HairWeaver: Photorealistic Hair Motion Synthesis with Sim-to-Real Physics Transfer Guided Video Diffusion

Di Chang[1,2], Ji Hou[1], Aljaz Bozic[1], Assaf Neuberger[1], Felix Juefei-Xu[1], Olivier Maury[1], Gene Wei-Chin Lin[1], Tuur Stuyck[1], Doug Roble[1], Mohammad Soleymani[2], and Stephane Grabli[1]

[1] Meta
[2] University of Southern California
https://boese0601.github.io/hairweaver/
dichang@usc.edu

Abstract. We present HairWeaver, a diffusion-based pipeline that animates a single human image with realistic and expressive hair dynamics. While existing methods successfully control body pose, they lack specific control over hair, and as a result, fail to capture the intricate hair motions, resulting in stiff and unrealistic animations. HairWeaver overcomes this limitation using two specialized modules: a **Motion-Context-LoRA** to integrate motion conditions and a **Style-Alignment-LoRA** to preserve the subject's photoreal appearance across different data domains. These lightweight components are designed to guide a video diffusion backbone while maintaining its core generative capabilities. By training on a specialized dataset of dynamic human motion generated from a CG simulator, HairWeaver affords fine control over hair motion and ultimately learns to produce highly realistic hair that responds naturally to movement. Comprehensive evaluations demonstrate that our approach sets a new state of the art, producing lifelike human hair animations with dynamic details.

1 Introduction

This work addresses the challenging problem of animating a single source image driven by a motion sequence. This technology holds immense promise for applications in virtual reality, next-generation gaming, and the film industry [10, 22]. While recent progress in generative models has enabled plausible human animation, a significant hurdle remains in synthesizing the intricate, non-rigid dynamics of secondary elements, most notably hair. The complex physics and fine-grained visual nature of hair motion are often overlooked by these models, leading to generated humans with fluid-like and unnatural hair motions, shattering the illusion of realism. Building on prior research in human animation [6, 15, 36, 41],

we aim to explicitly model and generate expressive and physically plausible hair dynamics in concert with accurate body pose transfer.

Recent breakthroughs in human image animation [6–8,15,26,31,35,35,36,38, 45,49,50] are largely driven by controlled image-to-video diffusion models [13,47]. These methods typically disentangle appearance and motion by conditioning a pretrained video diffusion model on a reference image for appearance and a sequence of poses (e.g., Pose skeletons [6,15,26,35,38,45,49] or DensePose [41]) for motion control. Appearance is often preserved using attention-based mechanisms that inject features from the reference image into the generation process [4, 5]. Despite achieving impressive control over body posture and identity preservation, these models fundamentally struggle with secondary dynamics. Their network architectures and training paradigms are optimized for overall postural correctness, treating hair as a static texture mapped to the head.

Alternatively, 3D-based models [23,24,32,39,40,46], offer a potential solution to generate diverse hair motions with controllability. However, these methods often lack the photorealism and fine details, producing results that appear synthetic and cannot be directly used to train a diffusion model as previous human video animation methods. We argue that these animation methods falter for a common reason: a critical lack of high-fidelity paired data that contain photorealistic appearance, and diverse, controllable, physically-plausible hair geometry conditions.

To solve this, we propose HairWeaver, a framework that achieves both fine-grained control and photorealism by addressing the data problem head-on. We first generate a small-scale synthetic dataset using a physics-based method [19], providing rich, physically-grounded hair motion and body pose conditions. We then propose a two-stage simulation-to-real physics transfer training strategy to transfer this motion control to a powerful, pre-trained photorealistic video diffusion transformer (DiT) [28] ***without*** transferring the CG domain's non-photorealistic style. This strategy involves two lightweight components: a **Motion-Context-LoRA** module to inject the new motion conditions, and a temporary **Style-Alignment-LoRA** module to overcome the domain gap between CG simulated and photorealistic videos. The Style-Alignment-LoRA is first trained to adapt the DiT to the CG domain. Then, it is frozen, and the Motion-Context-LoRA is trained to learn the mapping from motion conditions to video. Crucially, the Style-Alignment-LoRA is discarded entirely during inference. This strategy allows HairWeaver to leverage the precise motion physics from the CG data while retaining the full photorealistic power of the original foundation model.

As a result, HairWeaver achieves a unique combination of capabilities: the fine-grained, controllable hair physics, and the high-fidelity photorealism, while being trained on only a few samples (around 1k videos). We conduct comprehensive evaluations against state-of-the-art baselines on challenging video benchmarks. HairWeaver consistently outperforms existing methods in both quantitative metrics and qualitative user studies, particularly in the realism of motion and the preservation of fine details. In summary, our main contributions are:

- A novel diffusion-based pipeline, HairWeaver, specifically designed to synthesize expressive and dynamic hair motion for realistic simulation-to-real physics transfer human video animation, with a synthetic guiding signal from a simulator as training data.
- An efficient and effective **Motion-Context-LoRA** that injects hair motion control as additional attention context, preserving the generative power of the video diffusion backbone.
- A two-stage training strategy that uses a temporary **Style-Alignment-LoRA** to learn motion control from synthetic data, which is then discarded at inference to ensure photorealistic generation.

Fig. 1: Sim-to-real hair animation pipeline. From left to right: a CG-rendered reference, its photorealistic counterpart obtained via image stylization, and frames from the resulting animation generated by HairWeaver with physics-based hair dynamics.

2 Related Work

2.1 Diffusion Models for Human Video Animation

The animation of human subjects from a single image has been revolutionized by latent diffusion models [30]. A prevalent paradigm involves conditioning a model on both appearance and motion. To preserve the subject's identity, methods often employ a ReferenceNet architecture or cross-attention mechanisms to inject appearance features from a source image into the generation process [5, 6, 15, 43, 53]. For motion control, pose information derived from a driving video is typically supplied through a spatial guidance module like ControlNet [6, 41, 47].

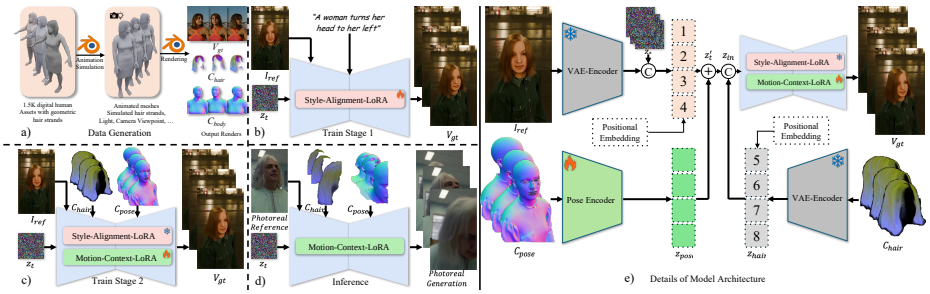

Fig. 2: Overview of HairWeaver pipeline. **a)** We use CG simulation to generate data including human videos with motions $\mathbf{V}_{gt}$, static reference image $\mathbf{I}_{ref}$ (a frame from $\mathbf{V}_{gt}$), pose condition $\mathbf{C}_{pose}$, and hair condition $\mathbf{C}_{hair}$. **b)** During training stage 1, we leverage the a diffusion transformer [28] (DiT) as the backbone model and pretrain the Style-Alignment-LoRA. This training process is conducted in Image-to-Video manner with $\mathbf{I}_{ref}$ and text prompt for $\mathbf{V}_{gt}$. **c)** During training stage 2, we freeze the Style-Alignment-LoRA and finetune the Motion-Context-LoRA with $\mathbf{C}_{pose}$, and hair condition $\mathbf{C}_{hair}$ as additional guidance. **d)** During inference, the Style-Alignment-LoRA is discarded and the trained model generates photorealistic human videos with hair and body motions with photorealistic reference and CG conditions $\mathbf{C}_{pose}, \mathbf{C}_{hair}$ as input. **e)** Details of the model architecture presented in (c). The Pose Encoder integrates the body motions as a trainable residual to the noisy latent. The hair motions are encoded as additional attention context to the DiT blocks by a frozen VAE-Encoder. The only trainable modules are the Pose Encoder and the Motion-Context-LoRA.

Before video models became ubiquitous, many methods followed a two-stage training process: first training a static image generator and then adding a temporal module for video consistency [6, 37]. However, the advent of powerful, large-scale video foundation models has enabled a more streamlined approach. Recent state-of-the-art works, such as MimicMotion [49] and UniAnimate-DiT [38], directly fine-tune Video Diffusion models like Stable Video Diffusion (SVD) [3] or Wan2.1 [34]. By leveraging the strong temporal priors of these pretrained models, they achieve high-fidelity motion transfer and identity preservation. Despite their success in replicating overall body movements, these frameworks are not explicitly designed to capture fine-grained secondary dynamics, a limitation that becomes particularly apparent in complex materials like hair.

2.2 Hair Synthesis

Synthesizing highly convincing, photorealistic human hair — particularly in dynamic scenarios — remains a significant challenge. Over the past several decades, the Visual Effects (VFX), Animation, and Video Game industries have refined the craft of hair synthesis using Computer Graphics (CG), both in offline rendering and real-time settings. CG-based hair offers numerous advantages. By representing hair as hundreds of thousands of individual geometric strands, artists gain a high degree of control and flexibility in the creative process. Skilled artists achieve compelling hair renders by leveraging advanced grooming and simulation

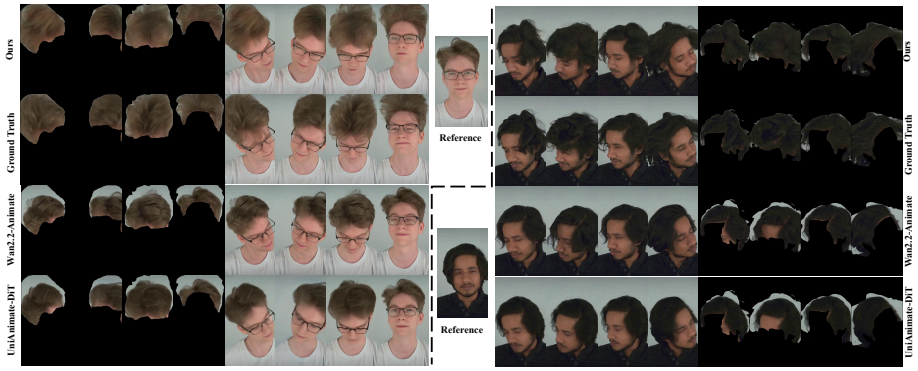

Fig. 3: Visualization of comparison between HairWeaver and the previous state-of-the-art baselines [34, 38]. Our model generates more realistic and diverse hair motions.

tools [1], sophisticated hair shading models [9], and physically-based path tracing algorithms [29]. Despite the relative ease of producing high-quality CG hair, achieving true photorealism remains elusive. Only top-tier studios can generate truly realistic video sequences, and doing so incurs substantial costs. One of our goals is to harness the controllability of CG hair combined with video diffusion models, to achieve photorealistic results on par with, or surpassing, high-end productions, while significantly reducing costs.

As previously discussed, while video diffusion models excel at producing realistic results and provide some degree of control over body pose, they generally lack explicit mechanisms for directing hair synthesis. This limitation poses a significant challenge when creating movies that require detailed and controllable hair dynamics. One notable exception is the recent ControlHair [25] paper, which finetunes a video diffusion backbone [34] to offer directable hair motion. ControlHair utilizes HairStep's hair direction neural extractor [51] to generate its driving signals. However, unlike CG data — which offers perfect alignment between conditioning and target signals, albeit with a domain gap — these extracted maps are only approximate. This limited accuracy ultimately constrains the quality and expressiveness of the resulting model. We choose to rely more thoroughly on CG data and describe our solution to handling its domain gap in section 3.4.

Recently, 3D Gaussian Splatting (3DGS) [17] has emerged as an innovative 3D representation, offering a promising shortcut to achieving photorealism. However, this approach trades off versatility for visual fidelity: the photorealistic appearance is typically derived from static multi-view captures and baked into the representation. As a result, animating or relighting objects represented with 3DGS remains a challenge. Notably, recent work such as Gaussian Haircut [44] has sought to bridge this gap by integrating 3DGS with traditional strand-based hair models, aiming to restore the flexibility inherent in classic approaches. Nevertheless, animating hair within the 3DGS framework continues to be an open research problem.

3 Method

Our method involves two main stages. First, we generate a synthetic dataset by rendering a set of animated digital human assets, where the hair is represented as geometric strands and animated using physics-based simulation, providing rich and physically-grounded hair motion signals paired with body pose, as detailed in Section 3.2. Second, we propose a novel two-step training strategy to adapt a large-scale Video DiT [28] for this task without inheriting the synthetic data's non-photorealistic artifacts. This strategy involves Motion-Context-LoRA, a lightweight module for injecting our dual motion conditions (described in Section 3.3), and Style-Alignment-LoRA, a temporary LoRA that bridges the domain gap during training (described in Section 3.4). Crucially, Style-Alignment-LoRA is discarded at inference, allowing Motion-Context-LoRA to drive the original photorealistic backbone. This results in a model with the fine-grained controllability of our CG data and the high-fidelity realism of the pre-trained DiT. The pipeline of HairWeaver is depicted in Fig. 2.

3.1 Preliminaries

The video diffusion transformer (DiT) [28] represents a significant advancement in video generation, utilizing fully transformer-based models in diffusion frameworks.

The core principle of DiT involves a forward diffusion process that adds Gaussian noise to the data, and a reverse denoising process that learns to reconstruct the original video from the noisy input. Given a video sequence $\mathbf{x}_0 \in \mathbb{R}^{T\times H\times W\times C}$, where T is the number of frames, H and W are spatial dimensions, and C is the number of channels, the forward process is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where β_t and $\mathbf{x}_t$ are the variance schedule controlling the noise level and current data sample with added noise at timestep t, respectively. The transformer backbone in DiTs processes spatiotemporal tokens through multi-head self-attention layers. Specifically, the DiT backbone utilizes a 3D patch embedding strategy, where the input video is divided into non-overlapping spatiotemporal patches. These patches are then linearly projected and augmented with positional embeddings before being fed into transformer blocks. The denoising objective is formulated as:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})\|^2\right], \tag{2}$$

where ϵ is the added noise, ϵ_θ is the learned denoising network parameterized by θ, and $\mathbf{c}$ represents conditioning information. This formulation allows the model to learn a direct mapping from noisy inputs to predicted noise, which can then be subtracted to recover clean video frames.

3.2 Synthetic Hair Motion Generation

While datasets of real subjects with hair motion exist and can be acquired, the low accuracy of hair labeling techniques would hinder the training of diffusion models. Therefore, to construct the hair-specific dataset necessary to train our model, we turn to Computer Graphics (CG) and physics-based simulation [19] and rendering. This approach enables us to create a large corpus of videos with physically plausible hair dynamics, driven by simulated forces, while simultaneously extracting precise, per-pixel ground-truth motion conditions. Each data sample in our dataset is a quadruplet $\{\mathbf{I}_{ref}, \mathbf{V}_{gt}, \mathbf{C}_{pose}, \mathbf{C}_{hair}\}$.

- $\mathbf{I}_{ref} \in \mathbb{R}^{H\times W\times 3}$ is the static reference image, which serves as the appearance condition. In our setup, this is typically the first frame of the video.
- $\mathbf{V}_{gt} \in \mathbb{R}^{T\times H\times W\times 3}$ is the ground truth video sequence, representing the target output. It is rendered using Blender Cycles' physically-based path tracer.
- $\mathbf{C}_{pose} \in \mathbb{R}^{T\times H\times W\times 3}$ is the body pose condition, represented as a sequence of camera-space normal renders augmented with a set of 68 facial landmarks. In these renders, the hair geometry is hidden, making the head and body fully visible. This provides dense structural and orientation information for the body, head and face, guiding the overall motion.
- $\mathbf{C}_{hair} \in \mathbb{R}^{T\times H\times W\times 3}$ is the hair motion condition, represented as a sequence of renders in which the hair is rasterized as a U, V, W buffer, where the U, V values are the texture coordinates of the scalp at the hair strand root locations and W is the normalized arc-length parameter along the hair curve. It is a dense, per-pixel representation that effectively captures the intricate deformations and flow of hair strands, providing a much richer signal than sparse keypoints or optical flow.

This dataset forms the basis for our two-stage training. It provides the explicit signals necessary to train a conditional generation model with motion control that is superior to standard I2V models. Furthermore, our training pipeline (detailed in Section 3.4) is designed specifically to leverage this data without the realism drawbacks common to 3D-based animation methods.

3.3 Motion-Context-LoRA

The Motion-Context-LoRA module is our core contribution for motion control. It is a lightweight adapter designed to inject dual motion conditions into the DiT backbone, ϵ_θ, without altering its pre-trained weights. It features two distinct pathways to process body and hair motion, reflecting the different nature of these signals.

Let the noisy video at timestep t be $\mathbf{x}_t$. This is first encoded by the frozen VAE encoder $\mathcal{E}$ and patchified into spatiotemporal tokens $\mathbf{z}_t = \text{Patchify}(\mathcal{E}(\mathbf{x}_t))$. Our Motion-Context-LoRA module injects conditions $\mathbf{C}_{pose}$ and $\mathbf{C}_{hair}$ directly into this token space as follows:

Body Motion Pathway: The body pose condition $\mathbf{C}_{pose}$ (normal map) provides structural guidance. We use a dedicated Pose Encoder, $\mathcal{E}_{pose}$, composed of convolutional layers, as introduced in UniAnimate-DiT [38]. $\mathcal{E}_{pose}$ is initialized with the weights of the VAE encoder $\mathcal{E}$ but is fine-tuned during training. The pose condition is encoded and patchified into pose tokens:

$$\mathbf{z}_{pose} = \text{Patchify}(\mathcal{E}_{pose}(\mathbf{C}_{pose})). \tag{3}$$

These pose tokens are injected via element-wise addition to the noisy tokens:

$$\mathbf{z}'_t = \mathbf{z}_t + \mathbf{z}_{pose}. \tag{4}$$

This additive injection modulates the features of the noisy input, effectively biasing the denoising process towards the target pose.

Hair Motion Pathway: The hair condition $\mathbf{C}_{hair}$ (UVW map) provides dense, fine-grained motion. This signal is processed using the ***frozen*** VAE encoder $\mathcal{E}$, inspired by DreamActor-M1 [26], and patchified to match the token dimensions:

$$\mathbf{z}_{hair} = \text{Patchify}(\mathcal{E}(\mathbf{C}_{hair})). \tag{5}$$

Unlike the pose condition, the hair tokens are injected by ***concatenating*** them with the modulated noisy tokens along the sequence length dimension:

$$\mathbf{z}_{in} = \text{Concat}([\mathbf{z}'_t, \mathbf{z}_{hair}]). \tag{6}$$

This approach, inspired by unified attention mechanisms [26, 35, 52], allows the DiT's self-attention layers to directly access and utilize the explicit hair motion information as part of the input sequence. The parameters of $\mathcal{E}_{pose}$ and the LoRA weights applied to the DiT's attention blocks constitute the trainable parameters of Motion-Context-LoRA, denoted ϕ_P.

3.4 Style-Alignment-LoRA and Training

Training ϕ_P directly on the CG dataset risks domain overfitting, where the model learns the non-photorealistic style of the simulator. To circumvent this, we propose a two-stage training strategy that separates the learning of domain-specific features from motion-control features.

Stage 1: Style-Alignment-LoRA Pre-training. We first pre-train a Style-Alignment-LoRA LoRA module, ϕ_D, on our synthetic dataset. This stage uses a standard Image-to-Video (I2V) objective, where the model learns to generate video frames $\mathbf{V}_{gt}$ from Reference Image $\mathbf{I}_{ref}$ and text prompts $\mathbf{T}_{prompt}$, describing the content which is captioned by Qwen-2.5-VL [2] from $\mathbf{V}_{gt}$. The objective is to capture the general motion dynamics and visual characteristics of the CG domain within ϕ_D. The loss is:

$$\mathcal{L}_{domain} = \mathbb{E}_{\mathbf{x}_0 \sim \mathbf{V}_{gt}, \epsilon, t, \mathbf{T}_{prompt}} \left[\|\epsilon - \epsilon_{\theta, \phi_D}(\mathbf{x}_t, t, \mathbf{T}_{prompt})\|^2 \right]. \tag{7}$$

This stage adapts the DiT backbone θ with LoRA weights ϕ_D to the synthetic domain.

Stage 2: Motion-Context-LoRA Training. In the second stage, we ***freeze*** both the DiT backbone θ and the pre-trained Style-Alignment-LoRA LoRA ϕ_D. We then introduce our Motion-Context-LoRA module ϕ_P (which includes the Pose Encoder $\mathcal{E}_{pose}$ and its own LoRA weights). The model is now trained on the full quadruplet $\{\mathbf{I}_{ref}, \mathbf{V}_{gt}, \mathbf{C}_{pose}, \mathbf{C}_{hair}\}$, with $\mathbf{I}_{ref}$ serving as the appearance condition. The objective for ϕ_P is to predict the noise ϵ given the rich conditional inputs:

$$\mathcal{L}_{pose} = \mathbb{E}_{\mathbf{x}_0 \sim \mathbf{V}_{gt}, \epsilon, t, \mathbf{c}} \left[\| \epsilon - \epsilon_{\theta, \phi_D, \phi_P}(\mathbf{x}_t, t, \mathbf{c}) \|^2 \right], \tag{8}$$

where $\mathbf{c} = \{\mathbf{I}_{ref}, \mathbf{C}_{pose}, \mathbf{C}_{hair}\}$. Because ϕ_D is frozen, it provides a stable, CG-domain-adapted feature space, forcing ϕ_P to focus exclusively on learning the mapping from the motion conditions ($\mathbf{C}_{pose}$, $\mathbf{C}_{hair}$) to the desired output.

Inference. At inference time, we ***discard*** the Style-Alignment-LoRA LoRA ϕ_D entirely. Inference is performed using only the original DiT backbone θ and our trained Motion-Context-LoRA module ϕ_P:

$$\epsilon_{pred} = \epsilon_{\theta, \phi_P}(\mathbf{x}_t, t, \mathbf{c}). \tag{9}$$

This strategy allows HairWeaver to leverage the precise, fine-grained motion control learned by ϕ_P from the synthetic data, while completely shedding the non-photorealistic domain artifacts captured by ϕ_D. The final generation is thus guided by ϕ_P but rendered using the original, photorealistic generative priors of the DiT backbone.

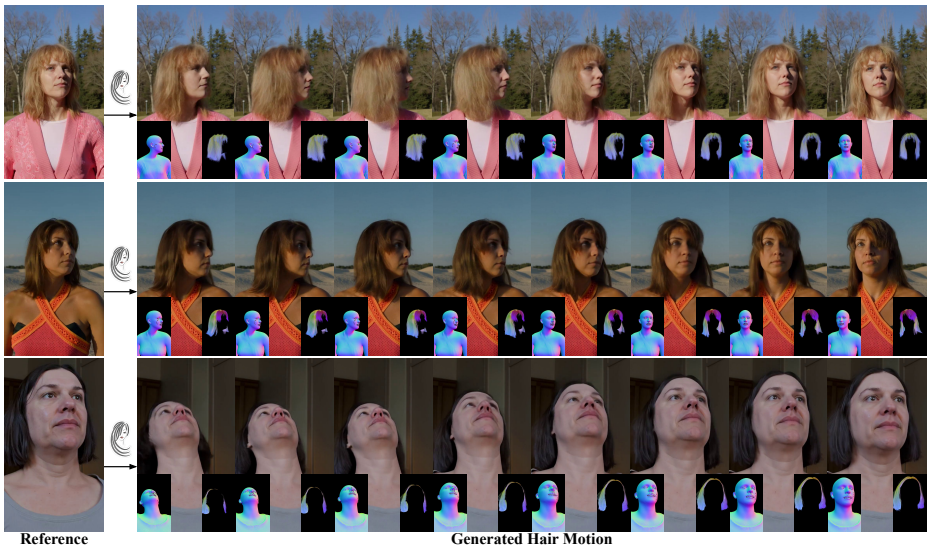

Fig. 4: Photorealistic motions generated by HairWeaver. The reference images are photorealistic human subjects generated by Flux [20].

4 Experiments

4.1 Implementation Details

Dataset Our model is trained on a synthetic human video dataset, which contains 83 minutes of synthetic videos (1,500 videos of 100 frames each with FPS=30) generated from the CG simulator. This dataset was specifically created to provide clean examples of complex hair dynamics, which are critical for our task. All videos and conditions were processed to a resolution of 896 (height) $\times$ 512 (width). This curated dataset ensures our model is exposed to a broad distribution of high-quality motions and expressive details. We test our model on two test sets. We first used the model trained on the synthetic data and tested it on the self-collected hair motion test-set, which is also generated from the simulator but with different (synthetic) human subjects than those in the training data. This test set comprises 10 randomly selected videos featuring diverse hair and body motions. We also evaluate the result on the NeRSemble [18] test-set. This test set comprises 30 videos featuring diverse photorealistic hair and body motions with human subjects facing the camera viewpoint, which are randomly sampled from the hair subset of the whole NeRSemble v2 Dataset. Note that since UVW maps and body normal maps are unavailable for the NeRSemble dataset, we fine-tuned a Motion-Context-LoRA using the alpha map as the hair condition and densepose map [12] as the body condition in this case. The purpose of such an experiment is to demonstrate our framework's flexibility: while it achieves peak precision with UVW maps in CG environments, it can successfully utilize coarser real-world signals while retaining the physical priors learned during synthetic pre-training.

Model Training and Inference Our framework is built upon the LTX-Video-0.9.8 [14], whose DiT backbone weights remain frozen during all training phases. For initialization, both the pose encoder for body and the VAE-Encoder for hair inherit the weights from the pretrained LTX-Video VAE-Encoder, while the layers in Motion-Context-LoRA and Style-Alignment-LoRA are initialized with zeros. We employ a two-stage training strategy. In the first stage, we train only the Style-Alignment-LoRA for 10,000 steps in a standard image-to-video manner with reference image and text prompt captioned by Qwen-2.5-VL [2] as input. In the second stage, we freeze the Style-Alignment-LoRA and train the Motion-Context-LoRA and Pose Encoder for 10,000 steps on our primary task of motion-conditioned image-to-video animation. For all training, we use the AdamW optimizer with a learning rate of $2e^{-4}$. The model is trained on video clips of 97 frames. All experiments were conducted on 8 NVIDIA H200 GPUs with a batch size of 8.

4.2 Evaluations and Comparisons

Metrics. We evaluate our model against state-of-the-art methods using a suite of standard metrics that assess different aspects of generation quality.

Table 1: Comparison to state-of-the-art video animation and video-to-video translation methods on our self-collected hair motion test-set.

Method	Hair					Full Body					Avg Infer Time(s)↓
	SSIM↑	PSNR↑	LPIPS↓	FID↓	cd-FVD↓	SSIM↑	PSNR↑	LPIPS↓	FID↓	cd-FVD↓	
LTX-Video-0.9.8-13B [14]	0.9642	30.3565	0.0468	156.8218	827.5890	0.7885	19.9854	0.2751	113.9513	866.8544	**56**
Wan-2.2-14B [34]	0.9153	24.8174	0.1326	172.6361	802.6791	0.5535	12.7273	0.6345	89.8836	868.1351	476
LTX-Video-ICLora [14]	0.9700	30.6511	0.0365	108.1872	652.3275	0.8665	24.1449	0.1446	45.4541	423.4040	58
UniAnimate-DiT [38]	0.9761	35.4174	0.0304	66.1997	678.9257	0.8724	25.6764	0.1407	45.0817	448.8855	870
Wan-2.2-Animate-14B [34]	0.9758	35.3174	0.0299	59.6534	587.4605	0.8400	25.2877	0.1803	49.4041	461.6013	312
Wan-VACE-[16] + Hair + Body	0.9724	34.6198	0.0348	57.9023	463.2543	0.8514	26.4512	0.1547	51.3914	412.5809	314
HairWeaver	**0.9794**	**37.0347**	**0.0263**	**50.5935**	**434.4682**	**0.8848**	**27.7980**	**0.1127**	**43.2788**	**401.1928**	42

Table 2: Comparison to state-of-the-art video animation methods on NeRSemble [18] test-set.

Method	Hair					Portrait				
	SSIM↑	PSNR↑	LPIPS↓	FID↓	cd-FVD↓	SSIM↑	PSNR↑	LPIPS↓	FID↓	cd-FVD↓
LTX-Video-0.9.8-13B [14]	0.9131	23.40	0.1020	120.74	891.18	0.6871	17.83	0.3878	73.50	703.35
Wan-2.2-14B [34]	0.9150	21.65	0.1051	113.11	652.18	0.6865	17.28	0.3830	51.43	503.07
LTX-Video-ICLora [14]	0.9397	24.95	0.0710	44.11	384.41	0.7635	20.99	0.2683	28.42	277.77
UniAnimate-DiT [38]	0.9350	25.33	0.0810	66.29	477.13	0.7545	20.32	0.3059	52.44	379.12
Wan-2.2-Animate-14B [34]	0.9544	28.75	0.0599	20.48	328.67	0.8011	23.62	0.2030	24.32	314.65
HairWeaver	**0.9670**	**34.34**	**0.0477**	**17.79**	**286.25**	**0.8291**	**26.47**	**0.1763**	**19.43**	**212.61**

- **Reconstruction and Perceptual Quality:** We measure frame-level similarity to the ground truth using Peak Signal-to-Noise Ratio (**PSNR**), Structural Similarity Index Measure (**SSIM**), **L1** distance, and the Learned Perceptual Image Patch Similarity (**LPIPS**) [48].
- **Video Realism and Temporal Consistency:** To evaluate the overall quality and realism of the generated video distribution, we report the Fréchet Inception Distance (**FID**) and the Content-Debiased Fréchet Video Distance (**cd-FVD**) [11]. The cd-FVD metric provides a more robust measure of temporal quality than the standard FVD [33] by disentangling content from motion.

These metrics are widely adopted in recent human video animation research [6, 37, 41, 49], providing a standardized basis for comparison.

Table 4: 30 participants used a forced-choice rating to select videos generated by six methods from eight identities. HairWeaver generates the most realistic videos.

Method	**Average**
LTX-Video-0.9.8-13B [14]	6.9%
Wan-2.2-14B [34]	8.2%
LTX-Video-ICLora [14]	9.0%
UniAnimate-DiT [38]	9.5%
Wan-2.2-Animate-14B [34]	16.4%
HairWeaver	**49.9%**

Quantitative Comparison We benchmark our method, HairWeaver, against several state-of-the-art video generation models that represent distinct architectural approaches. These baselines include: (1) Image-to-Video methods, Wan2.2 [34] and LTX-Video-0.9.8 [14], and (2) Human Video Animation methods with additional control, such as LTX-Video-0.9.8-ICLora [14], UniAnimate-DiT [38], and Wan2.2-Animate [34]. Our primary goal is to improve the quality of dynamic hair generation. While our model is able to leverage a driving signal for the hair that is unavailable to other models, we still find it relevant to conduct quantitative evaluations against those models to 1) validate that,

Table 3: Ablation Analysis. **DiT+Pose Encoder** denotes the DiT backbone is trained with Pose Encoder only without Style-Alignment-LoRA and Motion-Context-LoRA. **DiT+Pose-IC-LoRA** denotes the DiT backbone is trained with In-Context-Lora, the same as LTX-Video-ICLora [14], without Style-Alignment-LoRA and Motion-Context-LoRA. **w/o Style-Alignment-LoRA** denotes the Style-Alignment-LoRA module is removed from the HairWeaver pipeline. **Wan-VACE+UVW Map** denotes we finetune Wan-VACE with the same UVW Map as HairWeaver. **HairWeaver+Alpha Map** denotes we replace the UVW Map with Alpha Map as hair condition and replace the normal map with DensePose as body condition.

Method	Hair					Full Body				
	SSIM↑	PSNR↑	LPIPS↓	FID↓	cd-FVD↓	SSIM↑	PSNR↑	LPIPS↓	FID↓	cd-FVD↓
DiT+Pose Encoder	0.9158	27.4486	0.1214	94.0048	654.8677	0.5623	12.9256	0.6481	91.1008	431.4920
DiT+Pose-IC-LoRA	0.9700	30.6511	0.0365	108.1872	652.3275	0.8665	24.1449	0.1446	45.4541	423.4040
w/o Style-Alignment-LoRA	0.9693	36.7183	0.0236	**49.0808**	447.7628	0.8828	26.5544	0.1184	**38.4230**	416.9271
Wan-VACE+UVW Map	0.9724	34.6198	0.0348	57.9023	463.2543	0.8514	26.4512	0.1547	51.3914	412.5809
HairWeaver+Alpha Map	0.9693	36.7297	0.0236	49.0856	436.9389	0.8710	26.9807	0.1340	46.1248	410.2243
HairWeaver	**0.9794**	**37.6347**	**0.0233**	50.5938	**434.1582**	**0.8948**	**27.7903**	**0.1127**	43.2786	**407.1929**

regardless of control signals available or not, our model is the only one able to get close to ground truth and 2) to set a baseline for future research. Our quantitative results are presented in Table 1, where we compare performance on the self-collected hair motion CG test-set. Note that we use the hair segmentation mask from Matte-Anything [42] to segment the videos and calculate the metrics for hair area. The results demonstrate that HairWeaver consistently and significantly outperforms all baseline models across the relevant metrics. This indicates that our proposed method successfully generates animations with more vivid and expressive dynamics. We also compare the efficiency of HairWeaver to previous methods by reporting average inference time (in seconds) per video sample. Our method is more efficient compared to state-of-the-art animation methods [34, 38]. As mentioned in Section 4.1, we also test the model on the photorealistic NeRSemble [18] test-set. The result is presented in Table 2. We observe that the hair motion quality of HairWeaver outperforms all baselines across the relevant metrics as well.

Qualitative Comparison We qualitatively compare the dynamic hair and body motion generation of HairWeaver on NeRSemble test-set with the animation baselines in Figure 3. We also provide a comparison to an image-conditioned video-to-video method, Wan-VACE [16], with the same conditions as ours in Table 1. This demonstrates that our architectural contributions (Motion-Context-LoRA + Sim2Real strategy) provide genuine improvement beyond simply having richer conditioning. Visual comparison in Figure 7 shows that

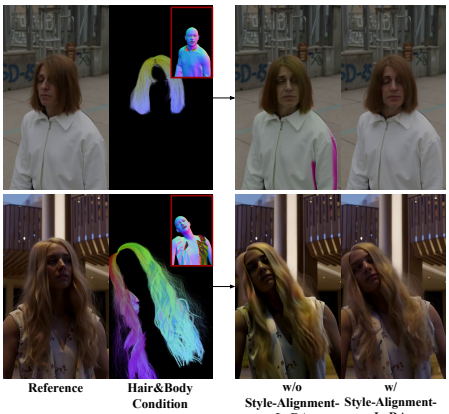

Fig. 5: Ablation analysis of Style-Alignment-LoRA. We visualize the generation without (w/o) and with (w/) Style-Alignment-LoRA. The one w/o such a module cannot preserve reference's appearance when it's a photorealistic image.

Wan-VACE, struggles to maintain identity consistency while translating fine-grained hair details into realistic motions.

We also provide more photorealistic visualizations in Figure. 4, where the reference images are generated by Flux [21], and in Figure 6 where the reference is real-world image. The photorealistic quality depends on the realism of the reference image; with authentic photographs, our model maintains photorealism through the Style-Alignment-LoRA, which prevents CG artifacts from leaking into the generation. More video visualizations are presented in the project page of the supplementary materials.

User Study We conducted a user study to compare HairWeaver with previous work [14, 34, 38]. We collect reference images, ground truth videos, and animation results from previous works and HairWeaver for 8 video subjects. For each subject, we visualize the reference, animation results, and ground truth videos side by side, with the animation results anonymized and randomly ordered. We ask 50 users to choose the **best** methods according to the following two criteria: (1) follows the hair motions and head poses in the Real Video and (2) preserves the identity and appearance of the reference image (person). We present the result in Table 4. We observe that the users prefered HairWeaver more than baseline methods. For more details, please refer to the supplementary material. We also provide more video visualizations in our supplementary website for more results with long hair. These videos can also be directly viewed in subfolder "hairweaver_page/static/flux_more_viz_compressed", e.g., 011.mp4; 027.mp4; 039.mp4; 043.mp4; 044.mp4.

Fig. 6: Motions generated by HairWeaver. The reference image is collected from in-the-wild photoreal source.

Ablation Analysis We provide the ablation analysis in Table 3. Since the self-collected test set has ground-truth video from the CG simulator domain, it is challenging to observe the improvement from Style-Alignment-LoRA in overcoming the domain gap between the CG training data and the photorealistic test data. To this end, we present a visualization of the effectiveness of such a module in Figure. 5. The proposed Style-Alignment-LoRA ability to preserve appearance is evident. To quantify the benefit of richer conditioning, we also provide a comparison on the CG test set between the UVW-conditioned model and the alpha-conditioned variant. This confirms that UVW maps provide superior fine-grained control, while alpha maps still produce results that significantly outperform all baselines — validating that simulation-quality dynamics are learned and expressed even from coarse conditioning.

4.3 End-to-end practical application

In this section, we provide the practical details of our pipeline for generating high quality photorealistic animations with fine-grained control from CG animations. Given a CG animation, we render it using the same simulation and

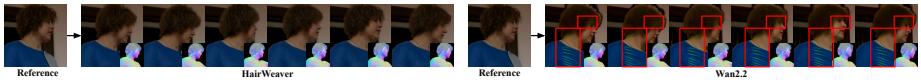

Fig. 7: Comparison to an image-conditioned video-to-video method, Wan-VACE [34], with same hair and body conditions as HairWeaver.

rendering pipeline that was employed previously for generating the CG training dataset. Specifically, we use Blender Cycles' physically-based path-tracer to render 1) the shaded images $\mathbf{V}_{gt}$ and 2) the normal and UVW images ($\mathbf{C}_{pose}$ and $\mathbf{C}_{hair}$ respectively) required for guiding the generation. We then take the first shaded frame and transform it into a more photorealistic reference image using an image-to-image pipeline [27] based on Flux [20] — the resulting image becoming our $\mathbf{I}_{ref}$. Since it is desirable for our image-to-image process to preserve the alignment between the source and the result as much as possible, we add only a small amount of noise to the source image and we provide a detailed prompt which thoroughly describes it. In our implementation, we set the noising timestep to 0.35 and perform 100 denoising steps following the Euler integration scheme. We use the Qwen-2.5 Vision Language Model [2] to generate the detailed prompt from the source image.
In this case, our approach and model are immediately useful for sim2real applications, e.g. human-centric synthetic data generation. UVW maps and body normal maps are readily available from the CG simulator and the input reference can be stylized as photorealistic identities using the flux-based approach, e.g., Figure. 4 illustrates how to achieve the photorealism enhancement. With such synthetic data, users can train models for other downstream tasks, e.g. creating digital avatars with punctual hairstyle editing or try-on.

4.4 Limitations and Future Works

The model's performance can degrade in scenarios with extreme deviations between the driving pose and the reference image. For instance, when the driving video involves significant zooming that leads to large changes in scale, the preservation of the subject's appearance and identity can be compromised. Furthermore, like many generative models, our method struggles to render certain details, particularly when generating consistently accurate and realistic hands. We attribute these challenges primarily to the scope of our training data and the limitations of the backbone model's (LTX-Video [14]) generation ability. Future work could address these by incorporating more diverse training data and employing more advanced backbone models.

5 Conclusion

In this paper, we introduce **HairWeaver**, a novel diffusion-based framework designed to address a critical limitation in human video animation: the synthesis of expressive and realistic hair motion. We identified that existing methods,

while proficient at transferring body pose, often fail to model secondary dynamics, resulting in characters with static, unnatural hair that undermines realism. Our solution leverages a pre-trained video DiT, enhanced with two key components. The **Motion-Context-LoRA** seamlessly integrates hair motion control by adding additional context to the input latent, effectively guiding the animation while preserving the rich generative priors of the backbone model. Concurrently, the **Style-Alignment-LoRA** improves the model's flexibility, enabling generalization to diverse photorealistic identities. By training on a synthetic dataset with simulated videos rich in hair dynamics, our model learns to generate motion that is not only temporally coherent but also physically plausible.

6 Ethics Statement

We clarify that, except for those in the NeRSemble [18] dataset, all characters in this paper are fictional. We strongly condemn any misuse of generative artificial intelligence that could harm individuals or disseminate misinformation. While we acknowledge the potential for misuse in human-centered animation generation, we are dedicated to upholding the highest ethical standards in our research. This commitment includes strict adherence to legal frameworks, respect for privacy, and a focus on promoting the generation of positive and constructive content. We have submitted the legal review for the usage of such data in our work and it's still pending approval from the legal team of our organization.

Acknowledgement

The authors would like to thank Michelle Hill, Lindsey Pollock, Jennie Antonio, and Steffi Liem for their help setting up and operating the motion capture sessions, and Nicky He for processing the motion capture data. We also thank Priyamvad Ravindra Deshmukh and the Meta Metasim team for providing the digital human asset library and pipeline. Soleymani's work was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-25-2-0040. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

1. Houdini, https://www.sidefx.com/, accessed: 2025-11-12
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
5. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023)
6. Chang, D., Shi, Y., Gao, Q., Xu, H., Fu, J., Song, G., Yan, Q., Zhu, Y., Yang, X., Soleymani, M.: Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In: Forty-first International Conference on Machine Learning (2023)
7. Chang, D., Xu, H., Xie, Y., Gao, Y., Kuang, Z., Cai, S., Zhang, C., Song, G., Wang, C., Shi, Y., et al.: X-dyna: Expressive dynamic human image animation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5499–5509 (2025)
8. Chen, Z., Xu, H., Song, G., Xie, Y., Zhang, C., Chen, X., Wang, C., Chang, D., Luo, L.: X-dancer: Expressive music to human dance video generation. arXiv preprint arXiv:2502.17414 (2025)
9. Chiang, M.J.Y., Bitterli, B., Tappan, C., Burley, B.: A practical and controllable hair and fur model for production path tracing. In: ACM SIGGRAPH 2015 Talks, pp. 1–1 (2015)
10. Epic Games: Creating believable digital humans with metahuman and unreal engine 5. https://www.unrealengine.com/en-US/metahuman (2024), accessed: 2025-07-17
11. Ge, S., Mahapatra, A., Parmar, G., Zhu, J.Y., Huang, J.B.: On the content bias in fréchet video distance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
12. Güler, R.A., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: CVPR (2018)
13. Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
14. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., Panet, P., Weissbuch, S., Kulikov, V., Bitterman, Y., Melumian, Z., Bibi, O.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
15. Hu, L.: 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. pp. 8153–8163 (2024)

16. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17191–17202 (2025)
17. Kerbl, B., Kopanas, G., Leimk"uhler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. In: ACM SIGGRAPH 2023 Transactions (2023)
18. Kirschstein, T., Qian, S., Giebenhain, S., Walter, T., Nießner, M.: Nersemble: Multi-view radiance field reconstruction of human heads. ACM Trans. Graph. **42**(4) (jul 2023). https://doi.org/10.1145/3592455, https://doi.org/10.1145/3592455
19. Kugelstadt, T., Schömer, E.: Position and orientation based cosserat rods. In: Symposium on Computer Animation. vol. 11, pp. 169–178 (2016)
20. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
21. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), https://arxiv.org/abs/2506.15742
22. Lee, J.S., Park, M.J., Chen, W.: Photorealistic avatars for immersive social interaction in the metaverse: A survey. In: Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 215–224. IEEE, Shanghai, China (2023). https://doi.org/10.1109/VR55154.2023.000XX
23. Li, Y., Lin, G.W.C., Larionov, E., Bozic, A., Roble, D., Kavan, L., Coros, S., Thomaszewski, B., Stuyck, T., Chen, H.Y.: Self-supervised learning of latent space dynamics. Proceedings of the ACM on Computer Graphics and Interactive Techniques **8**(4), 1–18 (2025)
24. Lin, G.W.C., Larionov, E., Chen, H.y., Roble, D., Stuyck, T.: Neuralocks: Real-time dynamic neural hair simulation. arXiv preprint arXiv:2507.05191 (2025)
25. Lin, W., Li, H., Zhu, Y.: Controlhair: Physically-based video diffusion for controllable dynamic hair rendering. arXiv preprint arXiv:2509.21541 (2025)
26. Luo, Y., Rong, Z., Wang, L., Zhang, L., Hu, T.: Dreamactor-m1: Holistic, expressive and robust human image animation with hybrid guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11036–11046 (2025)
27. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: SDEdit: Guided image synthesis and editing with stochastic differential equations. In: International Conference on Learning Representations (2022)
28. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
29. Pharr, M., Jakob, W., Humphreys, G.: Physically based rendering: From theory to implementation. MIT Press (2023)
30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
31. Song, G., Xu, H., Zhao, X., Xie, Y., Gu, T., Li, Z., Zhang, C., Luo, L.: X-unimotion: Animating human images with expressive, unified and identity-agnostic motion latents. arXiv preprint arXiv:2508.09383 (2025)
32. Stuyck, T., Lin, G.W.C., Larionov, E., Chen, H.y., Bozic, A., Sarafianos, N., Roble, D.: Quaffure: Real-time quasi-static neural hair simulation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 239–249 (2025)
33. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)

34. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
35. Wang, L., Xia, Z., Hu, T., Wang, P., Wei, P., Zheng, Z., Zhou, M., Zhang, Y., Gao, M.: Dreamactor-h1: High-fidelity human-product demonstration video generation via motion-designed diffusion transformers. arXiv preprint arXiv:2506.10568 (2025)
36. Wang, Q., Jiang, Z., Xu, C., Zhang, J., Wang, Y., Zhang, X., Cao, Y., Cao, W., Wang, C., Fu, Y.: Vividpose: Advancing stable video diffusion for realistic human image animation. arXiv preprint arXiv:2405.18156 (2024)
37. Wang, T., Li, L., Lin, K., Lin, C.C., Yang, Z., Zhang, H., Liu, Z., Wang, L.: Disco: Disentangled control for referring human dance generation in real world. arXiv preprint arXiv:2307.00040 (2023)
38. Wang, X., Zhang, S., Gao, C., Wang, J., Zhou, X., Zhang, Y., Yan, L., Sang, N.: Unianimate: Taming unified video diffusion models for consistent human image animation. Science China Information Sciences (2025)
39. Wang, Z., Nam, G., Stuyck, T., Lombardi, S., Cao, C., Saragih, J., Zollhöfer, M., Hodgins, J., Lassner, C.: Neuwigs: A neural dynamic model for volumetric hair capture and animation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8641–8651 (2023)
40. Wang, Z., Nam, G., Stuyck, T., Lombardi, S., Zollhöfer, M., Hodgins, J., Lassner, C.: Hvh: Learning a hybrid neural volumetric representation for dynamic hair performance capture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6143–6154 (2022)
41. Xu, Z., Zhang, J., Liew, J.H., Yan, H., Liu, J.W., Zhang, C., Feng, J., Shou, M.Z.: Magicanimate: Temporally consistent human image animation using diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1481–1490 (2024)
42. Yao, J., Wang, X., Ye, L., Liu, W.: Matte anything: Interactive natural image matting with segment anything model. Image and Vision Computing p. 105067 (2024)
43. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023)
44. Zakharov, E., Sklyarova, V., Black, M., Nam, G., Thies, J., Hilliges, O.: Human hair reconstruction with strand-aligned 3d gaussians. In: European Conference on Computer Vision. pp. 409–425. Springer (2024)
45. Zhang, C., Li, Z., Xu, H., Xie, Y., Zhao, X., Gu, T., Song, G., Chen, X., Liang, C., Jiang, J., et al.: X-actor: Emotional and expressive long-range portrait acting from audio. arXiv preprint arXiv:2508.02944 (2025)
46. Zhang, J.X., Zhu, J., Chen, H., Marschner, S.: Hairformer: Transformer-based dynamic neural hair simulation. arXiv preprint arXiv:2507.12600 (2025)
47. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
48. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)

49. Zhang, Y., Gu, J., Wang, L.W., Wang, H., Cheng, J., Zhu, Y., Zou, F.: Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. arXiv preprint arXiv:2406.19680 (2024)
50. Zhao, X., Xu, H., Song, G., Xie, Y., Zhang, C., Li, X., Luo, L., Suo, J., Liu, Y.: X-nemo: Expressive neural motion reenactment via disentangled latent attention. arXiv preprint arXiv:2507.23143 (2025)
51. Zheng, Y., Jin, Z., Li, M., Huang, H., Ma, C., Cui, S., Han, X.: Hairstep: Transfer synthetic to real using strand and depth maps for single-view 3d hair modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12726–12735 (2023)
52. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. arXiv preprint arXiv:2408.11039 (2024)
53. Zhu, S., Chen, J.L., Dai, Z., Xu, Y., Cao, X., Yao, Y., Zhu, H., Zhu, S.: Champ: Controllable and consistent human image animation with 3d parametric guidance. arXiv preprint arXiv:2403.14781 (2024)