PAI-Actor: Cinematic Multi-Character Replacement in Dynamic Scenes

Bangxun Tang[1,2*] **Heyuan Gao**[1,3,*] **Yiren Song**[1,4,*,‡]
Guian Fang[1,4] Zijian He[1] Jie Yang[1] Mike Zheng Shou[4,†]
[1]Utopai Studios, USA [2]University of California, Irvine, USA
[3]Nanyang Technological University, Singapore
[4]Show Lab, National University of Singapore, Singapore



Figure 1: Given a source video and a reference image per character, PAI-Actor replaces characters in cinematic dynamic scenes while preserving background, camera motion, pose-accurate reenactment at 1080P resolution. Readers can click and play the video clips in this figure using **Adobe Acrobat**.

Abstract

We present PAI-Actor, a cinematic multi-character animation framework for character replacement in dynamic movie scenes. Unlike conventional animation systems that mainly drive a single static image or a single subject, our goal is to replace and animate multiple characters within real video clips while preserving the original scene dynamics, camera motion, and background content. This setting is particularly challenging because the generated characters must remain consistent with the source performance in motion and interaction, while also matching the surrounding background in lighting, shadow, composition, and overall cinematic appearance. To address this, we formulate multi-character animation as a structure-guided human recovery problem and build a movie-driven training pipeline from high-quality film data. Furthermore, to support practical cinematic production, we introduce a bidirectional-to-autoregressive distillation framework: we first train a bidirectional diffusion transformer for high-quality short-clip generation at 1080P resolution, and then distill it into an autoregressive video-to-video model for efficient inference and longer video generation. Experiments show that PAI-Actor enables high-fidelity multi-character animation with strong scene consistency, cinematic visual quality, and efficient long-form generation. *Project page:* `https://github.com/showlab/PAI-Actor`.

*Equal contribution. ‡ Project leader. † Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Feature comparison between PAI-Actor and existing character-replacement approaches.

Feature	Wan-Animate	MultiAnimate	SCAIL	DOMO	Viggle	Kling O1	MoCha	Ours
1080P	✗	✗	✗	✗	✓	✓	✓	✓
Multi-Character	✗	✓	✓	✓	✓	✓	✗	✓
Long-Video Output (>10 s)	✗	✗	✗	✓	✓	✓	✗	✓
Pose-Driven Control	✓	✓	✓	✓	✗	✓	✗	✓
Background Preservation	✓	✗	✗	✓	✓	✓	✓	✓
Occlusion Handling	✗	✗	✗	✗	✗	✗	✗	✓

1 Introduction

Human animation has achieved impressive progress in recent years, yet cinematic multi-character animation remains a highly challenging problem. In real film production, character replacement is a common and practical demand: for example, motion-capture actors may perform in front of a green screen and later be transformed into cinematic characters or virtual figures. The goal is therefore not merely to animate a single static portrait or reenact one isolated subject. Instead, one often needs to replace and animate one or more characters inside an existing video clip, while preserving the original scene structure, camera movement, actor interactions, and background dynamics. This setting introduces a much more demanding form of controllable generation, where multiple characters must be re-rendered in a way that remains faithful to the source performance and visually coherent with the surrounding movie scene.

This task is exceptionally challenging. Generated characters must preserve the original motion, pose evolution, and multi-person interactions, while strictly maintaining scene consistency—matching illumination, shadows, and composition without altering the background. Furthermore, practical cinematic applications demand high-resolution, temporally stable, long-form videos rather than short clips. As summarized in Table 1, existing pipelines struggle to jointly satisfy high-resolution multi-character generation, long-form scalability, and background preservation. Most notably, they uniformly fail to handle complex occlusions, a critical requirement for realistic dynamic scenes.

In this work, we present PAI-Actor, a movie-driven framework for cinematic multi-character animation. Our key idea is to cast the task as a structure-guided human recovery problem. Given a movie clip, we segment and mask out the original human performers, extract structured motion signals through body pose association and face association, and render the corresponding skeletons on the masked regions as input. The model is then trained to recover the complete human performers in the original video space. This formulation allows the model to learn how humans should be re-rendered inside complex movie scenes, while preserving scene layout, background continuity, and interaction dynamics. More importantly, it naturally supports downstream multi-character replacement, where the original performers can be replaced by new target characters within the same dynamic scene.

To make the framework practical for cinematic production, we further introduce a bidirectional-to-autoregressive distillation strategy, which is a central component of our method. We first train a bidirectional video diffusion transformer as a teacher model to obtain high-quality generation results for short clips at 1080P resolution, typically around four seconds. While this bidirectional model provides strong visual fidelity and temporal modeling capacity, it is still expensive for long-video generation. We therefore distill it into an autoregressive video-to-video generator that supports much longer outputs. In this way, the distilled model inherits the quality of the teacher while significantly improving generation efficiency and temporal scalability. This design makes PAI-Actor suitable not only for high-quality short-clip synthesis, but also for practical long-form cinematic animation.

Our contributions are summarized as follows:

- We propose PAI-Actor, the first movie-driven framework for cinematic multi-character replacement and animation that enables high-resolution, long-video character replacement in dynamic scenes with strong motion fidelity and scene consistency.
- We formulate multi-character animation as a structure-guided human recovery problem, and further introduce a bidirectional-to-autoregressive distillation framework that enables efficient inference and scalable long-form video generation.

- Our framework supports cinematic-quality multi-character animation with high resolution and scalable long-video generation, providing a practical solution for real-world character replacement and virtual production.

2 Related Work

Video Diffusion Models. Diffusion models [20, 55] have become a dominant paradigm for visual generation, with latent diffusion reducing the computational burden of high-resolution synthesis [50] and diffusion transformers further improving scalability [47]. Extending diffusion models from images to videos requires modeling both spatial fidelity and temporal coherence. Early video diffusion systems explore pixel-space or cascaded generation for text-to-video synthesis [22, 21, 54], while subsequent latent-space approaches improve efficiency and resolution through temporal modules, hierarchical generation, or cascaded video latent diffusion [83, 18, 4, 65, 8, 3]. Another line adapts powerful image diffusion priors to video through one-shot tuning or zero-shot temporal attention mechanisms [68, 30]. More recent models introduce space-time architectures, latent diffusion transformers, window attention, autoregressive token modeling, or expert transformers to enhance motion quality and long-range coherence [2, 36, 17, 31, 74]. Flexible conditioning mechanisms further extend diffusion transformers with efficient control interfaces [80]. Despite rapid progress, most video diffusion models are designed for open-ended generation or coarse image-to-video synthesis. They do not directly solve cinematic character replacement, where multiple actors must be re-rendered with pose-accurate motion, stable identities, preserved camera motion, and unchanged dynamic backgrounds.

Animation and Motion-Driven Generation. Animation and motion-driven generation aim to synthesize videos conditioned on a driving signal such as keypoints, skeletons, dense motion, 3D parameters, audio, or a reference video. Classical image animation methods transfer motion from a driving video to a source image using learned keypoints, dense motion fields, articulated regions, or geometric warping [52, 7, 53, 81], while portrait animation methods use semantic or 3D facial parameters to obtain more disentangled control [49, 77]. With diffusion priors, recent human animation methods achieve stronger realism and temporal stability by combining reference appearance encoders, pose guiders, motion modules, disentangled foreground/background control, or 3D body guidance [63, 38, 29, 16, 69, 24, 84, 78]. Recent work also investigates fine-grained portrait, regional, and motion-driven animation, as well as cross-embodiment generation, selective subject representations, and relation-aware control [39, 41, 42, 56, 71, 58, 79, 14]. Motion customization further shows that motion and appearance can be partially decoupled in text-to-video diffusion models [82]. However, most existing animation systems focus on single-person reenactment, talking-head animation, fashion videos, or human dance. They typically animate one reference identity against a simple or weakly changing background, and therefore struggle with multi-actor interactions, occlusions, identity separation, and scene-level consistency in dynamic movie clips.

Video Editing. Video editing with generative models seeks to modify an existing video while preserving its layout, motion, and temporal coherence. Early text-driven editing methods demonstrate that local visual changes can be represented through layered edits, attention control, instruction-following diffusion, inversion, or additional spatial conditions [1, 19, 5, 43, 75]. Built upon these image-editing priors, diffusion-based video editors either train dedicated video diffusion models for structure/content-guided editing [44, 12] or adapt image diffusion models to videos through inversion, cross-frame attention, feature propagation, patch matching, canonical video representations, and image-to-video propagation [48, 34, 64, 6, 72, 13, 46, 32]. Recent studies additionally explore video inpainting, handle-based editing, feature-correspondence-based consistency, layered RGBA video generation, style consistency, and alpha-composited scene decomposition [40, 37, 62, 9, 57, 66]. These methods have substantially improved prompt-based stylization, object replacement, and local attribute editing. Nevertheless, they are not specifically designed for production-level multi-character replacement: edits are often prompt-driven rather than identity-specific, temporal consistency may rely on short clips or propagation assumptions, and multiple independently controlled actors can easily suffer from identity leakage, foreground-background mismatch, or occlusion artifacts. In contrast, our task requires structure-guided recovery of multiple performers inside dynamic cinematic scenes, where the model must preserve background, camera motion, interactions, and lighting while rendering new actors consistently over long videos.

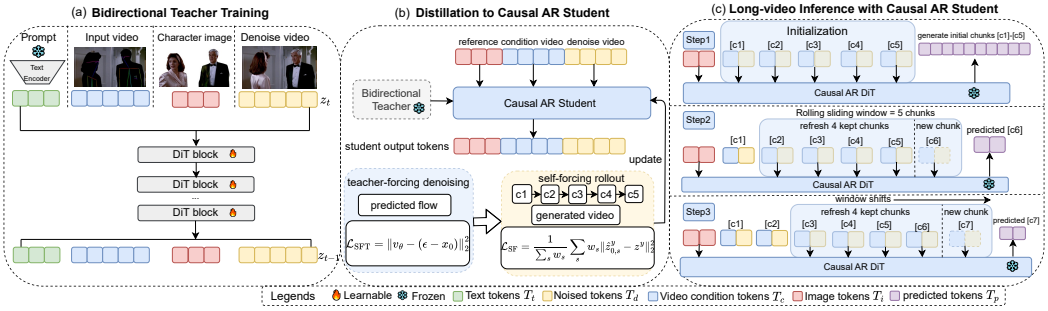

Figure 2: Overview of PAI-Actor. (a) **Bidirectional Teacher**: Denoises target videos conditioned on text, control signals, and references. (b) **Causal Distillation**: A frozen teacher guides the AR student via two-stage teacher- and self-forcing rollout losses. (c) **Inference**: Autoregressive long-video generation using a 5-chunk sliding window.

3 Method

We present **PAI-Actor**, a video-driven framework that preserves source motion and scene dynamics while rendering target actor appearances. As illustrated in Figure 2, our method follows a teacher-to-student design organized as follows: overall architecture formulation (Sec. 3.1), a bidirectional in-context teacher for high-quality short clips (Sec. 3.2), a two-stage autoregressive distillation for scalable long-video generation (Sec. 3.3), KV-cache-based inference procedures (Sec. 3.4), and a structure-guided training-data construction pipeline (Sec. 3.5).

3.1 Overall Architecture

PAI-Actor treats cinematic character animation as a reference-conditioned video editing problem. The input video provides motion, expression, interaction, and scene dynamics, while the reference images specify the actor appearances. Using references from the source actor naturally supports video-driven reenactment; replacing them with references from another identity turns the same pipeline into character replacement.

The overall pipeline has two stages. First, we train a bidirectional diffusion transformer teacher with an in-context formulation. The teacher receives actor reference tokens, a rendered condition video, and noisy target-video tokens in one shared sequence, allowing full spatiotemporal attention across the short clip. This unified reference-conditioned design is also related to visual in-context learning for personalized generation [26]. This design is expensive in memory and computation, but is important for 1080P cinematic quality because the model can jointly reason about actor boundaries, background continuity, occlusion, and temporal context. In practice, full in-context training at 1080P is limited to short clips of about 97 frames, i.e., roughly four seconds, even on large-memory GPUs.

Second, we distill the bidirectional teacher into a causal autoregressive student. The primary motivation for this distillation is to overcome the temporal limitations of the bidirectional model and enable scalable long-form video generation. The student generates the video chunk by chunk, reuses previous chunks through a KV cache, and avoids computing full bidirectional attention over the entire sequence. Consequently, this autoregressive design provides the essential temporal scalability and computational memory efficiency required for practical cinematic production.

3.2 Bidirectional In-Context Teacher

The teacher is trained as a structure-guided human recovery model. For each source video, we remove or degrade the original actor regions, and render body skeletons, face landmarks, and actor masks back onto the same video canvas. The resulting condition video preserves the original background and camera motion, while explicitly indicating where and how each actor should be reconstructed.

This condition design is deliberately video-like rather than a separate sparse control signal. By encoding the masked video, structural guidance, and scene content together, the model can better learn foreground-background fusion, occlusion handling, and boundary-consistent human recovery.

This is especially important for cinematic scenes, where the actor is not isolated from the environment but interacts with lighting, shadows, camera motion, and other characters.

Actor references are encoded separately by a visual image encoder. For actor k, the reference image r^k is mapped to appearance tokens $\mathbf{a}^k$. The teacher receives the concatenated in-context sequence

$$\mathbf{Z}_{\text{in}} = [\mathbf{a}^1, \ldots, \mathbf{a}^K \mid \mathbf{z}^c \mid \mathbf{z}_t^y], \tag{1}$$

where $\mathbf{z}^c$ is the latent condition video and $\mathbf{z}_t^y$ is the noisy target-video latent. We optimize the teacher with the rectified-flow objective:

$$\mathcal{L}_{\text{teacher}} = \mathbb{E}\left[\|F_\phi(\mathbf{z}_t^y, \mathbf{z}^c, \mathbf{A}, t) - \mathbf{v}^*\|_2^2\right], \tag{2}$$

where $\mathbf{A} = \{\mathbf{a}^k\}_{k=1}^K$ and $\mathbf{v}^*$ is the target flow. Full bidirectional attention allows the teacher to use both past and future frames, which improves short-clip quality under occlusion, fast motion, and multi-character interaction.

3.3 Two-Stage Distillation for Long-Video Generation

The teacher produces high-quality short clips, but its full-attention in-context design cannot scale to long cinematic videos. We therefore convert it into a causal autoregressive student through two-stage distillation. Stage 1 adapts the architecture and timestep distribution for causal generation; Stage 2 trains the student on its own rollout distribution to reduce long-horizon drift.

Stage 1: Causal Adaptation. We initialize the student from the bidirectional teacher and replace full temporal attention with block-causal attention. The video is divided into temporal chunks. Tokens inside the current chunk can still attend bidirectionally to preserve local temporal quality, while chunk n can only attend to the reference tokens, condition tokens, and previous chunks. During inference, previous chunks are stored in a persistent KV cache, so the model can extend the video without recomputing the full history.

Causal generation also changes the noise structure. In autoregressive inference, past chunks are already clean and cached, while the current chunk is still noisy. To match this mixed state, Stage 1 samples independent timesteps per frame or per chunk, including clean states with $t = 0$. The student is trained with the same rectified-flow objective under the block-causal mask:

$$\mathcal{L}_{\text{stage1}} = \mathbb{E}\left[\|F_\theta(\mathbf{z}_\mathbf{t}^y, \mathbf{z}^c, \mathbf{A}, \mathbf{t}) - \mathbf{v}_\mathbf{t}^*\|_2^2\right]. \tag{3}$$

This stage gives the student a causal architecture and exposes it to heterogeneous noise states, but the inputs are still derived from ground-truth latents.

Stage 2: On-Policy Self-Forcing. Stage 2 further aligns training with autoregressive inference. Starting from the Stage-1 checkpoint, we roll out the student chunk by chunk over S denoising steps with timesteps $t_s \in [\epsilon, 1]$ (ϵ=10^{-3}; the index s is distinct from the actor index k used above). At step s, the model predicts a clean latent $\hat{\mathbf{z}}_{0,s}^y$, which is re-noised to form the next-step input:

$$\mathbf{z}_{t_{s+1}}^y = (1 - t_{s+1})\hat{\mathbf{z}}_{0,s}^y + t_{s+1}\epsilon_{s+1}. \tag{4}$$

Thus, the student is trained on inputs produced by itself, including KV caches built from its own previous predictions. Since PAI-Actor is a conditional video-to-video task, each condition and reference set has a paired target video. We can therefore directly supervise the rollout with the ground-truth latent:

$$\mathcal{L}_{\text{stage2}} = \frac{1}{\sum_s w_s} \sum_{s=1}^{S} w_s \left\|\hat{\mathbf{z}}_{0,s}^y - \mathbf{z}^y\right\|_2^2, \qquad w_s = \frac{(1 - t_s)^2}{(t_s + \epsilon)^2}. \tag{5}$$

The SNR weight (with the same small ϵ for numerical stability) emphasizes low-noise steps where the clean target is more reliable. We backpropagate through the re-noising trajectory, so early predictions are optimized not only to match the target, but also to produce intermediate states that later steps can refine. This reduces timestep drift and cache-induced chunk drift during long-video generation.

3.4 Inference

At inference time, we first convert the source video into a rendered condition video. We estimate actor masks, body skeletons, and face landmarks, dilate the masks by 5–20 pixels, remove the original actor regions, and render the structural signals into the editable regions. The mask dilation provides a boundary buffer, allowing the model to adapt to reference actors with different body contours, hairstyles, or clothing silhouettes.

The reference images are encoded into appearance tokens and concatenated with the condition-video tokens and noisy target tokens. The causal student then generates the output video autoregressively. For the first chunk, the model denoises from noise conditioned on the reference tokens and the first condition chunk. After each chunk is generated, we perform a clean $t = 0$ cache-writing pass to store its keys and values. Subsequent chunks attend to the global reference tokens, the current condition chunk, and 4 previously cached chunks, forming a sliding temporal window of 5 chunks in total.

This inference process is the same for character animation and character replacement. The difference lies in the reference images: using references consistent with the source performer yields driven animation, while using references from a new identity yields character replacement. The bidirectional teacher can be used for highest-quality short clips, while the causal student is used for long-form generation.

3.5 Training Data Construction

We construct our training data from real-world movie footage. This ensures the model learns from natural character appearances, complex background dynamics, and authentic lighting conditions.

For a given original video sequence $\mathbf{V}$, we first extract the character masks $\mathbf{M}$ and the corresponding skeletal poses $\mathbf{P}$. To formulate the training condition, we mask out the character in the original video to create a black silhouette, effectively removing the original identity while perfectly preserving the background context. The colored skeletal pose is then spatially overlaid onto this masked sequence to construct the control video $\mathbf{V}_{ctrl}$. Formally, each training triplet is $\mathcal{T} = (I_{ref}, V_{ctrl}, V)$, where I_{ref} provides the identity prior (typically sampled from the same video). The control video $V_{ctrl} = (1 - M) \odot V \oplus P$ is constructed via mask-inverted element-wise multiplication ($\odot$) and spatial pose overlay ($\oplus$).

Crucially, during the extraction process, the character masks $\mathbf{M}$ are randomly dilated by 5–20 pixels before generating the control video. This dilation forces the model to learn how to seamlessly inpaint uncertain person-background boundaries and adapt the generated silhouette to novel reference identities with potentially different body shapes. By training on these real-world triplets $\mathcal{T}$, PAI-Actor learns robust, structure-guided character recovery under diverse and challenging cinematic conditions. Because $\mathbf{V}_{ctrl}$ completely obscures the source identity, the model must learn to reconstruct the character conditioned entirely on $\mathbf{I}_{ref}$. This self-supervised formulation naturally enables zero-shot character replacement at inference simply by substituting $\mathbf{I}_{ref}$ with a novel identity.

4 Experiments

4.1 Experimental Setup

Implementation Details. Our model is trained on the proposed **PAI-Actor Dataset** ($\sim$30K movie clips with diverse multi-character interactions, motions, and scenes), and evaluated on the **PAI-Benchmark**. During inference, outputs are generated at 1080P (1920×1056) and 24 FPS, comprising 97 frames ($\sim$4s) for short-clip and extended sequences for autoregressive testing using a temporal chunk size of 5 latent frames and a sliding window of 5 chunks. We finetune the base diffusion transformer in BF16 on 8 H200 GPUs (batch size 1) using LoRA [23] (rank 80) and AdamW [35] (lr=5×10^{-5}). The bidirectional teacher is trained for $\sim$8,000 steps. The causal autoregressive student is then distilled via two stages: (i) teacher-forcing causal SFT ($\sim$1,600 steps), and (ii) on-policy self-forcing with the frozen teacher ($\sim$950 steps). Appendix E details wall-clock and memory costs.

Baseline Methods. We compare PAI-Actor against leading commercial platforms (**Kling O1** [33], **Domo** [11], **Viggle** [61]) evaluated under their default APIs, and state-of-the-art open-source methods

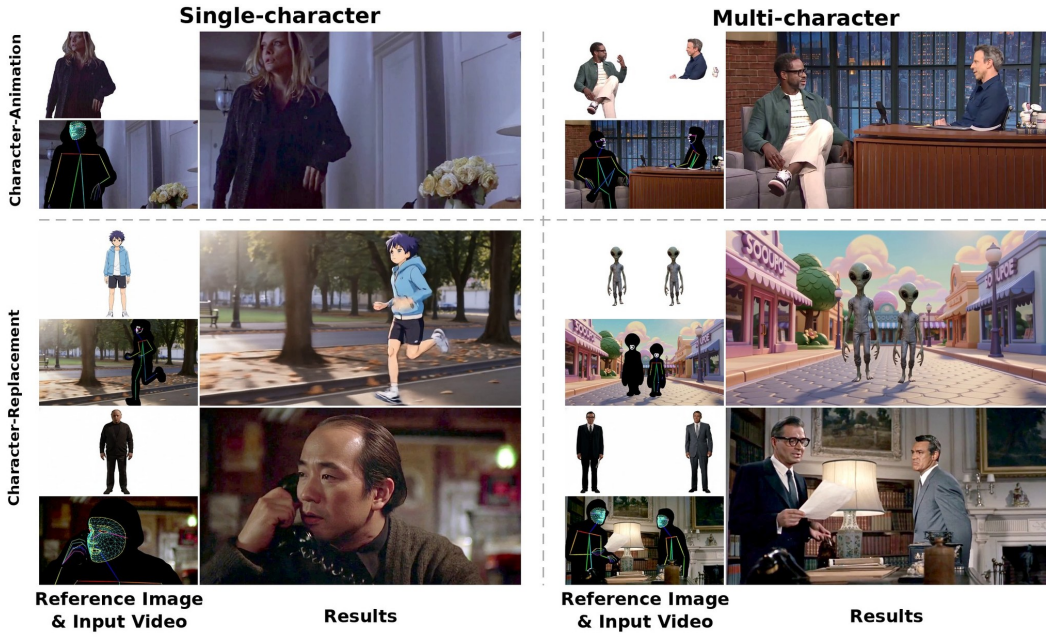

Figure 3: Generation results of PAI-Actor on two tasks. Readers can click and play the video clips in this figure using **Adobe Acrobat**.



Figure 4: Comparison against Wan-Animate, MultiAnimate, SCAIL, DOMO, Viggle, and Kling O1. Readers can click and play the video clips in this figure using **Adobe Acrobat**.

for controllable animation (**Wan-Animate** [10][2], **VACE** [27], **MoCha** [67],). Where applicable, we also include task-specific baselines (**MultiAnimate** [25], **SCAIL** [70]) to comprehensively evaluate multi-character task and foreground-background fusion under comparable conditions. To ensure a fair comparison, all evaluated models are inferred under their respective optimal settings.

Benchmark. We construct the **PAI-Benchmark** comprising three subsets: (1) **Character-Animation Subset** evaluates self-reconstruction on 100 movie clips (disjoint from training). The reference image matches the source actor, providing natural ground-truth to measure motion and appearance fidelity. (2) **Character-Replacement Subset** evaluates zero-shot generalization on 100 unseen movie clips. Here, source actors are replaced with novel reference identities, testing the model's robustness under complex multi-character interactions where no ground-truth exists. For each

[2]Wan-Animate supports single-character task only; for multi-character task, we use it by replacing one character at a time and feeding the output as the next round's driving video. The corresponding rows are marked with * in Tables 2 and 4.

Table 2: Quantitative comparison on the Character-Animation Subset under both single-character and multi-character settings. The last four metrics are evaluated via LLM-based perceptual scoring.

Method	MSE ↓ ($\times 10^{-3}$)		SSIM ↑		LPIPS ↓		Pose-Dist ↓ ($\times 10^{-2}$)		Pose-Match ↑		Motion Cons. ↑		BG Cons. ↑		FG-BG Fusion ↑		Illum. Harmony ↑		Char. Cons. ↑	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
Domo	17.0	19.1	0.647	0.623	0.335	0.360	4.65	4.09	0.674	0.857	0.732	0.791	9.52	9.02	7.63	7.28	7.72	7.85	5.00	3.91
Viggle	12.8	16.6	0.718	0.700	0.263	0.312	4.66	5.18	0.637	0.710	0.627	0.714	8.98	9.15	5.35	4.85	6.10	5.83	3.73	2.63
Kling O1	7.88	8.85	0.712	0.733	0.211	0.227	3.12	2.79	0.807	0.910	0.716	0.759	8.29	8.36	8.43	8.46	7.84	7.97	5.04	5.34
Wan-Animate*	5.30	9.94	0.832	0.766	0.170	0.251	2.65	3.21	0.833	0.898	0.796	0.814	8.88	9.30	8.41	8.31	**8.28**	8.05	5.00	4.44
SCAIL	52.9	47.6	0.852	0.779	0.161	0.217	4.74	4.00	0.755	0.835	0.749	0.736	1.77	1.60	5.20	5.25	6.23	6.66	4.98	4.28
MultiAnimate	26.2	32.6	**0.905**	**0.868**	0.126	0.170	3.78	3.06	0.766	0.891	0.524	0.518	1.67	1.33	4.90	4.35	6.56	6.19	5.46	5.23
VACE	3.22	4.77	0.833	0.836	0.169	0.188	2.63	2.63	0.862	0.907	0.751	0.774	9.77	9.75	8.50	8.31	8.05	7.95	4.94	5.16
Ours-Bidir	**1.80**	**2.85**	0.877	0.864	**0.102**	**0.120**	**2.20**	**2.15**	**0.879**	**0.935**	**0.833**	**0.862**	**9.83**	**9.98**	**8.65**	**9.05**	8.04	**8.66**	**5.72**	**6.10**

Table 3: Quantitative results of `Ours-Causal` on the Long-Video Subset.

Method	MSE ↓ ($\times 10^{-3}$)		SSIM ↑		LPIPS ↓		Pose-Dist ↓ ($\times 10^{-2}$)		Pose-Match ↑		Motion Cons. ↑		BG Cons. ↑		FG-BG Fusion ↑		Illum. Harmony ↑		Char. Cons. ↑	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
Ours-Causal	4.77	4.48	0.880	0.860	0.118	0.129	4.90	1.56	0.857	0.988	0.856	0.903	9.88	9.43	6.25	6.22	7.25	7.97	5.38	4.33

target identity, we first generate a full-body standing portrait and use Gemini 2.5 Flash Image [15] to re-pose it into a configuration approximating a representative pose of the corresponding character in the source video. To ensure a strictly fair comparison, these **co-posed** images are provided as the standard reference input for all evaluated models, including all baselines. (3) **Long-Video Subset** evaluates temporal scalability using 40 movie clips (20 single-character and 20 multi-character), consisting of long-form sequences of $\sim$417 frames ($\sim$17 s at 24 FPS) per clip.

Evaluation Metrics. We evaluate along three dimensions: (1) **Reconstruction Quality:** On the Character-Animation Subset, we report full-frame MSE, SSIM, and LPIPS [76] against the ground truth. For baselines that intentionally alter backgrounds (MultiAnimate, SCAIL), we compute metrics exclusively on masked character regions to ensure fair comparison. (2) **Motion Fidelity:** We extract whole-body keypoints (DWPose [73] fused with NLF-pose [51]) to compute *Pose-Dist Mean* and *Pose-Match*. We also use CoTracker3 [28] to track dense point grids, reporting the cosine similarity of displacement trajectories (*Motion Cons.*). (3) **LLM-based Perceptual Evaluation:** GPT-5.1 [45] and Gemini 3 Pro [59] score *BG Cons.*, *FG-BG Fusion*, *Illum. Harmony*, and *Char. Cons.* on a 1–10 scale. Final scores average both judges' 3-query medians, demonstrating strong agreement (mean Spearman $\rho = 0.75$).

4.2 Results

The generation results are shown in Figure 3. PAI-Actor demonstrates strong temporal stability and identity consistency across frames, even in complex multi-character scenarios. Unlike existing approaches that suffer from identity drift or inconsistent rendering, our method maintains precise character appearance while faithfully reproducing motion dynamics. Our model also can generate diverse character types across various scenes, including anime, aliens, and realistic humans. Additional results are provided in the **appendix**.

Furthermore, our framework scales effectively to multi-character settings. Multiple characters can be independently controlled within the same scene without identity mixing or structural artifacts. This is particularly evident in interaction-heavy scenarios, where baseline methods often fail.

Finally, the autoregressive distillation design enables long-form video generation with stable temporal coherence. Compared to diffusion-based baselines limited to short clips, PAI-Actor produces extended sequences without noticeable degradation on background and motion consistency.

4.3 Comparison and Evaluation

Quantitative Evaluation. Tables 2 and 4 report quantitative comparisons. PAI-Actor achieves state-of-the-art performance across the majority of metrics. While MultiAnimate obtains higher SSIM due to an easier foreground-only evaluation (necessitated by its background-altering task objective), our method is strictly evaluated on the full frame. Despite this harder setting, PAI-Actor maintains highly competitive SSIM and consistently outperforms all baselines in motion fidelity, background consistency, and multi-character identity preservation. Furthermore, on the Long-Video

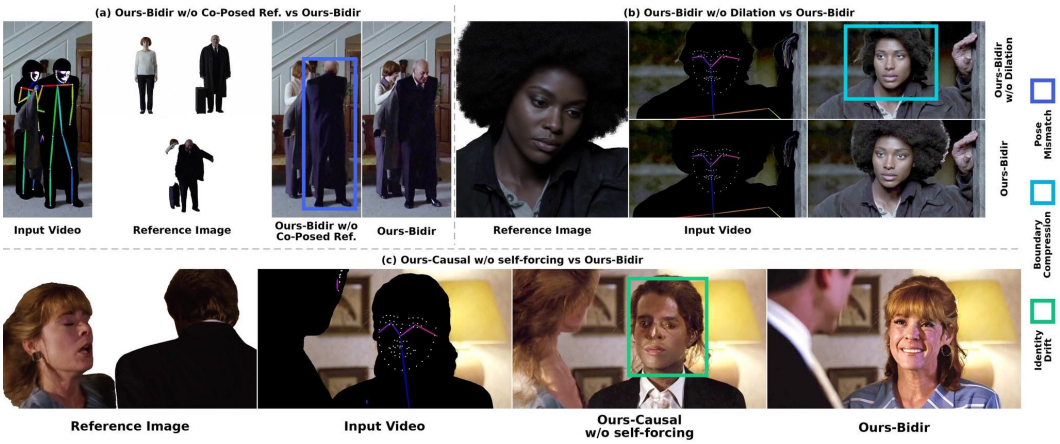

Figure 5: Ablation study on the Character-Animation Subset. The rightmost column is our full model. Readers can click and play the video clips in this figure using **Adobe Acrobat**.

Table 4: LLM-based evaluation (averaged GPT-5.1 & Gemini 3 Pro) on the Character-Replacement Subset.

Method	BG Cons. ↑		FG-BG Fusion ↑		Illum. Harmony ↑		Char. Cons. ↑	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi
Domo	9.33	8.78	7.22	7.30	6.94	7.35	6.39	6.09
Viggle	8.24	8.46	5.04	3.59	6.17	5.54	5.09	3.91
Kling O1	8.41	8.81	8.08	8.20	7.43	8.10	6.53	6.94
Wan-Animate*	8.80	8.84	8.16	8.17	**7.93**	7.47	5.76	4.34
MoCha	8.71	9.25	**8.19**	8.21	7.43	**8.29**	5.85	6.00
Ours-Bidir	**9.72**	**9.85**	8.18	**8.80**	7.79	8.06	**6.76**	**7.03**

Table 5: Ablation study on the Character-Animation subset. Each component was evaluated under single-character and multi-character settings

Method	MSE ↓ ($\times 10^{-3}$)		SSIM ↑		LPIPS ↓		Pose-Dist ↓ ($\times 10^{-2}$)		Pose-Match ↑		Motion Cons. ↑	
	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi	Single	Multi
Ours-Bidir w/o Dilation	2.36	3.14	0.830	0.834	0.117	0.130	2.48	2.29	0.862	0.914	0.809	0.847
Ours-Bidir w/o Co-Posed Ref.	6.24	14.9	0.812	0.765	0.164	0.218	2.99	3.39	0.804	0.848	0.816	0.824
Ours-Causal w/o self-forcing	2.76	3.84	**0.878**	0.860	0.121	0.139	2.73	2.35	0.819	0.917	0.826	0.842
Ours-Bidir	**1.80**	**2.85**	0.877	**0.864**	**0.102**	**0.120**	**2.20**	**2.15**	**0.879**	**0.935**	**0.833**	**0.862**

Subset (Table 3), our `Ours-Causal` student maintains competitive performance even on ∼417-frame sequences (∼17 s at 24 FPS), demonstrating that the proposed bidirectional-to-autoregressive distillation transfers cinematic-quality short-clip generation to scalable long-form video synthesis.

In particular, strong identity consistency scores demonstrate that our method effectively preserves character appearance across frames. Motion fidelity results confirm accurate transfer of motion and expressions. Moreover, superior performance in multi-character consistency validates the effectiveness of our unified in-context formulation.

Qualitative Evaluation. Figure 4 presents qualitative comparisons for the Character-Animation and Character-Replacement Subset. Existing methods commonly suffer from motion misalignment, background degradation, and identity drift. In contrast, PAI-Actor accurately adheres to the driving pose while maintaining strict identity consistency and preserving the original background context.

Notably, our method excels in highly challenging scenarios where baselines typically fail. It robustly handles complex occlusions and seamlessly retains unmasked handheld objects, which are often erroneously erased by other approaches. Furthermore, PAI-Actor successfully captures complex physical effects from the source video, naturally integrating the generated character with dynamic lighting conditions and volumetric light, resulting in highly realistic and immersive outputs.

4.4 Ablation Study

We conduct ablation experiments on the Character-Animation subset (Table 5), where the full `Ours-Bidir` achieves the best or near-best performance across metrics. Qualitatively (Figure 5), dropping the co-posed reference (`w/o Co-Posed Ref.`) deprives the model of aligned spatial guidance; as shown in (a), it struggles to infer the correct body orientation, resulting in pose misalignment. Second, removing mask dilation (`w/o Dilation`) restricts generation to the source silhouette, causing boundary artifacts such as unnaturally cropped hair (b). Finally, the comparison

in (c) reveals that the purely causal model prior to distillation (`Ours-Causal w/o self-forcing`) can occasionally suffer from noticeable facial degradation, whereas the full `Ours-Bidir` maintains better visual consistency.

5 Conclusion

In this work, we present **PAI-Actor**, a cinematic video-driven framework for multi-character animation and character replacement in dynamic scenes. Unlike prior systems limited to single-character or short-clip generation, PAI-Actor targets demanding film-level settings: preserving source motion, actor interactions, camera trajectories, and background dynamics while rendering target identities at 1080P resolution. We achieve this via a structure-guided data engine from real movie footage. We train a bidirectional diffusion teacher for high-quality short clips, which is then distilled into a causal autoregressive student to enable scalable, KV-cache-efficient long-form generation. Built upon the open-source Wan2.2-TI2V-5B backbone, extensive evaluations show that PAI-Actor surpasses both recent open-source approaches and commercial systems in motion fidelity, identity consistency, background preservation, and multi-character controllability. These results demonstrate that high-quality cinematic character replacement is achievable through real-world structure-guided recovery, reference-conditioned in-context modeling, and long-video autoregressive distillation.

Acknowledgments and Disclosure of Funding

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-030).

The authors declare no other competing interests.

References

[1] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022.

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[5] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[6] Duygu Ceylan, Chun-Hao Paul Huang, and Niloy J. Mitra. Pix2video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23206–23217, 2023.

[7] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5933–5942, 2019.

[8] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[9] Xuewei Chen, Zhimin Chen, and Yiren Song. Transanimate: Taming layer diffusion to generate RGBA video. *arXiv preprint arXiv:2503.17934*, 2025.

[10] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025.

[11] DomoAI. Domoai. `https://www.domoai.app/`, 2024. Accessed: 2026-05.

[12] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.

[13] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *International Conference on Learning Representations*, 2024.

[14] Yan Gong, Yiren Song, Yicheng Li, Chenglin Li, and Yin Zhang. Relationadapter: Learning and transferring visual relation with diffusion transformers. *arXiv preprint arXiv:2506.02528*, 2025.

[15] Google DeepMind. Gemini 2.5 Flash Image. `https://deepmind.google/technologies/gemini/`, 2025. Image generation and editing model (codename "Nano Banana").

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

[17] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *European Conference on Computer Vision*, pages 393–411, 2024.

[18] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

[19] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[21] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[22] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pages 8633–8646, 2022.

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

[24] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024.

[25] Yingcheng Hu, Haowen Gong, Chuanguang Yang, Zhulin An, Yongjun Xu, and Songhua Liu. Multianimate: Pose-guided image animation made extensible. *arXiv preprint arXiv:2602.21581*, 2026.

[26] Yuxin Jiang, Yuchao Gu, Yiren Song, Ivor Tsang, and Mike Zheng Shou. Personalized vision via visual in-context learning. *arXiv preprint arXiv:2509.25172*, 2025.

[27] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17191–17202, 2025.

[28] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6013–6022, 2025.

[29] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22623–22633, 2023.

[30] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[31] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation. In *International Conference on Machine Learning*, volume 235, pages 25105–25124, 2024.

[32] Max Ku, Cong Wei, Weiming Ren, Harry Yang, and Wenhu Chen. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024.

[33] Kuaishou Technology. Kling ai. `https://kling.ai/`, 2024. Accessed: 2026-05.

[34] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8599–8608, 2024.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

[36] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[37] Yue Ma, Xiaodong Cun, Yingqing He, Chenyang Qi, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Magicstick: Controllable video editing via control handle transformations. *arXiv preprint arXiv:2312.03047*, 2023.

[38] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 4117–4125, 2024.

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

[40] Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. Follow-your-creation: Empowering 4d creation through video inpainting. *arXiv preprint arXiv:2506.04590*, 2025.

[41] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Leqi Shen, Chenyang Qi, Jixuan Ying, Chengfei Cai, Zhifeng Li, Heung-Yeung Shum, et al. Follow-your-click: Open-domain regional image animation via motion prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6018–6026, 2025.

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

[43] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

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

[45] OpenAI. GPT-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[46] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8089–8099, 2024.

[47] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[48] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023.

[49] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021.

[50] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[51] István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 140032–140065, 2024.

[52] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[53] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13653–13662, 2021.

[54] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[55] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[56] Yiren Song, Cheng Liu, Weijia Mao, and Mike Zheng Shou. Mitty: Diffusion-based human-to-robot video generation. *arXiv preprint arXiv:2512.17253*, 2025.

[57] Yiren Song, Cheng Liu, and Mike Zheng Shou. Omniconsistency: Learning style-agnostic consistency from paired stylization data. *arXiv preprint arXiv:2505.18445*, 2025.

[58] Yiren Song, Xiyao Deng, Pei Yang, Yihan Wang, and Mike Zheng Shou. Omnihumanoid: Streaming cross-embodiment video generation with paired-free adaptation. *arXiv preprint arXiv:2605.12038*, 2026.

[59] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[60] Wan Team, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[61] Viggle AI. Viggle ai. `https://viggle.ai/`, 2024. Accessed: 2026-05.

[62] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. Cove: Unleashing the diffusion feature correspondence for consistent video editing. *arXiv preprint arXiv:2406.08850*, 2024.

[63] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9336, 2024.

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

[65] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025.

[66] Zitong Wang, Hang Zhao, Qianyu Zhou, Xuequan Lu, Xiangtai Li, and Yiren Song. Diffdecompose: Layer-wise decomposition of alpha-composited images via diffusion transformers. *arXiv preprint arXiv:2505.21541*, 2025.

[67] Cong Wei, Bo Sun, Haoyu Ma, Ji Hou, Felix Juefei-Xu, Zecheng He, Xiaoliang Dai, Luxin Zhang, Kunpeng Li, Tingbo Hou, et al. Mocha: Towards movie-grade talking character synthesis. *arXiv preprint arXiv:2503.23307*, 2025.

[68] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[69] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024.

[70] Wenhao Yan, Sheng Ye, Zhuoyi Yang, Jiayan Teng, ZhenHui Dong, Kairui Wen, Xiaotao Gu, Yong-Jin Liu, and Jie Tang. Scail: Towards studio-grade character animation via in-context learning of 3d-consistent pose representations. *arXiv preprint arXiv:2512.05905*, 2025.

[71] Pei Yang, Hai Ci, Yiren Song, and Mike Zheng Shou. X-humanoid: Robotize human videos to generate humanoid videos at scale. *arXiv preprint arXiv:2512.04537*, 2025.

[72] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. *ACM Transactions on Graphics*, 42(6), 2023.

[73] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2023.

[74] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[76] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[77] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8652–8661, 2023.

[78] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

[79] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8069–8078, 2024.

[80] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer. *arXiv preprint arXiv:2503.07027*, 2025.

[81] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3657–3666, 2022.

[82] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *European Conference on Computer Vision*, pages 273–290, 2024.

[83] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022.

[84] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Qingkun Su, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162, 2024.

Appendix for PAI-Actor: Cinematic Multi-Character Replacement in Dynamic Scenes

The appendix is organized as follows:

1. Sec. A: details of the user study, including the interface, evaluation criteria, and randomization.
2. Sec. B: prompt templates and metric definitions for our automated evaluation on the Character-Animation and Character-Replacement Subset.
3. Sec. C: protocol and the scale-invariant trajectory cosine similarity used to measure motion fidelity with CoTracker3.
4. Sec. D: statistical correlation between the automated scores and human preference.
5. Sec. E: computational cost and resources for training and inference.
6. Sec. F: additional results and visualizations.
7. Sec. G: limitations, future work, and analysis of typical failure cases.
8. Sec. H: cumulative-prefix reconstruction analysis for long autoregressive videos.

A User Study Details

We conduct a perceptual user study to compare PAI-Actor against representative open-source and commercial baselines. The study is organized into four tasks that jointly cover both directions of our setting and both single- and multi-character regimes:

- **single-character** $\times$ **character-animation**,
- **single-character** $\times$ **character-replacement**,
- **multi-character** $\times$ **character-animation**,
- **multi-character** $\times$ **character-replacement**.

Each task contains 10 video groups, giving 40 groups per participant in total.

Interface. For each group, the actor reference image and the source performance video are displayed at the top of the page. Below them, the outputs from all compared methods are shown side-by-side as anonymized videos labelled A, B, C, ... The order of the methods is randomized so that participants do not know which video belongs to which method. Synchronized playback controls (*Play All*, *Pause All*, *Replay All*) are provided so that participants can compare temporal consistency across methods.

Evaluation Criteria. For each group, participants answer four single-choice questions, each asking them to pick the single best video for one criterion:

1. **Overall Preference**: which video is the most realistic and visually pleasing overall, given the reference video and the actor reference image.
2. **Motion Fidelity**: which video best reproduces the body pose, gestures, and dynamics of the reference video.
3. **Character Identity**: which video best preserves the actor's face, clothing, and overall appearance from the reference image.
4. **Background Consistency**: which video most faithfully preserves the background of the reference video.

Participants and Statistical Analysis. A total of 16 participants with prior experience in video editing or visual content creation completed all 40 groups, yielding 160 evaluation choices per questionnaire across the four perceptual criteria. Aggregated per-method win rates across the four task settings $\times$ four perceptual criteria are visualized in Figure 6. Our method achieves the highest vote share in every (task, criterion) cell, confirming consistent human preference across all settings and dimensions. Statistical agreement between user-study votes and our LLM-based perceptual

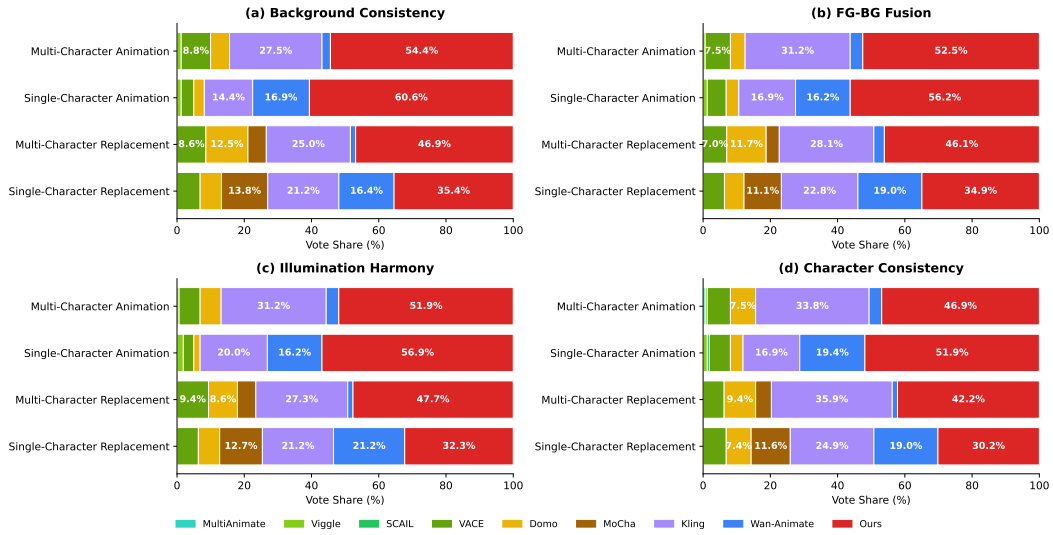

Figure 6: User study results across the four perceptual criteria (**Background Consistency**, **FG-BG Fusion**, **Illumination Harmony**, **Character Consistency**). Each subplot shows the vote share (%) per method on the four task settings (single/multi $\times$ animation/replacement). Values are aggregated over the 16 qualified participants. Our method achieves the largest vote share on every (task, criterion) cell.

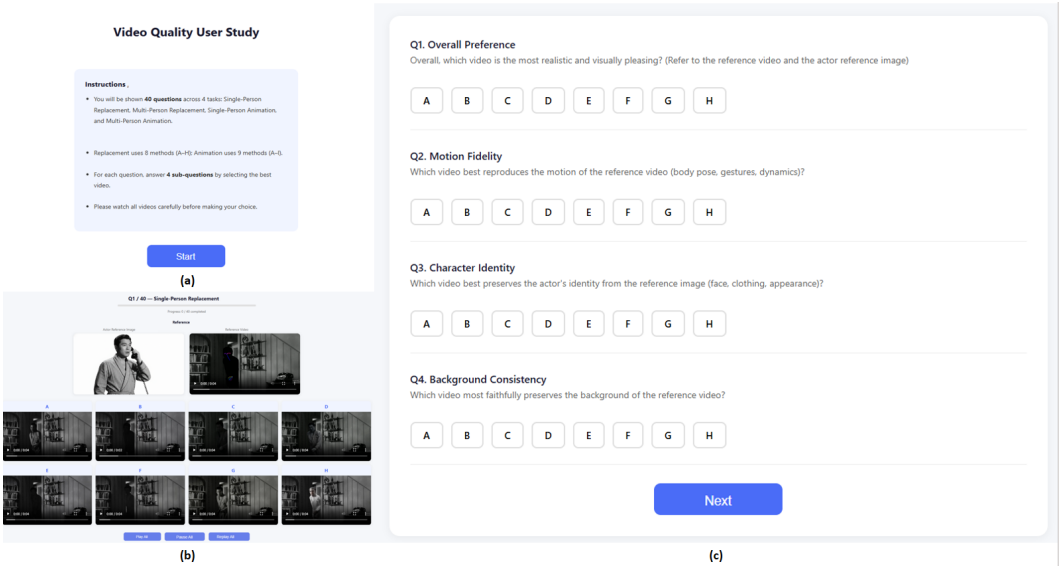

Figure 7: **User study interface and workflow.** (a) Study instructions, (b) side-by-side presentation of a test case with input references and anonymized results, and (c) the questionnaire with four evaluation criteria.

scores is reported in Section D (Table 6), with average Spearman $\rho = 0.554$ on single-character and $\rho = 0.724$ on multi-character settings (all $p < 0.01$), confirming that our automated evaluation protocol agrees with human consensus.

B LLM-based Evaluation Details

We employ two strong multimodal large language models, GPT-5.1 [45] and Gemini-3-Pro [59], as automated perceptual judges for both subsets. To reduce single-judge bias and scoring variance, each

LLM Evaluation Prompts (Part 1/2): BG_Cons & FG-BG_Fusion

Common Setup. *Inputs:* For each metric, the judge receives a sequence of frames sampled uniformly from the generated video and (when applicable) the source / reference. *Output:* JSON `{''score'': <1-10>, ''reason'': ''<sentence>''}`.

Background Consistency (`BG_Cons`) *Role: Expert evaluator for background preservation.*

Inputs: (1) Control sequence: frames from the source video where the foreground subject(s) are replaced with a black silhouette overlaid with colored skeleton lines; (2) Generated sequence: frames from the method output (foreground is the rendered character).

Task: Judge how consistent the generated background is compared to the source video's background. The judge must *ignore the foreground in both sequences* (the silhouette+skeleton on the control side and the rendered character on the generated side) and only evaluate non-foreground background regions.

Criteria:

- Scene identity: same room / street / environment as the control
- Structure & layout: spatial relationships, furniture, walls, no missing or added elements
- Motion & camera: same camera pan/tilt/parallax as the control
- Temporal consistency: stable across frames, no flicker, warping, or texture crawling

Acceptable (do not penalize): minor compression artifacts at mask boundaries; global color/lighting/tone shifts.

Scoring: 10: virtually identical background | 8–9: very faithful, minor structural diff | 6–7: clearly same scene, some changes | 4–5: visible structural differences | 2–3: major changes (missing/added objects) | 1: unrecognizable.

Foreground–Background Fusion (`FG-BG_Fusion`) *Role: Expert evaluator for compositing realism.*

Input: A sequence of frames sampled uniformly from one generated video.

Task: Evaluate foreground–background fusion realism (anterior–posterior fusion).

Criteria:

- Scale & perspective: subject fits the scene geometry and camera viewpoint
- Contact & depth cues: feet contact ground; occlusion boundaries look correct
- Shadows / reflections: consistent with background and subject placement
- Color & tone match: foreground grading matches background (contrast, saturation, grain, sharpness)
- Temporal integration: no floating, jitter, or background leaking onto the subject

Scoring: 10: seamless, looks like one real shot | 8–9: very good, minor grading mismatch | 6–7: moderate artifacts but plausible | 4–5: strong composite feel | 1–3: clearly fake.

Figure 8: **LLM evaluation prompt templates (part 1/2).** Both GPT-5.1 and Gemini-3-Pro receive the same prompts and return JSON. `BG_Cons` compares background regions of the control and generated videos while ignoring the foreground in both. `FG-BG_Fusion` judges compositing realism of the generated video alone.

metric is queried three times per model to obtain a median score, and the final reported score is the average of the medians from both GPT-5.1 and Gemini-3-Pro.

Inputs. For every video, the judges receive (i) the source performance video, (ii) the generated video, and (iii) the actor reference image(s). They are asked to return JSON of the form `{''score'': <1-10>, ''reason'': ''<sentence>''}` for each of four metrics: **Background Consistency** (`BG_Cons`), **Foreground–Background Fusion** (`FG-BG_Fusion`), **Illumination Harmony** (`Illum_Harmony`), and **Character Consistency** (`Char_Cons`). The full prompt templates used for both judges are shown in Figure 8 and Figure 9. Aggregated results are reported in Table 2 and Table 4 of the main paper.

C Details on Motion Fidelity Evaluation

To evaluate the temporal motion consistency between the generated videos and the ground truth (GT) on the **Character-Animation Subset**, we employ CoTracker3 [28] to establish dense, long-term point correspondences. The resulting score is reported as **Motion Cons.** in Tab. 2 of the main paper.

Foreground-only Tracking. Several baselines do not faithfully preserve the original background; comparing trajectories on those regions would unfairly penalize them for an aspect that is unrelated to character motion. To make the motion-fidelity comparison strictly about the character, we therefore restrict the evaluation to the **foreground human region**. Concretely, we apply the binary character mask of the GT video frame-by-frame to both the GT and the generated videos (background pixels are zeroed), and we further constrain the CoTracker3 query points so that they only lie inside the

LLM Evaluation Prompts (Part 2/2): Illum_Harmony & Char_Cons

Illumination Harmony (`Illum_Harmony`) *Role: Expert evaluator for lighting consistency.*

Input: A sequence of frames sampled uniformly from one generated video.

Task: Evaluate illumination harmony between the foreground character and the background.

Criteria:

- Lighting direction: key light on subject matches scene cues
- Color temperature: warm/cool balance matches the background
- Intensity / exposure: subject brightness matches scene exposure (no "cutout" contrast)
- Shadow plausibility: shadows on the subject (and any cast shadow) fit background lighting
- Temporal consistency: lighting does not flicker or change unnaturally across frames

Scoring: 10: perfect lighting match | 8–9: minor color-temp / contrast mismatch | 6–7: noticeable but not catastrophic | 4–5: strong contradiction (direction/temp/exposure clearly off) | 1–3: destroys realism.

Character Consistency (`Char_Cons`) *Role: Strict evaluator for identity preservation.*

Inputs: (1) Target reference image(s); (2) Frames sampled uniformly from the generated video.

Task: Judge how faithfully the generated character matches the target identity. The judge is instructed to be skeptical: passing requires clear, frame-by-frame evidence that the identity is preserved. Sharing only broad casting traits (gender, ethnicity, rough age) is **not** sufficient.

Inspection checklist (across $\geq$3 frames):

- Face geometry: face shape, jawline, cheekbones, eyes, nose, lips, eyebrows
- Skin tone, apparent age, hair color / length / style / hairline
- Facial hair / marks (beard, moles, scars, tattoos)
- Outfit: every garment's type, color, pattern, fit; accessories (glasses, hats, jewelry)
- Build / proportions and temporal stability of identity across all frames

Hard caps (apply the lowest triggered): face geometry differs $\rightarrow$ 4; skin tone or age clearly differs $\rightarrow$ 3; outfit clearly differs $\rightarrow$ 5; both face and outfit differ $\rightarrow$ 3; identity drifts across frames $\rightarrow$ 4; obviously a different person $\rightarrow$ 2; broken/distorted $\rightarrow$ 2; wrong number of people $\rightarrow$ 2; original (unswapped) character $\rightarrow$ 1.

Acceptable (do not penalize): motion blur, mild expression changes, minor texture artifacts, slight color/lighting shifts from scene relighting.

Scoring: 10: indistinguishable from reference | 8–9: unambiguously the same person, only subtle deviations | 6–7: recognizable but with one clear attribute mismatch | 4–5: mixed signals (one major attribute wrong) | 2–3: obviously a different person or severely mis-rendered | 1: completely different / broken / unswapped.

Figure 9: **LLM evaluation prompt templates (part 2/2).** `Illum_Harmony` judges lighting consistency between the rendered character and the scene. `Char_Cons` compares the generated character against the target reference image(s) under a strict-evaluation protocol with hard caps on each attribute mismatch.

foreground region of the first frame. This way, all methods are scored only on how well their characters move, regardless of how their backgrounds look.

Frame Alignment and Sampling. For each video pair, we take the first 80 frames from both the generated and GT videos, and align them to the same canvas of 1920×1056 via cover-resize and center-crop. Before tracking, both videos are then resized to a common compute resolution (short side 384 px). Query points are sampled inside the foreground mask of the first frame (with a uniform 20×20 grid as a fallback when the mask is empty).

Joint Visibility Mask. CoTracker3 returns trajectory coordinates and per-frame visibility confidences for the GT and the generated videos independently. To prevent occluded or out-of-bounds points from skewing the evaluation, a point's frame is considered valid only if its visibility confidence exceeds 0.5 in *both* videos simultaneously, and a point is included only if it has at least two such jointly-visible frames.

Scale-Invariant Trajectory Comparison. A key challenge when evaluating motion fidelity across diverse cinematic scenes is the variation in subject depth, focal length, and absolute motion magnitude. To ensure a fair, normalized comparison across heterogeneous scenes, we evaluate the relative displacement trajectories using **Trajectory Cosine Similarity**. Let $P_t^{(i)} \in \mathbb{R}^2$ and $\hat{P}_t^{(i)} \in \mathbb{R}^2$ be the 2D coordinates of the i-th tracked point at frame t for the GT and generated videos, respectively. The displacement relative to the first frame is $\Delta P_t^{(i)} = P_t^{(i)} - P_0^{(i)}$ (and analogously $\Delta \hat{P}_t^{(i)}$). Let $V^{(i)}$ be the set of jointly-visible frames for point i. We flatten the per-frame 2D displacements over $V^{(i)}$

Table 6: Spearman rank correlation between human preferences (user study) and our LLM-based perceptual scores, across 14 method×task tuples per metric. Higher values indicate stronger agreement.

Metric	n	**Single** ρ	**Multi** ρ
BG Cons.	14	$+0.451$	$+0.611$
FG-BG Fusion	14	$\mathbf{+0.788}$	$\mathbf{+0.871}$
Illum. Harmony	14	$+0.650$	$\mathbf{+0.811}$
Char. Cons.	14	$+0.326$	$+0.604$
Average	–	$\mathbf{+0.554}$	$\mathbf{+0.724}$

into single vectors and take their cosine similarity:

$$S^{(i)} = \frac{\sum_{t \in V^{(i)}} \langle \Delta P_t^{(i)},\, \Delta \hat{P}_t^{(i)} \rangle}{\sqrt{\sum_{t \in V^{(i)}} \|\Delta P_t^{(i)}\|_2^2}\, \sqrt{\sum_{t \in V^{(i)}} \|\Delta \hat{P}_t^{(i)}\|_2^2}}, \tag{6}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. The final motion consistency score is the average over all N_{valid} valid points,

$$\text{Motion Cons.} = \frac{1}{N_{\text{valid}}} \sum_{i=1}^{N_{\text{valid}}} S^{(i)}, \tag{7}$$

which evaluates the directional and structural fidelity of each motion path while remaining invariant to the absolute pixel distance traveled.

D Correlation between Human Preference and Automated Evaluation

To verify that our LLM-based perceptual evaluation (Sec. B) agrees with human consensus, we compute the Spearman rank correlation between the user-study votes and the LLM-based scores at the level of (method × metric × {single, multi}) cells. For each cell we obtain (i) the human win-rate (the percentage of times each method is voted best by participants) and (ii) the average LLM score from Table 2 and Table 4, and pair the two across 14 method×task tuples (six replacement methods plus eight animation methods). Table 6 reports the per-metric Spearman ρ separately for the single- and multi-character settings.

The two strongest agreements are obtained on **FG-BG Fusion** ($\rho = 0.788$ / 0.871 for single / multi) and **Illumination Harmony** (0.650 / 0.811), confirming that the LLM judges share the human notion of compositing realism and lighting plausibility. Background and character-identity consistency, while weaker, still show clear positive correlation with human ratings. Averaged over the four metrics, Spearman $\rho = 0.554$ on the single-character setting and $\rho = 0.724$ on the multi-character setting, indicating that our automated LLM-based protocol provides a faithful proxy for human perceptual preference.



Figure 10: Additional qualitative results for the **Character-Replacement Subset**, showcasing precise identity and background preservation in complex narrative scenes.

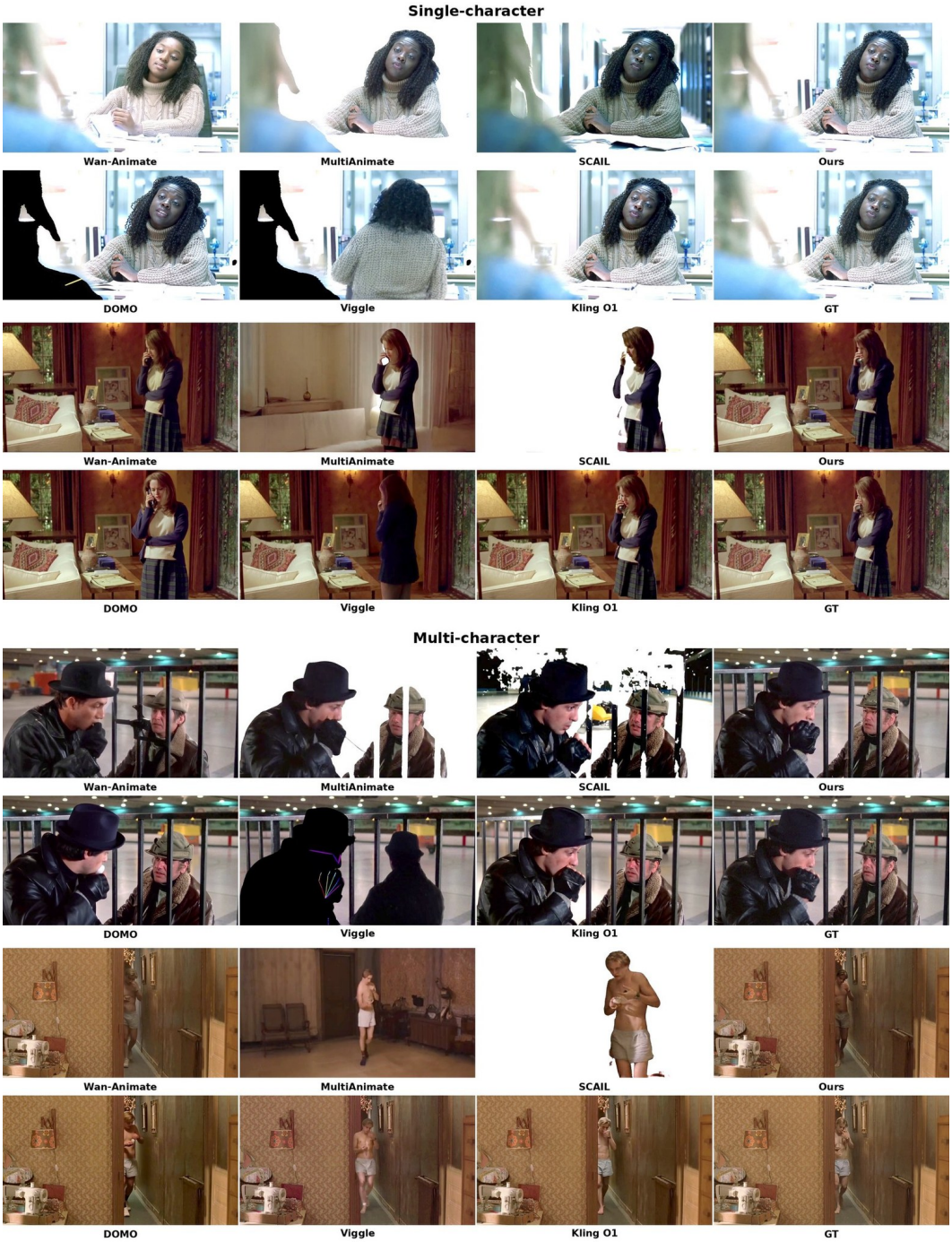

Figure 11: Additional qualitative results for the **Character-Animation Subset**, demonstrating high-fidelity motion replication across both single and multi-character scenarios.

Table 7: Inference efficiency, tested on one NVIDIA H200.

Method	**Resolution** (W×H)	**Steps**	**FPS** ↑
SCAIL	896×512	51	0.15
MultiAnimate	832×480	50	0.13
Wan-Animate	1280×720	20	0.22
MoCha	832×480	50	0.04
VACE	1280×704	50	0.18
Ours-Bidir	1920×1056	50	0.06
Ours-Causal w/o self-forcing	1920×1056	50	0.10

E Computational Cost and Resources

This section reports the additional cost details that are not already covered by Table 7 and the experimental setup of the main paper.

Training Cost. Following the experimental setup, all components are finetuned on top of the open-source Wan2.2-TI2V-5B [60] backbone using LoRA [23] (rank 80) and AdamW [35] (learning rate 5×10^{-5}) in BF16 precision, on 8 NVIDIA H200 GPUs (batch size 1 per GPU), at native 1920×1056 resolution and 97 frames. Beyond what the main paper reports, the breakdown across the three training stages is:

- Bidirectional in-context teacher: $8{,}000$ steps, $\sim$215 hours wall-clock, peak VRAM ≈ 136 GB per GPU.
- Stage 1 causal adaptation: $1{,}600$ steps, $\sim$40 hours, peak VRAM ≈ 92 GB per GPU.
- Stage 2 on-policy self-forcing: 950 steps, $\sim$134 hours, peak VRAM ≈ 114 GB per GPU.

Inference Cost. Inference latency and memory are measured on a single NVIDIA H200 GPU at the resolutions and frame counts reported in Table 7. For a 1920×1056, 97-frame clip:

- **Ours-Bidir** (50 steps): ≈ 27 min wall-clock per clip, peak VRAM ≈ 136 GB (FPS 0.06 in Tab. 7).
- **Ours-Causal w/o self-forcing** (causal student, 50 steps): ≈ 16 min wall-clock per clip, peak VRAM ≈ 98 GB (FPS 0.10 in Tab. 7).

For longer videos beyond a single short clip, **Ours-Causal** extends naturally via KV-cache rolling: a 417-frame ($\sim$17 s) clip at the same resolution costs ≈ 95 min wall-clock with peak VRAM ≈ 100 GB, whereas **Ours-Bidir** cannot scale to this length due to the quadratic cost of full bidirectional attention.

The bidirectional teacher prioritizes cinematic-level visual quality and strict temporal coherence within short clips. The autoregressive student, distilled via two-stage self-forcing, retains comparable per-clip fidelity at lower per-clip cost ($\sim 1.7\times$ faster wall-clock and $\sim 28\%$ less peak VRAM), and uniquely supports KV-cache-based long-video generation that the teacher cannot scale to.

F Additional Results

To further demonstrate the versatility and robustness of PAI-Actor, we provide more comprehensive qualitative results. Our evaluation spans four core tasks categorized into two primary applications: **Character-Replacement Subset** (shown in Fig. 10) and **Character-Animation Subset** (shown in Fig. 11). Each application includes both **single-character** and **multi-character** setups. These results highlight the model's ability to maintain strict identity consistency and temporal coherence, even in highly challenging scenarios featuring complex backgrounds and severe foreground occlusions.

G Limitations and Future Work

While PAI-Actor achieves robust cinematic character replacement, it still faces certain limitations under extreme or highly complex conditions, as illustrated in Figure 12. First, prioritizing high-fidelity generation currently requires extensive denoising steps, leading to higher inference latency compared



Figure 12: Representative failure cases of PAI-Actor. Readers can click and play the video clips in this figure using **Adobe Acrobat**.

to real-time systems. Second, **scale mismatch** between the reference and target views can cause generation artifacts: if the reference image is a full-body shot but the driving video features a close-up angle, the model may lack sufficient facial prior information, resulting in distorted or blurry faces. Third, **severe occlusions and complex multi-character interactions** can cause pose ambiguity. When characters heavily overlap, the extracted skeletal poses may become ambiguous, misguiding the generation process and leading to physically implausible limb rendering. Fourth, in scenarios with **high-speed dynamic motions**, fine details may suffer from slight degradation due to the difficulty of maintaining temporal consistency under extreme pixel shifts. Finally, the `Ours-Causal w/o self-forcing` faces difficulties with **out-of-distribution style transfer**, occasionally struggling to faithfully reproduce specific non-photorealistic artistic styles such as 2D anime. Addressing these efficiency and robustness challenges remains an important direction for our future work.

H Long-Horizon Reconstruction Analysis

We measure reconstruction drift over cumulative prefixes of eight long autoregressive sequences. Table 8 compares the first 25 and 55 seconds of the same sequences, averaging frame-wise metrics within each prefix and then across sequences. For this analysis, SSIM is computed with `pytorch-msssim`. Longer rollouts increase MSE and LPIPS and reduce SSIM, indicating accumulated reconstruction error. Maintaining long-horizon fidelity remains a limitation of the causal model.

Table 8: Cumulative-prefix reconstruction metrics on long autoregressive sequences. Each row averages all frames from the start up to the indicated horizon, rather than an isolated temporal window.

Prefix	MSE$\times 10^{-3}\downarrow$	SSIM $\uparrow$	LPIPS $\downarrow$
First 25 s	12.89	0.870	0.155
First 55 s	20.87	0.859	0.188

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 explicitly state the contributions (movie-driven multi-actor framework, structure-guided human-recovery formulation, bidirectional-to-autoregressive distillation), all of which are validated quantitatively in Section 4 and qualitatively in Figures 3 and 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section G ("Limitations and Future Work") of the supplementary material, including inference efficiency, scale discrepancies, severe occlusions, fast motions, and out-of-distribution style transfer.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper is purely empirical and presents no theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 discloses all training and inference details (LoRA rank, AdamW, BF16 precision, 8 H200 GPUs, 1920×1056 at 24 FPS). Specifically, it details the step counts for all three training stages: $\sim$8,000 for the teacher, and $\sim$1,600/950 for the two-stage student distillation. Section 3 describes the architecture and the distillation procedure. Benchmark splits and evaluation protocols are specified in Section 4.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We do not release code or data with this submission. PAI-Actor is intended for cinematic virtual production; we plan to evaluate release with appropriate documentation, licenses, and safeguards in a future release. Section 4.1 provides sufficient implementation details for an independent re-implementation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 specifies the benchmark splits (Character-Animation and Character-Replacement subsets), hyperparameters (LoRA rank 80, AdamW learning rate 5×10^{-5}, BF16 precision, batch size 1 per GPU), optimizer choice, training schedule (approximately 8,000 teacher steps, 1,600 causal SFT steps, and 950 self-forcing steps), and evaluation resolution / frame count.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: While we do not report per-method error bars over multiple training runs (re-training is computationally prohibitive for cinematic video generation at 1920×1056, 97 frames), several forms of statistical reliability information are reported. (i) For every (clip, method, metric), each judge is queried three times and the median is taken to mitigate single-sample variance. (ii) Two independent multimodal judges (GPT-5.1 and Gemini 3 Pro) are averaged at the per-cell level, and we report their inter-judge Spearman rank correlation (mean $\rho = 0.75$), establishing the reliability of the LLM-based evaluation protocol. (iii) In Section D of the supplementary material, we report per-metric and pooled Spearman correlations between user-study votes and LLM scores (Table 6; mean $\rho = 0.554$ single, 0.724 multi), with all reported correlations at $p < 0.01$.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.1 reports training on 8 NVIDIA H200 GPUs (batch size 1 per GPU) and gives the per-stage training schedule (approximately 8,000 teacher steps, 1,600 causal SFT steps, and 950 self-forcing steps). Section E of the appendix further reports per-stage wall-clock time and peak memory, and Table 7 reports per-method inference resolution, sampling steps, and FPS on a single NVIDIA H200.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The research conforms with the NeurIPS Code of Ethics. The work is conducted for research on cinematic virtual production and character animation. The human-subject component is limited to an anonymous, voluntary perceptual user study (Section A), where participants only compare generated videos along visual quality criteria; no personally identifiable information or sensitive participant data is collected. We also recognize the potential misuse risks of character-replacement technology, and therefore do not release the model, raw movie clips, or identity-related annotations with this submission.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Positive societal impacts of PAI-Actor are discussed in Section 1, including reducing production costs in cinematic virtual production, enabling efficient multi-actor character replacement for film and content creation, and broadening access to professional-grade animation tools for filmmakers. Potential negative impacts include the creation of non-consensual deepfake content, identity impersonation, or misleading media. We mitigate these risks by withholding the trained model, training data, and benchmark videos at submission time.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Given the misuse risks of character-replacement technology, the primary safeguard in this submission is non-release: we do not release the trained model, raw movie clips, identity-related annotations, or benchmark videos. The paper further states that any future release will require additional safeguards such as consent verification, restricted access, watermarking or provenance mechanisms, and compliance with applicable licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All baselines and external assets used in the paper are properly cited and used under their respective licenses. The backbone Wan2.2-TI2V-5B [60] (Apache 2.0) is used as the pre-trained foundation, and the algorithmic methods LoRA [23] and AdamW [35] are used following their original papers. Evaluation tools are used under their open-source licenses: CoTracker3 [28] (CC-BY-NC 4.0), LPIPS [76] (BSD-2-Clause), DWPose [73] (Apache 2.0), and NLF-pose [51] (MIT). Open-source baselines (Wan-Animate [10], VACE [27], MoCha [67], MultiAnimate [25], SCAIL [70]) are used under their respective repository licenses (Apache 2.0 or comparable permissive licenses). Commercial API baselines (Kling O1 [33], Domo [11], Viggle [61]) are accessed under their respective Terms of Service. The training videos are collected only from publicly available or properly licensed sources for research use. We do not release the raw movie clips or identity-related annotations, and the benchmark is distributed only in the form permitted by the corresponding source licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: We do not release new assets (model, dataset, or benchmark) with this submission, so the question is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: We conduct a perceptual user study, with full details in Section A of the appendix, including the study interface (Figure 7), evaluation criteria, randomization protocol, and participant population. Participation was voluntary and anonymous; participants were recruited from individuals with prior experience in video editing or visual content creation. Participants received no monetary compensation

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The paper describes the user study in Section A. The study only asks voluntary, anonymous participants to rate AI-generated and publicly available video clips on perceptual quality dimensions. No personally identifiable information or sensitive participant data is collected, and the task carries no foreseeable physical or psychological risk beyond ordinary online video viewing. Under the requirements of our institution, this minimal-risk perceptual rating task does not require IRB review.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: We use multimodal LLMs in two parts of the pipeline. First, for the Character-Replacement Subset, Gemini 2.5 Flash Image [15] is used to re-pose each generated full-body reference portrait into a pose closer to the corresponding source-video character. Second, GPT-5.1 and Gemini 3 Pro are used as multimodal perceptual judges for background consistency, foreground-background fusion, illumination harmony, and character consistency. This usage is described in Section 4.1 and the LLM-based scores are reported in Tables 2 and 4. In addition, LLMs are used solely for language polishing and grammar checking of the manuscript text; this writing-only use does not affect the methodology, results, or scientific rigor of the work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.