HUMANDIFF: ARTICULATED NOISE DIFFUSION FOR MOTION-CONSISTENT HUMAN VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite tremendous recent progress in human video generation, generative video diffusion models still struggle to capture the dynamics and physics of human motions faithfully. In this paper, we propose a new framework for human video generation, **HumANDiff**, which enhances the human motion control with three key designs: **1) Articulated motion-consistent noise sampling** that correlates the spatiotemporal distribution of latent noise and replaces the unstructured random Gaussian noise with 3D articulated noise sampled on the dense surface manifold of a statistical human body template. It inherits body topology priors for spatially and temporally consistent noise sampling. **2) Joint appearance-motion learning** that enhances the standard training objective of video diffusion models by jointly predicting pixel appearances and corresponding physical motions from the articulated noises. It enables high-fidelity human video synthesis, e.g., capturing motion-dependent clothing wrinkles. **3) Geometric motion consistency learning** that enforces physical motion consistency across frames via a novel geometric motion consistency loss defined in the articulated noise space. HumANDiff enables scalable controllable human video generation by fine-tuning video diffusion models with articulated noise sampling. Consequently, our method is agnostic to diffusion model design, and requires no modifications to the model architecture. During inference, HumANDiff enables image-to-video generation within a single framework, achieving intrinsic motion control without requiring additional motion modules. Extensive experiments demonstrate that our method achieves state-of-the-art performance in rendering motion-consistent, high-fidelity humans with diverse clothing styles.

1 INTRODUCTION

Human video generation has garnered significant attention in recent years, with widespread applications in areas such as AR/VR, animation, video games, and film production. Recently, with the rapid development of diffusion models, especially latent diffusion models Rombach et al. (2021); Blattmann et al. (2023a); Yang et al. (2024), video generation has achieved remarkable progress. However, generating realistic human motion videos is arguably challenging, as it requires not only high-quality appearance details but also smooth temporal consistency. Despite notable advances, existing methods for human video generation still struggle to capture the dynamics and physics of human motion accurately.

Within the domain of motion-controllable human video diffusion models, current approaches typically fall into three categories: (1) Learning a human-specific model from scratch Shao et al. (2024; 2025), which generally requires a tremendous amount of training data and expensive computing resources. (2) Training a specific pose module to guide a pre-trained image diffusion model (e.g., Stable Diffusion Rombach et al. (2021)) for human video generation Zhu et al. (2024); Hu et al. (2024a). However, the image diffusion model cannot accurately capture the temporal motion consistency, and the extra introduced pose module reduces inference efficiency. (3) Leveraging generic video diffusion models with text prompts Yang et al. (2024) or motion-conditioned noise for fine-grained motion control Burgert et al. (2025); Liu and Vahdat (2025), which preserves temporal consistency but struggles to faithfully capture the dynamics and physics of human motions due to the intricate human body structures.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

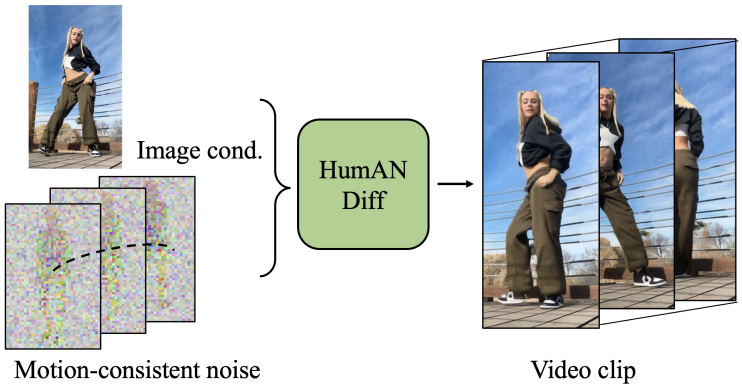


Figure 1: HumANDiff generates temporally consistent human videos from a single image by sampling motion-consistent noise on the UV manifold of a parametric body model (e.g., SMPL Loper et al. (2015)), enabling intrinsic motion control without explicit motion adapters.

To address these limitations, we propose a new method HumANDiff that leverages a video diffusion prior for generalizable human video synthesis which avoids the limitations of expensive training on a large volume of data. In addition, HumANDiff requires no modifications to the base model's inference architecture, i.e., it does not require a specific pose module at inference time, which allows HumANDiff to fully leverage the video diffusion prior for intrinsic motion control. Note that a lightweight motion decoder is introduced as an auxiliary training objective only and is discarded at inference, eliminating the need for additional modules during inference as shown in Fig. 1.

We achieve this by correlating the spatiotemporal distribution of latent noise, similar to the noise warping approaches Burgert et al. (2025); Liu and Vahdat (2025); Chang et al. (2025). However, we illustrate that the existing warping approaches (e.g.optical flow based warping Burgert et al. (2025)) do not work for human motions due to the intricate human body structure and non-rigid motions. Instead, we enhance the video diffusion model to capture the dynamics of human motion with three key designs: **1)** Articulated motion-consistent noise sampling; **2)** Joint appearance-motion learning; and **3)** Geometric motion consistency training.

Firstly, in contrast to existing video diffusion methods that learn human motions from unstructured random noises ignoring the human body structure and temporal motion consistency, we propose to model motions in a structured unified space, i.e., the surface manifold of a 3D parametric human body mesh (e.g.SMPL Loper et al. (2015)) which can be rendered as texture UV map in 2D space. We sample Gaussian noise on the texture noise UV map, which inherits articulated body topology priors for motion generalization and is shared across motion frames for temporal consistency. This allows us to replace the unstructured random Gaussian noise with structure-aware motion-conditioned and spatially and temporally consistent noise, making it easier to learn.

Secondly, different from conventional pixel reconstruction objective that biases models toward appearance fidelity at the expense of motion coherence Hu et al. (2024b); Chefer et al. (2025), we propose a joint appearance-motion learning strategy that encourages the video diffusion model to jointly predict pixel appearances and corresponding motions from the articulated noise space. In particular, in addition to the video decoder, we propose a motion decoder to predict physical motions from denoised latents. This physical motion learning enables high-fidelity human video synthesis. We illustrate that the joint appearance-motion learning strategy can improve the motion coherence for faithful motion control.

Thirdly, we enhance the standard training objective that focuses on the pixel reconstruction quality while failing to capture the temporal motion consistency across frames, by proposing a geometric motion consistency training strategy. In particular, we propose a novel geometric motion consistency loss defined in the articulated noise space, which enforces motion consistency across frames while maintaining per-frame pixel quality. Experimental results demonstrate that the motion consistency supervision can significantly improve the temporal motion consistency by a large margin.

The core contribution of HumANDiff lies in an **articulated noise representation** enabled by a new **diffusion formulation**. Instead of independently sampling Gaussian noise in image space and subsequently warping it, HumANDiff defines shared stochastic variables on a canonical articulated

human surface, i.e., the SMPL UV manifold. These shared variables are mapped to every target frame according to the driving SMPL motion, allowing the diffusion process itself to inherit **persistent semantic correspondences** and **3D articulated motion priors** before denoising begins.

HumANDiff is built upon a video diffusion model, and it leaves the inference architecture unchanged, matching the base model's memory usage and runtime. This enables scalable training for controllable human video generation.

In summary, our contributions are:

1) An articulated noise representation that maps shared Gaussian variables from the SMPL UV manifold to each frame, enabling model-agnostic, spatially and temporally consistent motion control.

2) A joint appearance-motion learning strategy that improves the motion coherence by jointly predicting pixel appearances and corresponding physical motions from the articulated noises in training.

3) A novel motion consistency training strategy that proposes a geometric motion consistency loss to enforce physical motion consistency while maintaining per-frame pixel quality.

4) Adapter-free inference with unchanged memory and runtime, validated by state-of-the-art results.

2 Related Work

Our approach is closely related to several subfields of visual computing. Below, we discuss some of the most relevant connections.

Diffusion-Based Image and Video Synthesis. Diffusion models have achieved remarkable results in high-quality image Ho et al. (2020); Song et al. (2020); Rombach et al. (2022); Zhang et al. (2023); Wang et al. (2024a) and video synthesis Lei et al. (2024); Shao et al. (2024); Zhu et al. (2024); Wang et al. (2024b); Işık et al. (2023); Zhang et al. (2025). Latent diffusion models, such as Stable Diffusion Rombach et al. (2021), improve efficiency by operating in a compact latent space rather than directly on high-resolution pixels.

Video generation requires modeling both spatial fidelity and temporal consistency. Recent methods address this by adding temporal layers to pre-trained image diffusion models Singer et al. (2022); Guo et al. (2023); Wu et al. (2023); Wang et al. (2024c); Blattmann et al. (2023b) or adopting transformer-based architectures for sequential video modeling Yan et al. (2021); Yu et al. (2023a). For example, Stable Video Diffusion (SVD) Blattmann et al. (2023b), CogVideoX Yang et al. (2024) and Wan2.1 Wan et al. (2025) extend latent diffusion to videos by integrating temporal modules and fine-tuning on large-scale video datasets, providing an open-source framework for high-fidelity video generation. CogVideoX Yang et al. (2024) and Wan2.1 Wan et al. (2025) adopt diffusion transformer architecture, and encode videos into a compact latent space using a 3D video VAE and incorporates temporal modeling, enabling efficient and scalable video synthesis while preserving visual quality and temporal coherence. In this work, we extend CogVideoX and Wan2.1 for motion-guided human video generation from single images with fine-grained motion control.

Human Animation Video Generation. Human video generation aims to animate static images into realistic, temporally coherent sequences. Methods can be broadly categorized into 3D rendering and 2D image-based approaches.

3D rendering approaches Peng et al. (2021b); Liu et al. (2021); Peng et al. (2021a); Noguchi et al. (2021); Su et al. (2021); Chen et al. (2021) combine geometry reconstruction with view synthesis using volume rendering, but these methods are computationally intensive. Hybrid 3D-GAN strategies Niemeyer and Geiger (2021); Zhou et al. (2021); Gu et al. (2021); Or-El et al. (2021); Hong et al. (2021); Chan et al. (2021); Hu et al. (2022; 2023) improve efficiency by combining geometry-aware rendering with generative models. Gaussian Splatting approaches Qiu et al. (2024; 2025) allow real-time rendering of human avatars.

2D approaches use GANs to map poses—represented as skeletal renderings Chan et al. (2019); Siarohin et al. (2018); Pumarola et al. (2018); Kratzwald et al. (2017); Zhu et al. (2019); Wang et al. (2018a), dense meshes Liu et al. (2019); Wang et al. (2018a); Liu et al. (2020); Sarkar et al. (2020); Neverova et al. (2018); Grigor'ev et al. (2019), or joint heatmaps Ma et al. (2017); Aberman et al. (2019); Ma et al. (2018)—to realistic images. Temporal stability can be improved by incorporating

SMPL priors Loper et al. (2015); Prokudin et al. (2021); Hu et al. (2021b). However, these methods do not explicitly model 3D geometry and struggle with self-occlusions.

Diffusion-based human video generation methods, such as Human4DiT Shao et al. (2024) and ISA4D Shao et al. (2025), achieve higher quality than GAN-based approaches but often require massive training data and computational resources. To reduce these requirements, some methods train pose modules to guide pre-trained image diffusion models (e.g., Stable Diffusion Rombach et al. (2021)). These methods use either 2D keypoints Hu et al. (2024a; 2025); Zhang et al. (2024) or 3D SMPL meshes Wang et al. (2024b); Xu et al. (2024); Zhu et al. (2024); Shao et al. (2024) as guidance. Keypoint-based methods, such as OpenPose Cao et al. (2019) and DWPose Yang et al. (2023), control motion with sparse skeletons, while SMPL-based methods better capture surface geometry, occlusions, and body contours Wang et al. (2024b); Zhu et al. (2024); Shao et al. (2024). While SMPL guidance improves realism, it may reduce generalization and introduce shape leakage. Furthermore, image-based diffusion models struggle with temporal consistency, and adding external pose modules reduces inference efficiency.

In contrast, we build our method on video diffusion models to leverage their temporal coherence and generalization while incorporating SMPL priors to improve pose-guided generation.

Noise Warping for Video Diffusion. Noise warping has been proposed to generate temporally consistent videos Burgert et al. (2025); Chang et al. (2025); Liu and Vahdat (2025). These methods control motion by modifying the noise input, e.g., Go-With-The-Flow Burgert et al. (2025) warps noise according to optical flow. However, optical-flow-based warping often fails for complex human motions due to the non-rigid and articulated nature of the human body. To address this, we propose articulated 3D noise warping, specifically designed for human motion, enabling motion-consistent and temporally coherent video generation.

3 METHODOLOGY

We conduct experiments on latent video diffusion models (LVDM) Yang et al. (2024). In this section, we first introduce the preliminaries of LVDM, and then introduce how to model human motions in a structured unified space by proposing articulated noise warping, which attaches Gaussian noise to the human body surface for motion-consistent noise sampling. Then, we describe how to fine-tune a video diffusion model with the articulated noise warping for controllable human video synthesis. For faithful motion control, we propose a joint appearance-motion learning strategy and motion consistency supervision as shown in Fig. 2.

3.1 PRELIMINARIES

Latent video diffusion models (LVDM) Yang et al. (2024); Blattmann et al. (2023a) learns a denoising process to simulate the probability distribution within the latent space. To improve computational efficiency, a video $\boldsymbol{x_0}$ is transformed from the pixel space into a latent space feature $\boldsymbol{z_0} = \mathcal{E}(\boldsymbol{x_0})$ with a 3D Variational Autoencoder (VAE) encoder Yang et al. (2024) $\mathcal{E}$. $\boldsymbol{x_0} \in \mathbb{R}^{f\times H\times W\times 3}$, $\boldsymbol{z_0} \in \mathbb{R}^{f\times h\times w\times C}$, where f is the number of frames, C is the number of channels, H, W and h, w are the height and width of input video and latent feature, respectively.

Gaussian noise is iteratively added to $\boldsymbol{z_0}$ at various timesteps t as $q(\boldsymbol{z_t} \mid \boldsymbol{z_{t-1}}) = \mathcal{N}(\boldsymbol{z_t}; \sqrt{1-\beta_t}\boldsymbol{z_{t-1}}, \beta_t I)$, where β represents a sequence schedule. The denoising process is defined as an iterative Markov chain that progressively denoises the initial Gaussian noise $z_T \sim \mathcal{N}(0, I)$ into a clean latent space z_0. The LVDM denoising function is commonly implemented with U-Net or Transformers, which is trained to minimize the loss of mean square error as

$$\mathcal{L}_{diff} = \mathbb{E}_{\boldsymbol{z_t},\boldsymbol{c},t,\epsilon\sim\mathcal{N}(0,I)}\left[\|\boldsymbol{\epsilon} - \epsilon_\theta(\boldsymbol{z_t}; \boldsymbol{c}, t)\|_2^2\right] \tag{1}$$

where ϵ_θ represents the parameterized diffusion model for predicting noise and c denotes an optional conditional input (e.g.image or text). Subsequently, the denoised latent feature is decoded into video frames (appearance) using the VAE Decoder $\mathcal{D}_A$.

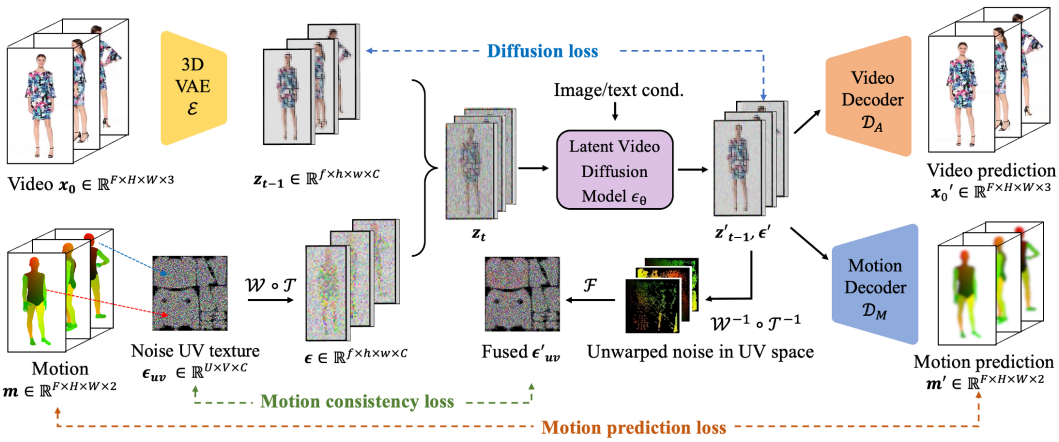


Figure 2: Training framework. During training, HumANDiff learns to generate motion-controllable human videos using a latent video diffusion model with four components: **1) Video Encoding**: we employ a 3D VAE to encode video segments $\boldsymbol{x_0}$ into latent space. **2) Articulated Motion Consistent Noise Sampling** (Sec. 3.2): We model motions on the surface UV manifold of SMPL Loper et al. (2015) and integrate motion control $\boldsymbol{m_0}$ as a warping field to sample noise $\boldsymbol{\epsilon}$ from a noise texture UV map. **3) Joint Appearance-Motion Learning** (Sec. 3.3): The diffusion model is enhanced to learn both video and motion for latent denoising to improve motion coherence. **4) Geometric Motion Consistency Learning** (Sec. 3.3): We further propose a motion consistency loss to enforce motion consistency across frames between the predicted noise $\boldsymbol{\epsilon'_{uv}}$ and sampled noise $\boldsymbol{\epsilon_{uv}}$ in UV space.

3.2 Articulated Motion-Consistent Noise Sampling

Structured Noise Texture UV Map. Existing video diffusion models learn human motions from unstructured random noises, ignoring the human body structure and motion consistency across frames, which leads to motion flickering for human video generation. Instead, we propose to model human motions in a structured unified space, i.e., on the texture UV map of a parametric body model, which is shared across different motion frames for consistent structure-aware motion modeling. We randomly sample a Gaussian noise texture map $\boldsymbol{\epsilon_{uv}} \in \mathbb{R}^{U\times V\times C}$ where C, U, V are the number of channels, the height and width of the texture UV map. Note that $\boldsymbol{\epsilon_{uv}}$is time-independent and does not vary with the length of the video sequence, which enables motion-consistent video generation.

Motion Representation. We further extract 3D human motions eas (2021) from a sequence of time-varying parametric posed body meshes (e.g.SMPL) obtained from training videos. We render these meshes and record the UV coordinate maps Güler et al. (2018) in image space as motion representations m for each time step t.

Motion-conditioned Articulated Noise Warping. We warp the noise texture map $\boldsymbol{\epsilon_{uv}}$ from UV space to image space via a geometric warping operation $\mathcal{W}$ that is pre-defined by the parametric body template (e.g., SMPL), which yields a warped noise tensor $\boldsymbol{\epsilon} = \mathcal{W}(\boldsymbol{m}, \boldsymbol{\epsilon_{uv}})$. $\boldsymbol{\epsilon} \in \mathbb{R}^{f\times h\times w\times C}$, same shape with $\boldsymbol{z_0}$. Note that besides human body, we also add Gaussian noise to the background to model diverse backgrounds in videos.

3.3 Model Fine-Tuning

Our video diffusion model is conditioned on the warped motion noise. In particular, we fine-tune the image-to-video (I2V) variant of a latent video diffusion model CogVideoX Yang et al. (2024). During training, HumANDiff uses the warped noise instead of regular Gaussian noise.

Different from existing noise warping methods (e.g., Go-with-the-Flow Burgert et al. (2025); Liu and Vahdat (2025)) that train the video diffusion model with normal finetuning objective, i.e., the mean squared error between the predicted noise and the sampled Gaussian noise, we notice that the standard training objective fails to generate temporally consistent videos with faithful motion controls. Because of articulated body structures and non-rigid movements, human motions are far more complex than the rigid object motions in Burgert et al. (2025). Moreover, as the human visual

system is highly sensitive to imperfections in human motion, human video generation typically demands high precision in motion control. To address these challenges, we further introduce several motion learning strategies described below.

Motion Noise Degradation. We use the strategy of noise degradation Burgert et al. (2025) to control the strength of motion conditioning at inference time. After calculating the clean warped noise, we then degrade it by a random degradation level $\gamma \in [0, 1]$, by first sampling uncorrelated Gaussian noise $\zeta \sim \mathcal{N}(0, 1)$ and modifying the warped noise via $\mathcal{T}$:

$$\mathcal{T}(\boldsymbol{\epsilon}) = \frac{(1-\gamma)\boldsymbol{\epsilon} + \zeta\gamma}{\sqrt{(1-\gamma)^2 + \gamma^2}} \tag{2}$$

As degradation level $\gamma \to 1$, $\boldsymbol{\epsilon}$ approaches an uncorrelated Gaussian, and as $\gamma \to 0$, $\boldsymbol{\epsilon}$ approaches clean warped noise. In inference, the user can control how strictly the resulting video should adhere to the input flow. Please see the appendix. for a qualitative depiction of the effect of γ.

Joint Appearance-Motion Learning (JAML). Despite tremendous recent progress, generative video models still struggle to capture the dynamics and physics of human motions. This limitation arises from the conventional pixel reconstruction objective, which biases models toward appearance fidelity at the expense of motion coherence Hu et al. (2024b); Chefer et al. (2025). To further enforce faithful motion control, we propose a joint appearance-motion learning strategy that encourages the video diffusion model to learn physical motions by jointly predicting appearance and motion from motion-conditioned noises during training. In addition to the latent diffusion supervision, we propose a motion decoder $\mathcal{D}_{\mathcal{M}}$ to predict motions $\boldsymbol{m'} = \mathcal{D}_{\mathcal{M}}(\boldsymbol{z'_{t-1}})$ from the predicted denoised latent $\boldsymbol{z'_{t-1}}$, and the prediction loss is defined as:

$$\mathcal{L}_{md} = \|\boldsymbol{m} - \boldsymbol{m'}\|_2 = \|\boldsymbol{m} - \mathcal{D}_{\mathcal{M}}(\boldsymbol{z'_{t-1}})\|_2 \tag{3}$$

where $\boldsymbol{z'_{t-1}}$ is the predicted denoised latent at time step t.

Note that similar joint appearance-motion learning strategies have been proposed in related papers, such as SurMo Hu et al. (2024b) and VideoJAM Chefer et al. (2025). However, ours is distinguished from these works: 1) Surmo is designed for 3D human-specific reconstruction. 2) VideoJAM models motions by optical flow and requires slow progressive inner-guidance during inference.

Geometric Motion Consistency Learning (GMCL). The standard training objective of LVDM mainly supervises the diffusion loss that focuses on the pixel reconstruction quality while failing to capture the motion consistency between consecutive frames, which causes motion flickering, especially for complex human motions. To tackle this issue, we propose geometric motion consistency supervision to encourage the LVDM to learn motion consistency during diffusion.

However, the 3D VAE of the CogVideoX model does not faithfully preserve the spatial information after video compression in the latent space, i.e.the denoised latents do not exhibit strict spatial alignment with the input motion. In contrast, the warped noise $\boldsymbol{\epsilon}$ and predicted added noise $\boldsymbol{\epsilon'}$ are motion-conditioned and spatially aligned with the input motion.

We first apply degradation reversal to the predicted noise $\boldsymbol{\epsilon'}$ as opposite to noise degradation,

$$\mathcal{T}^{-1}(\boldsymbol{\epsilon'}) = \frac{\sqrt{(1-\gamma)^2 + \gamma^2}\boldsymbol{\epsilon'} - \zeta\gamma}{(1-\gamma)} \tag{4}$$

$\boldsymbol{\epsilon'}$ is further unwarped into the UV space via inverse warping $\mathcal{W}^{-1}$ and then fused in UV space via average pooling $\mathcal{F}$, which yields $\boldsymbol{\epsilon'_{uv}} = \mathcal{F} \circ \mathcal{W}^{-1} \circ \mathcal{T}^{-1}(\boldsymbol{\epsilon'})$. The motion consistency loss is:

$$\mathcal{L}_{mc} = \|\boldsymbol{\epsilon_{uv}} - \boldsymbol{\epsilon'_{uv}}\|_2 = \|\boldsymbol{\epsilon_{uv}} - \mathcal{F} \circ \mathcal{W}^{-1} \circ \mathcal{T}^{-1} \circ \epsilon_\theta(\boldsymbol{z_t}; \boldsymbol{c}, t)\|_2 \tag{5}$$

where $\mathcal{W}^{-1}$, $\mathcal{T}^{-1}$, $\mathcal{F}$ are all differentiable. Note that we apply mask for $\boldsymbol{\epsilon_{uv}}$ and $\boldsymbol{\epsilon'_{uv}}$ in UV space when we calculate the loss.

Optimization. HumANDiff is trained end-to-end to optimize VLDM ϵ_θ and motion decoder $\mathcal{D}_{\mathcal{M}}$, whereas the 3D VAE $\mathcal{E}$ and video decoder $\mathcal{D}_{\mathcal{A}}$ are fixed. The total loss is defined as:

$$\mathcal{L} = \lambda_{diff}\mathcal{L}_{diff} + \lambda_{mc}\mathcal{L}_{mc} + \lambda_{md}\mathcal{L}_{md} \tag{6}$$

where the weights λ_{diff}, λ_{mc}, and λ_{md} are set experimentally. More details in the appendix.

3.4 Inference with Articulated Noise Sampling

We apply HumANDifffor image to video (I2V) generation, i.e., generating an animation sequence from the input image. In inference, given a single input image and a driving motion sequence, we generate motion-conditioned warped noise to guide the output video. The driving motion can be extracted from videos or obtained from external motion sources such as AMASS Mahmood et al. (2019). We estimate the SMPL body shape from the input image for motion transfer similar to Zhu et al. (2024). The warped noise is used to initialize the diffusion process of HumANDiff, and a deterministic sampling process is employed for latent denoising. Compared with existing human video generation methods (e.g., DreamPose Karras et al. (2023)) which trains a specific pose module to process poses, our method can fully utilize the video prior of the base model for intrinsic control, as it requires no adapters or modifications to the diffusion pipeline, while maintaining the same memory usage and runtime as the base model. Our method is agnostic to diffusion model design (e.g., CogVideoX-I2V and Wan2.1-I2V Wan et al. (2025)).

3.5 Implementation Details

We fine-tune the base video diffusion model, the I2V of CogVideoX-5B Yang et al. (2024) for image-to-video tasks. We also fine-tune on Wan2.1-I2V Wan et al. (2025) to show our method is model-agnostic. Unless otherwise specified, we adopt the variant trained on CogVideoX-5B same as GWTF Burgert et al. (2025) for fair comparisons. Refer to more details in the appendix.

4 Experiments

Data Processing. We evaluated our method on the UBC Fashion Zablotskaia et al. (2019) dataset, which contains 600 monocular human videos featuring natural fashion motions, with subjects wearing various styles, including dresses. We first estimate 3D SMPL parameters from the monocular videos using eas (2021), and then crop and scale each video to 480×720 px (the standard input format for CogVideoX) for training.

Evaluation Metrics. We evaluate our method using standard metrics widely adopted in prior work. For single-frame image quality, we report the L1 error, Structural Similarity Index (SSIM) Wang et al. (2004), Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. (2018), and Fréchet Inception Distance (FID) Heusel et al. (2017). Video fidelity is assessed using Fréchet Inception Distance with Fréchet Video Distance (FID-VID) Unterthiner et al. (2018) and Fréchet Video Distance (FVD) Unterthiner et al. (2018).

Baselines. We compare our approach with a range of state-of-the-art baselines. MRAA Siarohin et al. (2021) represents a leading GAN-based animation method that predicts optical flow from the driving sequence to warp the source image, followed by GAN-based inpainting to handle occluded regions. We also include Thin-Plate Spline Motion Model (TPSMM), the diffusion-based pose transfer method PIDM Bhunia et al. (2022), and the bidirectionally deformable motion modulation approach BDMM. DreamPose Karras et al. (2023) leverages the Stable Diffusion prior Rombach et al. (2021) for human animation generation. All these methods are trained on the UBC Fashion dataset for fair evaluation. UniAnimate Wang et al. (2025) adopt the Wan2.1 video base model for controllable human video synthesis.

In addition, we compare against generic video diffusion models, including Stable Video Diffusion (SVD-I2V) Blattmann et al. (2023a), our base model CogVideoX-I2V, and the motion-controllable video diffusion framework Go-With-The-Flow (GWTF) Burgert et al. (2025).

4.1 Quantitative Analysis

We test all models on the UBC Fashion test set, which consists of 100 fashion videos. For fair comparison, we scale our results to a 256 px resolution, the same as DreamPose Karras et al. (2023). For each video, we use the generated 49 frames for testing. The quantitative results are listed in Tab. 1, where the metrics of MRAA, TPSMM, PIDM, SVD-I2V, and DreamPose are reported in



Figure 3: Qualitative comparisons for fashion video synthesis. We show the results of each methods under similar novel poses. Our method can faithfully preserve the appearance details of the reference image, e.g., face identity and clothing details.

Table 1: Quantitative metrics for fashion video generation. "Dreampose*" denotes the result after sample finetuning. Note that pixel-level metrics (e.g., SSIM and L1) are more sensitive to human pose estimation accuracy, whereas perceptual metrics (e.g., LPIPS, FID, FVD) provide a more reliable evaluation for human generation tasks Liu et al. (2021).

	L1 ↓	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	FID-VID ↓
MRAA Siarohin et al. (2021)	.0857	.749	.212	23.42	253.65	–
TPSMM Zhao and Zhang (2022)	.0858	.746	.212	22.87	247.55	–
PIDM Bhunia et al. (2022)	.1098	.713	.288	30.28	238.75	–
BDMM Yu et al. (2023b)	.9180	–	**.048**	–	148.30	–
DreamPose Karras et al. (2023)	–	.8790	.111	–	279.60	–
DreamPose* Karras et al. (2023)	**.0256**	.8850	.068	13.04	238.75	–
SVD-I2V Blattmann et al. (2023a)	–	.8940	.095	-	175.4	–
CogVideoX5B-I2V Yang et al. (2024)	.0979	.827	.133	15.59	397.06	8.57
Go-With-The-Flow Burgert et al. (2025)	.1919	.756	.241	15.56	228.23	15.77
Ours	.0364	**.906**	.057	**7.94**	**84.10**	**4.68**

Karras et al. (2023); Hu et al. (2024a). The results indicate that our method outperforms existing baselines by a large margin on both image metrics (L1, SSIM, LPIPS, FID) and video metrics (FVD, FID-VID). This demonstrates that our method can generate not only high-quality images but also temporally consistent human videos. It should be mentioned that our model and Go-With-The-Flow (GWTF) are both built upon CogVideoX base model, whereas our model significantly outperforms the others for human video generation. UniAnimate was not trained on UBC Fashion videos, and instead we compare with UniAnimate on out-of-domain data for generalization analysis in Sec. 4.3.

4.2 Qualitative Analysis

We qualitatively compare our method to DreamPose Karras et al. (2023), TPSMM Zhao and Zhang (2022), BDMM Yu et al. (2023b), SVD-I2V Blattmann et al. (2023a), CogVideoX-5b-I2V Yang et al. (2024), and GWTF Burgert et al. (2025), as shown in Fig. 3. For DreamPose, BDMM, and SVD-I2V, note that the person identity, fabric folds, and fine patterns are lost in new poses, whereas our method can accurately retain those details even for complex texture styles.

Since GWTF and our model are both built upon CogVideoX and use the same video VAE model, they are able to preserve the appearance details of the reference image. However, neither CogVideoX nor GWTF supports faithful fine-grained human motion control, i.e., the rendered images are not fully aligned with the input poses. In contrast, thanks to articulated 3D noise sampling, joint



Figure 4: Comparisons against the pose-adapter-based UniAnimate Wang et al. (2025) on out-of-domain data using the same Wan2.1 backbone. Unlike UniAnimate, which relies on additional 2D pose adapters, our method performs articulated motion-consistent noise sampling to better utilize the video diffusion prior. Our method achieves stronger generalization and better appearance consistency across novel poses and viewpoints, particularly for accessories and back-view synthesis.

appearance-motion learning, and our geometric motion consistency learning strategy, our method can generate high-quality human motion videos. Refer to the appendix. for more details.

4.3 Generalization on Out-of-domain Data

Though our model was fine-tuned on studio fashion videos, it generalizes well on out-of-domain data. We evaluate the generalization capability on DeepFashion dataset Liu et al. (2016) and in-the-wild publicly available videos, as shown in Fig. 4.

We further compare our method with the pose-adapter-based UniAnimate Wang et al. (2025) on out-of-domain test data using the same Wan2.1 backbone. Unlike UniAnimate, which introduces additional 2D pose adapters for motion control, our method directly injects motion consistency into the diffusion process via articulated noise sampling on the 3D human body manifold. This enables better utilization of the intrinsic video diffusion prior for controllable human video generation.

As shown in Fig. 4, our method generalizes more effectively to unseen clothing styles, accessories, and viewpoints. In the first example, our method preserves accessories and scene structures, such as the chair and hat, while maintaining more consistent appearance across poses. In the second example, ours synthesizes faithful back-view appearances under large pose changes, whereas UniAnimate struggles with back-view rendering and produces inconsistent clothing details.

4.4 Ablation Study

We evaluate JAML and GMCL qualitatively and quantitatively. Additional ablations of 3D articulated versus 2D optical-flow noise warping and noise degradation are provided in the appendix. Refer to the appendix for more ablation studies about the **3D Articulated Noise Warping vs. 2D Optical Flow based Noise Warping** and **noise degradation**.

5 Conclusion

We presented HumANDiff, a controllable human video framework combining articulated motion-consistent noise sampling with joint appearance-motion and geometric consistency learning. It requires no inference-time motion modules and outperforms prior methods in fidelity and motion consistency. **limitation** are discussed in the appendix.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools solely to aid and polish the writing, including improving grammar, clarity, concision, and presentation. We did not use generative AI tools to formulate research ideas, develop the method, conduct experiments, analyze data, generate results, create figures, or select citations. All AI-assisted edits were reviewed by the authors to verify technical accuracy, preserve the intended scientific meaning, and ensure consistency with the reported methods and results. We take full responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

HumANDiff generates realistic human videos and could be misused for impersonation, misinformation, or the creation of non-consensual content. The method is intended for research and legitimate creative applications. Deployments should obtain consent from depicted individuals, clearly disclose AI-generated content, respect privacy and intellectual-property rights, and adopt safeguards such as provenance metadata or watermarking. Our experiments use established research datasets and publicly available videos for research evaluation. Because these data may not represent all identities, body types, clothing styles, or cultural contexts, the model may exhibit corresponding performance biases. We encourage future work to evaluate these limitations across more diverse populations and deployment settings.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

We provide the model formulation, articulated noise sampling procedure, training objectives, and inference process in Sec. 3.2 through Sec. 3.4. Implementation details, including model architecture, training configuration, hyperparameters, data processing, and evaluation protocols, are reported in Sec. 3.5 and the appendix. The datasets, baselines, and evaluation metrics are described in Sec. 4. Additional ablations and qualitative results are provided in the appendix and supplementary video to facilitate verification of the reported findings.

REFERENCES

Easymocap - make human motion capture easier. Github, 2021. URL https://github.com/zju3dv/EasyMocap.

Kfir Aberman, M. Shi, Jing Liao, Dani Lischinski, B. Chen, and D. Cohen-Or. Deep video-based performance cloning. *Computer Graphics Forum*, 38, 2019.

Ankan Kumar Bhunia, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5968–5976, 2022. URL https://api.semanticscholar.org/CorpusID:253761291.

A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, abs/2311.15127, 2023a. URL https://api.semanticscholar.org/CorpusID:265312551.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *arXiv preprint arXiv:2311.15127*, 2023b.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2501.08331.

Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. *ICCV*, pages 5932–5941, 2019.

Eric Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, S. Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. *ArXiv*, abs/2112.07945, 2021.

Pascal Chang, Jingwei Tang, Markus Gross, and Vinicius C. Azevedo. How i warped your noise: a temporally-correlated noise prior for diffusion models. *ArXiv*, abs/2504.03072, 2025. URL https://api.semanticscholar.org/CorpusID:270964689.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *ArXiv*, abs/2502.02492, 2025. URL https://api.semanticscholar.org/CorpusID:276107610.

Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, and Huchuan Lu. Animatable neural radiance fields from monocular rgb video. *ArXiv*, abs/2106.13629, 2021.

A. K. Grigor'ev, Artem Sevastopolsky, Alexander Vakhitov, and Victor S. Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. *CVPR*, pages 12127–12136, 2019.

Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *ArXiv*, abs/2110.08985, 2021.

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. *arXiv preprint arXiv:2307.04725*, 2023.

R. A. Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306, 2018.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33, 2020.

Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. Headnerf: A real-time nerf-based parametric head model. *ArXiv*, abs/2112.05637, 2021.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021a. URL https://api.semanticscholar.org/CorpusID:235458009.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate Anyone 2: High-Fidelity Character Image Animation with Environment Affordance. *arXiv preprint arXiv:2502.06145*, 2025.

T. Hu, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr: Hybrid volumetric-textural rendering for human avatars. *3DV*, 2022.

Tao Hu, Kripasindhu Sarkar, Lingjie Liu, Matthias Zwicker, and Christian Theobalt. Egorenderer: Rendering human avatars from egocentric camera images. In *ICCV*, 2021b.

Tao Hu, Hongyi Xu, Linjie Luo, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr++: Image and pose driven human avatars using hybrid volumetric-textural rendering. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–15, 2023. doi: 10.1109/TVCG.2023.3297721.

Tao Hu, Fangzhou Hong, and Ziwei Liu. Surmo: Surface-based 4d motion modeling for dynamic human rendering. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6550–6560, 2024b. URL https://api.semanticscholar.org/CorpusID:268819493.

Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. HumanRF: High-Fidelity Neural Radiance Fields for Humans in Motion. *ACM Transactions on Graphics*, 42(4):1–12, 2023.

Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22623–22633, 2023. URL https://api.semanticscholar.org/CorpusID:258078892.

Bernhard Kratzwald, Zhiwu Huang, Danda Pani Paudel, and Luc Van Gool. Towards an understanding of our world by GANing videos in the wild. arXiv:1711.11453, 2017. URL https://arxiv.org/abs/1711.11453.

Wentao Lei, Jinting Wang, Fengji Ma, Guanjie Huang, and Li Liu. A Comprehensive Survey on Human Video Generation: Challenges, Methods, and Insights. *arXiv preprint arXiv:2407.08428*, 2024.

Chao Liu and Arash Vahdat. On equivariance and fast sampling in video diffusion models trained with warped noise. 2025. URL https://api.semanticscholar.org/CorpusID:277781088.

Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 2019.

Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, 05 2020. doi: 10.1109/TVCG.2020.2996594.

Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *TOG*, 40, 2021.

Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016.

M. Loper, Naureen Mahmood, J. Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34:248:1–16, 2015.

Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NeurIPS*, pages 405–415, 2017.

Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *CVPR*, 2018.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Natalia Neverova, Riza Alp Güler, and Iasonas Kokkinos. Dense pose transfer. *ECCV*, 2018.

Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. *CVPR*, pages 11448–11459, 2021.

Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *IEEE/CVF ICCV*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024.

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. *ArXiv*, abs/2112.11427, 2021.

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021a.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. *CVPR*, 2021b.

Sergey Prokudin, Michael J. Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. *WACV*, 2021.

Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *CVPR*, June 2018.

Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, et al. Anigs: Animatable gaussian avatar from a single image with inconsistent gaussian reconstruction. *arXiv preprint arXiv:2412.02684*, 2024.

Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model from a single image in seconds. *arXiv preprint arXiv:2503.10625*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. In *ECCV*, 2020.

Ruizhi Shao, Youxin Pang, Zerong Zheng, Jingxiang Sun, and Yebin Liu. Human4DiT: 360-Degree Human Video Generation with 4D Diffusion Transformer. *ACM Transactions on Graphics*, 43 (6):1–13, 2024.

Ruizhi Shao, Yinghao Xu, Yujun Shen, Ceyuan Yang, Yang Zheng, Changan Chen, Yebin Liu, and Gordon Wetzstein. Isa4d: Interspatial attention for efficient 4d human video generation. *ACM Transactions on Graphics (TOG)*, 2025.

Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. Deformable GANs for pose-based human image generation. In *CVPR*, 2018.

Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and S. Tulyakov. Motion representations for articulated animation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13648–13657, 2021. URL https://api.semanticscholar.org/CorpusID:233388032.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-A-Video: Text-to-Video Generation without Text-Video Data. *arXiv preprint arXiv:2209.14792*, 2022.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. *arXiv preprint arXiv:2011.13456*, 2020.

Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, 2021.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023.

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

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. InstantID: Zero-shot Identity-Preserving Generation in Seconds. *arXiv preprint arXiv:2401.07519*, 2024a.

Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. DisCo: Disentangled Control for Realistic Human Dance Generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. In *NeurIPS*, 2018a.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, 2018b.

Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. MagicVideo-V2: Multi-Stage High-Aesthetic Video Generation. *arXiv preprint arXiv:2401.04468*, 2024c.

Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *Science China Information Sciences*, 2025.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

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 *International Conference on Computer Vision (ICCV)*, 2023.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. VideoGPT: Video Generation using VQ-VAE and Transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective Whole-body Pose Estimation with Two-stages Distillation. In *International Conference on Computer Vision (ICCV)*, 2023.

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

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. MAGVIT: Masked Generative Video Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023a.

Wing Yin Yu, Lai-Man Po, Ray C. C. Cheung, Yuzhi Zhao, Yu-Zhi Xue, and Kun Li. Bidirectionally deformable motion modulation for video-based human pose transfer. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7468–7478, 2023b. URL `https://api.semanticscholar.org/CorpusID:259937850`.

Polina Zablotskaia, Aliaksandr Siarohin, Leonid Sigal, and Bo Zhao. Dwnet: Dense warp-based network for pose-guided human video generation. In *British Machine Vision Conference (BMVC)*, 2019.

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4Diffusion: Multi-view Video Diffusion Model for 4D Generation. *Advances in Neural Information Processing Systems*, 37, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *International Conference on Computer Vision (ICCV)*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

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*, abs/2406.19680, 2024. URL `https://api.semanticscholar.org/CorpusID:270845480`.

Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3647–3656, 2022. URL `https://api.semanticscholar.org/CorpusID:247763055`.

Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *ArXiv*, 2021.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, 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 (ECCV)*, 2024.

Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, pages 2347–2356, 2019.

A Appendix: Implementation Details

We fine-tune the base video diffusion model, the I2V of CogVideoX-5B Yang et al. (2024) for image-to-video tasks, which generates 480×720px videos. We follow the same setup of CogVideoX-5B for video encoding, i.e., given a video $\boldsymbol{x_0}$ of size $(4f+1) \times 8h \times 8w \times 3$, the latent embedding $\boldsymbol{z_0}$ is encoded by the 3D VAE $\mathcal{E}$, where $\boldsymbol{z_0} \in \mathcal{R}^{(f+1)\times h \times w \times C}$. CogVideoX-5B generates 49 frames for each video, i.e., $f = 12$, $h = 60$, $w = 90$, $C = 16$.

We also fine-tune on Wan2.1-I2V Wan et al. (2025) to show our method is model-agnostic. Unless otherwise specified, we adopt the variant trained on CogVideoX-5B same as GWTF Burgert et al. (2025) for fair comparisons.

For noise warping, because the diffusion model works on latent embeddings, we estimate the 3D pose in image space via eas (2021) and warp noise in image space, and then downsample that noise into latent space, similar to the video downscaling of CogVideoX by a factor of 8×8 spatially and 4 temporally. We use nearest-neighbor interpolation along both the temporal axis and the two spatial axes.

During training, we randomly sample $\gamma \in [0, 1]$, and fix it to 0.1 at inference. During inference, we use the Wan2.1-I2V-14B-480P model with 50 denoising steps under the UniPC scheduler to generate videos.

Network Architecture. We fine-tune on different base video diffusion models for image to video generation, including CogVideox-I2V-5B Yang et al. (2024) and Wan2.1-I2V-14B Wan et al. (2025). For motion decoder, to accelerate training, we use a CNN-based architecture (Pix2PixHD Wang et al. (2018b)) with 2 Decoder blocks of [ReLU, ConvTranspose2d, BatchNorm].

Noise UV Texture Map. The size of the noise texture map is $128 \times 128 \times 16$.

Optimization. We used 8 NVIDIA H100 80GB GPUs over the course of 32 GPU days, for 26,000 iterations using a rank-2048 LoRA Hu et al. (2021a) with a learning rate of 1E-4 and a batch size of 6.

We set the weights $\lambda_{diff} = 1.0$, $\lambda_{mc} = 0.5$, and $\lambda_{md} = 0.5$ experimentally during training.

B Additional Experimental Results

B.1 Motion Fidelity and Novel-View Subject Consistency

Motion-Centric Evaluation. We additionally evaluate motion fidelity using four complementary pose-based metrics, including MPJPE, PCK, trajectory error, and acceleration error, on the standard UBC Fashion test set. MPJPE and PCK measure frame-wise pose accuracy, while trajectory and acceleration errors evaluate temporal motion consistency and smoothness. As shown in Table 2, our method consistently outperforms Go-With-The-Flow (GWTF) across pose accuracy, trajectory fidelity, and temporal acceleration metrics.

Subject Consistency Under Novel Viewpoints. We additionally evaluate back-view subject consistency using DINO-I, computed as the feature similarity between the front and back views. As shown in Table 3, HumANDiff achieves the highest score of **0.891**. This result indicates that the proposed articulated noise representation better preserves subject identity under novel viewpoints, particularly for challenging back-view synthesis with large pose changes.

Table 2: Motion-centric evaluation on paired ground-truth videos. The relative improvement is computed against Go-With-The-Flow (GWTF).

Metric	Ours	GWTF	Improvement
MPJPE ↓	**0.0274**	0.0342	19.6%
PCK@0.05 ↑	**0.864**	0.820	+5.4%
PCK@0.10 ↑	**0.948**	0.932	+1.7%
Trajectory Error ↓	**0.0119**	0.0136	12.4%
Acceleration Error ↓	**0.0133**	0.0180	26.1%

Table 3: Back-view subject consistency measured by DINO-I. Higher is better.

Method	DINO-I ↑
Ours	**0.891**
GWTF Burgert et al. (2025)	0.835
DreamPose Karras et al. (2023)	0.887
BDMM Yu et al. (2023b)	0.866

B.2 Generalization on Out-of-domain Data

Human Animation on DeepFashion. Our method generalizes well to unseen out-of-domain data, producing temporally consistent videos from a single image of humans in diverse clothing styles, as shown in Fig. 5a. More results can be found in the supplementary video.

Human Animation on In-the-Wild Videos. Our method generalizes well to unseen in-the-wild videos collected from publicly available online sources with diverse backgrounds, producing temporally consistent videos from a single image, as shown in Fig. 5b. More results can be found in the supplementary video.

B.3 Comparisons with Other Methods

We also compare our method with Animate Anyone Hu et al. (2024a) and Wan2.1 based UniAnimate Wang et al. (2025) for fashion video generation, and the qualitative results are shown in Fig. 6, which suggest that our method can generate high-quality clothing details (e.g., ①, ③), and high-quality face datails (②), whereas Animate Anyone cannot achieve the same level of realism. Note that Animate Anyone and our method were both fine-tuned on fashion videos, whereas the original UniAnimate was fine-tuned on generic videos. We compare against UniAnimate on out-of-domain data as presented in the main paper. Since Animate Anyone Hu et al. (2024a) is not publicly available, we directly downloaded their generated videos from the official website using similar input images and pose sequences to ensure fair comparisons.

Out-of-Domain Generalization. To assess generalization beyond the training distribution, we evaluate both our method and UniAnimate on 30 out-of-domain subjects drawn from DeepFashion Liu et al. (2016). Since no paired ground-truth animation exists for these subjects, standard reconstruction metrics (e.g., LPIPS, PSNR) are inapplicable. We instead report **DINO-I** Tumanyan et al. (2023), the average cosine similarity between DINOv2 Oquab et al. (2024) ViT-L/14 features of each generated frame and the reference image. DINO-I measures how faithfully the generated video preserves the appearance of the input subject, which is the primary objective of image-driven human animation.

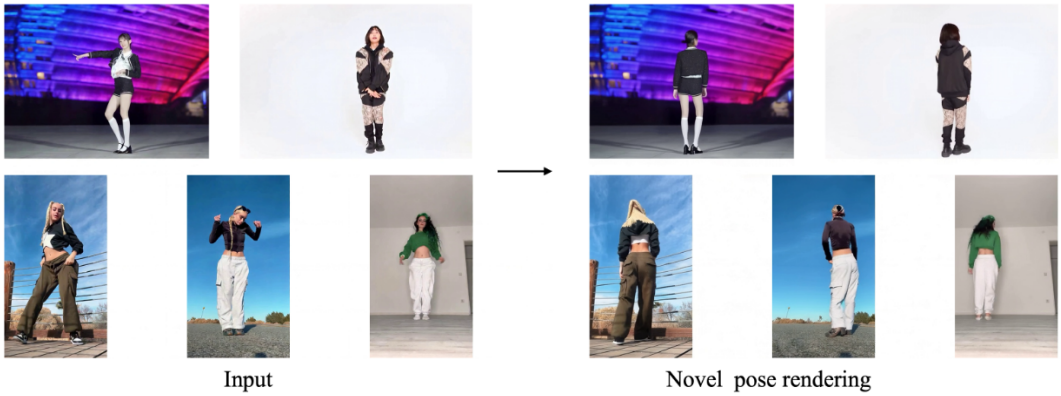
As shown in Table 4, our method achieves a mean DINO-I of $\mathbf{0.813 \pm 0.041}$, outperforming UniAnimate (0.693 ± 0.089) by a margin of $+\mathbf{0.120}$, demonstrating substantially better appearance preservation across diverse out-of-domain subjects. The smaller standard deviation of our method (0.041 vs. 0.089) further indicates more consistent fidelity when the input distribution shifts. Temporal consistency (DINO-TC) is comparable between the two methods (0.940 vs. 0.943), showing that the improvement in appearance fidelity does not come at the cost of temporal smoothness.

Table 4: DINO-I and DINO-TC scores on out-of-domain subjects. DINO-I measures cosine similarity between DINOv2 features of generated frames and the reference image (appearance preservation). DINO-TC measures cosine similarity between features of consecutive frames (temporal consistency). Both metrics are higher is better.

Method	DINO-I ↑	DINO-TC ↑
UniAnimate Wang et al. (2025)	0.693 ± 0.089	0.943 ± 0.022
Ours	$\mathbf{0.813 \pm 0.041}$	0.940 ± 0.017



(a) Generalization on out-of-domain data (DeepFashion Liu et al. (2016)).



(b) Generalization on in-the-wild videos.

Figure 5: Generalization results. Our method trained on fashion videos generalizes well to unseen out-of-domain data, and synthesizes human novel pose videos from a single image in diverse clothing styles (a), and diverse in-the-wild backgrounds (b).

B.4 Ablation Study

3D Articulated Noise Warping vs. 2D Optical Flow based Noise Warping. We further analyze two different noise warping methods: 2D optical flow based GWTF Burgert et al. (2025), and our 3D articulated noise warping. The qualitative comparisons are shown in Fig. 7, which indicates that our 3D noise warping is more suitable for human rendering, i.e., it enables more precise motion control, and generates more consistent rendering results even for complex clothing fashion style textures. More comparisons can be found in the supplementary video.

Joint Appearance-Motion Learning (JAML). We analyze the effectiveness of the JAML mechanism (Sec. 3.3) qualitatively and quantitatively. The qualitative comparisons are shown in Fig. 8, which suggests that the joint learning strategy improves motion coherence for faithful motion



Figure 6: Comparisons against Animate Anyone Hu et al. (2024a) and Wan2.1 based UniAnimate Wang et al. (2025). Our method can generate high-quality clothing details (①, ③), and face details (②).



Figure 7: Qualitative comparisons for appearance consistency of different noise warping methods, where two different frames are visualized for evaluations. Compared with 2D optical flow based noise warping GWTF Burgert et al. (2025), our 3D articulated noise warping enables more precise motion control, and generates more consistent rendering results for complex textures.

control, while the variant that was trained with a standard objective (w/o JAML) failed to generate accurate animation videos. The quantitative results listed in Tab. 5 also indicates that the joint learning strategy can improve the quantitative performance.

Geometric Motion Consistency Learning (GMCL). We also analyze the effectiveness of our proposed Geometric Motion Consistency Learning strategy. The qualitative comparisons are shown in Fig. 9, where we visualize two different frames for consistency comparison. Fig. 9 shows that the proposed GMCL can significantly improve appearance consistency. For example, the face identity and hand details are better preserved, e.g., ② vs. ①, and ④ vs. ③.

Noise Degradation. We use the noise degradation strategy Burgert et al. (2025) to control the strength of motion conditioning at inference time. We analyze the effect of the noise degradation parameter γ for human video generation on a subset of the UBC Fashion dataset. Table 6 sum-

Table 5: Ablation study results for JAML and GMCL. Note that perceptual metrics (e.g., LPIPS, FID, FVD) are generally more reliable than pixel-level metrics (e.g., SSIM and L1) for evaluating human generation tasks Liu et al. (2021).

	L1 ↓	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
Base + Articulated Noise Sampling	.0378	.903	.071	10.47	12.03
+ JAML	.0371	.905	.065	9.02	10.00
+ GMCL	**.0364**	**.906**	**.057**	**7.94**	**4.68**



Figure 8: Ablation study of Joint Appearance-Motion Learning (JAML). JAML improves motion coherence for faithful motion control, while the variant without JAML failed to generate accurate animation videos.

marizes the quantitative results, suggesting that our method achieves the best performance on both LPIPS and SSIM when $\gamma = 0.1$.

We also show qualitative results in Fig. 10. The generated videos strictly follow the motion signals when $\gamma = 0$, but the video quality is degraded. Therefore, we set $\gamma = 0.1$ in our experiments to achieve a balance between motion control and video quality.

C Discussion

HumANDiff is a **model-agnostic** framework that extends base video diffusion models for controllable human video generation without compromising efficiency.

• **Model Agnostic**. We fine-tune on different base video diffusion models including CogVideox-I2V-5B Yang et al. (2024) and Wan2.1-I2V-14B Wan et al. (2025) to show our method is model-agnostic, and the generation results are shown in Fig. 12. The model trained on Wan2.1-I2V-14B generates higher quality appearance details than that of CogVideox-I2V-5B.

• **Efficiency**. HumANDiff shares the same architecture as the base video diffusion model, and hence uses exactly the same amount of memory and runtime as the base model for both CogVideoX Yang et al. (2024) and Wan2.1 Wan et al. (2025). Each video (480×720, 81 frames, 50 steps) takes approximately 600s with a peak memory usage of approximately 46 GB on a single NVIDIA H100 GPU for Wan 2.1 model.

Limitations. Our model relies on the video VAE of the underlying video diffusion backbone to encode input videos, and the overall generation quality is therefore constrained by the VAE's reconstruction fidelity. In particular, the video VAE used in CogVideoX tends to lose fine-grained details—such as facial features and hand structures—during the encoding process. Consequently, models built upon CogVideoX struggle to synthesize high-quality facial and hand details, as illustrated in Fig. 11. A promising direction for future work is to explore stronger video diffusion backbones equipped with more expressive VAEs, or to jointly fine-tune the VAE together with the diffusion model, following prior work such as Karras et al. (2023).

Broader Impacts. HumANDiff enables controllable human video generation with applications in animation, AR/VR, film production, and digital content creation. However, like other genera-

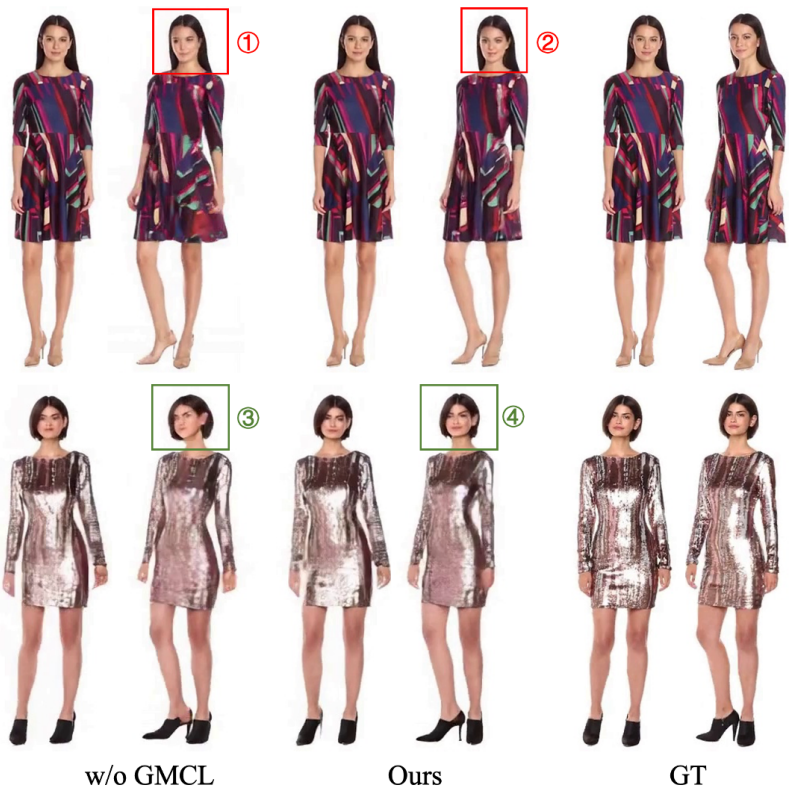


Figure 9: Ablation study of Geometric Motion Consistency Learning (GMCL), where two frames at different time steps are shown. GMCL improves appearance consistency, *i.e.*, it preserves the face identity details better (② vs. ①, and ④ vs. ③).



Figure 10: Ablation study of noise degradation. We set $\gamma = 0.1$ to achieve a balance between motion control and video quality.

tive video models, this technology carries potential risks of misuse, including the creation of non-consensual deepfakes or disinformation. We encourage responsible deployment and advocate for the use of content authentication and detection tools to mitigate such risks. Our method requires an explicit driving motion sequence as input, which limits fully unconstrained generation and provides a degree of interpretability. We do not release training data containing personal identities, and recommend that downstream users adhere to applicable regulations regarding consent and data privacy.

Table 6: Ablation study of noise degradation. γ measures how strictly the resulting video adheres to the motion input.

	LPIPS ↓	SSIM ↑
$\gamma = 0$	.0857	.858
$\gamma = 0.1$	**.0821**	**.860**
$\gamma = 0.2$	.0859	.860
$\gamma = 0.3$	.0898	.858
$\gamma = 0.4$	.1088	.841
$\gamma = 0.5$	.1184	.834
$\gamma = 0.6$	.1159	.836
$\gamma = 0.7$	.1031	.852
$\gamma = 0.8$	.1134	.841
$\gamma = 0.9$	.1055	.845
$\gamma = 1.0$	.1209	.832

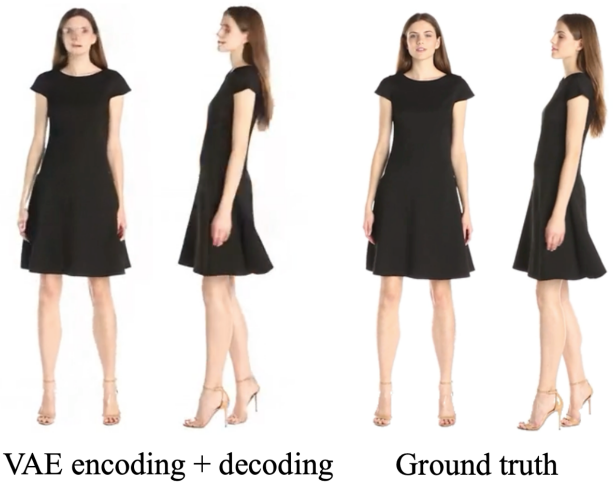


Figure 11: Illustration of video VAE encoding in CogVideoX. Fine details such as faces and hands are often lost during video encoding, which makes it challenging for our model to generate high-quality face and hand details.



Figure 12: Generation results of model variants trained on Wan2.1-I2V-14B and CogVideoX-I2V-5B. Our method is model agnostic, and the model trained on Wan2.1-I2V-14B generates higher quality appearance details (①②③).