Supplementary Materials: 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.

✉ Corresponding author

1 Related Works

Human Motion Transfer & Editing. *Image-to-video human motion transfer* has been a significant area of research (Chang et al. 2023; Zhai et al. 2024; Zhao and Zhang 2022; Karras et al. 2023) in computer vision. Early GAN-based methods such as LWGAN (Liu et al. 2019) and MRAA (Siarohin et al. 2021) focused on warping source images according to target motions. Recently diffusion-based methods have shown promising results in this domain. DisCo (Wang et al. 2023a) introduced disentangled control for pose and background. Champ (Zhu et al. 2024) leveraged the SMPL model to improve shape alignment and motion guidance. MagicAnimate (Xu et al. 2023) employed the Dense Pose (Güler, Neverova, and Kokkinos 2018) for precise guidance, while Animate Anyone (Hu et al. 2023) utilized a lightweight pose guider to maintain appearance consistency. Unlike these approaches that animate a single reference image, MotionEditor (Tu et al. 2024a) pioneered *video-to-video human motion editing* that aims to modify reference video motions while preserving the original appearance and scenarios. MotionFollower (Tu et al. 2024b) further introduced a lightweight score-guided diffusion model. Edit-Your-Motion (Zuo et al. 2024) apllied a spatio-temporal learning strategy to improve adaptability to unseen cases. However, these methods still exhibit limitations when handling complex human motions. To address this challenge, we introduce a new evaluation dataset that encompasses these challenging scenarios, along with a novel framework that effectively combines 2D and 3D guidance.

Diffusion-based Video Editing. Diffusion-based video editing models have rapidly evolved (Khachatryan et al. 2023; Wu et al. 2024; Ceylan, Huang, and Mitra 2023; Yang et al. 2023; Ku et al. 2024; Cheng, Xiao, and He 2023; Singer et al. 2025; Esser et al. 2023; Wang et al. 2023b), building upon the success of image diffusion models (Rombach et al. 2022; Mou et al. 2024). Tune-A-Video (Wu et al. 2023) introduced a one-shot video tuning technique. AnimateDiff (Guo et al. 2023) proposed motion modules that can be directly integrated into personalized diffusion models. ControlNet (Zhang, Rao, and Agrawala 2023) and its follow-up researches (Ma et al. 2024; Zhang et al. 2023; Men et al. 2024) have been adapted for video tasks, allowing more precise control over the generated content. Our work extends these foundations by introducing a dual-guidance framework that leverages both 2D and 3D information, specifically designed to handle intricate motion patterns.

2 Additional Experimental Results

2.1 Visual Data Samples & Objective Results

Figure 3 illustrates representative results from image-to-video human motion transfer methods, while Figure 4 demonstrates comparative outcomes from video-to-video human motion editing approaches. These visualizations provide comprehensive insights into the performance characteristics of both paradigms when handling challenging motion scenarios in our evaluation dataset.

2.2 Quantitative Results of Ablation Studies

In addition to Figure 5 in the **main body**, Table 1 reports ablation results on a validation set of 10 diverse videos to quantitatively analyze each component's contribution. The validation set encompasses various challenging scenarios including significant spatial movement, orientation changes, and complex poses.

Individual Guidance Performance. When comparing individual guidance types (Groups a & b), 3D guidance demonstrates slight advantages over 2D guidance across all metrics. Specifically, 3D guidance achieves better performance in structural metrics (SSIM: 0.68 vs. 0.65) and perceptual metrics (LPIPS: 0.45 vs. 0.48), suggesting its enhanced capability in capturing spatial relationships. However, both individual approaches show notable limitations, as evidenced by their relatively high FID scores (58.93 and 54.71).

Guidance Combination Effects. The naive combination of 2D and 3D guidance (Group c) yields measurable improvements, reducing the L1 error from 1.52E-04/1.43E-04 to 1.25E-04 and enhancing the PSNR from 27.84/28.15 to 28.92. This confirms the complementary nature of the two guidance types, though the improvement margin suggests room for better integration strategies.

[*]Email: jxzhong@pku.edu.cn. Work done while the author was an intern at ByteDance Inc.

Table 1: *Quantitative results of ablation studies on validation data.*

Exp. Group	Components				Metrics						
	2D Guidance	3D Guidance	Mutual Distill.	Weight Fusion	L1 ↓	SSIM ↑	PSNR ↑	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
(a)	✓				1.52E-04	0.65	27.84	0.48	58.93	45.82	465.31
(b)		✓			1.43E-04	0.68	28.15	0.45	54.71	42.36	442.45
(c)	✓	✓			1.25E-04	0.71	28.92	0.41	49.85	38.92	426.82
(d)	✓	✓	✓		9.84E-05	0.74	29.45	0.36	44.32	34.56	401.23
(e)	✓	✓	✓	✓	7.43E-05	0.75	29.85	0.34	41.53	31.28	385.42

Mutual Distillation Impact. The introduction of mutual distillation (Group d) brings substantial improvements across all metrics, with particularly notable gains in temporal consistency measures (FVD reduced by 6.0% from 426.82 to 401.23). This validates our hypothesis that better feature space alignment through distillation facilitates more effective guidance integration.

Weighting Fusion Benefits. Our complete model with weighting fusion (Group e) achieves the best performance across all metrics, demonstrating significant improvements over the baseline models. Most notably, it reduces the L1 error by 51.1% compared to the 2D-only baseline (7.43E-05 vs. 1.52E-04) and improves the FID-VID score by 31.7% compared to the naive combination approach (31.28 vs. 45.82). These substantial gains validate the effectiveness of our selective feature utilization strategy.

These quantitative results strongly correlate with our qualitative observations in the **main body**, demonstrating that our proposed components work synergistically to enhance generation quality, particularly in challenging scenarios involving complex spatial transformations and temporal dynamics.

3 Architecture Details

3.1 Architecture of Guidance Encoders

As illustrated in Figure 1, each modality of guidance (2D skeleton, 3D dense pose, 3D depth, and 3D normal) is processed through a dedicated encoder consisting of three main components: 1) Intra-modality Encoding Modules, 2) Inter-modality Projection Modules, and 3) Multi-modality Weighting Fusion Modules.

Intra-modality Encoding Modules. Following the previous practice (Zhu et al. 2024), we adopt an alternating structure of InflatedConv3d layers and transformer modules for feature extraction. The architecture progressively increases feature dimensions through multiple stages:

- Initial processing: InflatedConv3d ($3 \rightarrow 16$)
- Stage 1: InflatedConv3d ($16 \rightarrow 32$) + 3D-Transformer (32)
- Stage 2: InflatedConv3d ($32 \rightarrow 96$) + 3D-Transformer (96)
- Stage 3: InflatedConv3d ($96 \rightarrow 256$) + 3D-Transformer (256)
- Final stage: InflatedConv3d ($256 \rightarrow 320$)

Inter-modality Projection Modules $\mathcal{P}_{3D}$ & $\mathcal{P}_{2D}$. Taking $\mathcal{P}_{3D}$ as an example, 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:

$$\begin{aligned}\mathbf{Q}^{\mathcal{P}_{3D}} &= \mathbf{W}_Q^{\mathcal{P}_{3D}} F^r_{3D} \\ \mathbf{K}^{\mathcal{P}_{3D}} &= \mathbf{W}_K^{\mathcal{P}_{3D}} F^r_{3D} \\ \mathbf{V}^{\mathcal{P}_{3D}} &= \mathbf{W}_V^{\mathcal{P}_{3D}} F^r_{3D}\end{aligned}$$

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 spatial attention layers ($320 \rightarrow 640 \rightarrow 320$):

$${F'}^{r}_{3D} = \textit{attention}(\mathbf{Q}^{\mathcal{P}_{3D}}, \mathbf{K}^{\mathcal{P}_{3D}}, \mathbf{V}^{\mathcal{P}_{3D}}) \tag{1}$$

Multi-modality Weighting Fusion Modules. To effectively combine the information from 2D and 3D guidance, we propose a spatial-temporal fusion module that dynamically calculates fusion weights. The fusion weights are computed as:

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

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.

The weighting module $\mathcal{S}_{3D}$ implements a hierarchical spatio-temporal attention structure:

- Initial spatial cross-attention $\mathcal{S}_{3D}^{Spat.}$ expands features from 320 to 640 channels
- Temporal attention $\mathcal{S}_{3D}^{Temp.}$ processes the expanded features (640 channels)
- Secondary spatial attention reduces features from 640 back to 320 channels
- Final temporal attention processes the context-enriched features (320 channels)

The spatial cross-attention operation processes the 3D guidance representations F^r_{3D} in conjunction with the injected context features $F^r_{(\mathcal{I},\mathcal{U},\tau)}$ at the timestep τ:

$$\begin{aligned}\mathbf{Q}^{\mathcal{S}_{3D}^{Spat.}} &= \mathbf{W}_Q^{\mathcal{S}_{3D}^{Spat.}} F^r_{(\mathcal{I},\mathcal{U},\tau)} \\ \mathbf{K}^{\mathcal{S}_{3D}^{Spat.}} &= \mathbf{W}_K^{\mathcal{S}_{3D}^{Spat.}} F^r_{3D} \\ \mathbf{V}^{\mathcal{S}_{3D}^{Spat.}} &= \mathbf{W}_V^{\mathcal{S}_{3D}^{Spat.}} F^r_{3D}\,,\end{aligned} \tag{3}$$

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. The resulting feature $F_{3D-Spat.}^r$ is substantially fed into a temporal attention layer $\mathcal{S}_{3D}^{Temp.}$ based on learnable projection matrices $\mathbf{W}_Q^{\mathcal{S}_{3D}^{Temp.}}$, $\mathbf{W}_K^{\mathcal{S}_{3D}^{Temp.}}$, and $\mathbf{W}_V^{\mathcal{S}_{3D}^{Temp.}}$:

$$\begin{aligned}\mathbf{Q}^{\mathcal{S}_{3D}^{Temp.}} &= \mathbf{W}_Q^{\mathcal{S}_{3D}^{Temp.}} F_{3D-Spat.}^r \\ \mathbf{K}^{\mathcal{S}_{3D}^{Temp.}} &= \mathbf{W}_K^{\mathcal{S}_{3D}^{Temp.}} F_{3D-Spat.}^r \\ \mathbf{V}^{\mathcal{S}_{3D}^{Temp.}} &= \mathbf{W}_V^{\mathcal{S}_{3D}^{Temp.}} F_{3D-Spat.}^r \,.\end{aligned} \tag{4}$$

A parallel training procedure is implemented for $\mathcal{S}_{2D}$ with identical architectural design. The final fused features are computed through a weighted element-wise combination:

$$F^r = w_{3D}^r * F_{3D}^r + w_{2D}^r * F_{2D}^r \,, \tag{5}$$

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 trained with the reconstruction loss, enabling them to learn optimal fusion strategies for different spatial regions and temporal contexts.

3.2 Network Architecture of Shared Diffuser

The diffusion model $\mathcal{D} = \{\mathcal{U}, \mathcal{E}_r, \mathcal{I}, \mathcal{G}\}$ shared between 2D and 3D pathways comprises four main components, designed to leverage pre-trained weights from existing large-scale models. We detail the architecture of each component as follows:

U-Net Denoiser $\mathcal{U}$. The U-Net Denoiser follows a hierarchical 3D U-Net architecture (Guo et al. 2023; Zhu et al. 2024), which is detailed as follows:

- **Input Processing:**
  - Input tensor shape: (B, C, T, H, W)
  - Initial convolution: Conv3D($C_{in} = 4, C_{out} = 320, K = 3, P = 1$)
  - Time embedding: TimestepEmbedding($320 \to 1280$)
- **U-Net Down Blocks:** Four progressive downsampling stages with channel dimensions $(320, 640, 1280, 1280)$
  - First three blocks: CrossAttnDownBlock3D with cross-attention
  - Final block: Standard DownBlock3D
- **U-Net Mid Block:** UNetMidBlock3DCrossAttn with cross-attention
  - Input/Output channels: 1280
  - Combines spatial and temporal attention
- **U-Net Up Blocks:** Four progressive upsampling stages with mirrored channel dimensions
  - First block: Standard UpBlock3D
  - Following three blocks: CrossAttnUpBlock3D with cross-attention
- **Output Processing:**
  - GroupNorm: GroupNorm($n_{groups} = 32, n_{channels} = 320$)
  - SiLU activation
  - Final convolution: Conv3D($C_{in} = 320, C_{out} = 4, K = 3, P = 1$)

The denoiser processes inputs progressively through the encoder path, capturing multi-scale features, then reconstructs the signal through the decoder path while incorporating conditioning information via attention mechanisms. The model's architecture is specifically designed to handle both spatial and temporal dependencies in video data.

Reference Encoder $\mathcal{E}_r$. Based on CLIPVisionModelWithProjection, our reference encoder adopts a hierarchical architecture to extract appearance and background information from the reference video V^r. The model consists of:

- **Patch Embedding:**
  - Input: Reference video frames of shape (B, C, T, H, W)
  - Patchify: Conv2d(in_channels=3, out_channels=768, kernel_size=32, stride=32)
  - Position Embedding: Learnable embeddings for patches + CLS token
- **Transformer Layers:** 12 layers of Vision Transformer blocks
  - LayerNorm (eps=1e-5)
  - Self-Attention: - Hidden size: 768 - Number of heads: 12 - Head dim: 64
  - MLP layers: - Hidden size: 768 $\to$ 3072 $\to$ 768 - GELU activation
  - Residual connections after each sub-block
- **Visual Projection:**
  - Linear projection from 768 to projection dimension
  - L2 normalization of projected features
- **Temporal Self-Attention:**
  - Input: Sequence of frame features (B, T, D)
  - Multi-head attention across temporal dimension
  - Number of heads: 8
  - Query/Key/Value dimension: 96 per head
- **Temporal Position Encoding:**
  - Learned sinusoidal position encodings
  - Added to features before temporal attention
- **Feature Aggregation:**
  - Temporal feature pyramid pooling
  - Adaptive temporal average pooling
  - Skip connections to different denoising stages

The reference encoder first processes each frame independently through the vision transformer backbone, inheriting weights from pre-trained CLIP model. The extracted frame-level features are then enhanced with temporal context through the temporal self-attention mechanism. The model outputs multi-scale features that capture both spatial appearance details and temporal dynamics of the reference video.

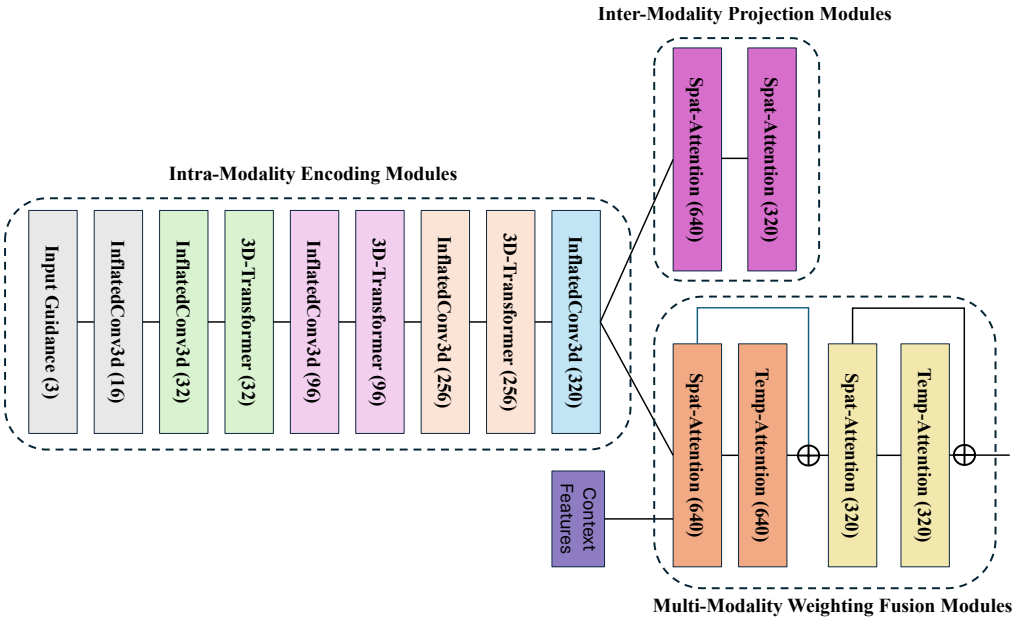

Figure 1: *Architecture of guidance encoder for each modality.* Different colors represent different layers. $\oplus$ denotes the skip connection in residual structures.

Layer-wise Injector $\mathcal{I}$. The Layer-wise Injector adopts a UNet-based structure to process temporal features from the reference encoder. Based on UNet2DConditionModel, it consists of the following components:

- **Input Processing**:
  - Input: Temporal features from $\mathcal{E}_r$ of shape (B, C, T, H, W)
  - Initial projection: Conv2d$(C_{in}, C_{out} = 320, K = 3, P = 1)$
- **Temporal Embedding:**
  - Positional: $\sin(\omega_k t)$ based encoding
  - Dimension: 1280 (4× channel width)
  - Time projection and embedding layers
- **Down Blocks:**
  - Channel dimensions: $(320, 640, 1280, 1280)$
  - Block types: [CrossAttnDownBlock2D × 3, DownBlock2D × 1]
  - Each block contains: - Two ResNet blocks with GroupNorm(32) - Cross-attention layers (where applicable) - Downsampling via strided convolution
- **Mid Block:**
  - Type: UNetMidBlock2DCrossAttn
  - Channels: 1280
  - Includes cross-attention and self-attention
  - Scale factor: 1.0
- **Up Blocks:**
  - Block types: [UpBlock2D × 1, CrossAttnUpBlock2D × 3]
  - Each block contains: - Three ResNet blocks - Cross-attention layers (where applicable) - Skip connections from corresponding down blocks - Upsampling via interpolation

The injector processes temporal features hierarchically while maintaining temporal coherence through inflated operations and attention mechanisms. Skip connections and layer-wise injection ensure effective feature integration at multiple scales.

Generation Decoder $\mathcal{G}$. Following the AutoencoderKL implementation from the diffusers library, the generation decoder adopts a hierarchical variational autoencoder (VAE) structure for high-quality image reconstruction. The architecture is defined as follows:

- **Input Processing:**
  - Input: Latent features (B, C, H, W)
  - Initial projection: Conv2d$(C_{in} = 4, C_{out} = 320, K = 3, P = 1)$
- **Decoder Blocks:**
  - Block types: [UpDecoderBlock2D × n]
  - Channel dimensions: $(320, 640, 1280, 1280)$ reversed
  - Each block contains: - ResNet blocks with GroupNorm(32) - Upsampling via transposed convolution
- **ResNet Block Structure:**
  - Two conv layers with GroupNorm and SiLU
  - Skip connections
  - Time embedding injection
- **Final Processing:**
  - GroupNorm with 32 groups

– SiLU activation
– Conv2d to RGB channels

The decoder efficiently transforms latent representations back to image space while maintaining high fidelity and smooth transitions. Its hierarchical structure and various optimization features make it particularly suitable for high-quality image generation tasks.

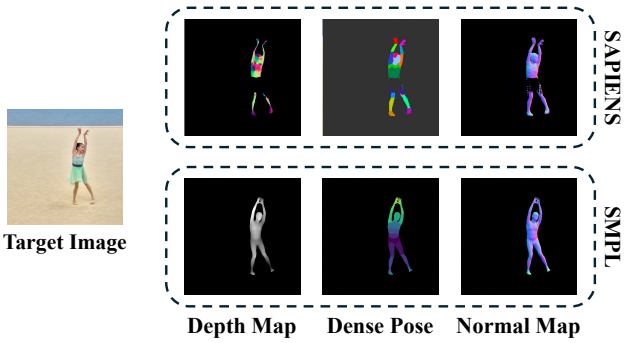

Figure 2: *Comparisons in the two methods of 3D guidance extraction.*

4 Implementation Details

Table 2: *Quantitative details of our evaluation dataset compared to existing benchmarks.* *Original resolution will also be preserved in the released dataset besides the 512×512 version.

Dataset	#Videos	Frame Rate	Resolution
MotionEditor	20	8fps	512×512
MotionFollower	100	Not Disclosed	512×512
Ours	130	30fps	512×512*

4.1 Datailed Settings

Leveraging the pre-trained weights from existing backbones (e.g., Stable Diffusion (Rombach et al. 2022)), our method requires training only on the reference video, similar to existing approaches such as MotionEditor (Tu et al. 2024a). However, unlike MotionEditor which necessitates case-specific fine-tuning during inference, our trained model can be directly applied to new inputs. The training process is conducted on a cluster of NVIDIA V100 GPUs. We set the learning rate to 2×10^{-5} and employ a batch size of 8 for each training stage. In the mutual distillation phase, the loss fusion hyperparameter λ is empirically set to 0.01.

4.2 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.

4.3 More Details on our Dataset

While Table 1 in the main paper provides initial comparisons with existing Video-to-video Motion Editing datasets, we provide more detailed quantitative comparisons in Table 2, including the requested frame rate information.

4.4 3D Guidance Extraction: SMPL *v.s.* SAPIENS

As illustrated in Figure 2, we conducted preliminary experiments comparing two approaches for extracting 3D guidance information: the SAPIENS model (Khirodkar et al. 2024) and the Skinned Multi-Person Linear model (SMPL) (Loper et al. 2015). Our analysis reveals distinct characteristics between these two methods in terms of their guidance generation:

- **SAPIENS Approach:**
  - Adopts a conservative prediction strategy
  - Exhibits selective output generation, particularly for heavily occluded or clothed regions
  - Produces incomplete but high-confidence 3D information
- **SMPL Approach:**
  - Provides comprehensive full-body 3D predictions
  - Generates complete guidance maps for depth, normal, and dense pose information
  - Maintains informative predictions across different clothing and occlusion conditions

The comparison demonstrates that while SAPIENS prioritizes prediction reliability by omitting uncertain regions, SMPL's holistic approach provides more comprehensive guidance signals for the diffusion model. The completeness of SMPL's predictions proves particularly advantageous in our motion editing framework, as it ensures informative 3D guidance across all body parts, facilitating more effective motion transfer and editing operations.

4.5 Evaluation Protocols

We employ both *objective and subjective* assessments to provide a comprehensive evaluation of our method. 1) For the objective evaluation, we split each video (featuring the same person and scenario) into two segments: the first 10 frames serve as the source video, while the remaining frames provide the target motion sequences. Following established protocols in the literature (Tu et al. 2024b; Zuo et al. 2024), we evaluate both single-frame image quality and video fidelity. The single-frame quality metrics include L1 error, SSIM (Wang et al. 2004), LPIPS (Zhang et al. 2018), PSNR (Hore and Ziou 2010), and FID (Heusel et al. 2017). Video fidelity is assessed using FID-FVD (Balaji et al. 2019) and FVD (Unterthiner et al. 2018). 2) For subjective evaluation, we conducted user studies involving 26 video clips

featuring different individuals and scenes. User Preference Ratio was used as the primary metric to gather subjective preferences about human appearance similarity and background consistency with the reference, and the motion alignment with the target video. To comprehensively evaluate our method's effectiveness, we conduct both objective and subjective assessments. These two evaluation paradigms are designed with distinct video pair settings to examine the performance of models.

Intra-video Objective Evaluation For objective evaluation, we employ *intra-video assessment* where the reference and target videos feature identical subjects and scenarios, enabling direct numerical comparison with ground truth.

Data Preparation. For each evaluation sequence of the 130 videos, we split the clip as follows:

- **Reference Segment**: The initial 10 frames are extracted to serve as the source video, providing appearance and background context
- **Target Segment**: The remaining frames constitute both:
  - Motion guidance sequence for video generation
  - Ground truth for quantitative metric computation

Single-frame Quality Metrics. We employ a comprehensive set of image quality metrics:

- **L1 Error**:
  - Measures absolute pixel-level discrepancy
  - Lower values indicate better reconstruction accuracy
- **SSIM (Structural Similarity Index Measure)** (Wang et al. 2004):
  - Evaluates structural similarity considering luminance, contrast, and structure
  - Range: [0,1], higher values indicate better quality
- **LPIPS (Learned Perceptual Image Patch Similarity)** (Zhang et al. 2018):
  - Utilizes deep feature representations for perceptual similarity
  - Aligns with human visual judgments
- **PSNR (Peak Signal-to-Noise Ratio)** (Hore and Ziou 2010):
  - Quantifies reconstruction quality in decibels
  - Standard benchmark for image fidelity
- **FID (Fréchet Inception Distance)** (Heusel et al. 2017):
  - Measures distributional similarity in feature space
  - Lower scores indicate better quality and diversity

Video Fidelity Metrics. To evaluate temporal aspects and video quality:

- **FID-VID (Frame-wise Fréchet Inception Distance)** (Balaji et al. 2019):
  - Extends FID to video domain through frame-wise computation
  - Captures per-frame generation fidelity
- **FVD (Fréchet Video Distance)** (Unterthiner et al. 2018):
  - Employs spatio-temporal features from I3D network
  - Assesses both temporal coherence and motion quality

Inter-video Subjective Evaluation For subjective assessment, we employ *inter-video evaluation* where reference and target sequences contain different subjects and scenarios, better reflecting real-world applications.

Dataset Preparation. We adopt 26 diverse videos for subjective evaluation under the following configuration:

- **Reference Segment**: A 10-frame clip serves as the source video, providing human foreground information and background context
- **Target Segment**: A clip from another video provides the motion guidance sequence for video generation with:
  - Varied performers with distinct appearances and physiques
  - Diverse environmental conditions and scene compositions
  - Complex motion patterns involving significant spatial movement, orientation changes, and intricate poses

References

Balaji, Y.; Min, M. R.; Bai, B.; Chellappa, R.; and Graf, H. P. 2019. Conditional GAN with discriminative filter generation for text-to-video synthesis. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 1995–2001.

Ceylan, D.; Huang, C.-H. P.; and Mitra, N. J. 2023. Pix2Video: Video Editing using Image Diffusion. In *ICCV*.

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*.

Cheng, J.; Xiao, T.; and He, T. 2023. Consistent video-to-video transfer using synthetic dataset. *arXiv preprint arXiv:2311.00213*.

Esser, P.; Chiu, J.; Atighehchian, P.; Granskog, J.; and Germanidis, A. 2023. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7346–7356.

Güler, R. A.; Neverova, N.; and Kokkinos, I. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7297–7306.

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*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*, 2366–2369. IEEE.

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*.

Karras, J.; Holynski, A.; Wang, T.-C.; and Kemelmacher-Shlizerman, I. 2023. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 22623–22633. IEEE.

Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; and Shi, H. 2023. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. *arXiv preprint arXiv:2303.13439*.

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*.

Ku, M.; Wei, C.; Ren, W.; Yang, H.; and Chen, W. 2024. Anyv2v: A plug-and-play framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*.

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, 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.

Men, Y.; Yao, Y.; Cui, M.; and Bo, L. 2024. MIMO: Controllable Character Video Synthesis with Spatial Decomposed Modeling. *arXiv preprint arXiv:2409.16160*.

Mou, C.; Wang, X.; Song, J.; Shan, Y.; and Zhang, J. 2024. 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.

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.

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.

Singer, U.; Zohar, A.; Kirstain, Y.; Sheynin, S.; Polyak, A.; Parikh, D.; and Taigman, Y. 2025. Video editing via factorized diffusion distillation. In *European Conference on Computer Vision*, 450–466. Springer.

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*.

Unterthiner, T.; Van Steenkiste, S.; Kurach, K.; Marinier, R.; Michalski, M.; and Gelly, S. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

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

Wang, W.; Xie, k.; Liu, Z.; Chen, H.; Cao, Y.; Wang, X.; and Shen, C. 2023b. Zero-Shot Video Editing Using Off-The-Shelf Image Diffusion Models. *arXiv preprint arXiv:2303.17599*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, B.; Chuang, C.-Y.; Wang, X.; Jia, Y.; Krishnakumar, K.; Xiao, T.; Liang, F.; Yu, L.; and Vajda, P. 2024. Fairy: Fast parallelized instruction-guided video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8261–8270.

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*.

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, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

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*.

Zhao, J.; and Zhang, H. 2022. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3657–3666.

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*.

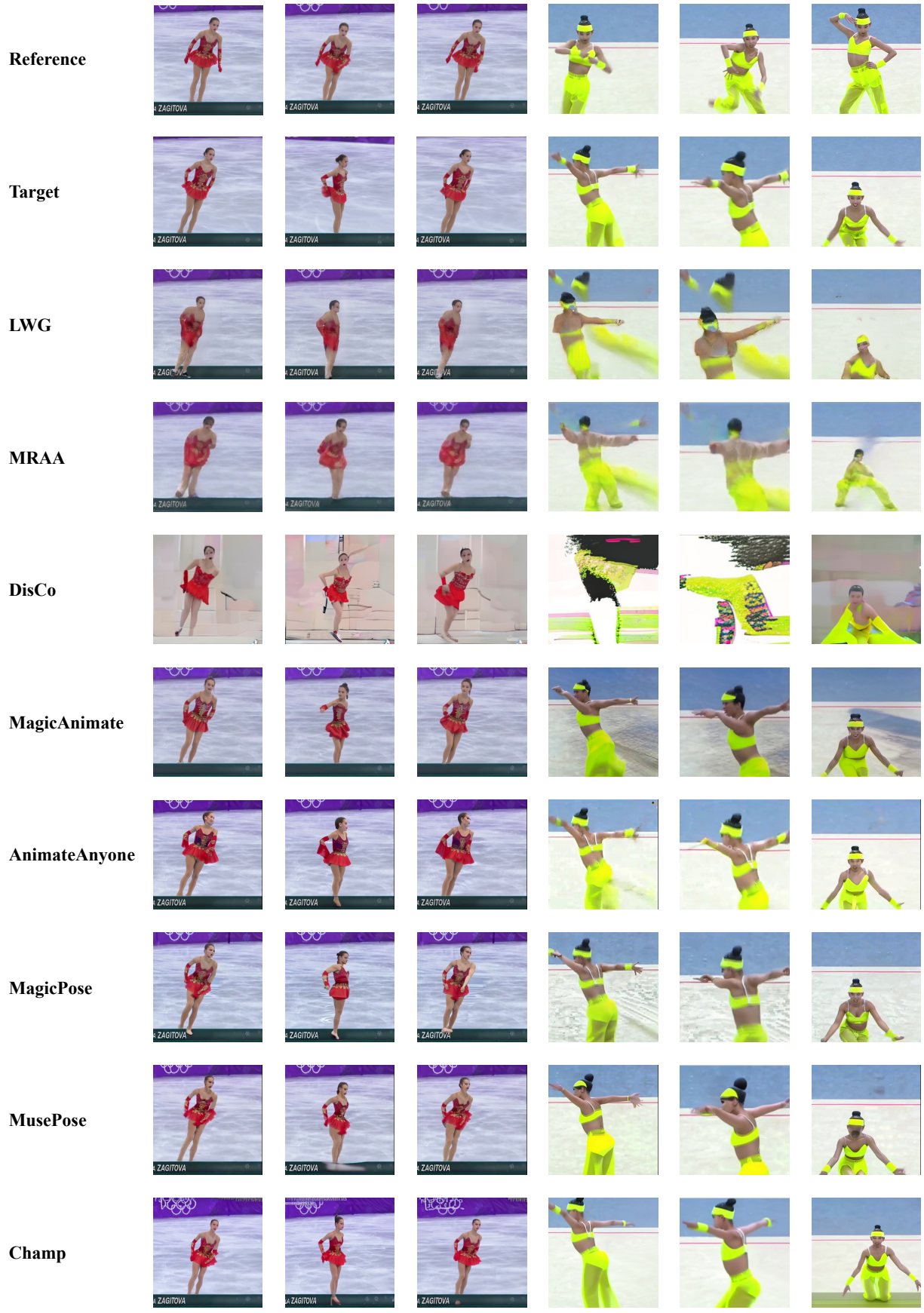

Figure 3: *Visual data samples and objective results: image-to-video human motion transfer methods.*

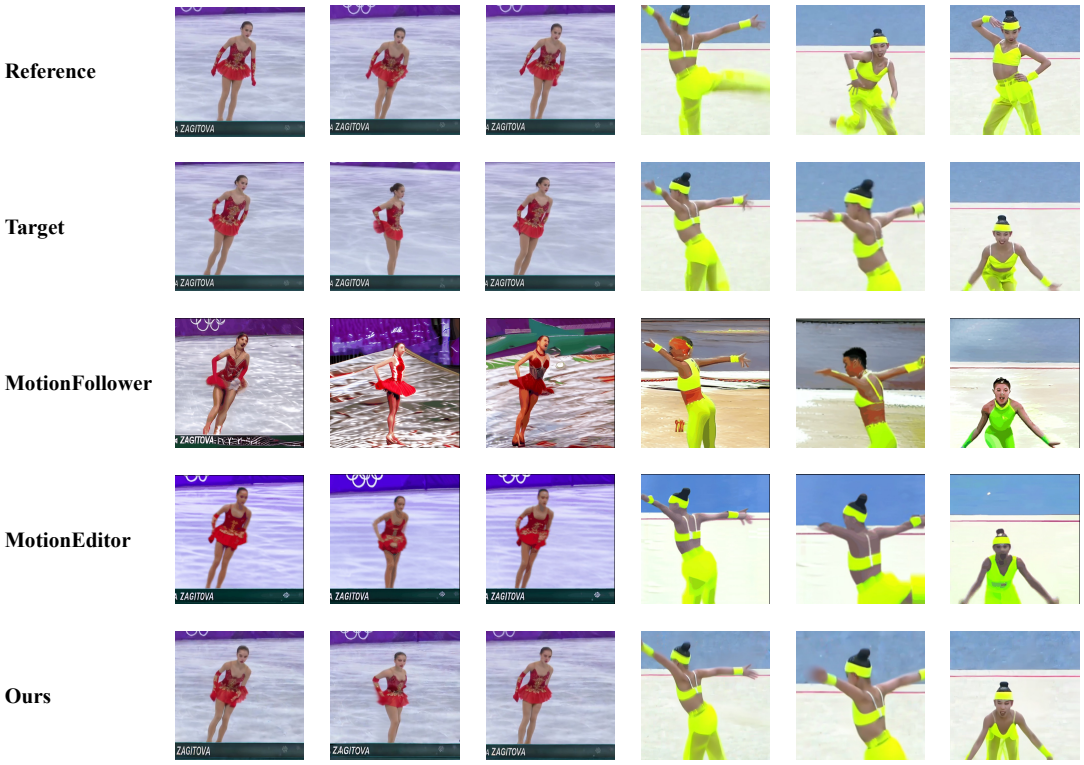

Figure 4: *Visual data samples and objective results: video-to-video human motion editing methods.*