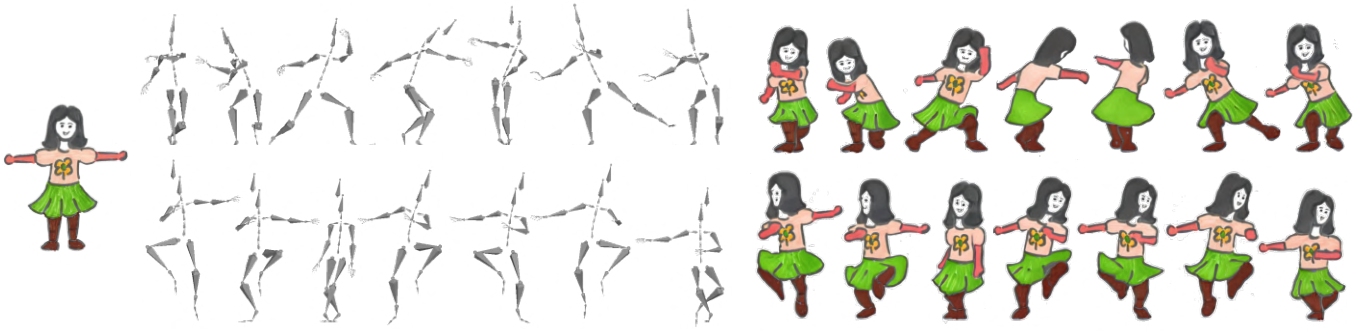
From Rigging to Waving: 3D-Guided Diffusion for Natural Animation of Hand-Drawn Characters

JIE ZHOU*, City University of Hong Kong, China
LINZI QU*, City University of Hong Kong, China
MIU-LING LAM†, City University of Hong Kong, China
HONGBO FU†, Hong Kong University of Science and Technology, China



(a) Reference (b) Target 3D motions (c) Generated animation frames

Fig. 1. Our system produces natural character animations (Right) given single input drawings (Left) and target 3D motions (Middle).

Hand-drawn character animation is a vibrant research area in computer graphics and presents unique challenges in achieving geometric consistency while conveying expressive motion details. Traditional skeletal animation methods maintain geometric consistency but often struggle with complex non-rigid elements like flowing hair and skirts, resulting in unnatural deformation and missing secondary dynamics. In contrast, video diffusion models effectively synthesize physically plausible dynamics, but exhibit real-human-like characteristics and geometric distortions when applied to stylized drawings due to the domain gap. In this work, we propose a novel hybrid animation system that integrates the strengths of skeletal animation and video diffusion priors. The core idea is to first generate coarse images from characters retargeted with skeletal animations for geometric consistency guidance, and then enhance these images in terms of texture details and secondary dynamics using video diffusion priors. We formulate the enhancement of coarse images as an inpainting task and propose a domain-adapted diffusion model to refine user-masked regions requiring improvement, particularly those involving secondary dynamics. To further enhance motion realism, we propose a Secondary Dynamics Injection (SDI) strategy during the denoising process to incorporate latent features from a pre-trained diffusion model enriched with human motion priors. Additionally, to address unnatural deformation artifacts caused by the integrated hair-body geometry in low-poly single-mesh character modeling, we introduce a Hair Layering Modeling (HLM) technique that employs segmentation maps to separate hair from the body in implicit fields, enabling more natural animation of challenging long-hair characters. Through extensive experiments, we demonstrate that our system outperforms state-of-the-art works in both quantitative and qualitative evaluations. Please refer to our project page (https://lordliang.github.io/From-Rigging-to-Waving) for the code and data for our method.

CCS Concepts: • **Computing methodologies** → **Animation**; *Non-photorealistic rendering*.

Additional Key Words and Phrases: Character Animation, Video Diffusion Model, Secondary Motion, Skeletal Animation

ACM Reference Format:
Jie Zhou*, Linzi Qu*, Miu-Ling Lam†, and Hongbo Fu†. 2025. From Rigging to Waving: 3D-Guided Diffusion for Natural Animation of Hand-Drawn Characters. 1, 1 (September 2025), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Hand-drawn character animation has emerged as a captivating research area in computer graphics. Breathing life into static hand-drawn characters opens doors to innovation in entertainment, education, and more, creating interactive and immersive experiences. Common techniques for this task can be categorized into two main approaches: traditional skeletal animation and pose-controllable video diffusion models. However, both face significant challenges in simultaneously maintaining geometric consistency and conveying expressive motion details.

Traditional skeletal animation methods [Hornung et al. 2007; Smith et al. 2023] embed hand-drawn characters as 2D meshes,

* Equal Contributions.
† Corresponding authors.
Authors' addresses: Jie Zhou*, jzhou67-c@my.cityu.edu.hk, City University of Hong Kong, China; Linzi Qu*, linziqu2-c@my.cityu.edu.hk, City University of Hong Kong, China; Miu-Ling Lam†, miu.lam@cityu.edu.hk, City University of Hong Kong, China; Hongbo Fu†, fuplus@gmail.com, Hong Kong University of Science and Technology, China.

ACM XXXX-XXXX/2025/9-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

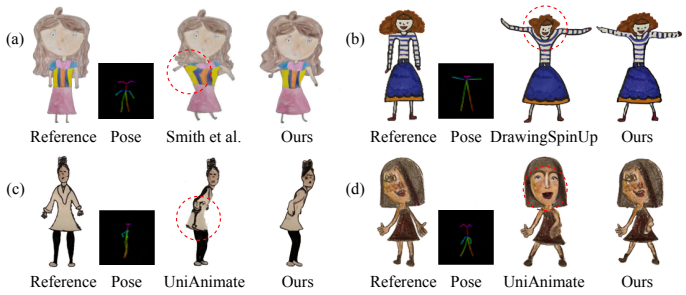


Fig. 2. Illustration of issues with existing methods: (a)-(b) unnatural deformations; (c) incorrect contours; (d) inconsistent identity.

rig them, and employ as-rigid-as-possible (ARAP) shape manipulation [Igarashi et al. 2005] to repose character meshes, ensuring identity consistency, but they are limited to 2D planar motions. DrawingSpinUp [Zhou et al. 2024] improves this process by reconstructing 3D models from character drawings prior to rigging, supporting 3D motions while maintaining geometric consistency. However, these methods struggle to animate characters with complex non-rigid elements, such as long hair or loose clothing. Due to the limitations of current single-to-3D generation methods [Liu et al. 2023; Long et al. 2024; Tang et al. 2024; Wang et al. 2024b], the 3D characters reconstructed from a single image often exhibit low-poly single-mesh geometry. Consequently, shoulder-length and longer hair frequently adheres to the neck and shoulders, leading to unnatural deformation (see Fig. 2 (a)-(b)). Additionally, skeletal animation focuses on primary motions and thus fails to convey realistic secondary dynamics (such as subtle movements of hair and clothing). While artists can achieve realistic visual details through more complex rigs, hierarchical modeling, or meticulous adjustments of material parameters, these manual processes are labor-intensive and demand specialized expertise.

Recently, pose-controllable video diffusion models [Hu 2024; Tan et al. 2025; Wang et al. 2024c; Zhu et al. 2024] have shown remarkable capabilities in generating dynamic motions for various identities within the domain of real human data. These models inherently learn physics-aware motion priors, effectively capturing both primary and secondary motions from extensive real human motion videos. Unfortunately, directly applying these models to hand-drawn character drawings often results in outputs that overlook specific artistic details, exhibit distorted appearances, particularly in regions close to the contour of the character drawing (see Fig. 2 (c)), and synthesize real-human-like characteristics (see Fig. 2 (d)). These issues stem from the domain gap (e.g., shape proportion, contour lines, exaggerated motions) between hand-drawn characters and real humans. Although training a video diffusion model specifically for hand-drawn character animation could be a potential solution, collecting a substantial dataset of hand-drawn character animations with realistic secondary motion is impractical. Building on these observations, we recognize a potential complementarity between skeletal animation and video diffusion models. We thus propose a novel hybrid animation system that integrates the strengths of both approaches, enabling the generation of diverse stylized hand-drawn character animations with natural motions.

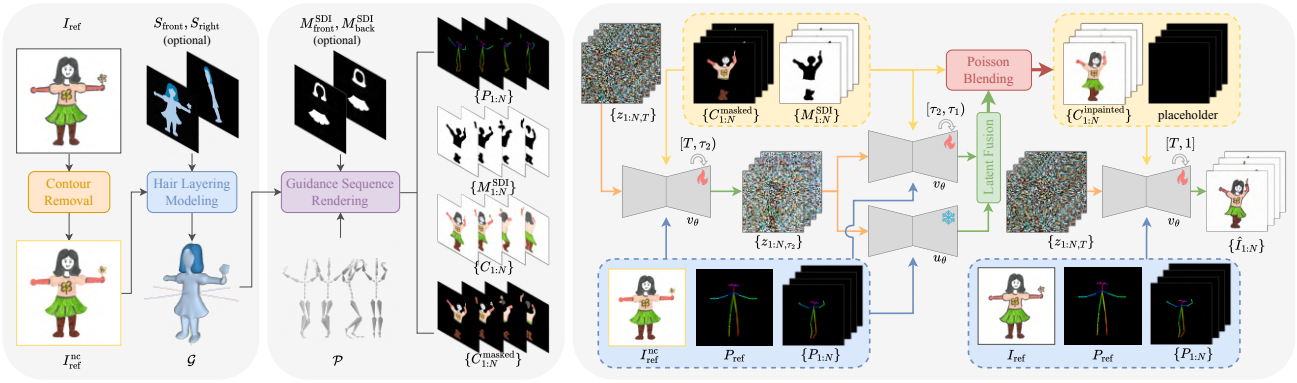
Specifically, we employ skeletal animation to generate coarse images from retargeted characters, ensuring geometric consistency for stable animation generation. We formulate the enhancement of coarse images as an inpainting task and introduce a domain-adapted diffusion model to refine them primarily from two aspects: appearance details and secondary dynamics. To bridge the gap between real humans and hand-drawn characters, we construct a small-scale animation dataset focusing on primary motions to enhance texture details and contour lines. However, since this dataset emphasizes primary motions, the domain-adapted diffusion model might neglect secondary motions during animation generation. To address this limitation, we propose a Secondary Dynamics Injection (SDI) strategy during the denoising process, which enhances the naturalness of motion by leveraging a pre-trained diffusion model to guide the denoising direction. This guidance is achieved by blending the latents from our domain-adapted diffusion model and the pre-trained one, using user-provided SDI masks. Additionally, to address unnatural deformation artifacts caused by the integrated hair-body geometry in low-poly single-mesh character modeling, we introduce a Hair Layering Modeling (HLM) method that uses segmentation maps to separate hair from the body in implicit fields, allowing our system to animate challenging long-hair characters naturally. The results of comprehensive experiments and a perceptual user study demonstrate that our system generates high-fidelity stylized animations that maintain geometric consistency while enhancing dynamic details. This underscores its superior performance compared to state-of-the-art methods. Furthermore, we highlight the potential of our system to facilitate a user-friendly application for editing existing animations.

2 RELATED WORK

2.1 Hand-Drawn Character Animation

The task of character drawing animation has been studied for a long time. 2D animation methods [Hornung et al. 2007; Smith et al. 2023] typically project 3D motions onto the image plane to animate a 2D character using As-Rigid-As-Possible (ARAP) deformation [Igarashi et al. 2005]. However, these methods are limited to generating results from a preset viewpoint. 3D animation methods [Weng et al. 2019; Zhou et al. 2024] typically reconstruct 3D geometries as proxies from drawings and perform skeletal animation. However, since these methods deform the entire character as a single 2D/3D mesh, they demonstrate only a single primary dynamic effect and struggle to animate complex characters with long hair or loose clothing.

Some multi-layered methods [Fan et al. 2018; Willett et al. 2017]add 2D secondary motions on hair and clothing by deforming a 2D layered puppet. To achieve 3D secondary dynamics, Jain et al. [2012] propose using 3D proxies to simulate physically driven clothes for a 2D hand-drawn character. This approach relies on a 3D simulation engine, such as Maya, and is notably time-consuming. Zhang et al. [2020] propose the concept of complementary dynamics to enhance rigged animations with detailed elastodynamics, though this incurs a computational cost. Benchekroun et al. [2023] further introduce a reduced-space elastodynamic solver to improve performance. However, these methods often result in a rubbery elasticity effect, suitable for flesh and skin but not for hair and fabrics. PhysAnimator [Xie et al. 2025] combines physics-based simulations with data-driven generative models to create intermediate frames between keyframes.



(a) Hair Layering Modeling (b) Guidance Sequence Rendering (c) Secondary Dynamics Injection

Fig. 3. An illustration of our pipeline, which consists of three main parts. (a) Given a hand-drawn character image I_{ref}, we first remove the contour to obtain I^{nc}_{ref} and then reconstruct a hair-layered model $\mathcal{G}$. (b) According to the target 3D motion $\mathcal{P}$ and user-specified SDI masks ($M^{\text{SDI}}_{\text{front}}$, $M^{\text{SDI}}_{\text{back}}$), we render several guidance sequences (pose $\{P_{1:N}\}$, SDI mask $\{M^{\text{SDI}}_{1:N}\}$, coarse color $\{C_{1:N}\}$, and masked coarse color $\{C^{\text{masked}}_{1:N}\}$) after rigging $\mathcal{G}$. (c) During the inference process, based on those guidance sequences, we begin denoising using $I^{\text{nc}}_{\text{ref}}$ via our domain-adapted model v_θ during the denoising steps $[T, \tau_2)$. Next, still using $I^{\text{nc}}_{\text{ref}}$, we fuse the latents from v_θ with those from the pre-trained model u_θ during the denoising steps $[\tau_2, \tau_1)$ to optimize the denosing direction with natural motion. At the denoising step τ_1, we estimate the output video and enhance $\{C^{\text{masked}}_{1:N}\}$ to create the inpainted coarse guidance $\{C^{\text{inpainted}}_{1:N}\}$ via Poisson Blending. Finally, guided by the $\{C^{\text{inpainted}}_{1:N}\}$ and I_{ref}, we further denoise from scratch to achieve the final results $\{\hat{I}_{1:N}\}$.

While it introduces expressive, data-driven dynamics, it is limited to 2D representations and cannot fully capture 3D effects. In contrast, our method integrates 3D skeletal animation into a domain-adapted video diffusion model to produce high-fidelity stylized animations with 3D geometric consistency and secondary dynamic details.

2.2 Pose-Controlled Video Diffusion Models

Recently, with the advancement of large generative models [Blattmann et al. 2023; Wang et al. 2024d], pose-controlled video diffusion models have achieved remarkable results in generating natural physical dynamics. DisCo [Wang et al. 2024a] pioneers this approach by incorporating a Pose ControlNet and a CLIP image encoder with a denoising UNet, separately guiding pose generation and encoding human semantics. To further preserve identity consistency, methods such as [Hu 2024; Xu et al. 2024; Zhu et al. 2024] extract fine-grained identity features using a reference UNet that is architecturally aligned with the denoising UNet. These features are then injected into the denoising UNet through attention mechanisms. UniAnimate [Wang et al. 2024c] simplifies the model architecture by mapping the reference image, pose guidance, and noise into a unified feature space within a single video diffusion framework, thereby effectively reducing model complexity. However, due to the domain gap between hand-drawn characters and real humans, most of the aforementioned approaches, trained on real-world human video datasets [Jafarian and Park 2021; Zablotskaia et al. 2019], fail to generalize effectively to diverse stylized characters. It is impractical to prepare a large-scale dataset for these stylized character animations to address the domain gap. Animate-X [Tan et al. 2025] attributes this limitation to an insufficient understanding of driving video motion patterns and misalignment with reference appearances. Thus, Animate-X introduces a Pose Indicator to extract motion patterns through both implicit and explicit mechanisms. Nevertheless, it overlooks the semantic understanding of reference images, and training on human videos struggles to establish correct contour relationships for characters in drawing styles. To bridge this gap, MikuDance [Zhang et al. 2024] proposes a mixed-control diffusion module that implicitly aligns the scale and body shape of stylized characters with motion guidance. However, its dataset is heavily biased toward a specific anime style, which limits its generalization to a wider range of stylized characters.

Unlike the above approaches, we construct an explicit 3D model for an input hand-drawn character and use it to render guidance for directing the diffusion model's generation process. The generation task is repurposed as an inpainting task, where the rendered guidance ensures consistency in identity and pose, even under complex 3D motions. Additionally, our approach reduces the reliance on large-scale datasets for model training.

3 METHOD

Given an input hand-drawn character as a reference image I_{ref} (we assume that the character in the reference image is approximately in a frontal A/T pose), manually-processed hair-body segmentation maps provided by the user (optional for long-hair cases) and a target 3D motion $\mathcal{P}$, our system is designed to generate a vivid 3D character animation $\{\hat{I}_{1:N}\}$. An overview of our system is shown in Fig. 3. During the image-to-3D process, we design a Hair Layering Modeling (HLM) method based on hair-body segmentation maps (S_{front}, S_{right}) to deal with challenging characters with long hair (Section 3.2). For characters without long hair, this step is optional. Users can optionally specify areas requiring improvement, particularly for dynamics enhancement, through SDI masks ($M^{\text{SDI}}_{\text{front}}$, $M^{\text{SDI}}_{\text{back}}$). Then our system reconstructs a 3D textured geometry $\mathcal{G}$ and renders three types of guidance sequences (Section 3.3), which

provide essential 3D geometric consistency for subsequent animation generation. We design a domain-adapted video diffusion model to translate rendered coarse color frames into stylized ones (Section 3.4). During inference, we employ a Secondary Dynamics Injection (SDI) strategy to further inject dynamic motions into the masked regions (Section 3.5).

3.1 Preliminaries

Considering the computation cost, recent diffusion models [Blattmann et al. 2023; Rombach et al. 2022] operate within the compressed latent space of a pre-trained variational autoencoder (VAE), denoted as $E(\cdot)$ and $D(\cdot)$. Starting with the encoded latent representation $z_0 = E(I)$ of an image I, the forward process $q(z_t|z_0, t)$ gradually corrupts the initial latent z_0 into Gaussian noise over T diffusion steps, following a Markov chain [Ho et al. 2020]. At each step t, noise $\epsilon \sim N(0, I)$ is added to the previous latent state z_{t-1} according to a predefined noise schedule$\{\alpha_t, \sigma_t\}$:

$$z_t = \alpha_t z_0 + \sigma_t \epsilon. \tag{1}$$

Conversely, the denoising process $p_\theta(z_{t-1}|z_t, t)$ iteratively removes noise over T steps using a learned denoising network. Recently, v-prediction [Salimans and Ho 2022] has been proposed in the denoising process for enhanced numerical stability. This approach parameterizes the denoising direction as a velocity v_t, defined as a linear combination of the signal z_0 and noise ϵ:

$$v_t = \alpha_t \epsilon - \sigma_t z_0. \tag{2}$$

Unlike the original ϵ-prediction, where the model predicts noise, the v-prediction model predicts the velocity. The training objective of the model is:

$$\mathcal{L}_\theta = \mathbb{E}_{z_0,t,\epsilon}[||v_t - v_\theta(z_t, t)||_2^2], \tag{3}$$

where θ is the model parameter. During sampling at the step t, the denoised latent estimate $\hat{z}_0^t$ can be recovered from the predicted velocity using the following function:

$$\hat{z}_0^t = \alpha_t z_t - \sigma_t v_\theta(z_t, t). \tag{4}$$

3.2 Hair Layering Modeling

Our method begins by generating a 3D textured geometry $\mathcal{G}$ from the reference image I_{ref}. Following DrawingSpinUp [Zhou et al. 2024], we first remove non-photorealistic contour lines from I_{ref} to be a contour-free image $I_{\text{ref}}^{\text{nc}}$. Next we use Wonder3D [Long et al. 2024] to generate multi-view images (I_{front}, I_{right}, ..., I_{back}) and reconstruct a neural implicit signed distance field $\mathcal{I}$ from multi-view images. To address the unnatural deformation artifacts for characters with long hair (see Fig. 2 (a)-(b)), we have developed a Hair Layering Modeling (HLM) method to separate hair from the body. As shown in Fig. 4, based on the manually-processed hair-body segmentation map provided by the user, we first manually segment the front-view and right-view foreground maps to be the hair-body segmentation maps S_{front} and S_{right}. Then we separate them into the hair segmentation maps ($S_{\text{front}}^{\text{hair}}$ and $S_{\text{right}}^{\text{hair}}$) and the body segmentation maps ($S_{\text{front}}^{\text{body}}$ and $S_{\text{right}}^{\text{body}}$). We also obtain an $S_{\text{back}}^{\text{hair}}$ by filling the internal region of $S_{\text{front}}^{\text{hair}}$ for the extraction of hair from the back of the head. Finally, the implicit field separation is formulated as:

$$\begin{aligned}
\mathcal{S}_{\text{hair}} &= S_{\text{front}}^{\text{hair}} \cup \left(S_{\text{back}}^{\text{hair}} \cap S_{\text{right}}^{\text{hair}}\right), \\
\mathcal{S}_{\text{body}} &= S_{\text{front}}^{\text{body}} \cap S_{\text{right}}^{\text{body}}, \\
\mathcal{I}_{\text{hair}} &= \mathcal{I} \odot \mathcal{S}_{\text{hair}}, \\
\mathcal{I}_{\text{body}} &= \mathcal{I} \odot \mathcal{S}_{\text{body}},
\end{aligned} \tag{5}$$

where $\odot$ represents the element-wise product. The individual geometries $\mathcal{G}_{\text{hair}}$ and $\mathcal{G}_{\text{body}}$ are reconstructed from their respective implicit fields $\mathcal{I}_{\text{hair}}$ and $\mathcal{I}_{\text{body}}$ via the Marching Cubes algorithm [Lorensen and Cline 1998], which are combined to form the final geometry $\mathcal{G}$. We then utilize the Mixamo [Inc. 2023] to create a rigged 3D character and retarget the target 3D motion onto it.

3.3 Guidance Sequence Rendering

We generate three types of guidance sequences (i.e., pose $\{P_{1:N}\}$, SDI mask $\{M_{1:N}^{\text{SDI}}\}$ and coarse color $\{C_{1:N}\}$) to provide geometric consistency for animation generation. We extract the pose sequence $\{P_{1:N}\}$ and reference pose P_{ref} from the animated character using the OpenPose [Cao et al. 2019] 18-keypoint format, excluding hand and feet joints due to the abstract nature of hand-drawn characters. As shown in Fig. 5, we allow users to provide optional SDI masks $M_{\text{front}}^{\text{SDI}}$ and $M_{\text{back}}^{\text{SDI}}$ to specify areas such as hair ends and skirts that require secondary dynamics enhancement (where a value of 1 indicates the regions to be enhanced and a value of 0 indicates the regions to be preserved). The SDI masks corresponding to all the examples used in this paper are shown in Supplementary Materials Fig. 1. We back-project the SDI masks onto the geometry $\mathcal{G}$ to recolor each vertex as white (value of 1) or black (value of 0). Subsequently, we render the vertex colors as SDI mask sequences $\{M_{1:N}^{\text{SDI}}\}$ to guide secondary dynamics injection. We render the animated 3D character back into the 2D image domain to generate a coarse color sequence $\{C_{1:N}\}$. We mask $\{C_{1:N}\}$ with $\{M_{1:N}^{\text{SDI}}\}$ to get a masked coarse color sequence $\{C_{1:N}^{\text{masked}}\}$ by $\{C_{1:N}^{\text{masked}}\} = \{C_{1:N}\} \cdot (1 - M_{1:N}^{\text{SDI}})$. Leveraging our domain-adapted video diffusion model, we subsequently refine the preserved regions and redraw the masked regions using our Secondary Dynamics Injection strategy.

3.4 Coarse Animation Refinement

Based on the three types of guidance sequences, our objective is to enhance the masked coarse color sequence by addressing two key aspects: insufficient appearance details and a lack of rich secondary motion. Ultimately, our domain-adapted model v_θ aims to generate a temporally coherent and realistic video sequence $\{\hat{I}_{1:N}\}$ from multiple conditions (i.e., I_{ref}, P_{ref}, $\{P_{1:N}\}$, $\{C_{1:N}^{\text{masked}}\}$ and $\{M_{1:N}^{\text{SDI}}\}$). We start with the state-of-the-art pose-controlled human animation method, UniAnimate [Wang et al. 2024c], which provides multiple natural motion priors. We then repurpose this generation task as an inpainting task tailored for the hand-drawn character domain. The rendered guidance sequences offer a stable, multi-view foundation during generation while significantly reducing the need for extensive tuning data.

3.4.1 Model Architecture. As shown in Fig. 6, the main architecture of our domain-adapted model v_θ is similar to UniAnimate, with

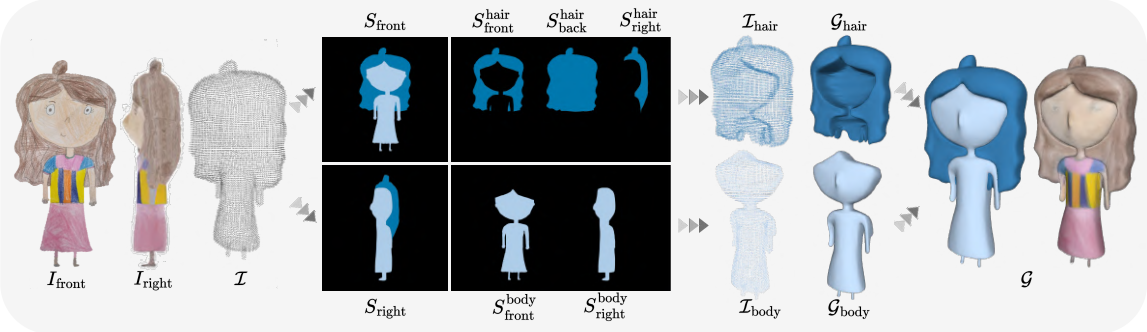


Fig. 4. An illustration of hair layering modeling.

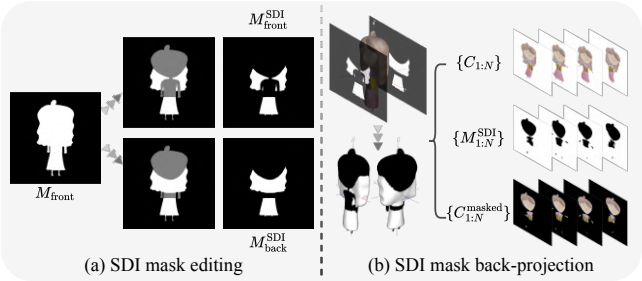


Fig. 5. An illustration of how we obtain the SDI mask sequence.

some minor modifications: we add a coarse prior encoder $\mathscr{E}_{\text{coarse}}$ to embed the masked coarse color sequence $\{C^{\text{masked}}_{1:N}\}$ and the SDI mask sequence $\{M^{\text{SDI}}_{1:N}\}$. $\mathscr{E}_{\text{coarse}}$ is similar to the lightweight STC-encoder [Wang et al. 2023], consisting of two 2D convolutional layers and a temporal Transformer layer to capture the spatial-temporal relations across all coarse frames. Specifically, $\{C^{\text{masked}}_{1:N}\}$ is concatenated with its corresponding $\{M^{\text{SDI}}_{1:N}\}$ along the channel axis and then fed into $\mathscr{E}_{\text{coarse}}$ to generate coarse prior embeddings. Such embeddings are subsequently concatenated with the input noise along the channel axis and then fed into the denoising UNet.

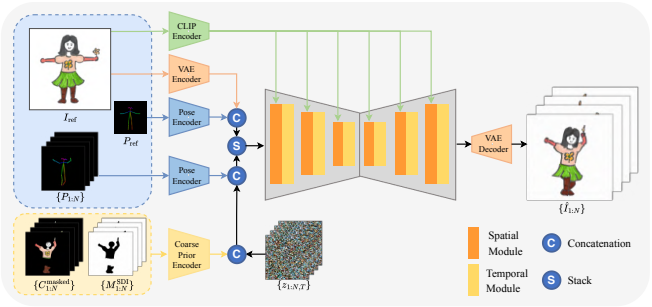


Fig. 6. An illustration of our domain-adapted diffusion model v_θ.

3.4.2 Model Tuning. We prepare a hand-drawn animated video dataset (Section 4.1) for model tuning to reduce the domain gap. Due to the significantly smaller size of our constructed dataset compared to the original training data for human dance videos (more than 10K), tuning the entire model on the limited motion variation and quantity poses a risk of catastrophic forgetting. Similar issues have been noted in ToonCrafter [Xing et al. 2024]. Following their strategy, we fine-tune the spatial layers to adapt the stylized appearance and freeze the temporal layers to maintain the real-world motion prior. Generally, the spatial layers primarily focus on appearance modeling, while the temporal layers aim to ensure motion coherence between frames. Additionally, we train the coarse prior encoder $\mathscr{E}_{\text{coarse}}$ to enhance its ability to represent the coarse context and the intended refinement direction. The spatial layers and the coarse prior encoder are trained simultaneously using the general diffusion loss (Eq. 3).

3.5 Secondary Dynamics Injection

In skeletal animation, skinning weights are usually determined by the distance between vertices and the nearest bone. While skinning-based deformation effectively generates primary motion, it often fails to capture realistic secondary motion, such as cloth swaying or hair flowing. This issue is also evident in our synthesized dataset. Although we freeze the temporal layers during training to preserve the natural motion priors inherent in real-world dynamics, the model's output after fine-tuning still more or less lacks the richness of secondary motion (Section 4.5.3). Inspired by prior works [Kim et al. 2025; Rout et al. 2025; Zhou et al. 2025], which leverage pre-trained diffusion models to guide generation for tasks such as super-resolution, image editing, and relighting, we propose Secondary Dynamics Injection to harness the real-world motion priors encoded in the native pre-trained UniAnimate, represented as u_θ (distinguished from our domain-adapted model v_θ). This strategy guides the denoising process in our framework, as detailed in Supplementary Materials Algorithm 1.

3.5.1 Blending Latent Estimates. To determine when and how to guide the denoising process, we begin by visualizing the frames decoded from the latent estimates $\{\hat{z}^{t,u_\theta}_{1:N,0}\}$, computed according to the equation 4, at each denoising step t, as illustrated in Fig. 7. From this figure, we observed: (1) The initial denoising steps primarily establish the spatial structure of the frames, while the subsequent steps progressively refine the texture details. (2) In the steps associated with spatial distribution, the primary motion is generated first, with secondary motion details being gradually introduced in the later steps. Based on these observations, we propose to blend the noise-free latent estimates $\{\hat{z}^{t,u_\theta}_{1:N,0}\}$ and $\{\hat{z}^{t,v_\theta}_{1:N,0}\}$ from the pre-trained model u_θ and our domain-adapted model v_θ separately,

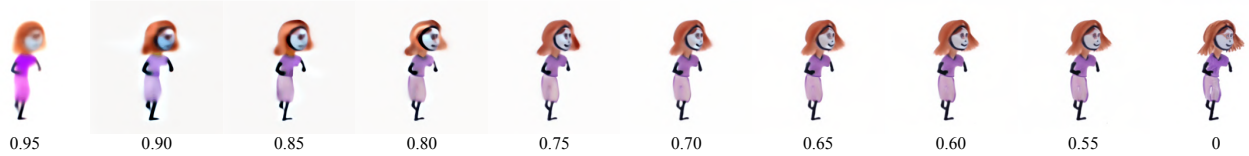


Fig. 7. Gradually decreasing t-step denoised latent estimation results. The numbers below the images represent the percentage of the denosing process.

during the early denoising steps to introduce more dynamic motion, as the estimated $\hat{z}_0^t$ indicates the denoising direction in v-prediction [Salimans and Ho 2022] process.

Specifically, we divide the generation process into three phases using two critical thresholds τ_2 and τ_1 ($T > \tau_2 > \tau_1$) expressed as percentages of the total timesteps ($\tau_2 = \alpha \cdot T$ and $\tau_1 = \beta \cdot T$ where $\alpha, \beta \in [0, 1]$). Here, τ_2 and τ_1 control the starting and ending step of secondary motion injection, respectively. First, during $[T, \tau_2)$, we employ only our model for denoising to ensure identity preservation and primary motion generation. Second, in $[\tau_2, \tau_1)$, we blend the two noise-free latent estimates with the downsampled masks $\{M^{\text{SDI}}_{1:N,\text{down}}\}$. We use the n-th frame as an example to illustrate the latent fusion function:

$$\hat{z}^{t,\text{blend}}_{n,0} = (1 - M^{\text{SDI}}_{n,\text{down}}) \cdot \hat{z}^{t,v_\theta}_{n,0} + M^{\text{SDI}}_{n,\text{down}} \cdot \hat{z}^{t,u_\theta}_{n,0}. \tag{6}$$

3.5.2 Reference Switching. However, when guiding hand-drawn characters, the pre-trained model u_θ presents two key challenges previously mentioned (Fig. 2 (c) and (d)): (1) Difficulty in accurately modeling the motion of contour lines. (2) Inability to fully capture the semantics of the reference drawings. To address these issues, during the initial denoising process, we utilize the contour-free reference image $I^{\text{nc}}_{\text{ref}}$ and leverage our domain-adapted model to produce initial latent estimates, which serve as a structurally aligned starting point that respects the reference style.

The combination of latent estimates will provide a new optimization direction. While replacing the contour-free reference $I^{\text{nc}}_{\text{ref}}$ with the original reference I_{ref} during later denoising steps $[\tau_1, 1]$ introduces contour structures, the results often fail to faithfully preserve the reference's contours (as analyzed in Section 4.5.4). Thus, in the third phase, we first augment the masked regions in $\{C^{\text{masked}}_{1:N}\}$ with the estimated video $D(\{\hat{z}^{\tau_1,\text{blend}}_{1:N,0}\})$, where $D(\cdot)$ is the VAE decoder, to obtain the inpainted coarse frames $\{C^{\text{inpainted}}_{1:N}\}$ via Poisson Blending [Pérez et al. 2003]. As illustrated in Fig. 8, compared with directly concatenating C^{masked}_n with the estimated video $D(\{\hat{z}^{\tau_1,\text{blend}}_{n,0}\})$, blending them using Poisson Blending achieves more natural stitching.

3.5.3 Re-denoising. Constrained with the updated coarse inputs with rich dynamics and the reference image I_{ref} with contours, we employ our domain-adapted model v_θ to re-denoise the initial noise $\{z_{1:N,T}\}$ from scratch. Since only early denoising steps influence the motion distribution, when $\tau_1 < 0.7$, the impact on motion distribution remains minimal; however, this reduction in τ_1 leads to an increase in processing time.

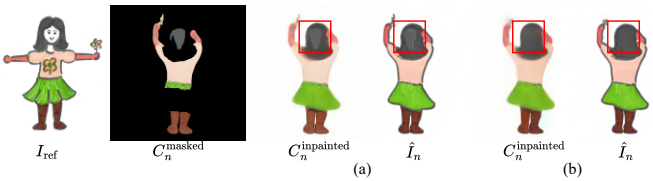


Fig. 8. An example illustrating Poisson Blending for inpainting the n-th masked coarse color image. (a) shows the inpainting of the n-th masked coarse color image C_n^{masked} by directly concatenating it with the estimated video $D(\{\hat{z}^{\tau_1,\text{blend}}_{n,0}\})$. (b) demonstrates the inpainting process by mixing them using Poisson Blending.

4 EXPERIMENTS

4.1 Dataset Construction

Due to the scarcity of hand-drawn animated video data that features plausible secondary motion, we created a small-scale primary animation dataset as a compromise. This dataset is used to help the model refine texture details and contour lines, reducing the gap between real-human and hand-drawn characters. We collect 174 high-quality character drawing images as training and evaluation data from two existing datasets: the Amateur Drawings Dataset [Smith et al. 2023] and the SketchAnim Dataset [Rai et al. 2024]. Since Mixamo's auto-rigging tool may not function properly if the character is significantly asymmetric or posed, before rigging, we applied rotation or local deformation to the severely asymmetric examples in these images to prevent rigging failures. Each character is assigned 1-2 motions selected from the Mixamo Animation Dataset, varying in length from a few dozen to several hundred frames, encompassing both simple and complex motions. To improve multi-view consistency, the training set also includes a 60-frame full rotation of the character in a rest pose. We use the stylized animation videos generated by DrawingSpinUp [Zhou et al. 2024] as ground truth to fine-tune our diffusion models. Ultimately, we obtained 428 high-quality drawing animation video clips, which were randomly divided into two sets: the training set contains 359 clips covering 124 different characters, while the evaluation set includes 69 clips with 50 different characters.

4.2 Implementation Details

The experiments were conducted using a single 80G NVIDIA A100 GPU. We optimized our diffusion model v_θ using an AdamW optimizer over 40k steps, with a learning rate of 2e-5 and a batch size of 4. For training videos, we sampled 16 frames from each video and randomly cropped a 768×512 region. In each iteration, we randomly decided whether to apply a mask, helping the model to refine textures in unmasked regions. We enhanced mask diversity by incorporating random masks and square masks, adapted from ProPainter [Zhou et al. 2023]. During inference, we adopted the DDIM sampler

[Song et al. 2021] with 20 steps. Based on experimental results, we recommend setting the hyperparameters: α in the range $[0.7, 0.95]$ and β in the range $[0.5, 0.7]$. For long video generation, we adopt the sliding window strategy proposed in [Xu et al. 2024; Zhu et al. 2024], synthesizing videos segment by segment with temporally overlapping frames. For the input conditions of overlapping frames, we leverage inpainted coarse frames $\{C_{1:N}^{\text{inpainted}}\}$ generated from previous segments to guide the generation of subsequent segments. Given a character drawing, it takes 3-5 minutes to generate a rigged 3D character (done only once for retargeting various animations). Our model generates a 32-frame video at a resolution of 768×512 in 45 seconds on a single A100 GPU without the Secondary Dynamics Injection strategy. When SDI is introduced, the total time increases to 72, 81, and 90 seconds for β values of 0.7, 0.6, and 0.5, respectively.

4.3 Qualitative Results

We evaluate our method against state-of-the-art animation methods, categorized into two paradigms: (1) traditional skeletal animation methods: Smith et al. [2023] and DrawingSpinUp [Zhou et al. 2024]; (2) diffusion-based methods: AnimateAnyone [Hu 2024], MikuDance [Zhang et al. 2024], and UniAnimate [Wang et al. 2024c]. Additionally, we implemented UniAnimate* by fine-tuning it on our synthesized dataset. For Smith et al. [2023] and DrawingSpinup [Zhou et al. 2024], we utilize their official implementations. For UniAnimate [Wang et al. 2024c], we also employ their official implementation. Additionally, we fine-tune UniAnimate on our synthesized dataset, following a similar tuning approach as described in Section 4.2, referred to as UniAnimate*. For AnimateAnyone [Hu 2024], we generate results using the publicly available reproduced code [MooreThreads 2023], as the official implementation was not accessible. For MikuDance [Zhang et al. 2024], we use the official code but disable scene motion tracking to ensure a fair comparison, given that our setup assumes a clean white background.

Ours vs Skeletal Animation Methods. Fig. 9 compares our method with two traditional skeletal animation techniques. Our observations are as follows. Both DrawingSpinUp and our method, supported by 3D modeling, yield more plausible 3D-aware animations that faithfully represent the target motions than Smith et al. Since skeletal animation can only generate primary motion, Smith et al. and DrawingSpinUp struggle to produce natural secondary movements. In contrast, our method effectively captures subtle dynamics, such as the swinging motion of the little girl's ponytails as she jumps (Row 1). In the results of Smith et al. and DrawingSpinUp, unnatural deformation artifacts occur regardless of whether a 2D or 3D mesh is used, as seen in the entanglement between the hair and shoulders of the characters (Row 2-3). In contrast, our method animates these challenging long-hair cases more naturally.

Ours vs Diffusion-Based Methods. Fig. 10 compares our method with several diffusion-based animation approaches. Our method demonstrates robust generalization across diverse character styles, from humanoid designs to highly stylized drawings, while faithfully preserving identity and delivering smooth, natural animations. Diffusion-based methods show limited ability for humanoid characters. They often struggle to maintain view-consistent contour lines due to the absence of such features in their training data (Row 2). Fine-tuning with our domain-specific data successfully achieves stylized appearance adaptation, as demonstrated in Fig. 10 (Column *UniAnimate* and *UniAnimate**). However, these methods frequently misinterpret the 2D skeleton, resulting in unnatural or erroneous animations when handling complex 3D motions (Rows 1 and 3). We hypothesize that the inherent limitations of 2D skeletons, particularly their inability to adequately represent occlusions and depth variations in intricate 3D movements, are the primary cause of these errors. To address this issue, our method leverages coarse input to enrich primary motion understanding, enabling more accurate and realistic animations in challenging scenarios. When applied to characters with styles that deviate significantly from typical human appearances, methods like UniAnimate and AnimateAnyone often misclassify parts, or even the entirety, of the character as background elements (Rows 1 and 3). This results in incomplete or entirely static animations, highlighting their inability to adapt to nonstandard character designs. Although MikuDance benefits from extensive training on large-scale anime datasets, it still struggles to generalize effectively to hand-drawn characters. Its reliance on specific data distributions limits adaptability to characters with unique stylistic elements or unconventional features.

4.4 Quantitative Results

We conducted a quantitative evaluation to assess the texture quality of the animation results. We measure LPIPS [Zhang et al. 2018] for texture consistency, FID [Heusel et al. 2017] for overall distribution similarity, and CLIP [Radford et al. 2021] similarity for assessing semantic alignment. These metrics are calculated between the generated frames and the reference images. Considering that UniAnimate, AnimateAnyone, and MikuDance fail to accurately animate the reference image with the 3D poses, they may directly maintain some texture in the reference images, as shown in Fig. 10. This may lead to biased results. Therefore, we omit comparisons with these methods, focusing instead on DrawingSpinUp and UniAnimate*, which are better at achieving accurate pose variation. Table 1 presents the quantitative results, and our method outperforms all compared approaches, demonstrating its ability to faithfully restore the texture of stylized hand-drawn characters across various 3D poses.

Method	LPIPS ↓	FID ↓	CLIP ↑
UniAnimate*	0.1792	158.1467	0.8964
DrawingSpinUp	0.1734	157.7452	0.8880
Ours	**0.1733**	**152.9022**	**0.9030**

Table 1. Quantitative comparisons with two pose animation methods.

4.5 Ablation Study

4.5.1 Ablation on HLM. We conducted an ablation study to verify the effectiveness of our Hair Layering Modeling (HLM). As shown in Fig. 11, the unnatural deformation in the face and armpits caused by the hair adhering to the shoulders adversely affects subsequent animation generation, degrading the final animation results.

4.5.2 Ablation on Training Components of v_θ. To validate the effectiveness of our domain-adapted diffusion model v_θ, we conducted

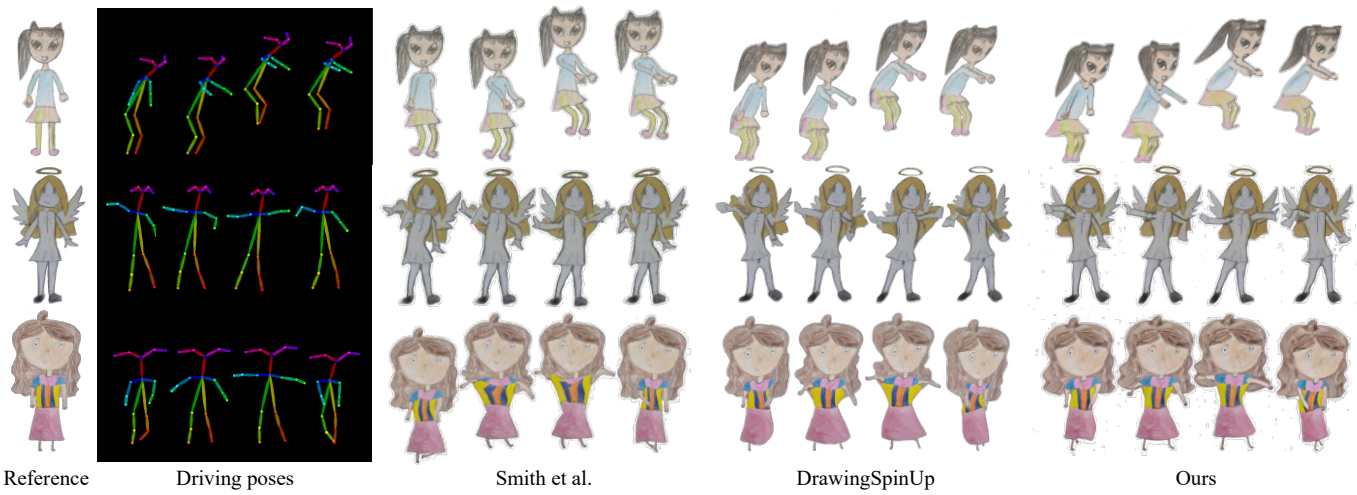


Fig. 9. Visual comparisons with two traditional skeletal animation methods.

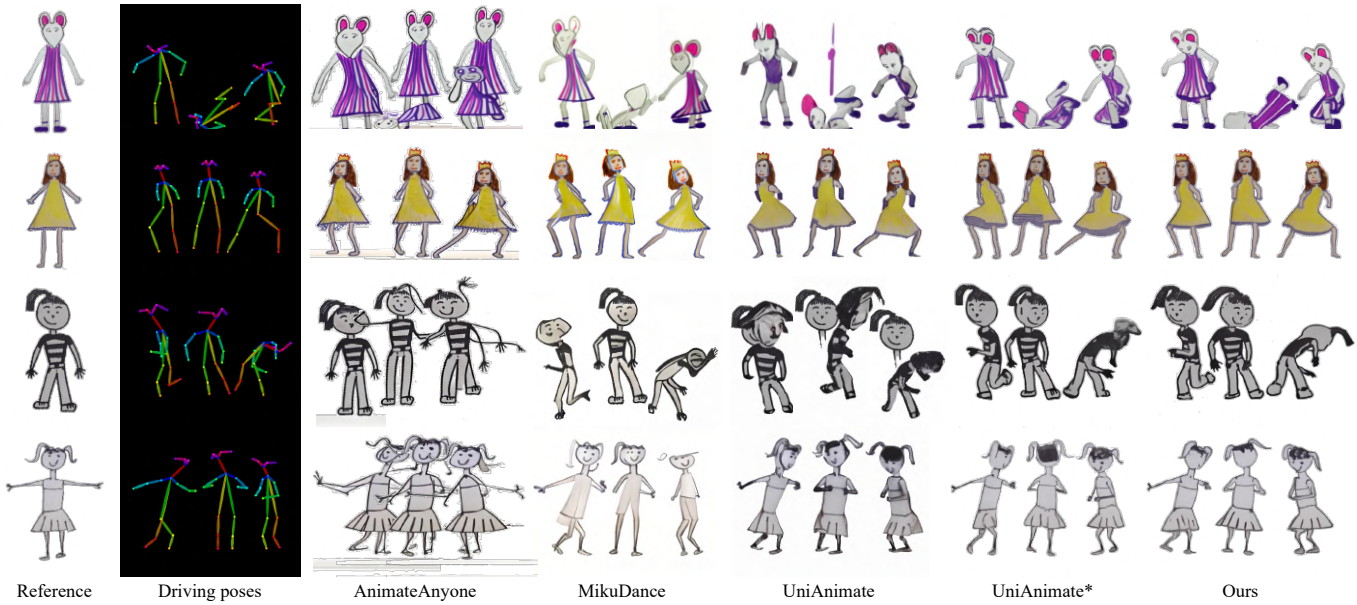


Fig. 10. Visual comparisons with four diffusion-based animation methods.

an ablation study to demonstrate the contributions of two key components, the coarse prior encoder $\mathscr{E}_{\text{coarse}}$ and the choice of spatial layer tuning (SLT). We compare the following configurations: (I) Ablating the coarse prior encoder $\mathscr{E}_{\text{coarse}}$; (II) Replacing spatial layer tuning (SLT) with temporal layer tuning (TLT); (III) Our method. As shown in Fig. 12, we can see that changing the key components significantly degrades performance, especially in 3D motion understanding. Although ablating $\mathscr{E}_{\text{coarse}}$ achieves stylized appearance adaptation, it fails to produce plausible results for complex poses involving occlusions or intricate 3D spatial relationships. This leads to limb confusion (e.g., left-right arm swaps) and the erroneous generation of occluded limbs that should be hidden. As mentioned in Section 4.3, the limitation of the 2D pose input is its inability to explicitly encode 3D spatial priors, such as depth ordering and occlusion cues, leading to geometric inconsistencies. $\mathscr{E}_{\text{coarse}}$ addresses this by introducing dense, structured 3D guidance, which disambiguates limb positions and enforces physically plausible occlusion reasoning. Replacing SLT with TLT achieves partial adaptation but fails to establish correct relationships between generated content and the reference image for masked regions. This is evidenced by artifacts such as hat discontinuities (Row 1) and erroneous texture propagation (e.g., using trouser patterns to synthesize hair) (Row 2). These results highlight that spatial layers play a more critical role in appearance modeling and should be prioritized during fine-tuning for robust adaptation.

4.5.3 Ablation on SDI. To evaluate the effectiveness of our Secondary Dynamics Injection (SDI) strategy, we performed an ablation study that compares scenarios with and without SDI. As shown in Fig. 13, compared with only using the domain-adapted model without SDI, our SDI strategy effectively generates secondary dynamics, extending beyond the primary dynamics.

Fig. 11. Ablation on our hair layering modeling (HLM) method.

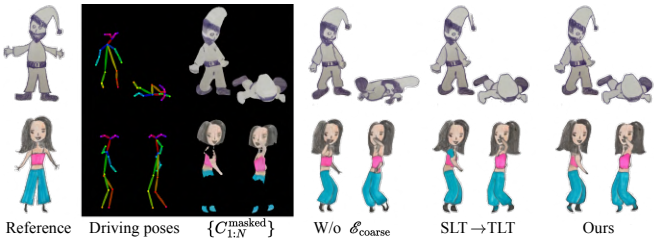


Fig. 12. Ablations on training components of v_θ.

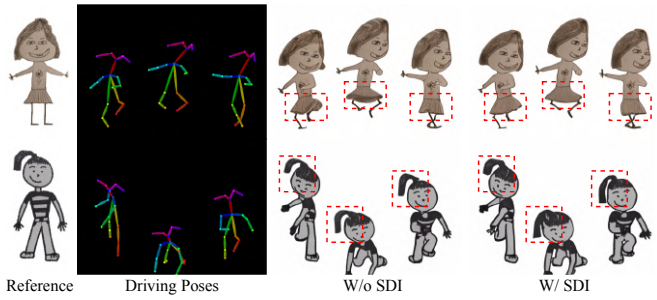


Fig. 13. Ablations on our Secondary Dynamics Injection (SDI) strategy.

4.5.4 Ablation on Key Components of SDI. To further evaluate the modular design of our Secondary Dynamics Injection (SDI) strategy, we conducted an ablation study to demonstrate the contributions of sub-modules: blending (Bld) with the pre-trained model u_θ, reference switching (RS) between the reference image I_{ref} and the contour-free reference image $I^{\text{nc}}_{\text{ref}}$, and re-denoising (RD). We compare the following configurations, with all comparisons using our domain-adapted model v_θ during denoising steps $[T, \tau_2)$: (I) Directly using v_θ conditioned on I_{ref} during steps $[\tau_2, 1]$ (*W/o SDI*); (II) Blending the latent estimates from v_θ and u_θ conditioned on I_{ref} during steps $[\tau_2, 1]$ (*Bld*); (III) Blending the latent estimates from v_θ and u_θ, initially conditioned on $I^{\text{nc}}_{\text{ref}}$ during steps $[\tau_2, \tau_1)$ and then switching to I_{ref} during $[\tau_1, 1]$ (*Bld+RS*); (IV) Blending the latent estimates from v_θ and u_θ, initially conditioned on $I^{\text{nc}}_{\text{ref}}$ during steps $[\tau_2, \tau_1)$ and then re-denoising from initial noise during steps $[T, 1]$ with I_{ref} (*Bld+RS+RD*).

As shown in Fig. 14, only using the domain-adapted model (Column *W/o SDI*) captures primary motion but fails to generate secondary dynamics. Blending with the pretrained model u_θ using contour references (Column *Bld*) often misplaces contour patterns into internal regions, primarily because of the incorrect texture propagation of the pretrained model u_θ. Initial use of contour-free references (Column *Bld+RS*) mitigates these artifacts. However, late-stage introduction of contour references at $[\tau_1, 1]$ leads to visible discrepancies in generated contours and inconsistent contour style preservation. Thus, our final solution (Column *Bld+RS+RD*) addresses these limitations through intermediate coarse video estimation and blending to ensure faithful motion injection and consistent style preservation.

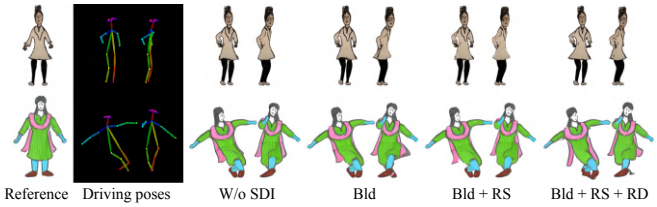


Fig. 14. Ablations on three key components of our Secondary Dynamics Injection (SDI) strategy. The following abbreviations are used: Bld for blending, RS for reference switching, and RD for re-denoising.

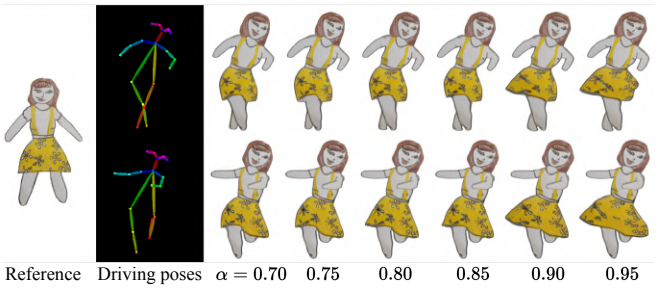


Fig. 15. The impact of different values of α between 0.70 and 0.95 on the secondary dynamics of the animation results, with $\beta = 0.60$.

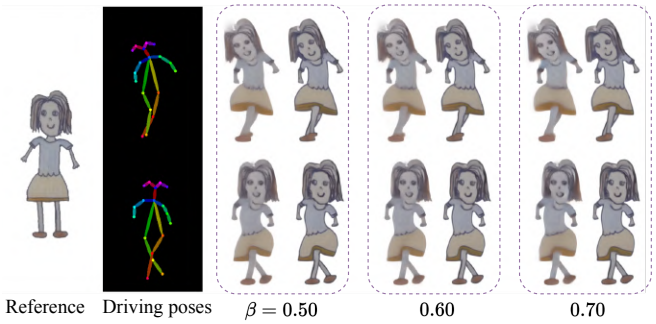


Fig. 16. The impact of different values of β (0.50, 0.60, 0.70) on the secondary dynamics of the animation results, with $\alpha = 0.95$.

4.5.5 Ablation on Key Parameters of SDI. To evaluate the effectiveness of our Secondary Dynamics Injection (SDI) strategy, we further examined the effects of varying the two key parameters, α and β, which define the thresholds τ_2 and τ_1, determining the transitions between different phases of the SDI process. We first vary α at values between 0.70 and 0.95 while keeping β fixed at 0.60. Then we set $\alpha = 0.95$ and vary β at values of 0.50, 0.60, and 0.70. As shown in Fig. 15, α significantly influences the initiation of the injection from the pretrained model u_θ. When α is at relatively low values (e.g., 0.7), the results are similar to those without SDI. As α increases, the motion intensity becomes more pronounced, until at very high values (e.g., 0.95), some artifacts begin to appear. Therefore, there is a reasonable range for the values of α, enabling users to achieve the desired effect by adjusting α. Fig. 16 illustrates the inpainted coarse color images $\{C^{\text{inpainted}}_{1:N}\}$ along with the corresponding results $\{\hat{I}_{1:N}\}$. It is observed that the results do not exhibit significant variations when β changes. This also demonstrates the robustness of our coarse prior encoder. From the perspective of minimizing runtime, we select 0.6 or 0.7 as an appropriate value for β.

5 APPLICATION

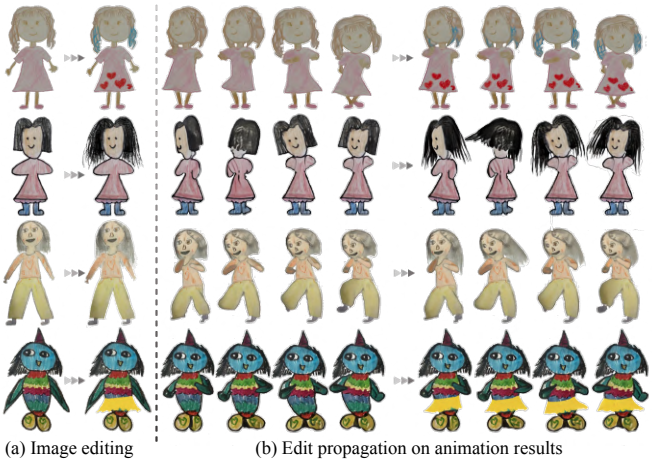


Fig. 17. Edit propagation. The reference frame (left) is edited, and the change is propagated to all animation frames.

Our system supports user-friendly editing, allowing users to change characters' appearances with just a few strokes. Users can locally erase and repaint on the reference image, and the system updates the SDI mask sequence accordingly. As shown in Fig. 17, users can edit character patterns and designs, as well as modify shape features like hairstyle and clothing. Edits are automatically propagated across all frames without regenerating the 3D geometry.

6 CONCLUSION

We introduced a hybrid system for animating hand-drawn characters with natural 3D motion by integrating skeletal animation and video diffusion priors. Our approach begins by generating coarse images via skeletal animation that maintain geometric consistency, followed by refinement using a domain-adapted diffusion model. Through our investigation of the denoising process in video diffusion models, we observed for the first time that different denoising steps are closely associated with distinct types of motion. This insight inspired the development of our novel Secondary Dynamics Injection (SDI) strategy. The SDI strategy enhances motion realism in the refined images by guiding the denoising process with a pretrained diffusion model, which leverages rich and realistic human motion priors. We also addressed unnatural deformation artifacts caused by the integrated hair-body single-mesh geometry with a Hair Layering Modeling (HLM) technique, enabling more natural animations for characters with long hair. Our experiments demonstrate that the proposed method produces visually compelling 3D animations while preserving the artistic style and conveying natural secondary dynamics.

In comparing our method to simulation-based secondary dynamics, it's important to note that the 3D models generated by our approach are rough geometries used as proxies, which do not meet the input requirements for simulation-based animation. Additionally, simulation methods are often labor-intensive and require specialized expertise, making them less accessible for novices. In contrast, our method prioritizes user-friendliness, enabling users with limited experience to create animations efficiently while tackling key challenges. The user-friendly editing applications also highlight the practicality and significance of our system.

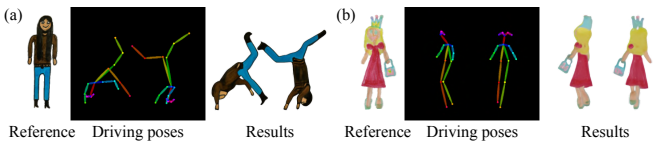


Fig. 18. Two examples for the limitations of our method.

There are still some limitations in our method. First, while the system supports inverted poses through skeletal retargeting (Fig. 18 (a)), the pre-trained human video diffusion model lacks sufficient priors for such motions, as they are rare in human video datasets. This limitation can lead to unrealistic secondary dynamics, such as hair failing to fall naturally under gravity. Second, to prevent the model from learning unnatural deformations caused by the integrated hair-body geometry (Fig. 2), such long-hair cases were excluded from training. However, this exclusion further limits the model's ability to generalize when inpainting hair regions in back views of long-hair characters. See the incorrect hair texture on the left shoulder in Fig. 18 (b). These two issues could potentially be addressed in the future by leveraging more powerful and realistic video generation models, capable of better handling complex motions. Lastly, our current framework does not support explicit facial expression control, which deserves future exploration.

REFERENCES

Otman Benchekroun, Jiayi Eris Zhang, Siddartha Chaudhuri, Eitan Grinspun, Yi Zhou, and Alec Jacobson. 2023. Fast Complementary Dynamics via Skinning Eigenmodes. *ACM Transactions on Graphics* 42, 4 (2023), 1–21.

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

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

Xinyi Fan, Amit H Bermano, Vladimir G Kim, Jovan Popović, and Szymon Rusinkiewicz. 2018. ToonCap: A Layered Deformable Model for Capturing Poses from Cartoon Characters. In *Proceedings of the Joint Symposium on Computational Aesthetics and Sketch-Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering*. 1–12.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Advances in Neural Information Processing Systems* 30 (2017).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems* 33 (2020), 6840–6851.

Alexander Hornung, Ellen Dekkers, and Leif Kobbelt. 2007. Character Animation from 2D Pictures and 3D Motion Data. *ACM Transactions on Graphics* 26, 1 (2007), 1–es.

Li Hu. 2024. Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8153–8163.

Takeo Igarashi, Tomer Moscovich, and John F Hughes. 2005. As-Rigid-As-Possible Shape Manipulation. *ACM Transactions on Graphics* 24, 3 (2005), 1134–1141.

Adobe Systems Inc. 2023. Mixamo. https://www.mixamo.com

Yasamin Jafarian and Hyun Soo Park. 2021. Learning High Fidelity Depths of Dressed Humans by Watching Social Media Dance Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12753–12762.

Eakta Jain, Yaser Sheikh, Moshe Mahler, and Jessica Hodgins. 2012. Three-Dimensional Proxies for Hand-Drawn Characters. *ACM Transactions on Graphics* 31, 1 (2012), 1–16.

Younghyun Kim, Geunmin Hwang, Junyu Zhang, and Eunbyung Park. 2025. Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 39. 4338–4346.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. 2023. One-2-3-45: Any Single Image to 3D Mesh in 45 Seconds without Per-Shape Optimization. *Advances in Neural Information Processing Systems* 36 (2023), 22226–22246.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. 2024. Wonder3D: Single Image to 3D using Cross-Domain Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9970–9980.

William E Lorensen and Harvey E Cline. 1998. Marching Cubes: A High Resolution 3D Surface Construction Alghrithm. In *Seminal Graphics: Pioneering Efforts that Shaped the Field*. 347–353.

MooreThreads. 2023. Moore AnimateAnyone. https://github.com/MooreThreads/Moore-AnimateAnyone. Accessed: 2025-05-02.

Patrick Pérez, Michel Gangnet, and Andrew Blake. 2003. Poisson Image Editing. *ACM Transactions on Graphics* 22, 3 (2003), 313–318.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. PmLR, 8748–8763.

Gaurav Rai, Shreyas Gupta, and Ojaswa Sharma. 2024. SketchAnim: Real-Time Sketch Animation Transfer from Videos. In *Computer Graphics Forum*, Vol. 43. Wiley Online Library, e15176.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. 2025. Semantic Image Inversion and Editing using Rectified Stochastic Differential Equations. In *International Conference on Learning Representations*.

Tim Salimans and Jonathan Ho. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. In *International Conference on Learning Representations*.

Harrison Jesse Smith, Qingyuan Zheng, Yifei Li, Somya Jain, and Jessica K Hodgins. 2023. A Method for Animating Children's Drawings of the Human Figure. *ACM Transactions on Graphics* 42, 3 (2023), 1–15.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, DanDan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. 2025. Animate-X: Universal Character Image Animation with Enhanced Motion Representation. In *International Conference on Learning Representations*.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2024. DreamGaussian: Generative Gaussian Splatting for Efficient 3D Content Creation. In *International Conference on Learning Representations*.

Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2024a. DisCo: Disentangled Control for Realistic Human Dance Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9326–9336.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. VideoComposer: Compositional Video Synthesis with Motion Controllability. *Advances in Neural Information Processing Systems* 36 (2023), 7594–7611.

Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. 2024c. Unianimate: Taming Unified Video Diffusion Models for Consistent Human Image Animation. *arXiv preprint arXiv:2406.01188* (2024).

Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. 2024d. A Recipe for Scaling up Text-to-Video Generation with Text-free Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6572–6582.

Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. 2024b. CRM: Single Image to 3D Textured Mesh with Convolutional Reconstruction Model. In *European Conference on Computer Vision*. Springer, 57–74.

Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. 2019. Photo Wake-Up: 3D Character Animation From a Single Photo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5908–5917.

Nora S Willett, Wilmot Li, Jovan Popovic, Floraine Berthouzoz, and Adam Finkelstein. 2017. Secondary Motion for Performed 2D Animation. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 97–108.

Tianyi Xie, Yiwei Zhao, Ying Jiang, and Chenfanfu Jiang. 2025. Physanimator: Physics-guided generative cartoon animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 10793–10804.

Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. 2024. ToonCrafter: Generative Cartoon Interpolation. *ACM Transactions on Graphics* 43, 6 (2024), 1–11.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. 2024. MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1481–1490.

Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. 2019. DwNet: Dense Warp-based Network for Pose-Guided Human Video Generation. *arXiv preprint arXiv:1910.09139* (2019).

Jiaxu Zhang, Xianfang Zeng, Xin Chen, Wei Zuo, Gang Yu, and Zhigang Tu. 2024. MikuDance: Animating Character Art with Mixed Motion Dynamics. *arXiv preprint arXiv:2411.08656* (2024).

Jiayi Eris Zhang, Seungbae Bang, David IW Levin, and Alec Jacobson. 2020. Complementary Dynamics. *ACM Transactions on Graphics* 39, 6 (2020), 1–11.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 586–595.

Jie Zhou, Chufeng Xiao, Miu-Ling Lam, and Hongbo Fu. 2024. DrawingSpinUp: 3D Animation from Single Character Drawings. In *SIGGRAPH Asia 2024 Conference Papers*. 1–10.

Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. 2023. ProPainter: Improving Propagation and Transformer for Video Inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10477–10486.

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, et al. 2025. Light-A-Video: Training-free Video Relighting via Progressive Light Fusion. *arXiv preprint arXiv:2502.08590* (2025).

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