arXiv:2508.19895v2 [cs.CV] 9 Nov 2025

PersonaAnimator: Personalized Motion Transfer from Unconstrained Videos

Ziyun Qian[1*], Runyu Xiao[1*], Shuyuan Tu[2], Wei Xue[1], Dingkang Yang[1†], Mingcheng Li[1], Dongliang Kou[1], Minghao Han[1], Zizhi Chen[1], Lihua Zhang[1‡]

[1]College of Intelligent Robotics and Advanced Manufacturing, Fudan University
[2]College of Computer Science and Artificial Intelligence, Fudan University
zyqian22@m.fudan.edu.cn

Abstract

Recent advances in motion generation show remarkable progress. However, several limitations remain: (1) Existing pose-guided character motion transfer methods merely replicate motion without learning its style characteristics, resulting in inexpressive characters. (2) Motion style transfer methods rely heavily on motion capture data, which is difficult to obtain. (3) Generated motions sometimes violate physical laws. To address these challenges, this paper pioneers a new task: Video-to-Video Motion Personalization. We propose a novel framework, PersonaAnimator, which learns personalized motion patterns directly from unconstrained videos. This enables personalized motion transfer. To support this task, we introduce PersonaVid, the first video-based personalized motion dataset. It contains 20 motion content categories and 120 motion style categories. We further propose a Physics-aware Motion Style Regularization mechanism to enforce physical plausibility in the generated motions. Extensive experiments show that PersonaAnimator outperforms state-of-the-art motion transfer methods and sets a new benchmark for the Video-to-Video Motion Personalization task.

Introduction

With the rapid rise of the metaverse and digital content creation, building digital avatars with a strong sense of identity has become a central requirement for interactive experiences. Motion style, as a highly distinctive representation of individual behavior, not only captures subtle kinematic variations but also conveys high-level semantic cues such as identity and emotion—for instance, their gait or signature gestures can often recognize a person. (Ribet, Wannous, and Vandeborre 2019; Abdul-Massih, Yoo, and Benes 2017; Shen et al. 2024) In computer vision, prior works (Zhao, Li, and Tang 2024; Wang et al. 2025a) have achieved basic motion generation, but key limitations remain: (1) Pose-guided character motion transfer methods (Tu et al. 2024a; Li et al. 2024a; Wang et al. 2025b) merely mechanically replicate motion without capturing personalized traits, severely limiting the expressiveness of digital characters in real-world applications. (2) Motion style transfer techniques (Aberman et al. 2020; Kim et al. 2024, 2025) can learn style representations. However, they rely wholly or partially on motion capture data, which is costly, labor-intensive, and often unavailable for historical figures or virtual characters. (3) Existing methods generally lack explicit physical constraints on the full-body skeleton, leading to distorted or implausible motion.

To address the above challenges, we propose the novel task of Video-to-Video Motion Personalization, which firstly enables personalized motion transfer from unconstrained videos, as illustrated in Fig. 1. Our PersonaAnimator framework makes two key breakthroughs: First, it is the first to learn personalized motion features entirely from ordinary video data, without relying on any motion capture data, making it feasible to model historical figures or virtual characters for whom such data is unavailable. Second, our framework preserves the structural integrity of the content motion while learning personalized style characteristics, generating personalized motions. Rather than performing simple, mechanical motion transfer, it establishes a new paradigm that shifts from "postural approximation" to "personalized dynamic essence" in motion generation. These innovations enable wide applications in digital content creation—for example, allowing stunt doubles to replicate iconic celebrity movements (*e.g.*, Michael Jackson's "Moonwalk") or letting users embed signature gestures (*e.g.*, a chin rest) into virtual avatars, offering critical support for building digital humans with authentic identities.

To train the video-to-video motion personalization model, we introduce PersonaVid, the first video-based personalized motion dataset, comprising 20 motion content categories and 120 motion style categories. We adopt a "one person, one style" paradigm, treating each individual as a unique style class to better capture personalized motion traits. Additionally, we propose the Physics-aware Motion Style Regularization (PMSR) mechanism, which enforces physically plausible motion through dynamic bone stability and body connectivity constraints. Extensive experiments demonstrate that PersonaAnimator achieves state-of-the-art (SOTA) results on both PersonaVid and mainstream human animation datasets (Jafarian and Park 2021; Zablotskaia et al. 2019). The main contributions are summarized as follows:

- We pioneer a novel task, Video-to-Video Motion Personalization, and propose a new framework, PersonaAnima-

*These authors contributed equally.
†Corresponding authors.
‡Corresponding authors.

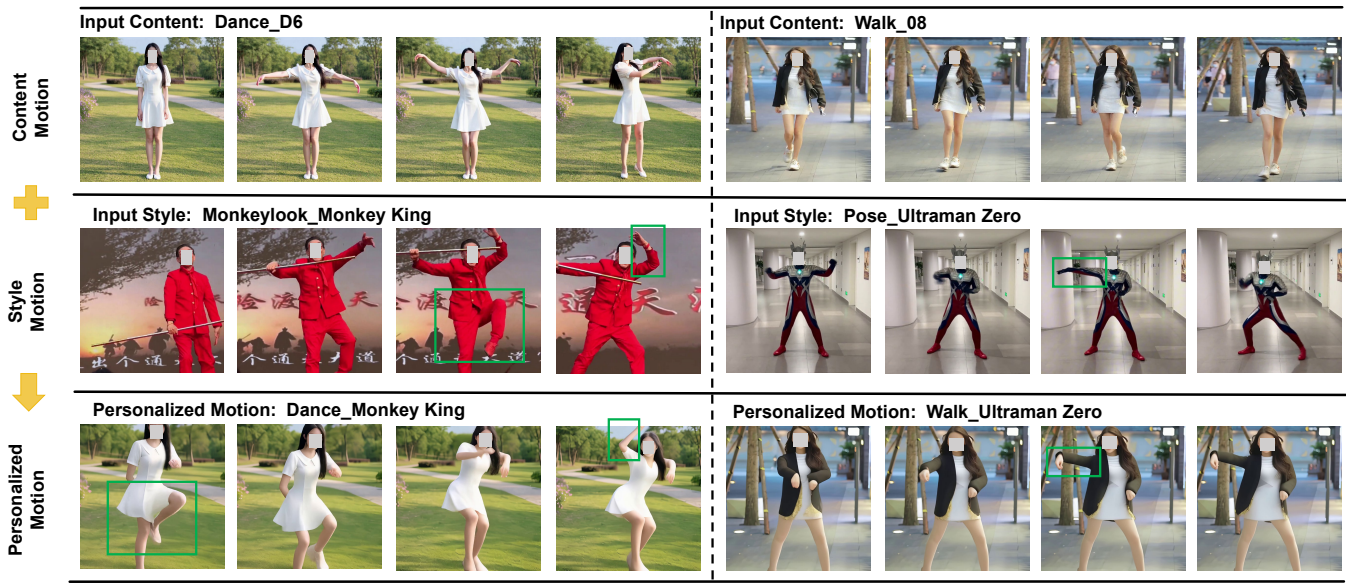


Figure 1: Video-to-Video Motion Personalization. Our method preserves the basic motion content while learning personalized motion characteristics from the style motion video, achieving personalized transfer. Learned personalized features are marked with green bounding boxes.

tor, which incorporates the SA-PMT module to enable personalized motion transfer from unconstrained videos.

- We introduce PersonaVid, the first video-based personalized motion dataset.
- We propose the PMSR mechanism to enforce physical plausibility in generated motions.
- Extensive experiments show that PersonaAnimator outperforms SOTA methods.

Related Work

Diffusion Models for Video Generation

Diffusion models are central to video generation, focusing on two directions: Text-to-Video (T2V) (Xing et al. 2024; Ho et al. 2022b,a; Lin et al. 2023; Fei et al. 2023) and unconditional generation (Mei and Patel 2023; Lu et al. 2023; Fei et al. 2024). VDM (Ho et al. 2022b) pioneered diffusion-based T2V by extending 2D U-Net to 3D. Imagen Video (Ho et al. 2022a) uses seven submodels in a cascade system. Recent work integrates LLMs: VideoDirectorGPT (Lin et al. 2023) for multiscene planning and Dysen-VDM (Fei et al. 2023) for scene graphs. In an unconditional generation, VIDM (Mei and Patel 2023) established a two-stream architecture. VDT (Lu et al. 2023) introduced Transformers for spatio-temporal modeling, and VITRON (Fei et al. 2024) combines LLMs with diffusion models.

Pose-guided Character Motion Transfer

Pose-Guided Character Motion Transfer synthesizes target character videos from pose signals, with research evolving from traditional explicit methods to modern diffusion-based approaches (Tu et al. 2025b; Tan et al. 2025; Li et al. 2024a; Tu et al. 2024a; Wang et al. 2025a; Tu et al. 2023, 2024b, 2025c,a). Early techniques relying on optical flow estimation and Thin-Plate Spline transformations with GAN inpainting frequently exhibited artifacts, including texture tearing and temporal flickering (Chan et al. 2019; Zhu et al. 2023). Contemporary diffusion models have demonstrated superior performance in motion accuracy and detail preservation. Recent innovations include MagicAnimate's (Xu et al. 2024) integration of DensePose with temporal attention modules for effective motion-identity decoupling and Animate Anyone's (Hu 2024) ReferenceNet combined with pose guidance for stable generation from single images. Champ (Zhu et al. 2024) advances the field by incorporating SMPL models with multimodal 3D signals (depth, normal, and semantic maps) to enhance alignment precision, while StableAnimator (Tu et al. 2025b) tackles identity drift through specialized facial preservation techniques. Despite these advancements, current methods remain limited to mechanical motion transfer without capturing personalized characteristics. This fundamental constraint results in rigid, expressionless animations that significantly hinder applications in film production and immersive digital environments, where nuanced character expressiveness is crucial.

Motion Style Transfer

Motion Style Transfer methods (Zhong et al. 2024; Li et al. 2024b; Xie et al. 2024) transfer stylistic features while preserving target motion content integrity. Recent advances show diversified approaches and improved results. Aerman *et al.* (Aberman et al. 2020) pioneer unpaired motion style transfer, extracting styles directly from videos for 3D animation. Finestyle (Song et al. 2023) enhances control through semantic-aware DIFF modules, while MCM-LDM (Song et al. 2024) achieves Arbitrary Motion Style Transfer via

trajectory-content-style decoupling. Most (Kim et al. 2024) enables cross-content transfer using part-aware modulators, and PersonaBooth (Kim et al. 2025) introduces text-guided personalized generation. However, current methods rely on motion capture or SMPL data (Loper et al. 2023), which limits applications for special subjects, creating demand for more flexible solutions.

METHODOLOGY

Pose Representation

We categorize the input motion videos of the PersonaAnimator framework into two types based on function: content motion video $\boldsymbol{V}_c \in \mathbb{R}^{F\times 3\times H\times W}$, which provides basic motion content (*e.g.*, dancing, walking), and style motion video $\boldsymbol{V}_s \in \mathbb{R}^{F\times 3\times H\times W}$, which offers personalized motion characteristics (*e.g.*, gait, wave amplitude). Here, c denotes motion content, s denotes motion style, F is the frame count, and H and W are the height and width of the video frames, respectively. Using DWpose (Yang et al. 2023), we perform pose estimation to obtain the content skeleton pose sequence $\boldsymbol{P}_c^m \in \mathbb{R}^{F\times J\times 2}$ and the style skeleton pose sequence $\boldsymbol{P}_s^n \in \mathbb{R}^{F\times J\times 2}$, where m and n denote the pose content of the content and style poses, $J = 20$ indicates the number of joints, and 2 represents the x, y coordinates of each joint. PersonaAnimator aims to learn personalized features while preserving the basic motion structure of the pose content m, thereby generating a personalized pose $\boldsymbol{P}_s^m \in \mathbb{R}^{F\times J\times 2}$ that combines both characteristics.

PersonaAnimator Framework Overview

Prior work in computer vision has achieved basic motion generation, but key limitations remain: (1) Pose-guided character motion transfer methods (Tan et al. 2025; Tu et al. 2025b; Li et al. 2024a) replicate motion mechanically without capturing personalized traits, severely limiting the expressiveness of digital characters. (2) Motion style transfer techniques (Kim et al. 2024, 2025). However, capable of modeling style, relies fully or partially on motion capture data, which is costly, labor-intensive, and often unavailable for historical figures or virtual characters. To address these issues, we propose a new task, Video-to-Video Motion Personalization, and introduce the PersonaAnimator framework, which requires no motion capture data and enables personalized motion transfer from unconstrained videos. This is critical for enhancing the realism of digital characters, as illustrated in Fig. 2.

We first use DWPose (Yang et al. 2023) to extract skeleton poses $\boldsymbol{P}_c^m$ and $\boldsymbol{P}_s^n$ from content video $\boldsymbol{V}_c$ and style video $\boldsymbol{V}_s$, respectively. These poses are encoded into content features $\boldsymbol{Z}_c^m$ and style features $\boldsymbol{Z}_s^n$ via dedicated encoders. Meanwhile, semantic content labels are tokenized into $\boldsymbol{Z}_{\text{sem}}$, which performs Hadamard product with $\boldsymbol{Z}_s^n$ to produce $\boldsymbol{Z}_{\text{sem,s}}^n$. This semantic-guided weighting enhances motion-relevant style features while suppressing those that are irrelevant. The appendix provides a detailed description of the semantic tokenizer.

In Section SA-PMT Module, we propose the Semantic-Aware Personalized Motion Transfer (SA-PMT) Module, extracting basic motion content and personalized characteristics under semantic guidance and then fuses them into $\boldsymbol{Z}_{pm}$ features to generate personalized poses $\boldsymbol{P}_s^m$.

In the personalized motion generation process, we first extract a content reference frame $\boldsymbol{I}^r$ from $\boldsymbol{V}_c$ and encode it using pre-trained CLIP (Radford et al. 2021) and VAE (Kingma, Welling et al. 2013) encoders to obtain appearance feature $\boldsymbol{f}_a^r$ and latent feature $\boldsymbol{f}_e^r$ respectively. These features are concatenated with personalized motion characteristics $\boldsymbol{Z}_{pm}$ and noise latent $\boldsymbol{Z}_t$ as input to the video diffusion model ϵ_θ. The model's spatial, motion, and temporal attention layers jointly capture human joint topology, motion dynamics, and inter-frame dependencies. The final personalized motion video $\boldsymbol{V}_s^m$ is generated through a VAE decoder (Kingma, Welling et al. 2013). In addition, we introduce the PMSR mechanism to enforce physically plausible personalized motion generation.

PersonaAnimator redefines motion generation by learning personalized motion features directly from videos, eliminating reliance on motion capture. Its decoupled architecture precisely preserves content motion structure while transferring personalized characteristics, establishing a novel "postural approximation-to-personalized dynamic essence" generation standard that sets a new benchmark for digital character animation.

Semantic-Aware Personalized Motion Transfer

Existing pose-guided motion transfer methods (Tan et al. 2025; Tu et al. 2025b) lack personalization, as they perform only rigid retargeting. Our Semantic-Aware Personalized Motion Transfer (SA-PMT) module learns both basic motion structures and personalized traits under semantic guidance, generating motions that preserve the advantages of both sources. Specifically, we use AdaIN (Huang and Belongie 2017) to align content and style poses statistically, enabling effective feature fusion to produce the personalized feature $\boldsymbol{Z}_{ada}$:

$$\boldsymbol{Z}_{ada} = \text{AdaIN}\left(\boldsymbol{Z}_c^m, \text{MLP}\left(\boldsymbol{Z}_{sem,s}^n\right)\right), \tag{1}$$

where $\boldsymbol{Z}_c^m$ is the content skeleton pose feature, while $\boldsymbol{Z}_{sem,s}^n$ denotes the semantic-conditioned personalized feature, with AdaIN$(\cdot)$ indicating the AdaIN structure. Subsequently, we encode $\boldsymbol{Z}_{ada}$, projecting it into a stable feature distribution space and incorporating positional encoding to preserve spatiotemporal information:

$$\boldsymbol{Z}_{pm}^c = \mathcal{E}_{SA-PMT}\left(\text{PE}\left(\boldsymbol{Z}_{ada}\right)\right), \tag{2}$$

where $\mathcal{E}_{SA-PMT}(\cdot)$ denotes the encoder of the SA-PMT Module, primarily composed of fully-connected layers, while PE$(\cdot)$ represents positional encoding. We employ a Multi-Head Attention (MHA) mechanism (Vaswani et al. 2017) to concurrently capture content-style relationships across different dimensions, yielding the final personalized motion feature vector $\boldsymbol{Z}_{pm}$, computed as:

$$\begin{aligned} \boldsymbol{Z}_{pm}^f &= \text{LN}\left(\boldsymbol{Z}_{pm}^c + \text{MHA}\left(\boldsymbol{Z}_{pm}^c\right)\right), \\ \boldsymbol{Z}_{pm} &= \text{LN}\left(\boldsymbol{Z}_{pm}^f + \mathcal{F}\left(\boldsymbol{Z}_{pm}^f\right)\right), \end{aligned} \tag{3}$$

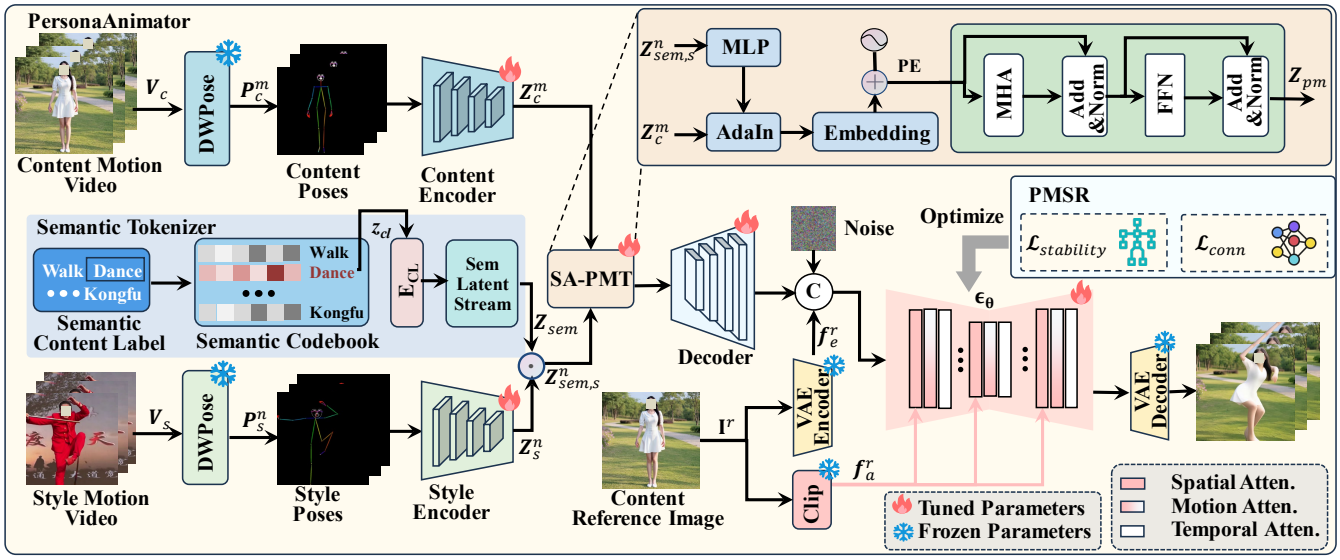


Figure 2: Architecture overview of PersonaAnimator framework. $\odot$ denotes the Hadamard product. $\oplus$ denotes the element-wise sum. C denotes the concatenation operation.

where LN$(\cdot)$ denotes Layer Normalization, $\mathcal{F}(\cdot)$. is the Feed-Forward Network, and $\boldsymbol{Z}^c_{pm}$, $\boldsymbol{Z}^f_{pm}$ are the module's personalized motion latent codes. The SA-PMT Module fuses features under semantic guidance, preserving motion content while learning subject-specific characteristics to generate personalized motions, thereby overcoming the limitations of current methods in rigid retargeting.

Physics-aware Motion Style Regularization

Current motion style transfer methods often struggle to simultaneously preserve physical plausibility and structural topology, leading to motion distortion. Prior works (Kim et al. 2024; Song et al. 2024) mitigate this issue to some extent by introducing velocity and foot contact constraints. However, they lack explicit physical constraints on the full-body skeleton and ignore repulsive supervision between non-connected joints, resulting in physically implausible motions. To address this, we propose the Physics-aware Motion Style Regularization (PMSR) mechanism, which enforces physically plausible motion generation through dynamic bone stability and body connectivity constraints.

Dynamic Bone Stability Constraints. We penalize the second-order derivative (acceleration) of bone length variations to prevent unnatural stretching and maintain dynamic stability. Here, $\boldsymbol{p}_{f,j} \in \mathbb{R}^2$ denotes the 2D coordinates of joint j in personalized pose $\boldsymbol{P}^m_s \in \mathbb{R}^{F\times J\times 2}$ at frame f. Let $\mathcal{B} = \{(i_k, j_k)\}_{k=1}^N$ represent the bone set, where (i_k, j_k) indicates the indices of two connected joints for the k-th bone, and N is the total bone count. For any given bone $(i,j) \in \mathcal{B}$, its length at frame f is defined as:

$$\boldsymbol{l}_{f,(i,j)} = \|\boldsymbol{p}_{f,i} - \boldsymbol{p}_{f,j}\|_2 . \tag{4}$$

Here, $\|\cdot\|_2$ is the Euclidean norm. We penalize the temporal second-order difference of bone length sequence $\boldsymbol{l}_{f,(i,j)}$, representing length variation acceleration. For frame $\boldsymbol{f} \in [2, \boldsymbol{F}-1]$, the second-order difference of bone (i,j)'s length is approximated as:

$$\Delta^2 \boldsymbol{l}_{f,(i,j)} = \boldsymbol{l}_{f+1,(i,j)} - 2\boldsymbol{l}_{f,(i,j)} + \boldsymbol{l}_{f-1,(i,j)} \tag{5}$$

Finally, we implement the dynamic bone stability constraint using the L2 norm, formulated as:

$$\mathcal{L}_{\text{stability}} = \frac{1}{\boldsymbol{F}-2} \sum_{f=2}^{F-1} \left\| \Delta^2 l_{f,(i,j)} \right\|_2^2 . \tag{6}$$

Body Connectivity Constraints. To preserve correct joint connectivity, we propose body connectivity constraints based on the adjacency matrix, which enhances the cohesion of connected joints while suppressing false proximity between unconnected joints. We define an adjacency matrix $\boldsymbol{A} \in \{0,1\}^{J\times J}$ where $\boldsymbol{A}_{ij} = 1$ indicates whether joint i and j should be connected. For frame f, we construct a Euclidean distance matrix $\boldsymbol{D}^{(f)} \in \mathbb{R}^{J\times J}$, $\boldsymbol{D}^{(f)}(i,j) \leftarrow \boldsymbol{l}_{f,(i,j)}$. To strengthen joint aggregation, we minimize distances between connected joints as follows:

$$\mathcal{L}^+_{\text{conn}} = \frac{1}{\|\boldsymbol{A}\|_0 \times \boldsymbol{F}} \sum_{\text{f}=1}^{\boldsymbol{F}} \sum_{\text{i,j}} \boldsymbol{A}_{\text{ij}} \cdot \boldsymbol{D}^{(\text{f})}(\text{i,j}). \tag{7}$$

Meanwhile, to prevent unconnected joints from being too close to each other, we enforce a minimum distance constraint between them:

$$\mathcal{L}^-_{\text{conn}} = \frac{1}{\|\boldsymbol{A}\|_0 \times F} \sum_{f=1}^{F} \sum_{i,j} (1 - \boldsymbol{A}_{ij}) \cdot \max\left(0, \delta - \boldsymbol{D}^{(f)}(i,j)\right) . \tag{8}$$

The complete body connectivity constraints, with $\delta = 0.1$ as the minimum distance threshold, are formulated as:

$$\mathcal{L}_{\text{conn}} = \mathcal{L}^+_{\text{conn}} + \mathcal{L}^-_{\text{conn}} . \tag{9}$$

Optimization Objectives. In addition, we adopt the $\mathcal{L}_{sc}$ proposed by Aberman *et al.* (Aberman et al. 2020) to enforce style consistency, and the mean squared error (MSE) loss $\mathcal{L}_d$ used in prior works (Tan et al. 2025) for appearance alignment and motion modeling. Further details are provided in the appendix. The final training loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{stability}} + \mathcal{L}_{\text{conn}} + \mathcal{L}_{sc} + \mathcal{L}_d. \quad (10)$$

PersonaVid Dataset

Existing datasets for the Video-to-Video Motion Personalization task exhibit notable limitations. Current motion capture datasets (Aberman et al. 2020; Mason, Starke, and Komura 2022; Xia et al. 2015) can train motion style transfer models but rely on professional equipment and controlled environments, rendering data collection costly and impractical for deceased individuals or fictional characters. Pose-guided character motion transfer datasets (Jafarian and Park 2021; Zablotskaia et al. 2019) contain numerous videos but often lack precise style and content labels, which limits their direct applicability. Many videos also present low-quality issues, such as ambiguous motion styles and severe occlusions, further constraining their utility.

To address this, we propose PersonaVid, the first video dataset for personalized motion learning. Constructed by collecting and annotating internet-sourced videos, it comprises 20 motion categories (*e.g.*, walking, running) and 120 distinct styles. It covers a diverse range of performers, including ordinary individuals, athletes, and animated characters, thereby enhancing diversity in both individual variation and style representation. Table 1 compares PersonaVid with existing datasets, while Figure 3 shows samples.

Current motion style transfer datasets (Xia et al. 2015; Aberman et al. 2020) use emotion categories (*e.g.*, angry, depressed) as style labels, which misses individual specificity and lacks person-centric annotations. PersonaVid addresses this with a novel "one-person-one-style" paradigm, treating each individual as a unique style category to better capture subject-specific motion characteristics. Videos are annotated as "Content_Style_Number," where "Content" denotes the basic motion (*e.g.*, walking, running), and "Number" represents the video serial number. The "Style" is named in two ways: (1) If the video features famous figures, "Style" is abbreviated as their name (*e.g.*, "Walk_Trump_05" for the fifth video of Trump walking). (2) For anonymous ordinary individuals, "Style" is denoted by the motion content abbreviation followed by a number (*e.g.*, "Dance_D10_01" where "D10" indicates the style of the tenth anonymous dancer, or "Skating_Sk3_01" where "Sk3" represents the style of the third anonymous skater). The complete dataset taxonomy is detailed in the supplementary material.

Experiments

Implementation

Implementation Details. We train and evaluate our model on the PersonaVid dataset. Training is conducted on an NVIDIA A800 using the RMSprop optimizer (Tieleman 2012) (lr=le-4).

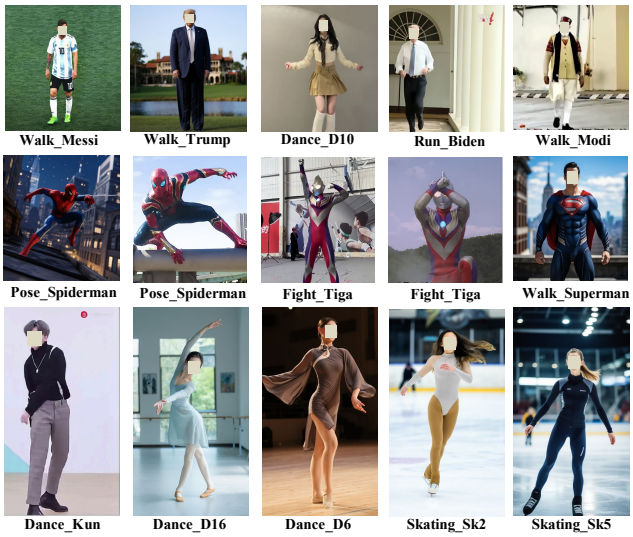


Figure 3: Examples of the PersonaVid dataset.

Dataset	Styles	Contents	MP4	Actor	Frames Number	PM
Xia (Xia et al. 2015)	8	6	✗	-	572	✗
BFA (Aberman et al. 2020)	16	9	✗	1	32	✗
BN-1 (Kobayashi et al. 2023)	15	17	✗	2	175	✗
BN-2 (Kobayashi et al. 2023)	7	10	✗	2	2902	✗
100Style (Mason, Starke, and Komura 2022)	100	8	✗	1	810	✗
PerMo (Kim et al. 2025)	34	10	✗	5	6610	✗
Tiktok (Jafarian and Park 2021)	0	0	✓	-	92961	✗
Fashion (Zablotskaia et al. 2019)	0	0	✓	-	268474	✗
PersonaVid (Ours)	**120**	**20**	✓	**120**	18867	✓

Table 1: Comparison with existing datasets. "PM" denotes whether the dataset contains personalized motion, while "-" indicates no information.

Baseline Models. As no prior work addresses the Video-to-Video Motion Personalization task, we compare our PersonaAnimator framework with SOTA methods (Hu 2024; Tu et al. 2025b; Tan et al. 2025) in pose-guided motion transfer to show its superiority. However, these methods cannot learn personalized features from video motion. To ensure a fair comparison, we modify the three baseline frameworks as follows: We first equip them with the same content encoder, style encoder, and semantic tokenizer as used in PersonaAnimator, enabling them to take content and style videos as input. Next, we incorporate the Dual Interactive Flow Fusion (DIFF) module from FineStyle (Song et al. 2023) to allow the baselines to learn personalized features. The DIFF module has a similar parameter size to our SA-PMT module. Finally, the baselines utilize the DIFF-generated personalized motion for transfer, and we compare the results with those obtained using our method. Detailed descriptions of the evaluation metrics and baseline models are provided in the appendix.

Qualitative Evaluation

We qualitatively compare PersonaAnimator with SOTA pose-guided motion transfer methods on style expressiveness and motion realism using the PersonaVid dataset. The model should preserve the basic motion content while learn-

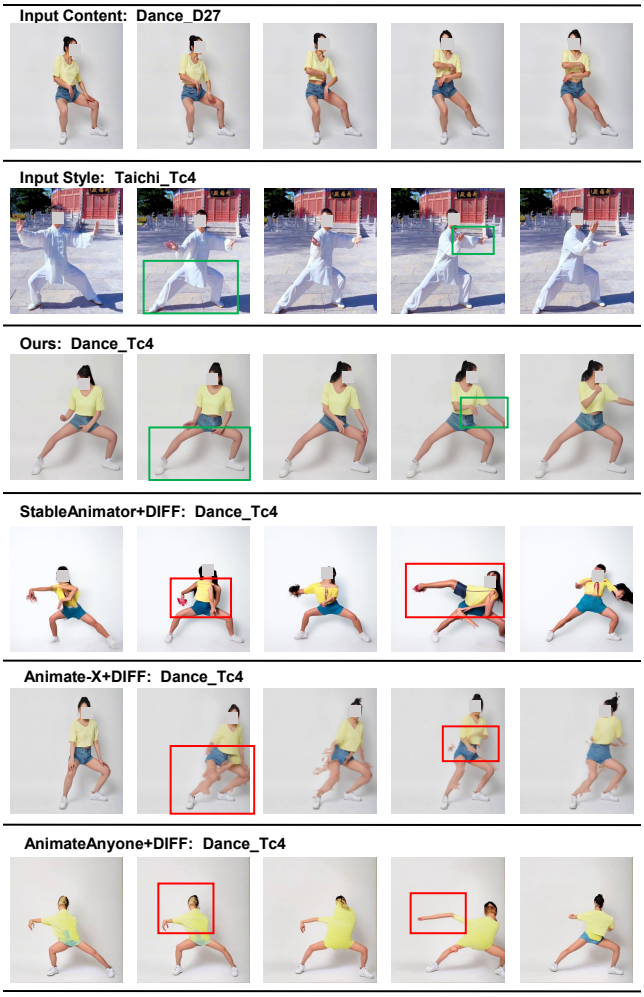


Figure 4: Qualitative comparison. Our framework outperforms SOTA methods, with green/red boxes marking learned/unlearned motion features, respectively. The results demonstrate the superior performance of our framework.



Figure 5: Generalization evaluation. We test different methods on an unseen style, with green/red boxes marking learned/unlearned motion features, respectively. Our method achieves better results than the SOTA methods.

ing personalized features to generate personalized motions, combining both aspects.

Visual comparison results are shown in Figure 4, where our PersonaAnimator framework achieves superior generation quality. We use a dancing video of a female as the content motion and a Taichi video of an anonymous individual as the style motion, aiming to learn this person's personalized Taichi features. Our method captures key style traits, such as a lunge pose—where the left leg is bent and the right leg is extended—and both hands sweeping from left to right, as highlighted in the green box. Importantly, the motion is not mechanically copied but reflects a personalized adaptation. In contrast, other methods produce distorted results: StableAnimator+DIFF (Tu et al. 2025b) and Animate-X+DIFF (Tan et al. 2025) capture the lunge pose but fail to reproduce hand movements accurately. AnimateAnyone+DIFF (Hu 2024) generates a dancer facing away from the camera, showing only one hand. None of the SOTA methods successfully learn the man's personalized Taichi features. These results stem from our SA-PMT module, which semantically guides the learning of motion features, and the PMSR mechanism, which enforces physical plausibility and prevents motion distortion.

Generalizability

Our framework extracts personalized features from arbitrary motions but may encounter unseen styles in practice. This generalization capability is crucial for practical applications. To evaluate this, we select a skating video of Russian figure skater Trusova from the PersonaVid dataset, exclude it from training, and attempt to transfer her skating style to a running female subject. We apply the same procedure to SOTA methods and compare the generated results.

As shown in Figure 5, even under unseen styles, our PersonaAnimator framework effectively captures Trusova's personalized motion—crossed arms gradually

opening—without mechanically copying the input (green box).In contrast, StableAnimator+DIFF (Tu et al. 2025b) and AnimateAnyone+DIFF (Hu 2024) produce unnatural poses, while Animate-X+DIFF (Tan et al. 2025) fails to learn any personalized characteristics. The generalizability comparison demonstrates that PersonaAnimator achieves superior generalization under unseen styles, enabling more realistic and vivid digital avatars in real-world scenarios.

Conclusion

This paper introduces the Video-to-Video Motion Personalization task, which enables personalized motion transfer from unconstrained videos, a crucial step in building identity-consistent digital avatars in the metaverse. To support this task, we introduce the large-scale PersonaVid dataset, which contains diverse motion content and style categories, along with the PersonaAnimator framework that achieves content-style decoupling and fusion through its SA-PMT module. Additionally, we design the PMSR mechanism to ensure physically plausible motion generation. Experimental results demonstrate that PersonaAnimator significantly outperforms existing methods, achieving SOTA performance in both qualitative and quantitative comparisons. The framework shows great potential for various applications, including virtual idol animation, film production, and game character animation generation.

References

Abdul-Massih, M.; Yoo, I.; and Benes, B. 2017. Motion style retargeting to characters with different morphologies. In *Computer Graphics Forum*, volume 36, 86–99. Wiley Online Library.

Aberman, K.; Weng, Y.; Lischinski, D.; Cohen-Or, D.; and Chen, B. 2020. Unpaired motion style transfer from video to animation. *ACM Transactions on Graphics (TOG)*, 39(4): 64–1.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5933–5942.

Fei, H.; Wu, S.; Ji, W.; Zhang, H.; and Chua, T.-S. 2023. Empowering dynamics-aware text-to-video diffusion with large language models. *arXiv preprint arXiv:2308.13812*, 5(6).

Fei, H.; Wu, S.; Zhang, H.; Chua, T.-S.; and Yan, S. 2024. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. *arXiv preprint arXiv:2412.19806*.

Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; and Salimans, T. 2022a. Imagen Video: High Definition Video Generation with Diffusion Models. arXiv:2210.02303.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8153–8163.

Huang, X.; and Belongie, S. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, 1501–1510.

Jafarian, Y.; and Park, H. S. 2021. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12753–12762.

Kim, B.; Jeong, H. I.; Sung, J.; Cheng, Y.; Lee, J.; Chang, J. Y.; Choi, S.-I.; Choi, Y.; Shin, S.; Kim, J.; et al. 2025. PersonaBooth: Personalized Text-to-Motion Generation. *arXiv preprint arXiv:2503.07390*.

Kim, B.; Kim, J.; Chang, H. J.; and Choi, J. Y. 2024. MoST: Motion Style Transformer between Diverse Action Contents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1705–1714.

Kingma, D. P.; Welling, M.; et al. 2013. Auto-encoding variational bayes.

Kobayashi, M.; Liao, C.-C.; Inoue, K.; Yojima, S.; and Takahashi, M. 2023. Motion capture dataset for practical use of AI-based motion editing and stylization. *arXiv preprint arXiv:2306.08861*.

Li, H.; Li, Y.; Yang, Y.; Cao, J.; Zhu, Z.; Cheng, X.; and Chen, L. 2024a. DisPose: Disentangling Pose Guidance for Controllable Human Image Animation. *arXiv preprint arXiv:2412.09349*.

Li, Z.; He, Y.; Zhong, L.; Shen, W.; Zuo, Q.; Qiu, L.; Dong, Z.; Yang, L. T.; and Yuan, W. 2024b. MulSMo: Multimodal Stylized Motion Generation by Bidirectional Control Flow. *arXiv preprint arXiv:2412.09901*.

Lin, H.; Zala, A.; Cho, J.; and Bansal, M. 2023. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Lu, H.; Yang, G.; Fei, N.; Huo, Y.; Lu, Z.; Luo, P.; and Ding, M. 2023. Vdt: General-purpose video diffusion transformers via mask modeling. *arXiv preprint arXiv:2305.13311*.

Mason, I.; Starke, S.; and Komura, T. 2022. Real-time style modelling of human locomotion via feature-wise transformations and local motion phases. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 5(1): 1–18.

Mei, K.; and Patel, V. 2023. Vidm: Video implicit diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, 9117–9125.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Ribet, S.; Wannous, H.; and Vandeborre, J.-P. 2019. Survey on style in 3d human body motion: Taxonomy, data, recognition and its applications. *IEEE Transactions on Affective Computing*, 12(4): 928–948.

Shen, C.; Yu, S.; Wang, J.; Huang, G. Q.; and Wang, L. 2024. A Comprehensive Survey on Deep Gait Recognition: Algorithms, Datasets, and Challenges. *IEEE Transactions on Biometrics, Behavior, and Identity Science*.

Song, W.; Jin, X.; Li, S.; Chen, C.; Hao, A.; and Hou, X. 2023. Finestyle: Semantic-aware fine-grained motion style transfer with dual interactive-flow fusion. *IEEE Transactions on Visualization and Computer Graphics*, 29(11): 4361–4371.

Song, W.; Jin, X.; Li, S.; Chen, C.; Hao, A.; Hou, X.; Li, N.; and Qin, H. 2024. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 821–830.

Tan, S.; Gong, B.; Wang, X.; Zhang, S.; Zheng, D.; Zheng, R.; Zheng, K.; Chen, J.; and Yang, M. 2025. Animate-x: Universal character image animation with enhanced motion representation.

Tieleman, T. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2): 26.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7882–7891.

Tu, S.; Dai, Q.; Wu, Z.; Cheng, Z.-Q.; Hu, H.; and Jiang, Y.-G. 2023. Implicit temporal modeling with learnable alignment for video recognition. In *Proceedings of the ieee/cvf international conference on computer vision*, 19936–19947.

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*.

Tu, S.; Pan, Y.; Huang, Y.; Han, X.; Xing, Z.; Dai, Q.; Luo, C.; Wu, Z.; and Jiang, Y.-G. 2025a. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*.

Tu, S.; Xing, Z.; Han, X.; Cheng, Z.-Q.; Dai, Q.; Luo, C.; and Wu, Z. 2025b. Stableanimator: High-quality identity-preserving human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21096–21106.

Tu, S.; Xing, Z.; Han, X.; Cheng, Z.-Q.; Dai, Q.; Luo, C.; Wu, Z.; and Jiang, Y.-G. 2025c. Stableanimator++: Overcoming pose misalignment and face distortion for human image animation. *arXiv preprint arXiv:2507.15064*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, B.; Wang, X.; Ni, C.; Zhao, G.; Yang, Z.; Zhu, Z.; Zhang, M.; Zhou, Y.; Chen, X.; Huang, G.; et al. 2025a. HumanDreamer: Generating Controllable Human-Motion Videos via Decoupled Generation. *arXiv preprint arXiv:2503.24026*.

Wang, X.; Zhang, S.; Gao, C.; Wang, J.; Zhou, X.; Zhang, Y.; Yan, L.; and Sang, N. 2025b. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*.

Xia, S.; Wang, C.; Chai, J.; and Hodgins, J. 2015. Realtime style transfer for unlabeled heterogeneous human motion. *ACM Transactions on Graphics (TOG)*, 34(4): 1–10.

Xie, Y.; Jampani, V.; Zhong, L.; Sun, D.; and Jiang, H. 2024. OmniControl: Control Any Joint at Any Time for Human Motion Generation. In *The Twelfth International Conference on Learning Representations*.

Xing, Z.; Feng, Q.; Chen, H.; Dai, Q.; Hu, H.; Xu, H.; Wu, Z.; and Jiang, Y.-G. 2024. A survey on video diffusion models. *ACM Computing Surveys*, 57(2): 1–42.

Xu, Z.; Zhang, J.; Liew, J. H.; Yan, H.; Liu, J.-W.; Zhang, C.; Feng, J.; and Shou, M. Z. 2024. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1481–1490.

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.

Zablotskaia, P.; Siarohin, A.; Zhao, B.; and Sigal, L. 2019. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*.

Zhao, K.; Li, G.; and Tang, S. 2024. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. *arXiv preprint arXiv:2410.05260*.

Zhong, L.; Xie, Y.; Jampani, V.; Sun, D.; and Jiang, H. 2024. SMooDi: Stylized Motion Diffusion Model. In *ECCV*.

Zhu, S.; Chen, J. L.; Dai, Z.; Dong, Z.; Xu, Y.; Cao, X.; Yao, Y.; Zhu, H.; and Zhu, S. 2024. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, 145–162. Springer.

Zhu, W.; Ma, X.; Ro, D.; Ci, H.; Zhang, J.; Shi, J.; Gao, F.; Tian, Q.; and Wang, Y. 2023. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4): 2430–2449.