-video human motion editing research.

References

Bar-Tal, O.; Ofri-Amar, D.; Fridman, R.; Kasten, Y.; and Dekel, T. 2022. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, 707–723. Springer.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22563–22575.

Chang, D.; Shi, Y.; Gao, Q.; Xu, H.; Fu, J.; Song, G.; Yan, Q.; Zhu, Y.; Yang, X.; and Soleymani, M. 2023. MagicPose: Realistic Human Poses and Facial Expressions Retargeting with Identity-aware Diffusion. In *Forty-first International Conference on Machine Learning*.

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 preprint arXiv:2310.19512*.

Guo, Y.; Yang, C.; Rao, A.; Wang, Y.; Qiao, Y.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.

Hinton, G. 2015. Distilling the Knowledge in a Neural Network. *arXiv preprint arXiv:1503.02531*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; et al. 2022. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*.

Hu, L.; Gao, X.; Zhang, P.; Sun, K.; Zhang, B.; and Bo, L. 2023. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*.

Khirodkar, R.; Bagautdinov, T.; Martinez, J.; Zhaoen, S.; James, A.; Selednik, P.; Anderson, S.; and Saito, S. 2024. Sapiens: Foundation for Human Vision Models. *arXiv preprint arXiv:2408.12569*.

Liu, W.; Piao, Z.; Min, J.; Luo, W.; Ma, L.; and Gao, S. 2019. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5904–5913.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16.

Ma, W.-D. K.; Lewis, J. P.; and Kleijn, W. B. 2023. Trailblazer: Trajectory control for diffusion-based video generation. *arXiv preprint arXiv:2401.00896*.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5): 4117–4125.

MooreThreads. 2025. GitHub - MooreThreads/Moore-AnimateAnyone: Character Animation (AnimateAnyone, Face Reenactment) — github.com. https://github.com/MooreThreads/Moore-AnimateAnyone. [Accessed 24-07-2025].

Mou, C.; Cao, M.; Wang, X.; Zhang, Z.; Shan, Y.; and Zhang, J. 2024a. ReVideo: Remake a Video with Motion and Content Control. *arXiv preprint arXiv:2405.13865*.

Mou, C.; Wang, X.; Song, J.; Shan, Y.; and Zhang, J. 2024b. Diffeditor: Boosting accuracy and flexibility on diffusion-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8488–8497.

Qi, C.; Cun, X.; Zhang, Y.; Lei, C.; Wang, X.; Shan, Y.; and Chen, Q. 2023. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15932–15942.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Shi, X.; Huang, Z.; Wang, F.-Y.; Bian, W.; Li, D.; Zhang, Y.; Zhang, M.; Cheung, K. C.; See, S.; Qin, H.; et al. 2024. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. *SIGGRAPH 2024*.

Siarohin, A.; Lathuilière, S.; Tulyakov, S.; Ricci, E.; and Sebe, N. 2019. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2377–2386.

Siarohin, A.; Woodford, O. J.; Ren, J.; Chai, M.; and Tulyakov, S. 2021. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13653–13662.

Tong, Z.; Li, C.; Chen, Z.; Wu, B.; and Zhou, W. 2024. MusePose: a Pose-Driven Image-to-Video Framework for Virtual Human Generation. *arxiv*.

Tu, S.; Dai, Q.; Cheng, Z.-Q.; Hu, H.; Han, X.; Wu, Z.; and Jiang, Y.-G. 2024a. MotionEditor: Editing Video Motion via Content-Aware Diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7882–7891.

Tu, S.; Dai, Q.; Zhang, Z.; Xie, S.; Cheng, Z.-Q.; Luo, C.; Han, X.; Wu, Z.; and Jiang, Y.-G. 2024b. MotionFollower: Editing Video Motion via Lightweight Score-Guided Diffusion. *arXiv preprint arXiv:2405.20325*.

Wang, J.; Zhang, Y.; Zou, J.; Zeng, Y.; Wei, G.; Yuan, L.; and Li, H. 2024a. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*.

Wang, T.; Li, L.; Lin, K.; Zhai, Y.; Lin, C.-C.; Yang, Z.; Zhang, H.; Liu, Z.; and Wang, L. 2023. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*.

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 2024 Conference Papers*, 1–11.

Wu, J. Z.; Ge, Y.; Wang, X.; Lei, S. W.; Gu, Y.; Shi, Y.; Hsu, W.; Shan, Y.; Qie, X.; and Shou, M. Z. 2023. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. *arXiv preprint arXiv:2212.11565*.

Wu, W.; Li, Z.; Gu, Y.; Zhao, R.; He, Y.; Zhang, D. J.; Shou, M. Z.; Li, Y.; Gao, T.; and Zhang, D. 2025. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, 331–348. Springer.

Xu, Z.; Zhang, J.; Liew, J. H.; Yan, H.; Liu, J.-W.; Zhang, C.; Feng, J.; and Shou, M. Z. 2023. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*.

Yang, Z.; Zeng, A.; Yuan, C.; and Li, Y. 2023. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4210–4220.

Zhai, Y.; Lin, K.; Li, L.; Lin, C.-C.; Wang, J.; Yang, Z.; Doermann, D.; Yuan, J.; Liu, Z.; and Wang, L. 2024. Idol: Unified dual-modal latent diffusion for human-centric joint video-depth generation. *arXiv preprint arXiv:2407.10937*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*.

Zhu, S.; Chen, J. L.; Dai, Z.; Su, Q.; Xu, Y.; Cao, X.; Yao, Y.; Zhu, H.; and Zhu, S. 2024. Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance. *arXiv preprint arXiv:2403.14781*.

Zuo, Y.; Li, L.; Jiao, L.; Liu, F.; Liu, X.; Ma, W.; Yang, S.; and Guo, Y. 2024. Edit-Your-Motion: Space-Time Diffusion Decoupling Learning for Video Motion Editing. *arXiv preprint arXiv:2405.04496*.