Supplemental Materials for "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

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

1 NOTATION TABLE

Table 1 lists the symbols used in this paper along with their corresponding descriptions for the reader's reference.

2 POSE GUIDANCE EXTRACTION

We describe how we extract the pose sequence $\{P_{1:N}\}$ and the reference pose P_{ref} from the animated character. We adopt the OpenPose [Cao et al. 2019] 18-keypoint format, excluding the hand and feet joints due to the abstract nature of hand-drawn characters. Since the animated 3D skeleton generated from Mixamo does not include facial keypoints, we employ an approximate method to locate them: we predict the 2D positions of 5 facial keypoints (nose, eyes, ears) from I_{ref} using X-Pose [Yang et al. 2024] and then back-project these points onto the surface of $\mathcal{G}$ to obtain their depth values. By this, we get the 3D positions of facial keypoints in the character's rest pose. Inspired by Astropulse's tool [Astropulse 2024], we assume that the character's head is a rigid body, and calculate the new positions of the facial keypoints based on the orientation of the head in each frame and relative positions between the head and facial keypoints. For the other 13 body keypoints, we directly extract their per-frame 3D position from the animated 3D skeleton. Finally, we sort the 17 limbs formed by these 18 keypoints according to depth to ensure that their 2D projections maintain the correct occlusion order.

* 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

Table 1. Notation Table

Notation	Description
I_{ref}	Reference image
$I_{\text{ref}}^{\text{nc}}$	Contour-free reference image
$\mathcal{P}$	Target 3D motion
$\{\hat{I}_{1:N}\}$	Generated 3D character animation
S_{front}	Hair-body segmentation map (front-view)
S_{right}	Hair-body segmentation map (right-view)
$M_{\text{front}}^{\text{SDI}}$	SDI mask (front-view)
$M_{\text{back}}^{\text{SDI}}$	SDI mask (back-view)
$\mathcal{G}$	3D textured geometry reconstructed from I_{ref}
$E(\cdot)$	VAE encoder
$D(\cdot)$	VAE decoder
$\{P_{1:N}\}$	Pose guidance sequence
$\{M_{1:N}^{\text{SDI}}\}$	SDI mask guidance sequence
$\{C_{1:N}\}$	Coarse color guidance sequence
$\{C_{1:N}^{\text{masked}}\}$	Masked coarse color guidance sequence
P_{ref}	Reference pose extracted from I_{ref}
$\mathcal{E}_{\text{coarse}}$	Coarse prior encoder
u_θ	Native pre-trained UniAnimate
v_θ	Our domain-adapted model
$\{\hat{z}_{1:N,0}^{t,u_\theta}\}$	Noise-free latent estimates from u_θ
$\{\hat{z}_{1:N,0}^{t,v_\theta}\}$	Noise-free latent estimates from v_θ
$\{\hat{z}_{1:N,0}^{t,\text{blend}}\}$	Blended noise-free latent estimates
T	The total number of denoising steps
τ_2	The starting step of SDI
τ_1	The ending step of SDI
α	The percentage to control τ_2 where $\tau_2 = \alpha \cdot T$
β	The percentage to control τ_1 where $\tau_1 = \beta \cdot T$

3 ALGORITHM FOR SDI

The algorithm 1 describes the process for generating the synthesized video using reference images and guidance sequences.

Algorithm 1 Generation Process with Secondary Dynamics Injection (SDI)

Input:
- Total timesteps T
- Thresholds τ_2, τ_1
- Reference image I_{ref}
- Contour-free reference image $I_{\text{ref}}^{\text{nc}}$
- Reference pose P_{ref}
- Pose guidance sequence $\{P_{1:N}\}$
- Masked coarse color guidance sequence $\{C_{1:N}^{\text{masked}}\}$
- SDI mask guidance sequence $\{M_{1:N}^{\text{SDI}}\}$

Initialization: Input random noise sequence $\{z_{1:N,T}\}$
Output: Synthesized video $\{\hat{I}_{1:N}\}$
Stage 1:
for each timestep $t \in [T, \tau_2)$ **do**
$\{z_{1:N,t-1}\} \leftarrow v_\theta(I_{\text{ref}}^{\text{nc}}, \{z_{1:N,t}\}, P_{\text{ref}}, \{P_{1:N}\}, \{C_{1:N}^{\text{masked}}\}, \{M_{1:N}^{\text{SDI}}\})$
end for
Stage 2:
Initialization: $\{z_{1:N,\tau_2}^{v_\theta}\} = \{z_{1:N,\tau_2}\}, \{z_{1:N,\tau_2}^{u_\theta}\} = \{z_{1:N,\tau_2}\}$
for each timestep $t \in [\tau_2, \tau_1)$ **do**
$\{z_{1:N,t-1}^{v_\theta}\} \leftarrow v_\theta(I_{\text{ref}}^{\text{nc}}, \{z_{1:N,t}^{v_\theta}\}, P_{\text{ref}}, \{P_{1:N}\}, \{C_{1:N}^{\text{masked}}\}, \{M_{1:N}^{\text{SDI}}\})$
$\{z_{1:N,t-1}^{u_\theta}\} \leftarrow u_\theta(I_{\text{ref}}^{\text{nc}}, \{z_{1:N,t}^{u_\theta}\}, P_{\text{ref}}, \{P_{1:N}\})$
Extract the noise-free latent estimates:
$$\{\hat{z}_{1:N,0}^{t-1,v_\theta}\} \leftarrow \{z_{1:N,t-1}^{v_\theta}\}, \quad \{\hat{z}_{1:N,0}^{t-1,u_\theta}\} \leftarrow \{z_{1:N,t-1}^{u_\theta}\}$$
Perform blending using the SDI mask:
$$\{\hat{z}_{1:N,0}^{t-1,\text{blend}}\} = (1 - \{M_{1:N,\text{down}}^{\text{SDI}}\}) \cdot \{\hat{z}_{1:N,0}^{t-1,v_\theta}\} + \{M_{1:N,\text{down}}^{\text{SDI}}\} \cdot \{\hat{z}_{1:N,0}^{t-1,u_\theta}\}$$
Guide the denoising direction:
$$\{z_{1:N,t-1}^{v_\theta}\} \leftarrow \{\hat{z}_{1:N,0}^{t-1,\text{blend}}\}, \{z_{1:N,t-1}^{u_\theta}\} \leftarrow \{\hat{z}_{1:N,0}^{t-1,\text{blend}}\}$$
end for
Inpaint coarse frames using Poisson blending:
$$\{C_{1:N}^{\text{inpainted}}\} \leftarrow \text{PoissonBlend}(\{C_{1:N}^{\text{masked}}\}, D(\hat{z}_{1:N,0}^{\tau_1,\text{blend}}))$$
$$\{M_{1:N}^{\text{black}}\} \leftarrow \text{Placeholder: Full Black Mask}$$
Stage 3:
for each timestep $t \in [T, 1]$ **do**
$\{z_{1:N,t-1}\} \leftarrow v_\theta(I_{\text{ref}}, \{z_{1:N,t}\}, P_{\text{ref}}, \{P_{1:N}\}, \{C_{1:N}^{\text{inpainted}}\}, \{M_{1:N}^{\text{black}}\})$
end for
Output: $\{\hat{I}_{1:N}\} \leftarrow D(\{z_{1:N,0}\})$

4 PERCEPTUAL USER STUDY

We performed a user study to assess the perceptual quality of our proposed method (Ours) against the baseline approaches. The evaluation was based on three key metrics: (1) Pose Consistency (PC): Measures the alignment between target poses and generated results. (2) Style Preservation (SP): Evaluates how well the original reference style is retained, especially for the contour regions. (3) Motion Naturalness (MN): Assesses the realism and plausibility of motion dynamics, with emphasis on secondary motion. The study was conducted through an online questionnaire, featuring 20 randomly ordered result sets spanning a diverse range of character styles (e.g., cartoon characters and anthropomorphic animals) and motion styles (e.g., casual walking, athletic jumps, and expressive dances) to ensure broad stylistic coverage. A total of 50 participants were recruited for the study. They were tasked with identifying and selecting all methods that aligned closely with each specific evaluation standard (PC, SP, MN). Table 2 presents the voting results, indicating that our method was consistently the most frequently selected among the three evaluation criteria. This preference, expressed by human subjects, highlights the effectiveness of our approach in animating stylized characters with natural motions and ensuring character-specific details. Since both our method and DrawingSpinUp are based on skeletal animation, the pose consistency metrics are comparable. However, when combined with diffusion priors, our method significantly surpasses DrawingSpinUp in terms of stylistic consistency and motion naturalness.

Method	PC ↑	SP ↑	MN ↑
AnimateAnyone	6.0	8.0	5.0
MikuDance	10.8	6.6	7.7
UniAnimate	10.8	9.9	8.8
UniAnimate*	21.5	20.7	20.1
DrawingSpinUp	25.4	26.7	24.4
Ours	**25.5**	**28.1**	**34.0**

Table 2. Summary of the voting results from the user study.

5 SDI MASKS

Fig. 1 shows the SDI masks corresponding to all the examples used in the paper. The ranges of the SDI masks are relatively flexible. Users can choose to mask only the hair ends or the entire hair, leading to different final redraws. The larger the range of the SDI mask, the greater the enhancement range of the diffusion model; however, this may also introduce increased geometric distortion. This creates a trade-off that needs to be balanced.

REFERENCES

Astropulse. 2024. mixamotoopenpose. https://github.com/Astropulse/mixamotoopenpose Last updated: November 2024.
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.
Jie Yang, Ailing Zeng, Ruimao Zhang, and Lei Zhang. 2024. X-Pose: Detecting Any Keypoints. In *European Conference on Computer Vision*. Springer, 249–268.

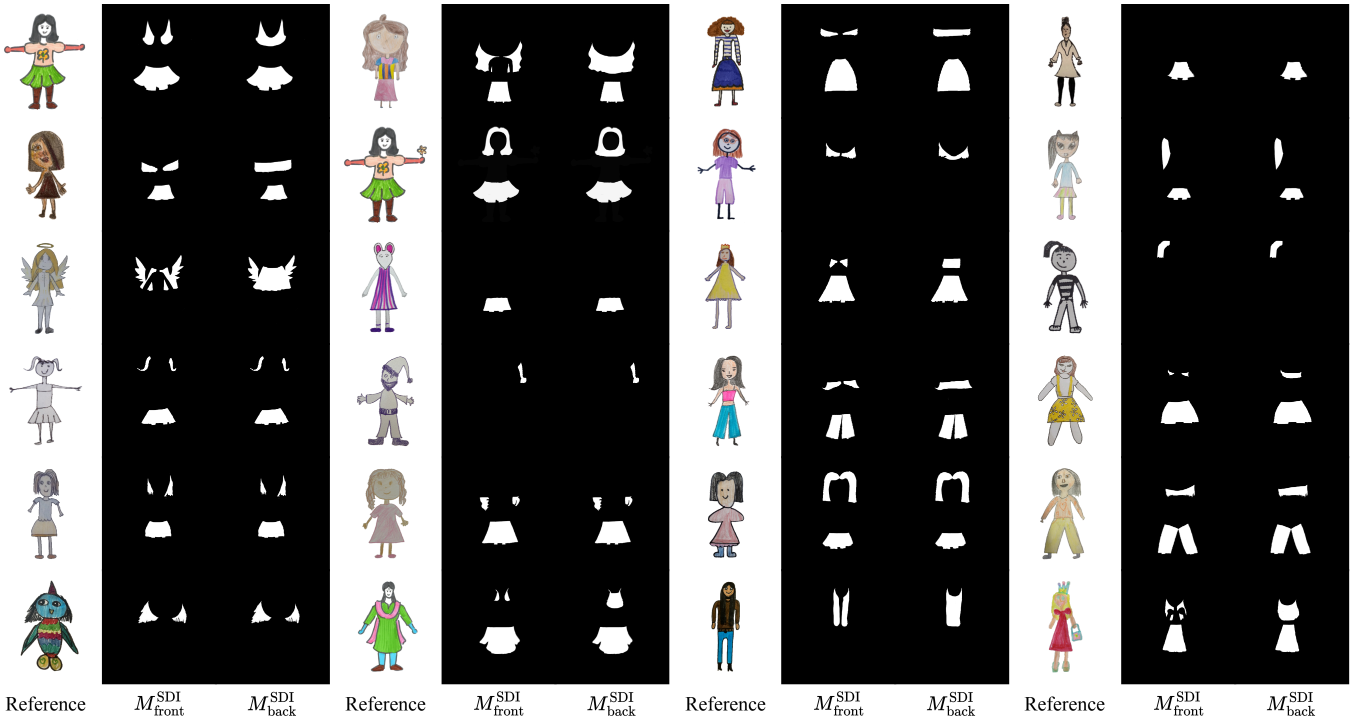

Fig. 1. The SDI masks corresponding to all the examples used in the paper.