Reflective Language-Orchestrated Motion Personalization in Semantically Aligned Arbitrary Scenes

Anonymous ACL submission

Abstract

Natural language prompts provide a common semantic interface for character motion generation, and motion style plays a central role in turning linguistic intent into visible motion while giving digital characters distinct identities. Existing methods learn character motion traits from videos, but they still have two main limitations: (1) attention-based global spatiotemporal aggregation often smooths high-frequency micro-variations across joints, which weakens fine-grained style expression; and (2) generated videos are usually tied to fixed scenes, reducing diversity and flexibility in practical creation. To address these limitations, this paper presents PersonaScape, a language-driven Motion Personalization framework for arbitrary scenes. PersonaScape introduces an Asymmetric Spatio-Temporal Decoupling (ASTD) module that separates motion into a global structural stream and a micro-variation detail stream. This design preserves the overall motion content while strengthening personalized style cues. The framework also includes a Multi-agent Collaboration-based Arbitrary Scene Alignment (MCASA) module, which aligns and verifies user prompts and motion semantics through progressive semantic planning and reverse-verifier-guided refinement, producing scene images that remain consistent with the semantics. Extensive experiments show that PersonaScape achieves SOTA performance in both fixed-scene and arbitrary-scene settings.

1 Introduction

Natural-language prompts (Han et al., 2025; Wu et al., 2025; Tan et al., 2026) are becoming the most general creative interface for multimodal content generation, allowing users to specify character attributes, motion intent, and scene settings in text. For character motion generation, this language-driven paradigm requires models to translate textual constraints on identity, behavior, and context into perceptible motion expression. Motion style (Qian et al., 2025b; Kim et al., 2025; Qian et al., 2025a) serves as the key semantic dimension linking textual intent to visual motion: it determines how a character performs the same motion and encodes character-specific motion habits, emotional states, and narrative identity. However, existing pose-guided character motion transfer methods (Tan et al., 2025; Tu et al., 2025; Li et al., 2024a; Wang et al., 2025a; Jiang et al., 2025) mostly replicate reference motions mechanically, making it difficult to generate individualized and expressive character motions or to respond to fine-grained textual requirements on behavior and scene semantics.

Recently, PersonaAnimator (Qian et al., 2025a) makes progress in this direction by enabling video-based personalized motion generation. However, it mainly relies on attention mechanisms for global spatiotemporal feature aggregation, which is essentially a normalized weighted average and may attenuate high-frequency micro-variation features across joints during long-range spatiotemporal modeling, leading to insufficient personalized style expression. Moreover, PersonaAnimator (Qian et al., 2025a) relies entirely on video input and lacks explicit textual control, making it difficult to meet diverse demands in open-ended creation.

On the other hand, character motion in practical creation never exists in isolation; its final effect is always tightly coupled with scene settings, visual style, and narrative context (Yu et al., 2025; Cheng et al., 2025; Agata and Igarashi, 2025). The same personalized motion may need to be placed in different background spaces, character appearances, and visual atmospheres under different textual descriptions. Therefore, a practical motion personalization framework should not only learn personalized motion style but also support open-ended generation for arbitrary scenes. It should adaptively generate plausible, diverse, and semantically consistent scenes according to user text descriptions

Figure 1: Semantically aligned arbitrary-scene motion personalization. We support two task settings. Setting 1 is motion personalization in fixed scenes. PersonaScape preserves the macroscopic motion structure of the content motion (*e.g.*, a girl walking) while learning personalized motion characteristics from the style motion video (*e.g.*, a skater's shoulder sway), generating personalized motion (*e.g.*, jumping with shoulder sway). Setting 2 is motion personalization in arbitrary scenes. After motion personalization, the generated motion is transferred to arbitrary scenes aligned with the user prompt and motion semantics (*e.g.*, a park walkway). Circles highlight the learned personalized characteristics.

and motion semantics, thereby unifying the expression of motion, character, and scene. However, existing methods (Qian et al., 2025a; Kim et al., 2025) mostly assume fixed scenes, making it difficult to satisfy real creative workflows that require scene variability, semantic adaptability, and visual richness. Although large-model-based domain-specific fine-tuning methods (OpenAI, 2025; Kling AI, 2025; Wang et al., 2025b; Decatur et al., 2025) can generate stylized scenes to some extent, they often depend on large-scale domain data and high computational cost. More importantly, most existing large models (OpenAI, 2025; Gemini Team and Google, 2025; Cao et al., 2025) adopt a one-way and static inference paradigm, typically converting user text prompts into generation conditions in a single pass. They lack intermediate verification and feedback correction mechanisms for consistency among motion semantics, user intent, and scene expression, and thus struggle to stably generate scene results that satisfy both motion and textual constraints.

To address these issues, this paper proposes PersonaScape, a language-driven personalized motion generation framework for arbitrary scenes. To mitigate the over-smoothing of fine-grained motion style, we propose the Asymmetric Spatio-Temporal Decoupling (ASTD) module, which explicitly decouples motion into a global structural motion flow and a micro-variation detail motion flow. Through asymmetric spatiotemporal modeling, ASTD preserves macroscopic content structure while enhancing personalized style features. Furthermore, to overcome fixed-scene constraints and achieve semantically consistent generation for arbitrary scenes, we propose the Multi-agent Collaboration-based Arbitrary Scene Alignment (MCASA) module. MCASA takes the user text prompt as the core input and adopts a tri-loop reflective multi-agent collaboration mechanism. It combines progressive semantic planning with reverse-verifier-guided refinement to conduct multi-round planning, verification, and iterative optimization of scene designs. It finally generates scene images that satisfy both user text prompts and motion semantics, thereby supporting motion personalization in open scenes.

Based on the above design, PersonaScape supports two complementary task settings, as shown in Fig. 1. The first is motion personalization in fixed scenes, where the model preserves the

macroscopic content structure from the content motion video (*e.g.*, a girl jumping) while learning and transferring personalized motion characteristics from the style motion (*e.g.*, Hatsune Miku's hands on hips), generating personalized motions with the target style (*e.g.*, jumping with hands on hips). The second is motion personalization in arbitrary scenes, where PersonaScape generates semantically aligned scenes (*e.g.*, a street in an ancient town) and personalized motions (*e.g.*, jumping with hands on hips) according to user text prompts. Extensive experiments demonstrate that PersonaScape achieves state-of-the-art (SOTA) performance in both settings.

The main contributions are as follows:

- We propose PersonaScape, a language-driven framework for arbitrary-scene motion personalization. It includes the ASTD module, which captures and preserves personalized motion characteristics in videos through asymmetric spatiotemporal decoupling.
- We propose the MCASA module, which generates scenes consistent with both prompt semantics and motion semantics via a tri-loop reflective collaboration mechanism.
- Extensive experimental results demonstrate the effectiveness and superiority of the proposed method.

2 Related Work

2.1 Pose-guided Character Motion Transfer

The field of pose-guided character motion transfer, which involves the generation of character videos from driving pose signals, has undergone a significant paradigm shift. Research has transitioned from traditional, explicit methodologies toward contemporary frameworks based on diffusion (Tu et al., 2025; Tan et al., 2025; Li et al., 2024a; Tu et al., 2024; Wang et al., 2025a). Early approaches, which relied heavily on optical flow, Thin-Plate Spline (TPS) transformations, and GAN-based inpainting, were frequently susceptible to visual artifacts such as texture shearing and temporal instability (Chan et al., 2019; Zhu et al., 2023). In contrast, modern diffusion models achieve substantially higher fidelity in both motion accuracy and the preservation of fine details. Recent advancements underscore the effectiveness of these frameworks. For instance, StableAnimator (Tu et al., 2025) mitigates identity drift through the employment of specialized facial-preservation techniques. To enhance alignment precision, Champ (Zhu et al., 2024) integrates SMPL models with diverse 3D multimodal signals, such as depth, normal, and semantic maps. MagicAnimate (Xu et al., 2024) combines DensePose with temporal attention mechanisms, facilitating a robust decoupling of motion and appearance. Furthermore, Animate Anyone (Hu, 2024) utilizes a ReferenceNet and pose guidance to enable stable animation from a single source image. Despite this progress, a critical limitation remains: current methods are predominantly restricted to the mechanical replication of reference motions and fail to encapsulate personalized or stylistic characteristics. This fundamental constraint results in animations that often lack expressiveness, which limits their utility in practical applications such as film production and immersive digital environments where nuanced character performance is essential.

2.2 Motion Style Transfer

Motion Style Transfer (MST) (Zhong et al., 2024, 2025; Li et al., 2024c; Xie et al., 2024) aims to integrate a target style into a source motion while preserving the integrity of its underlying content. Recent research in this domain has introduced a variety of increasingly effective methodologies. For instance, MCM-LDM (Song et al., 2024) facilitates arbitrary transfer by disentangling trajectory, content, and style features, while MOST (Kim et al., 2024) utilizes part-aware modulators to enable cross-content transfer. Furthermore, Aberman et al. (Aberman et al., 2020) introduced a framework for unpaired style extraction from video to 3D motion. To mitigate the information loss and feature misalignment often associated with two-stream designs, UMSD (Qian et al., 2025b) adopts a unified one-stream architecture that jointly models content and style. Subsequent advancements include Finestyle (Song et al., 2023), which incorporates semantic-aware DIFF modules, and PersonaBooth (Kim et al., 2025), which addresses personalized generation guided by text. However, a significant portion of these approaches remains heavily dependent on motion-capture data or SMPL models (Loper et al., 2023). This reliance restricts the applicability of such methods to unconventional subjects and highlights the necessity for more versatile solutions.

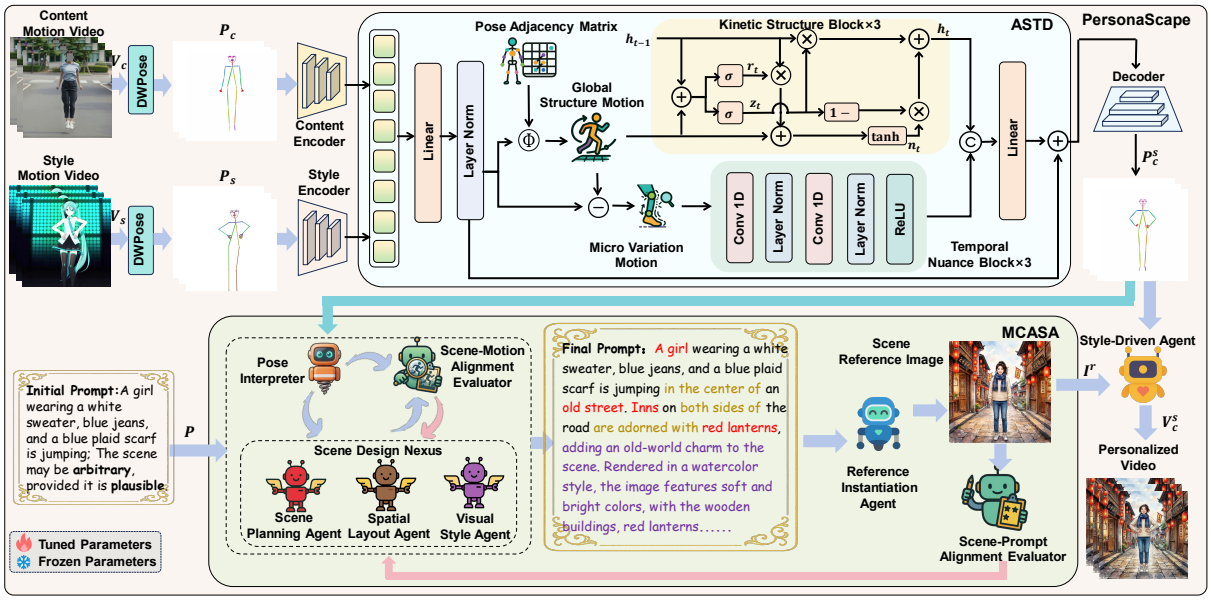

Figure 2: Overview of PersonaScape. Content and style videos are encoded and processed by ASTD to decouple global motion structure and micro-variation style, producing a personalized pose. Conditioned on this pose and the user prompt, MCASA performs multi-agent scene planning and alignment to generate a reference scene image, which is finally combined with the personalized pose by the style-driven agent to synthesize the personalized video. $\oplus$ and $\ominus$ denote the element-wise sum and element-wise minus. $\otimes$ denotes the element-wise multiplication. C denotes the concatenation operation. Φ is the neighborhood aggregation operation.

3 Methodology

3.1 Framework Overview

Personalized motion video generation requires preserving content motion while capturing individual style. Existing methods (Kim et al., 2024; Qian et al., 2025b,a) often rely on attention-based global spatiotemporal aggregation, which weakens high-frequency joint variations, and most are restricted to static scenes (Qian et al., 2025a; Tan et al., 2025; Tu et al., 2025). Even large-scale models (OpenAI, 2025; Cao et al., 2025; Kling AI, 2025) struggle to reconcile motion semantics, user intent, and scene expression. To address these issues, PersonaScape introduces ASTD for asymmetric spatiotemporal decoupling and MCASA for multi-agent motion-scene semantic alignment, as shown in Figure 2.

For motion style transfer, DWPose (Yang et al., 2023) extracts the content pose $\boldsymbol{P}c \in \boldsymbol{R}^{F\times J\times 2}$ from $\boldsymbol{V}c \in \boldsymbol{R}^{F\times 3\times H\times W}$ and the style pose $\boldsymbol{P}s \in \boldsymbol{R}^{F\times J\times 2}$ from $\boldsymbol{V}s \in \boldsymbol{R}^{F\times 3\times H\times W}$, where c and s denote content and style, F is the frame number, H and W are frame size, $J = 20$ is the joint number, and 2 denotes (x, y) coordinates. A content encoder and a style encoder feed features into ASTD (Sec. 3.2), which preserves global structure while capturing personalized details, producing $\boldsymbol{P}c^s \in \boldsymbol{R}^{F\times J\times 2}$. A style-driven agent then animates $\boldsymbol{I}^r$ with $\boldsymbol{P}c^s$ to generate $\boldsymbol{V}_c^s \in \boldsymbol{R}^{F\times 3\times H\times W}$.

For scene generation, the user prompt and personalized pose are fed into MCASA. A pose interpreter converts skeletons into textual priors, after which the Scene Design Nexus (SDN) coordinates a scene planning agent, spatial layout agent, and visual style agent to refine the design. As shown in Figure 5, scene-motion and scene-prompt evaluators accept the design only when both consistency constraints are satisfied; otherwise, feedback guides further optimization. This closed loop resolves semantic conflicts among motion, user intent, and scene expression before video generation.

3.2 Asymmetric Spatio-Temporal Decoupling

In stylized motion, content mainly encodes low-frequency macroscopic structure, whereas style is reflected by high-frequency inter-joint fluctuations, such as ankle-knee dragging in a "lazy" walk. Existing methods (Kim et al., 2024; Song et al., 2023; Aberman et al., 2020; Qian et al., 2025b; Kim et al., 2025) typically use GCNs (Jiang et al., 2019; Jang et al., 2022) or Transformers (Vaswani et al., 2017; Li et al., 2024b; Wang et al., 2024; Dosovitskiy, 2020) to fuse motions through joint dependencies and temporal attention, but neighborhood aggregation and softmax weighting tend to smooth fine-grained stylistic variations.

Motivated by this observation, we propose the Asymmetric Spatio-Temporal Decoupling (ASTD) module. As shown in Figure 2, ASTD explicitly decouples and refines encoded content/style features $\boldsymbol{Z}c$ and $\boldsymbol{Z}s$ to obtain the motion feature vector $\boldsymbol{Z}_m$:

$$\boldsymbol{Z}_m = \mathrm{LN}(\mathcal{F}([\boldsymbol{Z}_c; \boldsymbol{Z}_s])), \tag{1}$$

where [;] denotes concatenation, $\mathcal{F}(.)$ represents a Linear layer, and LN(.) represents Layer Norm. ASTD then decomposes $\boldsymbol{Z}_m$ into a global-structure stream $\boldsymbol{M}_g$ and a micro-variation stream $\boldsymbol{M}_{m1}$:

$$\begin{aligned} \boldsymbol{M}_g &= \Phi(\boldsymbol{Z}_m, \boldsymbol{M}), \\ \boldsymbol{M}_{m1} &= \boldsymbol{Z}_m \ominus \boldsymbol{M}_g, \end{aligned} \tag{2}$$

where $\boldsymbol{M}$ is the pose adjacency matrix, Φ denotes neighborhood aggregation, and $\ominus$ represents element-wise minus. This separation enables low-frequency content and high-frequency style to be modeled independently. In the global-structure stream, the kinetic structure block captures temporal macro-structure:

$$\begin{aligned} \boldsymbol{r}_t &= \sigma(\mathrm{GCN}_r(\boldsymbol{M}_g) + \mathrm{GCN}_r(\boldsymbol{h}_{t-1})), \\ \boldsymbol{z}_t &= \sigma(\mathrm{GCN}_z(\boldsymbol{M}_g) + \mathrm{GCN}_z(\boldsymbol{h}_{t-1})), \\ \boldsymbol{n}_t &= \tanh(\mathrm{GCN}_n(\boldsymbol{M}_g) + \boldsymbol{r}_t \odot \mathrm{GCN}_n(\boldsymbol{h}_{t-1})), \\ \boldsymbol{h}_t &= (1 - \boldsymbol{z}_t) \odot \boldsymbol{n}_t + \boldsymbol{z}_t \odot \boldsymbol{h}_{t-1}, \end{aligned} \tag{3}$$

where $\boldsymbol{h}_{t-1}$ is the hidden state from the previous time step, $\mathrm{GCN}_i(.)$ with $i \in \{r, z, n\}$ is a GCN layer, σ is the Sigmoid activation function, and $\odot$ is the Hadamard product. For $\boldsymbol{M}_{m1}$, the temporal nuance block uses local temporal receptive fields of 1D convolutions to enhance micro-variation details and produce $\boldsymbol{M}_{m2}$:

$$\boldsymbol{M}_{m2} = \theta(\mathrm{LN}(\phi(\mathrm{LN}(\phi(\boldsymbol{M}_{m1}))))), \tag{4}$$

where θ represents the ReLU layer, and ϕ represents a 1D convolutional layer. The two streams are fused into the personalized motion feature $\boldsymbol{Z}_{cs}$, which is decoded into the personalized skeleton pose $\boldsymbol{P}_{cs}$:

$$\boldsymbol{Z}_{cs} = \boldsymbol{Z}_m + \mathrm{Lin}\left([\boldsymbol{h}_t; \boldsymbol{M}_{m2}]\right). \tag{5}$$

By applying asymmetric temporal modeling to the two streams, ASTD preserves macro-motion content while strengthening granular style features. The training of the ASTD module is provided in the **Appendix E**.

3.3 MCASA Module

In film and game production, scenes and character modeling must adapt to narrative context, motion type, and visual style (Yu et al., 2025; Cheng et al., 2025; Agata and Igarashi, 2025). PersonaAnimator (Qian et al., 2025a) mainly assumes fixed scenes, while fine-tuning large models (OpenAI, 2025; Gemini Team and Google, 2025; Kling AI, 2025) is costly and data-intensive. Moreover, existing models (Wang et al., 2025b; Decatur et al., 2025; Cao et al., 2025) often use one-way semantic planning without verification, making it difficult to align motion semantics, user intent, and scene expression. We therefore propose the Multi-agent Collaboration-based Arbitrary Scene Alignment module (MCASA), which assigns specialized agents to generate realistic and expressive scenes consistent with both motion and text.

Given a user prompt, MCASA generates a reference scene image aligned with both prompt and motion semantics. It includes: the pose interpreter $\mathcal{PI}$, which converts the ASTD-generated personalized pose $\boldsymbol{P}_c^s$ into textual motion priors; the scene design nexus (SDN), where $\boldsymbol{A}_{SP}$ plans the theme, background, and objects, $\boldsymbol{A}_{SL}$ constrains spatial layout, and $\boldsymbol{A}_{VS}$ specifies visual style; evaluators ε_{SM} and ε_{SP} for motion and prompt consistency; a reference instantiation agent; and a style-driven agent for final video rendering. At round t, SDN produces:

$$S^{(t)} = \mathrm{SDN}(\mathcal{PI}(\boldsymbol{P}_c^s), P, \Delta_i^{(t-1)}), \tag{6}$$

where $\mathcal{PI}(\boldsymbol{P}_c^s)$ is the textual motion representation, P is the user prompt, and $\Delta_i^{(t-1)}$ denotes feedback from the previous round. Here $i \in \{SMA, SPA, SP, SL, VS\}$ indexes ε_{SM}, ε_{SP}, $\boldsymbol{A}_{SP}$, $\boldsymbol{A}_{SL}$, and $\boldsymbol{A}_{VS}$, with $\Delta_i^{(0)} = \varnothing$ for the initial round.

MCASA uses a tri-loop reflective collaboration mechanism:

Scene Design Nexus (SDN) internal loop: $\boldsymbol{A}_{SP}$, $\boldsymbol{A}_{SL}$, and $\boldsymbol{A}_{VS}$ jointly receive motion descriptions from $\mathcal{PI}$, the prompt, and feedback $\Delta_i^{(t-1)}$. At iteration t, they generate domain-specific suggestions $\Delta_{SP}^{(t)}$, $\Delta_{SL}^{(t)}$, and $\Delta_{VS}^{(t)}$, which are merged into $S^{(t)}$:

$$SDN(\cdot) = \mathrm{Merge}(\Delta_{SP}^{(t)}, \Delta_{SL}^{(t)}, \Delta_{VS}^{(t)}), \tag{7}$$

where $\mathrm{Merge}(\cdot)$ is performed by $\boldsymbol{A}_{SP}$. The loop stops when $\Delta_{SP}^{(t)} = \Delta_{SL}^{(t)} = \Delta_{VS}^{(t)} = \emptyset$, yielding $S^{(t)}$.

Scene-motion consistency loop: ε_{SM} checks whether $S^{(t)}$ matches the motion semantics $\mathcal{PI}(P_c^s)$:

$$(r_{SM}^{(t)}, \Delta_{SMA}^{(t)}) = \varepsilon_{SM}(\mathcal{PI}(P_c^s), S^{(t)}), \quad (8)$$

where $\boldsymbol{r}_{\text{SM}}^{(t)} \in \{0, 1\}$ indicates scene-motion consistency and $\boldsymbol{\Delta}_{\text{SMA}}^{(t)}$ provides refinement feedback. If $\boldsymbol{r}_{\text{SM}}^{(t)} = 0$, the feedback is returned to SDN (Eq. 6) for $\boldsymbol{S}^{(t+1)}$; otherwise, the loop terminates.

Scene-prompt consistency loop: Once motion consistency is satisfied, $\boldsymbol{A}_{RI}$ instantiates $Ir^{(t)}$ from $S^{(t)}$, and ε_{SP} evaluates its semantic alignment with the prompt P:

$$(r_{SP}^{(t)}, \Delta_{SPA}^{(t)}) = \varepsilon_{SP}(I_r^{(t)}, P), \quad (9)$$

where $\boldsymbol{r}_{SP}^{(t)} \in \{0, 1\}$ denotes scene-prompt consistency and $\boldsymbol{\Delta}_{SPA}^{(t)}$ gives image correction feedback. If $\boldsymbol{r}_{SP}^{(t)} = 0$, Eq. 6 is reused to obtain $\boldsymbol{S}^{(t+1)}$; otherwise, the loop outputs $\boldsymbol{I}^r$. The style-driven agent then generates $\boldsymbol{V}_c^s$ from I^r and $\boldsymbol{P}_c^s$.

Algorithm 1 in the **Appendix** summarizes the complete tri-loop execution. By separating planning, motion verification, and prompt verification, MCASA improves controllability: SDN proposes plausible scene hypotheses, the motion loop rejects motion-incompatible designs, and the prompt loop removes prompt-inconsistent images. Thus, the final reference scene is produced only after both semantic checks are satisfied, yielding more reliable scene control than one-way generation pipelines.

4 Experiments

4.1 Implementation

Implementation Details. All experiments are conducted on an NVIDIA A800 GPU with 80 GB memory, and the overall ASTD training process takes approximately 10 hours. ASTD is trained on the PersonaVid (Qian et al., 2025a), TikTok (Jafarian and Park, 2021), and Fashion (Zablotskaia et al., 2019) datasets for 120K iterations using RMSprop with a learning rate of 1×10^{-4} and a batch size of 8. We evaluate personalized video generation using three metric groups: FID-VID (Balaji et al., 2019) and FVD (Unterthiner et al., 2018) for video quality, CRA-V and SRA-V (Qian et al., 2025a) for motion personalization, and SMA and SPA for scene-motion and scene-prompt alignment. FID-VID and FVD measure the distribution gap between generated and real videos, while CRA-V and SRA-V evaluate content preservation and style recognition accuracy. SMA and SPA further assess whether the generated scene is compatible with the personalized motion and prompt text, as well as the coherence of the underlying logic. For these two scene metrics, ChatGPT-4o (OpenAI, 2024), ChatGPT-5 (OpenAI, 2025), and Gemini 2.5 (Gemini Team and Google, 2025) independently assign scores on a 100-point scale, and the final value is computed by averaging the three scores. Detailed prompts are provided in **Appendix J and K**.

Table 1: Quantitative comparisons on the PersonaVid dataset (Qian et al., 2025a), showing that our framework achieves SOTA results.

Methods	FID-VID↓	FVD↓	CRA-V↑	SRA-V↑	SMA↑	SPA↑
Animate-X (Tan et al., 2025)	73.85	757.31	55.32	55.43	64.23	59.62
StableAnimator (Tu et al., 2025)	80.37	821.42	23.46	12.32	61.64	52.96
VACE (Jiang et al., 2025)	47.91	632.32	57.45	38.15	54.21	41.44
PersonaAnimator (Qian et al., 2025a)	43.84	601.43	59.96	61.47	57.42	45.75
PersonaScape (Ours)	**35.21**	**531.42**	**76.31**	**84.78**	**94.32**	**91.42**

Datasets. We use the PersonaVid dataset (Qian et al., 2025a) for training and testing, which employs a "Content_Style" video annotation format. "Content" signifies the basic motion (e.g., running, dancing). "Style" naming is twofold: (1) For famous figures, it is a name abbreviation (*e.g.*, "Skate_Shcherbakova"); (2) For anonymous individuals, it is the content abbreviation plus a serial number (e.g., "Walk_W1" represents the style of walker 1). Additional dataset see the **Appendix D**.

Baseline Models. Following the protocol adopted in recent video style transfer studies (Qian et al., 2025a,b; Song et al., 2023), we conduct fair comparisons and systematic evaluations across competing methods. For both our framework and the baselines, the reference instantiation agent is implemented with Kling (Kling AI, 2025), the pose interpreter, scene design nexus, and the evaluators of scene-motion alignment and scene-prompt alignment are implemented with ChatGPT-5 (OpenAI, 2025), and the style-driven agent is implemented with Wan 2.2 (Wan et al., 2025). Detailed descriptions of the evaluation metrics and baseline models are provided in **Appendix B and C**.

4.2 Quantitative Evaluation

Table 1 reports quantitative comparisons on PersonaVid (Qian et al., 2025a). Our method outperforms state-of-the-art approaches on FID-VID

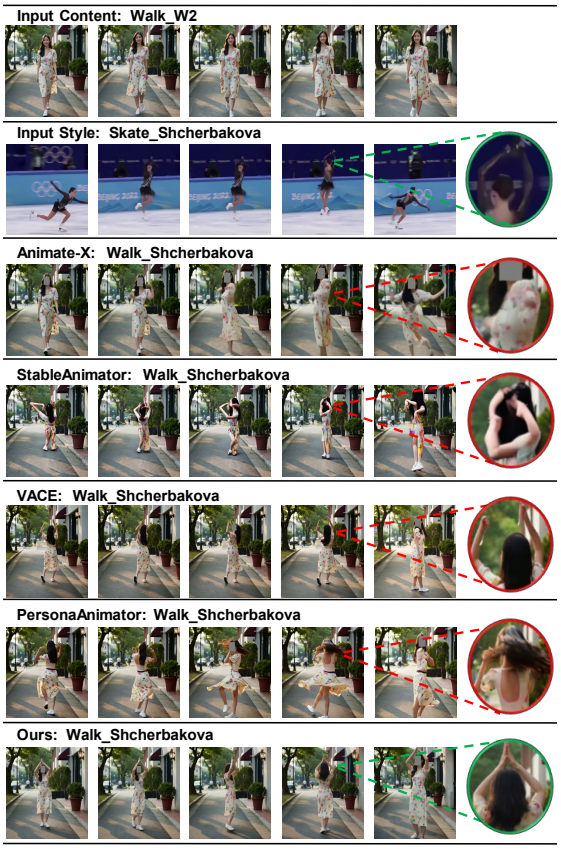

Figure 3: Qualitative comparison without scene variation. Green and red circles indicate correctly and incorrectly learned personalized features, respectively, demonstrating SOTA performance of our framework.

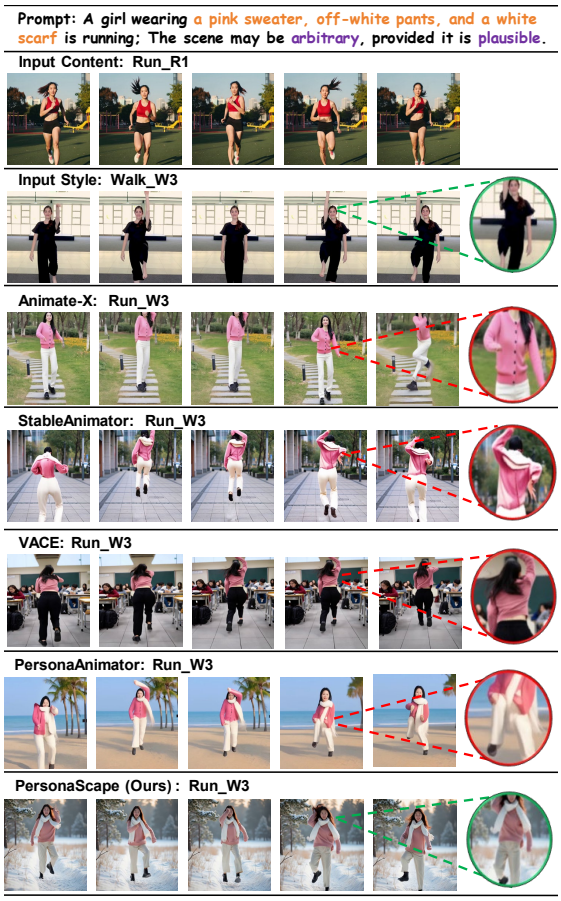

Figure 4: Qualitative comparison with scene variation. Green and red circles indicate correctly learned and incorrectly learned personalized features, respectively, demonstrating the SOTA performance of our framework.

(Balaji et al., 2019), FVD (Unterthiner et al., 2018), CRA-V, and SRA-V (Qian et al., 2025a), demonstrating strong video quality, content preservation, style recognition, and personalized motion modeling. These gains mainly come from ASTD, which disentangles motion and enhances micro-variations. For scene references, strong SMA and SPA results further show effective alignment with motion style and prompts, benefiting from MCASA's collaborative semantic planning and verifier-guided refinement. Results on TikTok (Jafarian and Park, 2021) and Fashion (Zablotskaia et al., 2019) are provided in **Appendix G**.

4.3 Qualitative Evaluation

4.3.1 Fixed-Scene Personalization

We first compare video-to-video motion personalization without scene transformation, as shown in Figure 3. Using a walking video as content motion and Anna Shcherbakova's skating video as style motion, this setting evaluates whether models can preserve the walking structure while transferring personalized skating characteristics. Our method successfully captures the key style features, including simultaneous upward hand movement and body rotation. In contrast, PersonaAnimator (Qian et al., 2025a) only captures partial rotation, VACE (Jiang et al., 2025) produces incomplete hand movement, and Animate-X (Tan et al., 2025) and StableAnimator (Tu et al., 2025) fail to transfer the target style. This advantage mainly comes from ASTD, which disentangles motion signals and enhances fine-grained high-frequency style details.

4.3.2 Arbitrary-Scene Personalization

We further evaluate motion personalization with scene transformation, as shown in Figure 4. This setting requires consistency among motion style, character appearance, and scene semantics. Our method preserves the running structure while accurately transferring the same-side arm-leg motion pattern. Other methods either fail to capture this personalized style or introduce severe artifacts,

Table 2: Quantitative ablation study of the ASTD.

Methods	FID-VID↓	FVD↓	CRA-V↑	SRA-V↑	SMA↑	SPA↑
w/o kinetic structure block	40.96	582.47	63.42	65.53	83.31	85.62
w/o temporal nuance block	37.42	552.68	68.25	62.52	86.42	89.36
PersonaScape (Full)	**35.21**	**531.42**	**76.31**	**84.78**	**94.32**	**91.42**

Table 3: Quantitative ablation study of the MCASA.

Methods	FID-VID↓	FVD↓	CRA-V↑	SRA-V↑	SMA↑	SPA↑
w/o Scene planning agent	42.58	601.25	68.21	70.34	75.16	78.42
w/o Spatial layout agent	40.47	589.43	70.13	73.22	78.64	82.37
w/o Visual style agent	43.84	605.31	67.32	68.47	74.23	77.96
w/o Pose interpreter	38.96	563.87	72.54	78.65	76.95	71.27
w/o Scene-motion alignment evaluator	36.43	545.62	75.83	82.97	72.35	90.24
w/o Scene-prompt alignment evaluator	37.85	553.48	74.58	80.25	88.21	63.97
PersonaScape (Full)	**35.21**	**531.42**	**76.31**	**84.78**	**94.32**	**91.42**

and StableAnimator even produces anatomically invalid motion. For scene coherence, existing methods often conflict with either the prompt or motion semantics, such as generating unsuitable scenes or missing required clothing elements. In contrast, our method aligns the scene, character, and motion style more effectively, benefiting from the tri-loop reflective multi-agent collaboration and iterative feedback in MCASA. Additional qualitative evaluation are provided in the **Appendix H**.

4.4 Ablation Studies

Importance of the ASTD Module. We conduct quantitative ablations on the internal components of ASTD. As shown in Table 2, removing either the kinetic structure block or the temporal nuance block degrades all metrics. The former mainly weakens content preservation, while the latter causes a larger drop in style recognition, confirming that ASTD requires both global structural modeling and fine-grained micro-variation enhancement.

Importance of the MCASA Module. We further evaluate each component of MCASA. Table shows that removing any planning agent or evaluator consistently reduces performance, especially on SMA and SPA. These results demonstrate that pose interpretation, scene planning, layout reasoning, visual style design, and verifier-guided feedback are all necessary for generating scenes aligned with both motion semantics and user prompts. Full ablation studies are provided in the **Appendix I**.

Table 4: **User study results**. Our PersonaScape outperforms SOTA methods in terms of Content Preservation (CP), Style Expressiveness (SE), Scene-Motion Coherence (SMC), and Scene-Prompt Coherence (SPC).

Methods	CP↑	SE↑	SMC↑	SPC↑
Animate-X (Tan et al., 2025)	$6.28^{\pm 0.43}$	$5.58^{\pm 0.96}$	$6.79^{\pm 0.75}$	$7.35^{\pm 0.58}$
StableAnimator (Tu et al., 2025)	$2.65^{\pm 0.03}$	$3.24^{\pm 0.15}$	$3.44^{\pm 0.27}$	$4.34^{\pm 0.69}$
VACE (Jiang et al., 2025)	$5.36^{\pm 0.71}$	$6.15^{\pm 0.66}$	$8.76^{\pm 0.83}$	$6.98^{\pm 0.24}$
PersonaAnimator (Qian et al., 2025a)	$7.64^{\pm 0.48}$	$7.75^{\pm 0.53}$	$6.56^{\pm 0.24}$	$7.47^{\pm 0.75}$
PersonaScape (Ours)	$\mathbf{9.54}^{\pm 0.87}$	$\mathbf{9.46}^{\pm 0.54}$	$\mathbf{9.58}^{\pm 0.51}$	$\mathbf{9.73}^{\pm 0.97}$

4.5 User Study

Beyond quantitative experiments, we conducted a user study to evaluate perceptual performance. 50 volunteers rated generated videos on a 1-10 scale across four metrics: Content Preservation (CP), Style Expressiveness (SE), scene-motion coherence (SMC), and scene-prompt coherence (SPC). Our method achieved significantly superior mean scores (95% confidence interval, Table 4) on all metrics. These results further validate our superior performance, demonstrating the ability to generate high-fidelity, personalized motion for arbitrary scenes and indicating greater realism and practical applicability. User study details are provided in the **Appendix L**.

5 Conclusion

This paper presents PersonaScape, a personalized motion video generation framework for semantically aligned arbitrary scenes, which addresses the limitations of existing methods in both personalized motion and scene generation. We design the ASTD module to explicitly disentangle motion into a global structural motion stream and a fine-grained detail motion stream, and model and enhance them separately, thereby mitigating the tendency of attention mechanisms to average out personalized styles. In addition, we propose the MCASA module, which enables arbitrary scene planning and semantic alignment through a tri-loop reflective collaboration mechanism. Extensive quantitative and qualitative results demonstrate that our method outperforms existing state-of-the-art approaches in personalized motion generation quality, scene semantic consistency, and overall visual fidelity, verifying the effectiveness of PersonaScape in generating semantically coherent personalized motion videos across arbitrary scenes.

Limitations

Although PersonaScape achieves strong performance in semantically aligned arbitrary-scene motion personalization, several limitations remain. First, the framework relies on pose estimation and skeleton-based motion representation; therefore, severe occlusions or rapid camera motion may occasionally introduce noisy pose signals and slightly affect fine-grained motion personalization. Second, the reflective multi-agent collaboration mechanism in MCASA improves semantic alignment and controllability, but it also increases inference time compared with one-pass generation pipelines. Future work will explore more robust motion representations and more efficient collaborative reasoning strategies for scalable personalized video generation.

References

Kfir Aberman, Yijia Weng, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. 2020. Unpaired motion style transfer from video to animation. *ACM TOG*, 39(4):64–1.

Naoki Agata and Takeo Igarashi. 2025. Motion control via metric-aligning motion matching. In *SIGGRAPH*, pages 1–12.

Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. 2019. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, volume 1, page 2.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Jae Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. Technical report, OpenAI.

Black Forest Labs. 2024. Introducing flux.1 tools. Official announcement for the FLUX.1 model family.

Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, Tiankai Hang, Duojun Huang, Jie Jiang, Zhengkai Jiang, Weijie Kong, Changlin Li, Donghao Li, Junzhe Li, Xin Li, and 5 others. 2025. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody dance now. In *ICCV*, pages 5933–5942.

Junhao Cheng, Yuying Ge, Yixiao Ge, Jing Liao, and Ying Shan. 2025. Animegamer: Infinite anime life simulation with next game state prediction. In *ICCV*, pages 10875–10885.

Dale Decatur, Thibault Groueix, Wang Yifan, Rana Hanocka, Vladimir Kim, and Matheus Gadelha. 2025. Reusing computation in text-to-image diffusion for efficient generation of image sets. *arXiv preprint arXiv:2508.21032*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gemini Team and Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Haonan Han, Xiangzuo Wu, Huan Liao, Zunnan Xu, Zhongyuan Hu, Ronghui Li, Yachao Zhang, and Xiu Li. 2025. Atom: Aligning text-to-motion model at event-level with gpt-4vision reward. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22746–22755.

Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, pages 8153–8163.

Yasamin Jafarian and Hyun Soo Park. 2021. Learning high fidelity depths of dressed humans by watching social media dance videos. In *CVPR*, pages 12753–12762.

Deok-Kyeong Jang, Soomin Park, and Sung-Hee Lee. 2022. Motion puzzle: Arbitrary motion style transfer by body part. *ACM TOG*, 41(3):1–16.

Bo Jiang, Ziyan Zhang, Doudou Lin, Jin Tang, and Bin Luo. 2019. Semi-supervised learning with graph learning-convolutional networks. In *CVPR*, pages 11313–11320.

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

Boeun Kim, Hea In Jeong, JungHoon Sung, Yihua Cheng, Jeongmin Lee, Ju Yong Chang, Sang-Il Choi, Younggeun Choi, Saim Shin, Jungho Kim, and 1 others. 2025. Personabooth: Personalized text-to-motion generation. *arXiv preprint arXiv:2503.07390*.

Boeun Kim, Jungho Kim, Hyung Jin Chang, and Jin Young Choi. 2024. Most: Motion style transformer between diverse action contents. In *CVPR*, pages 1705–1714.

Kling AI. 2025. Introducing the kling 2.1 lineup! Official release note for a commercial closed-source system.

Hongxiang Li, Yaowei Li, Yuhang Yang, Junjie Cao, Zhihong Zhu, Xuxin Cheng, and Long Chen. 2024a. Dispose: Disentangling pose guidance for controllable human image animation. *arXiv preprint arXiv:2412.09349*.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, and 1 others. 2024b. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. *NeurIPS*, 37:28515–28536.

Zhe Li, Yisheng He, Lei Zhong, Weichao Shen, Qi Zuo, Lingteng Qiu, Zilong Dong, Laurence Tianruo Yang, and Weihao Yuan. 2024c. Mulsmo: Multimodal stylized motion generation by bidirectional control flow. *arXiv preprint arXiv:2412.09901*.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2025. Introducing gpt-5. Official announcement for a commercial system.

Ziyun Qian, Runyu Xiao, Shuyuan Tu, Wei Xue, Dingkang Yang, Mingcheng Li, Dongliang Kou, Minghao Han, Zizhi Chen, and Lihua Zhang. 2025a. Personaanimator: Personalized motion transfer from unconstrained videos. *arXiv preprint arXiv:2508.19895*.

Ziyun Qian, Zeyu Xiao, Xingliang Jin, Dingkang Yang, Mingcheng Li, Zhenyi Wu, Dongliang Kou, Peng Zhai, and Lihua Zhang. 2025b. Umsd: High realism motion style transfer via unified mamba-based diffusion. In *ACM MM*, pages 7424–7433.

Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, and Xia Hou. 2023. Finestyle: Semantic-aware fine-grained motion style transfer with dual interactive-flow fusion. *IEEE Trans. Visualization Comput. Graph.*, 29(11):4361–4371.

Wenfeng Song, Xingliang Jin, Shuai Li, Chenglizhao Chen, Aimin Hao, Xia Hou, Ning Li, and Hong Qin. 2024. Arbitrary motion style transfer with multi-condition motion latent diffusion model. In *CVPR*, pages 821–830.

Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. 2025. Animate-x: Universal character image animation with enhanced motion representation.

Xiaofeng Tan, Hongsong Wang, Xin Geng, and Pan Zhou. 2026. Sopo: Text-to-motion generation using semi-online preference optimization. *Advances in Neural Information Processing Systems*, 38:60714–60747.

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

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. 2025. Stableanimator: High-quality identity-preserving human image animation. In *CVPR*, pages 21096–21106.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2018. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *NeurIPS*, 30.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, and 1 others. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Boyuan Wang, Xiaofeng Wang, Chaojun Ni, Guosheng Zhao, Zhiqin Yang, Zheng Zhu, Muyang Zhang, Yukun Zhou, Xinze Chen, Guan Huang, and 1 others. 2025a. Humandreamer: Generating controllable human-motion videos via decoupled generation. *arXiv preprint arXiv:2503.24026*.

Yuzheng Wang, Dingkang Yang, Zhaoyu Chen, Yang Liu, Siao Liu, Wenqiang Zhang, Lihua Zhang, and Lizhe Qi. 2024. De-confounded data-free knowledge distillation for handling distribution shifts. In *CVPR*, pages 12615–12625.

Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. 2025b. Designdiffusion: High-quality text-to-design image generation with diffusion models. In *CVPR*, pages 20906–20915.

Qi Wu, Yubo Zhao, Yifan Wang, Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. 2025. Motion-agent: A conversational framework for human motion generation with llms. In *International Conference on Learning Representations*, volume 2025, pages 48023–48043.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. 2024. Omnicontrol: Control any joint at any time for human motion generation. In *ICLR*.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2024. Magicanimate: Temporally consistent human image animation using diffusion model. In *CVPR*, pages 1481–1490.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. 2023. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, pages 4210–4220.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. 2025. Gamefactory: Creating new games with generative interactive videos. In *ICCV*, pages 11590–11599.

Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. 2019. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*.

Lei Zhong, Chuan Guo, Yiming Xie, Jiawei Wang, and Changjian Li. 2025. Sketch2anim: Towards transferring sketch storyboards into 3d animation. *ACM Transaction on Graphics (TOG)*, 44(4):1–15.

Lei Zhong, Yiming Xie, Varun Jampani, Deqing Sun, and Huaizu Jiang. 2024. Smoodi: Stylized motion diffusion model. In *ECCV*.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. 2024. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *ECCV*, pages 145–162. Springer.

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. 2023. Human motion generation: A survey. *IEEE TPAMI*, 46(4):2430–2449.

A Appendix Sections

In this technical appendix, we further detail the following aspects omitted in the main paper.

- **Section B:** Model Zoo.
- **Section C:** Additional Description of Evaluation Metrics.
- **Section D:** Dataset Description
- **Section E:** Loss Functions of the ASTD Module
- **Section F:** Tri-loop Reflective Collaboration Workflow.
- **Section G:** Additional Quantitative Evaluation.
- **Section H:** Additional Qualitative Evaluation
- **Section I:** Full Ablation Studies
- **Section J:** The Prompt Design of MCASA
- **Section K:** Prompts for SMA and SPA Evaluation
- **Section L:** User Study Details.
- **Section M:** Ethics, Privacy, and Responsible Use Statement
- **Section N:** Practical application scenarios

B Model Zoo

To demonstrate the superiority of the proposed method and the realism of its generated results, we compare it with representative state-of-the-art (SOTA) methods in the field of pose-guided character motion transfer, including PersonaAnimator (Qian et al., 2025a), VACE (Jiang et al., 2025), Animate-X (Tan et al., 2025), and StableAnimator (Tu et al., 2025). The details of the baselines are as follows:

- **PersonaAnimator** (Qian et al., 2025a) introduces a new task, Video-to-Video Motion Personalization, that aims to learn subject-specific motion styles from unconstrained videos rather than merely copying poses. To solve this problem, the authors propose PersonaAnimator. This framework combines motion content and personalized style through the SA-PMT module, while KTCR regularization enforces skeletal-topology consistency and reduces motion distortion. The paper also presents PersonaVid, the first video-based personalized motion dataset, containing 20 motion-content categories and 120 style categories. Experiments show that the method outperforms prior approaches in personalization, motion realism, and overall generation quality.
- **VACE** (Jiang et al., 2025) proposes an "All-in-one" framework for video creation and editing. Firstly, it designs the "Video Condition Unit (VCU)," a unified interface that organizes disparate task inputs such as reference images, editing instructions, and masks. Secondly, it uses a "Context Adapter" structure to inject the temporal and spatial concepts required for different tasks into the base model. This enables the model to flexibly handle a range of video synthesis tasks. As a result, this single, unified model handles multiple tasks—including reference-image generation, video-to-video editing, and mask-based editing—and supports flexible combinations of these tasks.
- **Animate-X** (Tan et al., 2025) is a framework for general character image animation. It addresses the limited generalisation ability of existing methods to diverse body structures, including anthropomorphic characters, by enhancing motion representation. Specifically,

it introduces a pose-guided module that combines implicit CLIP features with explicit pose transformations to improve adaptability across different body types. Its general animation capability is validated on relevant datasets.

- **StableAnimator** (Tu et al., 2025) is an end-to-end framework for high-quality identity-preserving human image animation. This work tackles the difficulty of ensuring identity consistency in character animation with current diffusion models. Integrating dedicated identity-preservation modules, including facial feature extrmotion and global/local techniques, synthesizes high-fidelity and identity-stable videos without relying on post-processing.

These baseline methods provide critical references and comparative foundations for this study, particularly in key technical aspects of digital human style transfer, such as pose transfer, appearance preservation, identity consistency, and cross-character generalization.

C Additional Description of Evaluation Metrics

To quantitatively evaluate the generation performance of different methods on the Video-to-Video Motion Personalization task, we use the following metrics to assess their generation quality. These metrics mainly cover three aspects: video quality metrics, personalized motion generation metrics, and consistency metrics. For evaluating video sequence dynamics and temporal aspects, we use the following metrics:

- **FID-VID (Fréchet Inception Distance for Video)** (Balaji et al., 2019) is a variant of the FID metric applied to video evaluation. It primarily measures the distance between the distributions of the generated and real video sets. Specifically, the method treats each frame in the video as an independent image, uses the Inception v3 network to extract features, and then computes the Fréchet distance between the generated and real sets in the feature space. A lower value indicates that the quality distribution of the generated videos is closer to that of the input videos.

- **Fréchet Video Distance (FVD)** (Unterthiner et al., 2018) is one of the most authoritative metrics for evaluating video generation quality. It employs a 3D CNN to simultaneously extract both the spatial and temporal features of the video. In the context of motion style transfer, FVD effectively measures both the visual quality and the motion fluency of the generated videos. It computes the distance between the spatiotemporal feature distributions of the generated and real video sets. A lower FVD value indicates that the generated videos are more realistic and better align with real videos in terms of motion coherence, speed, and physical laws.

To evaluate personalized motion generation performance, we use the CRA-V and SRA-V metrics proposed in PersonaAnimator (Qian et al., 2025a).

- **Content Recognition Accuracy-Video (CRA-V):** This metric uses a custom-trained content classifier to assess the preservation of content in motion videos generated by the model. A higher CRA value indicates that the content classifier correctly identifies more motion content in the generated video set, suggesting that the model-generated personalized motion better retains the original motion's underlying content.

- **Style Recognition Accuracy-Video (SRA-V):** This metric uses a custom-trained style classifier to assess the style recognizability of motion videos generated by the model. A higher SRA value indicates that the style classifier identifies the motion style in the generated video set more frequently, suggesting that the model's personalized motion effectively learns the stylistic features of the reference style.

To comprehensively evaluate the quality of the generated scenes in terms of semantic consistency and logical coherence, particularly to assess whether the generated scene images maintain a high degree of alignment in visual narrative with the personalized motion style and user prompt, we introduce the following two new quantitative evaluation metrics:

- **Scene-Motion Alignment (SMA):** This metric quantitatively evaluates the semantic coherence between the generated scene reference image and the target motion style, as well

as the scene's internal logical consistency. It aims to assess whether the background environment is semantically aligned with the specific motion's characteristics. To ensure evaluation accuracy, we use three large language models—ChatGPT4o (OpenAI, 2024), ChatGPT5 (OpenAI, 2025), and Gemini 2.5 (Gemini Team and Google, 2025)—to independently score the generated results on a 100-point scale, and the final SMA score is the average of the three ratings.

- **Scene-Prompt Alignment (SPA):** This metric evaluates the alignment between the generated scene reference image and the user-provided textual prompt, as well as the logical coherence of the scene content. It focuses on assessing whether the generated scene elements accurately reflect the user's descriptive instructions. Similar to SMA, SPA also employs three large language models—ChatGPT4o (OpenAI, 2024), ChatGPT5 (OpenAI, 2025), and Gemini 2.5 (Gemini Team and Google, 2025)—as evaluation agents. Each model independently scores the generated results on a 100-point scale, and the final SPA score is the average of the three ratings.

D Dataset Description

D.1 PersonaVid Dataset

The PersonaVid dataset (Qian et al., 2025a) defines 20 distinct categories of motion content (e.g., walking, running, dancing) and encompasses 120 unique motion styles, significantly expanding the diversity of motion personalization resources. The dataset features a wide array of subjects, ranging from ordinary individuals to sports stars, motion performers, and animated characters, with 93 styles belonging to ordinary individuals and 27 associated with celebrities or fictional characters. Each video sample is systematically annotated using a standardized "Content_Style_Number" format to distinguish between fundamental motion categories and specific personalized stylistic nuances, ensuring efficient data retrieval. By adopting a "One Person, One Style" paradigm, PersonaVid provides a robust, large-scale benchmark designed to facilitate the learning of fine-grained, identity-specific motion features from unconstrained videos for the Motion Personalization task

D.2 TikTok Dataset

The TikTok dataset (Jafarian and Park, 2021) comprises over 300 single-person dance sequences sourced from social media platforms, capturing diverse poses, clothing styles, identities, and motion dynamics. Sampled at 30 FPS, it contains over 100K RGB images, each annotated with human foreground masks via semantic segmentation and UV surface coordinates using DensePose. Designed for dynamic human geometry reconstruction, the dataset emphasizes fine-grained cloth wrinkle and local deformation modeling, significantly surpassing existing static 3D scanned model datasets in scale and diversity. The dataset and code are publicly released, providing a large-scale, weakly supervised resource for single-view high-fidelity human depth estimation.

D.3 Fashion Dataset

The Fashion dataset (Zablotskaia et al., 2019) comprises 500 training and 100 test high-resolution videos, each containing approximately 350 frames of a single human subject captured under static camera conditions, with diverse clothing and texture variations. It is publicly released to support pose-guided human video generation research.

E Loss Functions of the ASTD Module

The ASTD module in PersonaScape performs motion style transfer, generating personalized motion that preserves the macro-structure of the content motion while capturing the characteristics of the style motion. We adopt the style consistency loss proposed by Aberman *et al.* (Aberman et al., 2020) for training. A skeleton pose sequence is denoted by $\boldsymbol{P}_i^j$, where i indicates motion content and j indicates motion style. Given a content pose $\boldsymbol{P}_c^a$ and a style pose $\boldsymbol{P}_b^s$, the generated personalized pose is $\boldsymbol{P}_c^s$, which retains content c from $\boldsymbol{P}_c^a$ and inherits style s from $\boldsymbol{P}_b^s$.

The generated personalized pose is obtained by combining the content representation of the content skeleton pose with the style representation of the style skeleton pose. The process is formulated as:

$$\boldsymbol{P}_c^s = \boldsymbol{D}\left(\boldsymbol{E}_C(\boldsymbol{P}_c^a) \mid \boldsymbol{E}_S(\boldsymbol{P}_b^s)\right), \qquad (10)$$

where $\boldsymbol{D}$ denotes the decoder in the motion transfer framework, $\boldsymbol{E}_C$ and $\boldsymbol{E}_S$ denote the content encoder and the style encoder, respectively, and $\mid$ denotes the fusion of the encoded content and style features.

If the content skeleton pose $\boldsymbol{P}_c^a$ and the style skeleton pose $\boldsymbol{P}_b^s$ share the same motion content, *i.e.*, $c = b$, then the generated personalized pose $\boldsymbol{P}_c^s$ should be as close as possible to the style skeleton pose $\boldsymbol{P}_b^s$, because in this case the two inputs mainly differ in motion style. Therefore, the style consistency loss is defined as:

$$\mathcal{L}_{\text{sc}} = \mathbb{E}_{\boldsymbol{P}_c^a,\, \boldsymbol{P}_b^s \sim \mathcal{M},\, c=b} \left[\| \boldsymbol{D}(\boldsymbol{E}_C(\boldsymbol{P}_c^a) \mid \boldsymbol{E}_S(\boldsymbol{P}_b^s)) - \boldsymbol{P}_b^s \|_1 \right], \tag{11}$$

where $\| \cdot \|_1$ denotes the L_1 norm, $\mathcal{M}$ is the PersonaVid dataset (Qian et al., 2025a), $\boldsymbol{D}$ is the decoder in the motion transfer framework, and $\boldsymbol{E}_C$ and $\boldsymbol{E}_S$ denote the content and style pose encoders, respectively.

Following the same rationale, we also employ a content consistency loss. If the content skeleton pose $\boldsymbol{P}_c^a$ and the style skeleton pose $\boldsymbol{P}_b^s$ share the same motion style, *i.e.*, $a = s$, then the generated personalized pose $\boldsymbol{P}_c^s$ should be as close as possible to the content skeleton pose $\boldsymbol{P}_c^a$, because in this case the two inputs mainly differ in motion content. The loss is written as:

$$\mathcal{L}_{\text{cc}} = \mathbb{E}_{\boldsymbol{P}_c^a,\, \boldsymbol{P}_b^s \sim \mathcal{M},\, a=s} \left[\| \boldsymbol{D}(\boldsymbol{E}_C(\boldsymbol{P}_c^a) \mid \boldsymbol{E}_S(\boldsymbol{P}_b^s)) - \boldsymbol{P}_c^a \|_1 \right]. \tag{12}$$

In addition, we adopt the dynamic bone stability constraint $\mathcal{L}_{\text{stability}}$ and the body connectivity constraint $\mathcal{L}_{\text{conn}}$ proposed in PersonaAnimator (Qian et al., 2025a). The overall training loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sc}} + \mathcal{L}_{\text{cc}} + \mathcal{L}_{\text{stability}} + \mathcal{L}_{\text{conn}}. \tag{13}$$

F Tri-loop Reflective Collaboration Workflow.

The Algorithm 1 and Fig. 5 jointly illustrate the execution logic of MCASA. The pseudocode formalizes the iterative control flow, including scene proposal, motion-semantic verification, prompt-semantic verification, and feedback-driven refinement. Fig. 5 provides an intuitive visualization of this process, showing how the pose interpreter, scene planning agent, spatial layout agent, visual style agent, and two alignment evaluators collaborate in a closed loop. Together, they demonstrate that MCASA does not generate scenes through one-shot prompting, but progressively refines scene designs until both motion consistency and prompt consistency are satisfied.

Algorithm 1: Tri-loop reflective collaboration procedure of the MCASA module

Input : User prompt P, personalized skeleton pose $\boldsymbol{P}_c^s$
Output : Final scene design scheme $\boldsymbol{S}^\star$, final reference scene image $\boldsymbol{I}^r$

$\Delta_i^{(0)} \leftarrow \varnothing$
$t \leftarrow 1$
while *true* **do**
  $\mathcal{T}^{(t)} \leftarrow \mathcal{PI}(\boldsymbol{P}_c^s)$ `// fine-grained textual representation decoded from the personalized pose`
  `// Loop 1: Scene Design Nexus (SDN) internal loop`
  while *true* **do**
    $\Delta_{SP}^{(t)} \leftarrow \boldsymbol{A}_{SP}(\mathcal{T}^{(t)}, P, \Delta_i^{(t-1)})$
    $\Delta_{SL}^{(t)} \leftarrow \boldsymbol{A}_{SL}(\mathcal{T}^{(t)}, P, \Delta_i^{(t-1)})$
    $\Delta_{VS}^{(t)} \leftarrow \boldsymbol{A}_{VS}(\mathcal{T}^{(t)}, P, \Delta_i^{(t-1)})$
    $\boldsymbol{S}^{(t)} \leftarrow \text{Merge}(\Delta_{SP}^{(t)}, \Delta_{SL}^{(t)}, \Delta_{VS}^{(t)})$
    if $\Delta_{SP}^{(t)} = \varnothing$ ***and*** $\Delta_{SL}^{(t)} = \varnothing$ ***and*** $\Delta_{VS}^{(t)} = \varnothing$ **then**
      break
  `// Loop 2: scene-motion consistency loop`
  $(\boldsymbol{r}_{SM}^{(t)}, \boldsymbol{\Delta}_{SMA}^{(t)}) \leftarrow \varepsilon_{SM}(\mathcal{T}^{(t)}, \boldsymbol{S}^{(t)})$
  if $\boldsymbol{r}_{SM}^{(t)} = 0$ **then**
    $\Delta_i^{(t)} \leftarrow \boldsymbol{\Delta}_{SMA}^{(t)}$
    $t \leftarrow t + 1$
    continue
  `// Loop 3: scene-prompt consistency loop`
  $\boldsymbol{I}_r^{(t)} \leftarrow \boldsymbol{A}_{RI}(\boldsymbol{S}^{(t)})$
  $(\boldsymbol{r}_{SP}^{(t)}, \boldsymbol{\Delta}_{SPA}^{(t)}) \leftarrow \varepsilon_{SP}(\boldsymbol{I}_r^{(t)}, P)$
  if $\boldsymbol{r}_{SP}^{(t)} = 1$ **then**
    $\boldsymbol{S}^\star \leftarrow \boldsymbol{S}^{(t)}$
    $\boldsymbol{I}^r \leftarrow \boldsymbol{I}_r^{(t)}$
    return $\boldsymbol{S}^\star, \boldsymbol{I}^r$
  else
    $\Delta_i^{(t)} \leftarrow \boldsymbol{\Delta}_{SPA}^{(t)}$
    $t \leftarrow t + 1$

G Additional Quantitative Evaluation

As shown in Tables 5 and as shown in Tables 6, PersonaScape delivers the best overall performance on both the TikTok and Fashion datasets, which demonstrates that the proposed framework generalizes effectively across different data domains. On TikTok, our method achieves the highest CRA-V, SRA-V, SMA, and SPA, while also obtaining the lowest FID-VID and FVD, indicating clear advantages in motion content preservation, personalized style transfer, scene semantic alignment, visual fidelity, and temporal coherence. A similar trend is observed on Fashion, where PersonaScape again

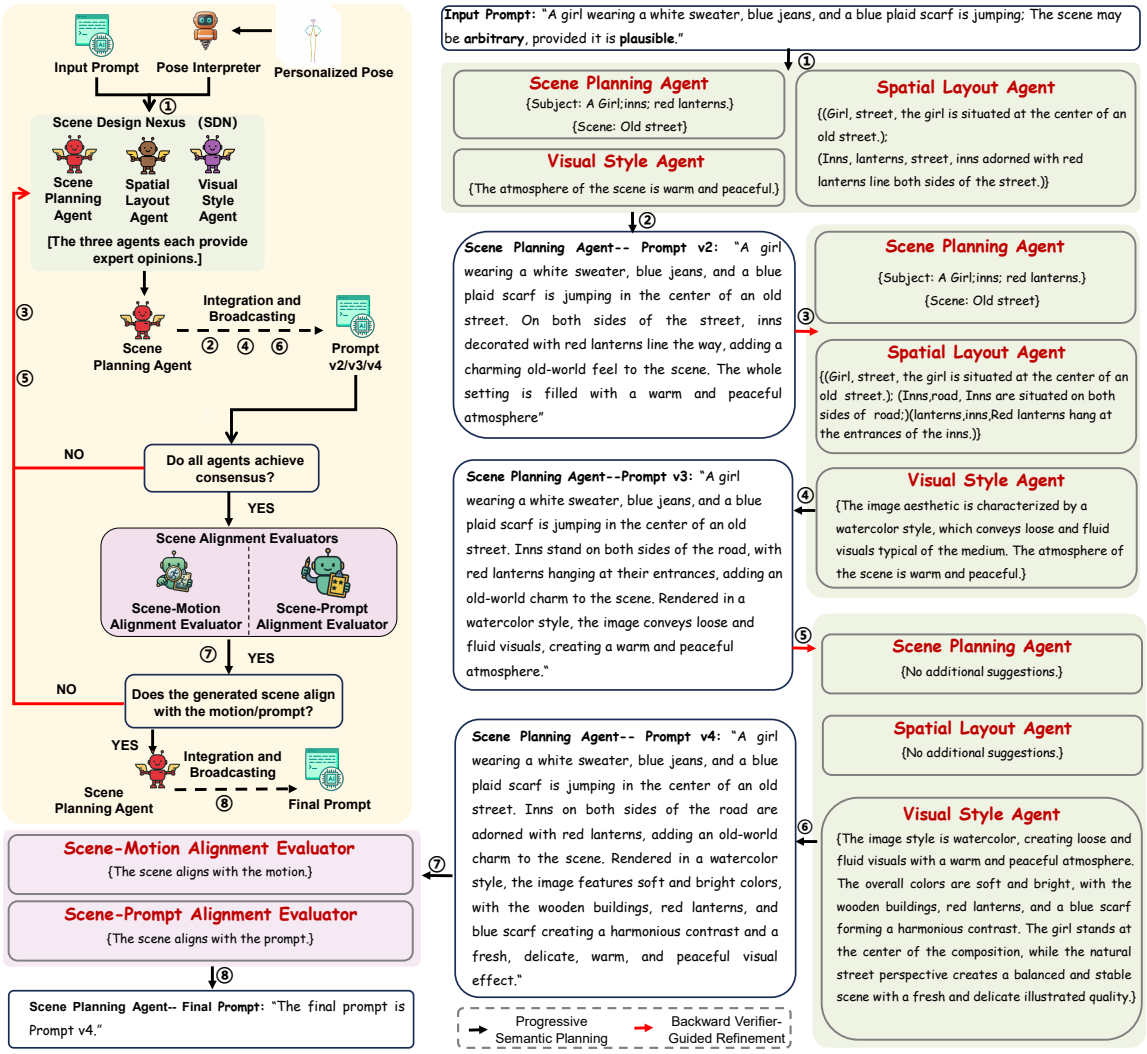

Figure 5: Workflow of the tri-loop reflective collaboration mechanism in the MCASA module. Given the input prompt and the personalized pose from the pose interpreter, the SDN enables progressive semantic planning across the scene-planning, spatial-layout, and visual-style agents, iteratively refining the scene description. Their outputs are integrated into updated prompts (*e.g.*, Prompt v2–v4) via the scene planning agent. The resulting scene scheme is then subjected to backward verifier-guided refinement through the scene-motion and scene-prompt alignment evaluators to ensure consistency with motion semantics and the original prompt. The final prompt is accepted only if both evaluations are satisfied; otherwise, feedback is returned to the SDN for further refinement, forming a closed-loop process of planning, verification, and optimization before producing the final scene reference.

ranks first on all six metrics and surpasses previous methods by a clear margin. These results show that the ASTD module can preserve macro motion structure while enhancing fine-grained stylistic cues, and that the MCASA module further improves consistency between generated scenes, motion semantics, and text prompts, thereby enabling more realistic, semantically coherent personalized motion generation.

Table 5: Quantitative comparisons on the Tiktok dataset (Jafarian and Park, 2021). The results show the superiority of our framework over SOTA methods.

Method	CRA-V↑	SRA-V↑	FID-VID↓	FVD↓	SMA↑	SPA↑
PersonaAnimator (Qian et al., 2025a)	57.43	65.58	16.57	276.96	43.54	56.32
VACE (Jiang et al., 2025)	45.64	36.95	13.38	214.47	54.67	46.76
StableAnimator (Tu et al., 2025)	28.30	19.88	13.14	386.63	25.86	67.64
Animate-X(Tan et al., 2025)	61.26	49.16	11.26	766.32	67.52	72.86
PersonaScape (Ours)	**76.54**	**78.65**	**10.54**	**166.77**	**83.72**	**87.97**

Table 6: Quantitative comparisons on the Fashion dataset (Zablotskaia et al., 2019). The results show the superiority of our framework over SOTA methods.

Method	CRA-V↑	SRA-V↑	FID-VID↓	FVD↓	SMA↑	SPA↑
PersonaAnimator (Qian et al., 2025a)	67.54	57.54	13.76	83.68	45.75	49.64
VACE (Jiang et al., 2025)	46.43	46.75	11.43	81.45	37.96	56.34
StableAnimator (Tu et al., 2025)	28.30	31.49	9.92	75.97	36.75	53.97
Animate-X (Tan et al., 2025)	51.20	48.10	11.62	78.74	64.97	68.62
PersonaScape (Ours)	**78.53**	**85.95**	**9.14**	**67.54**	**89.86**	**95.86**

H Additional Qualitative Evaluation

H.0.1 Comparison between the MCASA Module and SOTA T2I methods

To demonstrate the superiority of the MCASA module within the PersonaScape framework in fine-grained attribute control and scene logical coherence, we conduct a qualitative comparison with current state-of-the-art text-to-image (T2I) models (Black Forest Labs, 2024; Kling AI, 2025; Cao et al., 2025; Betker et al., 2023; OpenAI, 2024). The results are shown in Figures 6 and 7.

In Figure 6, the prompt requires generating a woman wearing "a beige knitted sweater paired with a dark brown skirt and a red plaid scarf." Although all T2I methods correctly reproduce the basic clothing items, they show clear deficiencies in fine-grained details. DALL-E-3 (Betker et al., 2023) and FLUX1.11-pro (Black Forest Labs, 2024) completely ignore the key detail of the "red plaid scarf," while Kling 2.1 (Kling AI, 2025) generates a scarf but incorrectly renders it in blue, contradicting the instruction. More critically, there are semantic conflicts in scene logic: HunyuanImage3.0 (Cao et al., 2025) places the character wearing a thick sweater on a tropical beach with palm trees, and ChatGPT-4o (OpenAI, 2024) positions her beside a swimming pool. Such seasonally inconsistent combinations between attire and environment create strong logical dissonance, undermining the realism of the visual narrative.

Similarly, in Figure 7, the prompt explicitly requires the character to wear a thick necklace. However, DALL-E-3 (Betker et al., 2023), HunyuanImage3.0 (Cao et al., 2025), and FLUX1.1-pro (Black Forest Labs, 2024) again fail to capture this detail and do not generate the necklace. In terms of scene coherence, Kling 2.1 (Kling AI, 2025) places a character in a short-sleeved top and skirt in a snowy environment, which conflicts with basic physical common sense. ChatGPT-4o (OpenAI, 2024) generates a realistic library scene, but it does not match the motion of "dancing," since dancing

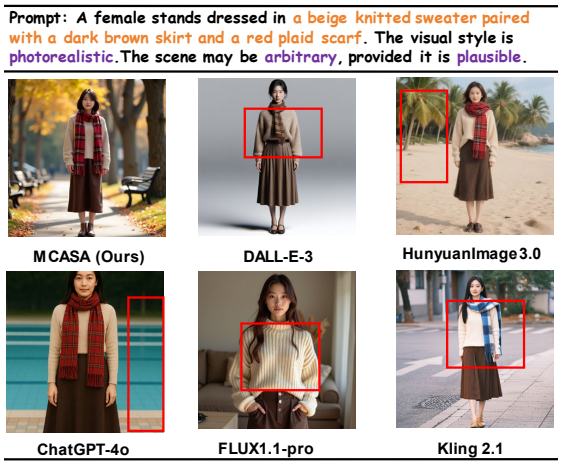

Figure 6: Qualitative comparison between the MCASA module and SOTA T2I methods (Case 1). Red boxes indicate regions with visible defects in the generated images.

Figure 7: Qualitative comparison between the MCASA module and SOTA T2I methods (Case 2). Red boxes indicate regions with visible defects in the generated images.

in a quiet reading space is clearly inappropriate.

The above comparative experiments provide strong evidence that general-purpose T2I models lack deep reasoning capabilities and effective feedback mechanisms, making it difficult to achieve great detail fidelity and logical consistency within scenes simultaneously. In contrast, our MCASA module, through a tri-loop reflective multi-agent collaboration mechanism, effectively resolves semantic conflicts and ensures that the generated results are highly aligned with user intent in both detail and logic.

H.0.2 Comparison of scene-level consistency rates

To further evaluate scene-level semantic reliability, we conduct an additional consistency-rate study on

Table 7: Comparison of scene-level consistency rates over 50 cases.

Method	SMCR↑	SPCR↑
DALL-E-3 (Betker et al., 2023)	58%	66%
HunyuanImage3.0 (Cao et al., 2025)	60%	68%
ChatGPT-5 (OpenAI, 2025)	78%	72%
FLUX1.1-pro (Black Forest Labs, 2024)	58%	62%
Kling 2.1 (Kling AI, 2025)	54%	60%
PersonaScape (Ours)	**98%**	**96%**

50 test cases, and the results are reported in Table 7. Specifically, we use Scene-Motion Consistency Rate (SMCR) to measure the proportion of cases in which the generated scene is semantically consistent with the target motion, and Scene-Prompt Consistency Rate (SPCR) to measure the proportion of cases in which the generated scene remains semantically aligned with the input prompt. MCASA achieves 98% SMCR and 96% SPCR, substantially outperforming all compared text-to-image baselines. The margin is especially pronounced for scene-motion consistency. The strongest baseline, ChatGPT-5 (OpenAI, 2025), achieves 78%, which is still 20 percentage points lower than our method, whereas the remaining models range from 54% to 60%. A similar trend is observed for scene-prompt consistency, where MCASA achieves 96%, compared with 72% for ChatGPT-5 (OpenAI, 2025) and 60%–68% for the other methods (Kling AI, 2025; Cao et al., 2025; Betker et al., 2023; Black Forest Labs, 2024). These results show that generic T2I models may occasionally satisfy either motion plausibility or prompt fidelity, but they struggle to maintain both consistently across diverse cases. In contrast, the clear advantage of MCASA indicates that its tri-loop reflective collaboration mechanism more effectively filters out semantically incompatible scene candidates while preserving alignment with user intent, thereby producing scene generation with stronger semantic coherence.

I Full Ablation Studies

I.0.1 Importance of the ASTD Module

The quantitative ablation in Table 2 shows that the full ASTD design yields the best results on all six metrics, which indicates that both motion structure modeling and motion detail enhancement are necessary rather than redundant. Compared with removing the kinetic structure block, the full model improves FID-VID from 40.96 to 35.21 and FVD from 582.47 to 531.42, while raising CRA-V from 63.42 to 76.31 and SRA-V from 65.53 to 84.78. Compared with removing the temporal nuance block, the full model also improves FID-VID from 37.42 to 35.21 and FVD from 552.68 to 531.42, while increasing CRA-V from 68.25 to 76.31 and SRA-V from 62.52 to 84.78. A more fine-grained comparison further reveals complementary functional roles between the two branches. Removing the kinetic structure block causes the largest degradation in CRA-V, suggesting that this branch is more critical for preserving the macro motion pattern and content structure. By contrast, removing the temporal nuance block causes the lowest SRA-V score among the ablated variants, indicating that this branch is more closely tied to capturing personalized style cues. The qualitative ASTD ablation in Figure 8 supports the same conclusion: the full model preserves the girl's walking structure while still absorbing Marchand's characteristic shoulder sway and same-side limb coordination, whereas removing either component makes the generated motion visibly unnatural and weakens the intended personalized traits. Taken together, these results show that ASTD improves motion personalization through a clear structure-detail division, where one branch stabilizes content inheritance and the other strengthens fine-grained style expressiveness.

I.0.2 Importance of the SDN and pose interpreter

The quantitative ablation in Table 4.4 shows that removing any planning-related component degrades not only scene metrics but also overall generation quality, which suggests that scene design quality and final video quality are tightly coupled. Among the SDN components, removing the visual style agent causes the worst visual degradation, with FID-VID increasing from 35.21 to 43.84 and FVD increasing from 531.42 to 605.31, while SRA-V also drops from 84.78 to 68.47. Removing the scene planning agent leads to a similarly severe decline, with FID-VID 42.58, FVD 601.25, SMA 75.16, and SPA 78.42, showing that high-level scene conception is essential for semantic alignment and downstream fidelity. Removing the spatial layout agent causes a relatively smaller but still clear deterioration, lowering SMA and SPA to 78.64 and 82.37, which indicates that plausible object arrangement remains indispensable for scene coherence. The pose interpreter also plays a substantial role: after removing it, CRA-V, SRA-V, SMA, and SPA fall to 72.54, 78.65, 76.95, and

Figure 8: Qualitative ablation study of the ASTD module. The full model preserves the walking structure while effectively capturing Marchand's personalized shoulder-sway style. Removing either component makes the generated motion less natural and weakens the personalized characteristics.

71.27, respectively, implying that insufficient motion understanding weakens motion-scene matching and prompt consistency. The qualitative ablation in Figure 9 explains these numbers more concretely. Without the scene planning agent, the generated scene collapses into an overly monotonous composition dominated by a single wall, and the visual expressiveness is clearly reduced. Without the pose interpreter, the scene loses surrounding objects and retains mainly the person itself, indicating that motion semantics are no longer effectively converted into scene priors. Without the spatial layout agent, scene elements are placed illogically, such as a streetlamp standing in the middle of the road, which directly breaks scene coherence. Without the visual style agent, the oil-painting prompt no longer yields a coherent artistic appearance, and the image instead exhibits an abnormal moldy green and harsh orange mixture. Therefore, the SDN and the pose interpreter do not merely decorate the output; they jointly determine whether the scene is semantically rich, spatially plausible, and visually compatible with the target motion.

I.0.3 Importance of the alignment evaluators

The two evaluator agents produce the most targeted and interpretable metric changes in Table

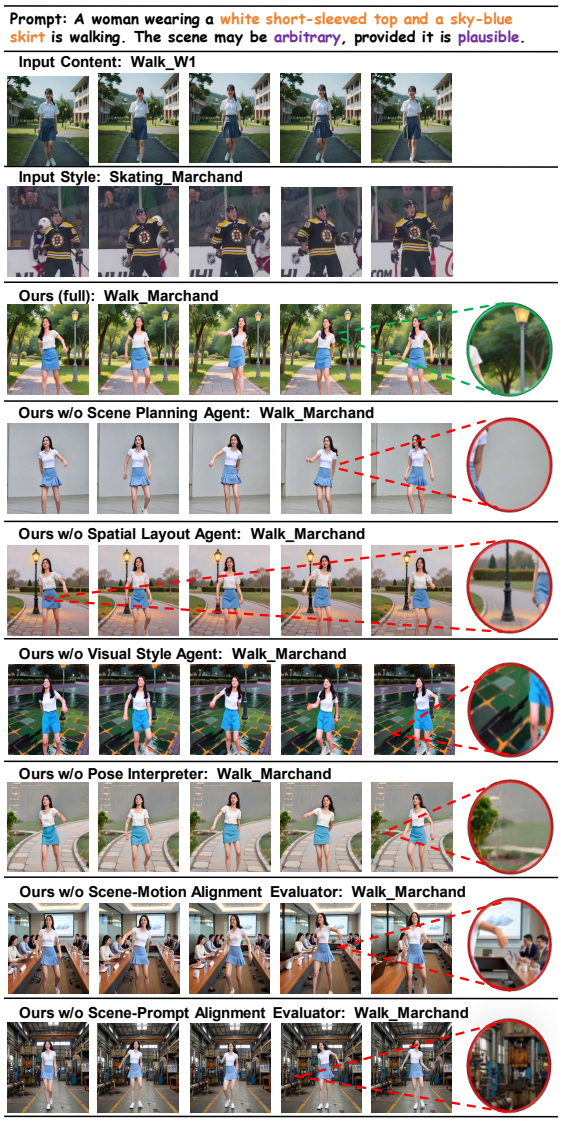

Figure 9: Qualitative ablation study of the MCASA module. The full model generates a scene that is semantically coherent with both the motion and the prompt. Removing any agent leads to less reasonable scene generation and weaker semantic consistency.

4.4, which shows that they supervise two different semantic axes. When the scene-motion alignment evaluator is removed, SMA drops sharply from 94.32 to 72.35, which is the largest decrease on this metric among all ablations, whereas SPA remains relatively high at 90.24. This pattern indicates that the main failure lies in motion-scene compatibility rather than prompt satisfaction. In contrast, when the scene-prompt alignment evaluator is removed, SPA drops from 91.42 to 63.97, which is the largest decrease on prompt alignment, while SMA remains 88.21. This asymmetry shows that the two evaluators are not interchangeable, because each one constrains a distinct type of semantic consistency. The qualitative ablation in Figure 9 is fully consistent with this metric-level diagnosis. Without the

scene-motion alignment evaluator, the generated scene becomes a meeting room in which people are sitting on both sides of the woman, creating a serious social context that is semantically incompatible with a swaying walking motion. Without the scene-prompt alignment evaluator, the generated scene becomes a heavy industrial factory, even though the character wears a short-sleeved top and a short skirt, producing an obvious conflict between scene conditions and prompt-specified appearance. Therefore, the evaluator pair is critical not because it marginally refines the image, but because it explicitly filters out two different categories of semantic contradiction: incompatibility between motion and scene, and incompatibility between prompt and scene.

J The Prompt Design of MCASA

In this section, we provide the detailed descriptions and prompt implementation templates for Pose Interpreter, Scene-Motion Alignment Evaluator, Scene-Prompt Alignment Evaluator, and all agents in Scene Design Nexus. **Note that the text enclosed within the curly braces denotes placeholders that will be dynamically populated during runtime based on the input text prompt and the output of the agent.**

J.1 Pose Interpreter

The task of the pose interpreter is to convert the personalized skeleton pose into a fine-grained textual representation that describes the motion content, motion tendency, and distinctive style cues. This textual prior provides motion-aware guidance for subsequent scene planning, helping the downstream agents design a scene that is semantically compatible with the personalized motion. The prompt template of the pose interpreter is shown in Table 8.

Table 8: The prompt template for the pose interpreter.

Pose Interpreter
[INST]<SYS> As the pose interpreter, your task is to convert the personalized skeleton pose into a fine-grained textual representation. You should describe the motion content, motion tendency, body dynamics, and distinctive personalized motion cues that may influence scene design. Your output should provide motion-aware semantic priors for subsequent scene planning. Focus on motion semantics rather than image appearance. Your output format is {text description}. </SYS>
<USER> The personalized skeleton pose is described as {personalized pose}. Based on this pose, please provide a fine-grained textual representation of the motion, including the main motion category, dynamic tendency, limb coordination, posture characteristics, and any distinctive personalized style cues that are relevant to scene planning. </USER>[INST]

J.2 Scene Design Nexus

Scene Planning Agent: The task of the scene planning agent is to determine the overall scene concept based on the input prompt, the motion prior from the pose interpreter, and the feedback from other agents or evaluators. It is responsible for proposing the overall theme, background category, and candidate object sets, while also integrating and summarizing suggestions from the spatial layout agent and the visual style agent into an updated scene description. The prompt template of the scene planning agent is shown in Table 9.

Spatial Layout Agent: The task of the spatial layout agent is to model the spatial relationships and layout constraints among the objects proposed in the scene plan. It ensures that the scene is structurally reasonable and physically plausible by specifying object placement, relative positions, composition, and environmental arrangement. The prompt template of the spatial layout agent is shown in Table 10.

Visual Style Agent: The task of the visual style agent is to assign an appropriate visual style to the scene based on the input prompt, such as oil painting, anime, or realism. This ensures that the generated scene is expressive and adaptable to various multimedia domains. The prompt template of the visual style agent is shown in Table 11.

J.3 Scene-Motion Alignment Evaluator

The task of the scene-motion alignment evaluator is to determine whether the current scene description is semantically compatible with the motion prior decoded from the personalized pose. It checks whether the scene context, object configuration, and environmental setting support the intended mo-

Table 9: The prompt template for the scene planning agent.

Scene Planning Agent
[INST]<SYS> As the scene planning agent, your task is to determine the overall scene concept based on the input prompt, the motion prior, and any feedback from other agents or evaluators. You should propose the scene theme, background category, and candidate object sets that are semantically compatible with both the prompt and the motion. In addition, you are responsible for integrating useful suggestions from the other agents into an updated prompt. Your output format is {text description}. If no further modification is needed, please output {no objection}. </SYS>
<USER> The input prompt is {input prompt}. The motion prior is {motion prior}. The previous feedback is {feedback}. Based on this information, please determine the overall scene concept, including the scene theme, background category, and candidate object sets, so that the scene is plausible and semantically aligned with both the prompt and the motion. If the current plan is already appropriate and no further modification is needed, please output no objection. </USER>[INST]

Table 10: The prompt template for the spatial layout agent.

Spatial Layout Agent
[INST]<SYS> As the spatial layout agent, your task is to model the spatial relationships and layout constraints among scene objects. You should determine where the main subject and supporting objects are placed, how they are spatially related, and whether the arrangement is structurally reasonable and physically plausible. Focus on composition, object positions, relative distances, and environmental arrangement. Your output format is {text description}. If the current scene layout is already reasonable and no additional adjustment is needed, please output {no objection}. </SYS>
<USER> The input prompt is {input prompt}. The motion prior is {motion prior}. The current scene plan is {scene plan}. The previous feedback is {feedback}. Based on this information, please propose the spatial layout of the scene, including subject placement, supporting object arrangement, relative positions, composition constraints, and any necessary layout corrections to ensure scene rationality. If no further modification is needed, please output no objection. </USER>[INST]

tion semantics, and provides targeted refinement suggestions when inconsistencies are detected. The prompt template of the scene-motion alignment evaluator is shown in Table 12.

J.4 Scene-Prompt Alignment Evaluator

The task of the scene-prompt alignment evaluator is to determine whether the generated scene image or scene result remains semantically aligned with the original user prompt. It checks whether key subject attributes, required objects, scene conditions, and prompt-level constraints are faithfully preserved, and it provides correction suggestions when deviations occur. The prompt template of the scene-prompt alignment evaluator is shown in Table 13.

K Prompts for SMA and SPA Evaluation

To ensure reproducible evaluation of scene-level consistency, we use exactly the same evaluation prompts for ChatGPT-4o (OpenAI, 2024), ChatGPT-5 (OpenAI, 2025), and Gemini 2.5 (Gemini Team and Google, 2025), with temperature set to 0 and top-p set to 1. The three large models score each sample independently on a 0–100 scale, and the final SMA or SPA score is obtained by averaging the three scores. Table 14 presents the unified prompt for Scene-Motion Alignment (SMA), while Table 15 presents the unified prompt for Scene-Prompt Alignment (SPA).

Importantly, the evaluation does not stop at superficial semantic overlap. It explicitly emphasizes the detection of deeper logical contradictions. For example, if the target motion is “running” but the generated scene is a classroom, library, or examination room, this should be regarded as a scene–

Table 11: The prompt template for the visual style agent.

Visual Style Agent

[INST]<SYS>
As the visual style agent, your task is to assign an appropriate visual style to the scene based on the input prompt, such as oil painting, anime, or realism. This ensures that the generated scene is expressive and suitable for a wide range of multimedia domains. Your output format is {text description}. If the existing prompt has already specified the visual style and no additional information is needed, please output {no objection}.
</SYS>

<USER>
The input prompt is described as {input prompt}. Based on the input prompt, please assign an appropriate visual style to the scene, such as oil painting, anime, or realism, to make the generated scene expressive and suitable for various multimedia domains. If the existing prompt has already specified the visual style and no additional information is needed, please output no objection.
</USER>[INST]

Table 12: The prompt template for the scene-motion alignment evaluator.

Scene-Motion Alignment Evaluator

[INST]<SYS>
As the scene-motion alignment evaluator, your task is to determine whether the current scene description is semantically compatible with the motion prior. You should examine whether the environment, objects, and scene context support the intended motion semantics and body dynamics. If the scene is compatible with the motion, output {The scene aligns with the motion.}. Otherwise, output specific refinement suggestions in the format {text description}.
</SYS>

<USER>
The motion prior is {motion prior}. The current scene description is {scene description}. Based on this information, please determine whether the scene is semantically compatible with the motion. Pay attention to whether the motion can naturally happen in the given environment and whether the scene context supports the intended action tendency and style cues. If aligned, output The scene aligns with the motion. Otherwise, provide concrete revision suggestions.
</USER>[INST]

motion conflict. Likewise, if the prompt specifies short sleeves or a short skirt but the generated scene is snowy or winter-like, this should be regarded as a scene–prompt conflict. Such deep inconsistencies should lead to substantial score reductions.

L User Study Details

Due to space limitations in the main paper, we provide additional details of the user study in this section. We conduct a user study to qualitatively evaluate semantically aligned arbitrary-scene motion personalization results generated by different methods. The questionnaire includes 10 different style transfer combinations, where personalized motion videos produced by PersonaScape and other state-of-the-art methods are presented for volunteer comparison. Participants are asked to evaluate each result according to four criteria: Content Preservation, which measures whether the generated motion preserves the source motion content; Style Expressiveness, which assesses whether personalized motion characteristics are successfully captured; Scene-Motion Coherence (SMC), which evaluates whether the generated scene is coherent with the motion; and Scene-Prompt Coherence (SPC), which measures whether the scene is semantically consistent with the user-provided prompt. Each result is rated on a scale from 1 (not achieved at all) to 10 (fully achieved). The questionnaire interface is shown in Figure 10. Notably, if a generated motion is merely a mechanical copy of the reference motion, it is not considered to have learned the target motion style.

We recruit 50 volunteers, including 25 males and 25 females. In terms of age, 35 participants are between 18 and 25 years old, 8 are between 26 and 30 years old, 2 are under 18, and 5 are over 30. Regarding occupation, 42 participants are students, while the remaining participants include 3 teachers, 2 doctors, 1 lawyer, and 2 delivery workers. After collecting all questionnaires, we compute the mean scores of the 50 volunteers under a 95% confidence level. The results are shown in Table 4. Our method significantly outperforms other methods across all

Table 13: The prompt template for the scene-prompt alignment evaluator.

Scene-Prompt Alignment Evaluator

[INST]<SYS>
As the scene-prompt alignment evaluator, your task is to determine whether the generated scene result is semantically aligned with the original user prompt. You should examine whether the subject attributes, required objects, scene conditions, and explicit prompt constraints are faithfully preserved. If the scene aligns with the prompt, output {The scene aligns with the prompt.}. Otherwise, output specific correction suggestions in the format {text description}.
</SYS>

<USER>
The original input prompt is {input prompt}. The generated scene result is {scene image or scene description}. Based on this information, please determine whether the generated result is semantically aligned with the prompt. Pay attention to subject identity, clothing attributes, required accessories or objects, scene conditions, and any explicitly specified constraints. If aligned, output The scene aligns with the prompt. Otherwise, provide concrete correction suggestions.
</USER>[INST]

Table 14: Unified SMA prompt for ChatGPT-4o, ChatGPT-5, and Gemini 2.5 (temperature=0, top-p=1).

SMA Evaluation Prompt

[SYSTEM]
You are a strict scene-motion evaluator. Score SMA from 0–100 using the motion semantics and generated scene.
Goal: assess semantic, physical, and commonsense consistency between the scene and target motion. Penalize unsuitable motion-scene pairs, implausible physical/social settings, and deep conflicts beyond keyword matches.
Examples include running in classrooms, libraries, examination rooms, or narrow offices; dancing in serious meeting rooms or quiet reading spaces; open-space motions in crowded/obstructed scenes; and contradictions among function, norms, and motion.
Scoring criteria:
90–100: consistent, no conflict;
70–89: mostly consistent, minor unnaturalness;
40–69: noticeable inconsistency or commonsense issue;
0–39: severe conflict, unsuitable scene.
Briefly reason, then output exactly:
Score: <integer from 0 to 100>
Reason: <one-sentence explanation>

[USER]
Motion semantics: {motion semantics}
Generated scene/image: {generated scene or image description}
Score SMA by checking logical and commonsense consistency, not word overlap.

metrics, further validating its ability to generate realistic personalized motions and transfer them to arbitrary scenes with strong semantic coherence and practical value for multimedia production.

M Ethics, Privacy, and Responsible Use Statement

This work studies personalized motion generation and semantically aligned scene synthesis for research purposes. Although the proposed framework is designed to improve motion expressiveness, scene coherence, and controllability, we recognize that technologies for personalized human video generation may also introduce ethical and privacy risks, especially when source materials involve real individuals or recognizable public figures. Accordingly, this work should be understood and used only in a responsible research context.

We emphasize that this research is not intended to support identity impersonation, deceptive media generation, fraud, defamation, harassment, or any other harmful use of synthetic human content. The goal of the method is to model motion characteristics and scene consistency rather than to reconstruct personal identity or infer sensitive personal attributes. Any use of this technology to create misleading, harmful, or non-consensual content is strictly against the intended purpose of this work.

To reduce privacy risks, only information necessary for motion modeling and evaluation should be retained in practical use. Personally identifiable cues unrelated to motion generation, such as account information, external profile links, audio identifiers, or other unnecessary metadata, should not be collected, preserved, or redistributed. When real-person data are involved, de-identification measures such as removing direct identifiers and weakening indirect identity cues should be adopted

Table 15: Unified SPA prompt for ChatGPT-4o, ChatGPT-5, and Gemini 2.5 (temperature=0, top-p=1).

SPA Evaluation Prompt

[SYSTEM]
You are a strict scene-prompt evaluator. Score SPA from 0–100 using the user prompt and generated scene.
Goal: assess whether the scene satisfies subject attributes, clothing, environment, visual style, and prompt semantics while avoiding deep conflicts. Check identity, colors, accessories, constraints, style, and logic beyond partial element matching.
Penalize light clothing in snowy scenes, winter outfits on tropical beaches, missing/incorrect key accessories, and local keyword matches that violate narrative, seasonal, or environmental logic.
Scoring criteria:
90–100: faithful, complete details, no conflict;
70–89: generally faithful, minor omissions/unnaturalness;
40–69: detail errors, missing attributes, or scene mismatch;
0–39: serious inconsistency or major conflict.
Briefly reason, then output exactly:
Score: <integer from 0 to 100>
Reason: <one-sentence explanation>

[USER]
User prompt: {user prompt}
Generated scene/image: {generated scene or image description}
Score SPA by checking detail fidelity and logic, not only prompt-element appearance.

whenever possible.

We also stress that datasets and generated results should not be used for re-identification. It is inappropriate to match motion data, generated videos, or derived representations with external information to recover real identities. Similarly, generated outputs should not be presented in ways that falsely imply authenticity, endorsement, or real behavior by a specific person. These safeguards are particularly important because improved controllability and realism may lower the barrier to producing deceptive synthetic media.

In addition, copyright, platform policies, and data-use restrictions must be respected throughout data collection, model training, evaluation, and result dissemination. Original videos, if subject to copyright or platform constraints, should not be redistributed in ways that violate ownership rights or usage terms. Any future release, reuse, or extension of this research should therefore follow strict privacy protection, responsible access control, and clear prohibitions against harmful identity-based misuse.

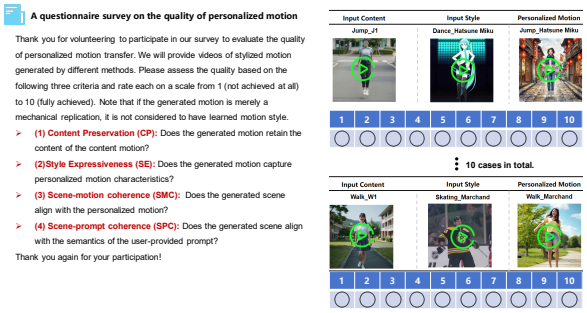

Figure 10: Questionnaire interface of the user study. The volunteers evaluate the results from four aspects: Content Preservation (CP), Style Expressiveness (SE), Scene-Motion Coherence (SMC), and Scene-Prompt Coherence (SPC).

Overall, we believe that personalized motion generation can benefit animation, games, virtual humans, and creative content production, but these benefits should be pursued only under clear ethical boundaries. Future work should continue to strengthen safeguards against misuse, improve privacy-preserving data processing, and promote responsible governance of synthetic human video generation systems.

N Practical application scenarios

N.1 Film and animation production

PersonaScape is highly applicable to film and animation production, where creators often need a character to preserve a distinctive motion style while appearing in different narrative scenes, costumes, and visual atmospheres. Traditional pipelines typically require repeated manual adjustments to motion, background, and character presentation for each shot, which is costly and time-consuming. By learning personalized motion traits from style videos and simultaneously generating scenes that remain consistent with both motion semantics and textual prompts, the proposed framework can support more efficient character animation, storyboard realization, and scene-adaptive performance generation. This is especially valuable for producing expressive digital characters whose behavior matches the story context rather than merely reproducing generic poses.

N.2 Game development and virtual world content creation

In-game production, the same character often needs to perform style-consistent actions across multiple levels, maps, and interaction settings. PersonaScape can serve as a useful tool for generating personalized character motions that remain coherent as the environment changes, which is important for open-world, role-playing, and story-driven interactive experiences. Its ability to jointly control motion style, scene plausibility, and prompt alignment can reduce the burden of hand-crafted animation assets and improve the scalability of content creation. This makes it particularly suitable for dynamic character generation, NPC behavior design, and scene-aware animation synthesis in large virtual worlds.

N.3 Virtual humans and digital avatar generation

The framework is also well-suited to virtual humans, digital idols, and personalized avatars. In these scenarios, users usually expect a virtual character not only to move, but to exhibit recognizable personal style, habitual motion patterns, and semantically appropriate visual presentation. PersonaScape can help generate videos in which a virtual avatar inherits the motion characteristics of a reference performer while adapting to new scenes and prompts. This has direct value for digital hosts, virtual streamers, online education presenters, and branded virtual spokespersons, where consistent behavior style and controllable scene composition are both important.

N.4 Short-video creation and social media content generation

Short-video platforms demand content that is visually appealing, easily producible, and highly personalized. PersonaScape provides a practical solution for this scenario by generating personalized motions from ordinary videos and placing them in plausible new scenes under textual control. Creators can therefore reuse motion patterns, redesign backgrounds, and produce diverse visual narratives without reshooting every sequence. This is particularly meaningful for entertainment videos, dance-style transfer, creator tools, fan-made content, and personalized visual storytelling, where users care about both vivid motion expression and flexible scene editing.

N.5 Advertising and commercial media production

In advertising and commercial content creation, brands often require digital characters to perform customized actions within scenes that align with product themes, visual identity, and campaign narratives. PersonaScape is well aligned with this need because it can preserve the style and expressiveness of a target motion while generating scene references that satisfy prompt-level requirements. For example, a digital model or spokesperson can be made to perform a specific, personalized action within a product-compatible environment. At the same time, clothing details, scene conditions, and the overall visual atmosphere remain semantically coherent. This can improve production efficiency for promotional videos, e-commerce display media, and personalized marketing campaigns.

N.6 AR, VR, and metaverse experiences

Immersive media applications require character motions and surrounding scenes to remain mutually consistent; the user experience can quickly become unrealistic or visually disruptive. PersonaScape is promising for AR, VR, and metaverse environments because it supports semantically aligned motion personalization under arbitrary scenes. Its scene-planning and alignment mechanism can help ensure that generated environments match both the intended action and the textual description. At the same time, its motion modeling module preserves fine-grained style cues. This is useful for interactive avatars, immersive performances, virtual exhibitions, and social virtual spaces where realism, controllability, and personalized expression must coexist.

N.7 Intelligent story visualization and creative AI tools

Another important application lies in intelligent content authoring systems. Writers, designers, and creative practitioners increasingly need tools that can convert narrative descriptions into controllable multimodal results. PersonaScape can function as a core module in such systems by bridging motion personalization and scene generation. It enables a creator to specify what a character should wear, how the character should move, and in what type of environment the action should take place. At the same time, the generated result remains semantically consistent across these factors. This makes

the framework valuable for AI-assisted storytelling, previsualization, concept development, and interactive media authoring.

N.8 Digital cultural production and educational simulation

The framework also has potential in digital cultural communication and educational visualization. In cultural production, it can be used to recreate stylized performances, character-specific motion behaviors, or historically inspired visual scenes in a more controllable and expressive manner. In education and training, it can support the generation of demonstrative motion videos in semantically matched environments, which may benefit fields such as arts education, sports movement presentation, and virtual instruction. Because the model jointly considers motion content, personalized style, and scene semantics, it is better suited than purely pose-driven systems for scenarios that require both realism and contextual coherence.