One-to-All Animation: Alignment-Free Character Animation and Image Pose Transfer

Shijun Shi[1*] Jing Xu[2*] Zhihang Li[3] Chunli Peng[4] Xiaoda Yang[5]
Lijing Lu[3] Kai Hu[1†] Jiangning Zhang[5†]
[1]Jiangnan University [2]University of Science and Technology of China
[3]Chinese Academy of Sciences [4]Beijing University of Posts and Telecommunications
[5]Zhejiang University

https://ssj9596.github.io/one-to-all-animation-project/

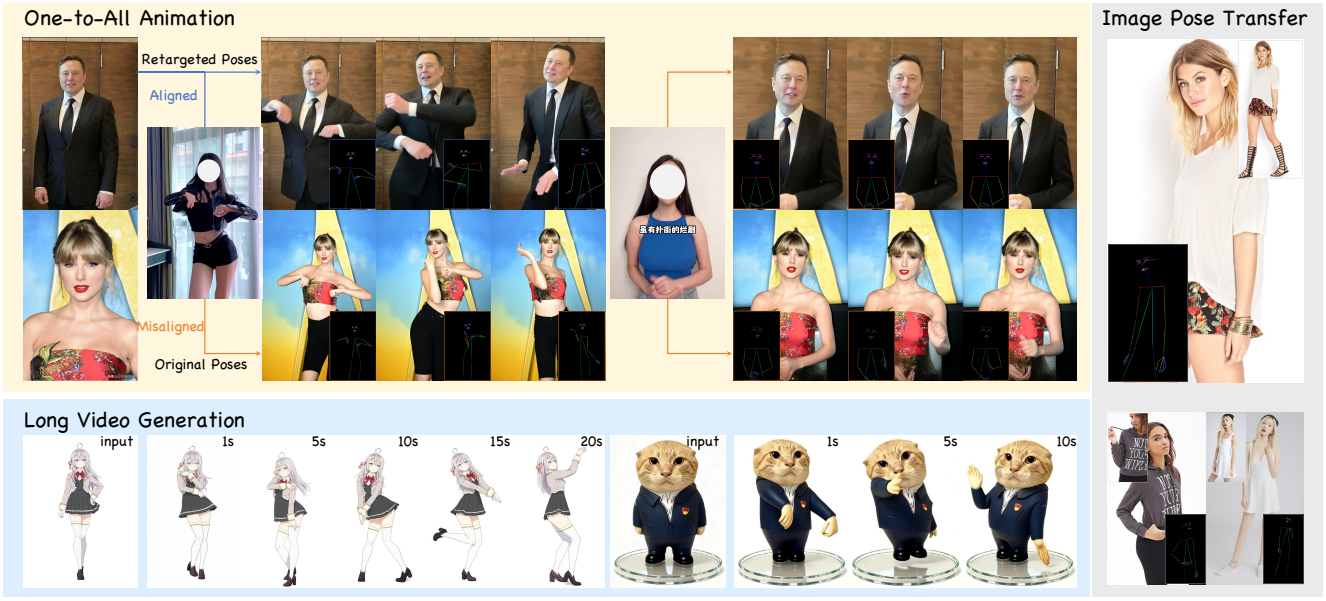


Figure 1. We introduce *One-to-All Animation*, a unified framework for pose-driven personalized generation. Unlike prior methods that require both spatially-aligned references and pose retargeting, our framework supports: (1) cross-scale video animation with either retargeted or original driving motion, (2) cross-scale image pose transfer, and (3) temporally coherent long video generation.

Abstract

Recent advances in diffusion models have greatly improved pose-driven character animation. However, existing methods are limited to spatially aligned reference-pose pairs with matched skeletal structures. Handling reference-pose misalignment remains unsolved. To address this, we present One-to-All Animation, a unified framework for high-fidelity character animation and image pose transfer for references with arbitrary layouts. First, to handle spatially misaligned reference, we reformulate training as a self-supervised outpainting task that transforms diverse-layout reference into a unified occluded-input format. Second, to process partially visible reference, we design a reference extractor for comprehensive identity feature extraction. Further, we integrate hybrid reference fusion attention to handle varying resolutions and dynamic sequence lengths. Finally, from the perspective of generation quality, we introduce identity-robust pose control that decouples appearance from skeletal structure to mitigate pose overfitting, and a token replace strategy for coherent long-video generation. Extensive experiments show that our method outperforms existing approaches. The code and model are available at https://github.com/ssj9596/One-to-All-Animation.

*Equal contribution: ssj180123@gmail.com, xujing0@mail.ustc.edu.cn
†Corresponding author: hukai_wlw@jiangnan.edu.cn, 186368@zju.edu.cn

1. Introduction

In recent years, diffusion-based models have revolutionized the field of visual content generation, achieving significant improvements in both visual fidelity and controllability. These advances have opened new opportunities across film production, digital advertising, immersive virtual avatars, and other creative industries. Building on this progress, a growing line of research [4, 17, 46, 47, 51, 62] has focused on character animation, which generates animated videos of a given character by transferring motion from a driving video. Recent approaches [6, 8, 52] have leveraged large-scale pre-trained video foundation models built on the Diffusion Transformer (DiT) architecture [33] to achieve better temporal consistency and visual realism.

However, a critical challenge limits the practical use of character animation: the **misalignment** between the reference image and the driving video. Specifically, "misalignment" refers to the inconsistency of the target subject's presentation between two inputs, which manifests in two main aspects: (1) **Spatial layout mismatch**. The reference image and driving video differ in pose scale, or body part coverage (e.g., the reference shows a half-body close-up while the driving video shows full-body dancing), and (2) **Facial inconsistency**. The reference and driving subjects differ in facial skeletal structure and geometry, particularly in the distances and proportions between facial features (eyes, nose, mouth, neck). Such misalignment often causes severe artifacts in the generated animation, including distorted body shapes and mismatched appearance.

To address these issues, an ideal solution must satisfy two requirements: spatial flexibility to handle arbitrary layout variations, and identity robustness to preserve appearance consistency despite skeletal and facial geometry differences. However, existing methods fall short in both aspects. Current approaches adopt a self-driven reconstruction strategy during training, inherently ensuring shared layout and skeleton between reference and driving. To maintain such alignment at inference, they impose two constraints: explicitly requiring spatial-matched reference images and strong dependence on pose retargeting to align driving poses.

As shown in Fig. 2, when facing mismatched inputs, previous works [47, 62] completely lose appearance information, generating results with wrong identity. Recent DiT-based methods [6, 52] trained on large-scale data achieve facial preservation but still exhibit noticeable visual artifacts. Moreover, since identity preservation relies heavily on accurate pose retargeting, its failure directly causes identity drift (See Fig. 9). These limitations inspire us to rethink the training strategy. Instead of relying on perfect alignment during training, we ask: **can we train the model to handle misalignment directly?**

In this paper, we propose *One-to-All Animation*, a unified framework for pose-driven personalized generation from arbitrary reference images. Our key insight is to reformulate training as an outpainting problem with a **unified occluded-input format**, enabling the model to learn generation from diverse spatial layouts.

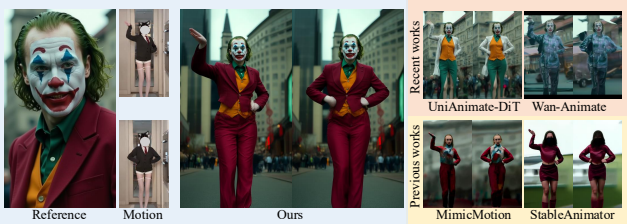


Figure 2. Visual comparison under spatial-misaligned inputs. Previous methods such as MimicMotion [62] and StableAnimator [47] fail to preserve identity. Recent approaches including UniAnimate [52] and Wan-Animate [6] show degraded quality. Our method maintains robust appearance consistency.

Built on this insight, we introduce three key components. First, we introduce a Reference Extractor for multi-level identity feature extraction and a hybrid reference fusion attention module to handle variable resolutions and dynamic sequence lengths. Second, we decouple identity from skeletal structure through face region enhancement and reference-guided pose control. The former randomly replaces facial structures in driving poses to break identity-skeleton coupling, while the latter balances standard and enhanced pose signals using reference appearance features. Third, we adopt a token replace strategy for long-video generation that ensures smooth cross-segment transitions.

Moreover, we extend to hybrid image-video training, naturally enabling image pose transfer as a complementary capability. By solving the misalignment problem, One-to-All Animation enables a single reference to adapt to multiple scenarios without alignment constraints. As illustrated in Fig. 1, a single arbitrary reference can drive video generation with different poses (retargeted or original), and supports cross-scale image pose transfer. Our main contributions are summarized as follows:

- We present One-to-All Animation, the first unified framework for pose-driven personalized image and video generation. It supports various cross-scale applications from any reference image, including character animation and image pose transfer.
- We design a Reference Extractor with hybrid fusion attention that enables robust identity preservation under arbitrary spatial layouts and partial visibility conditions.
- We introduce identity-robust pose control that decouples identity from skeletal structure to address the facial pose overfitting problem.
- We propose token replace training strategy for seamless long video generation.

2. Related Works

2.1. Diffusion Models

Diffusion models [15, 45] have demonstrated remarkable generative capabilities in both image [7, 21, 34, 36, 39, 40, 59] and video generation [2, 11–13, 20, 24, 42, 44, 49, 56, 63], transforming visual synthesis from a research topic into practical applications. Among these, Diffusion Transformers (DiT) [33] have become increasingly popular due to their better scalability and ability to capture long-range dependencies. On the image side, DiT-based models like SD3 [7] and Flux [21] achieve state-of-the-art quality in text-to-image generation. On the video side, recent foundation models [11, 20, 49, 56] follow a similar pipeline: a 3D VAE [19] first compresses video into spatio-temporal latents, which are then processed by stacked Transformer blocks for joint spatial–temporal modeling. Notably, many of these models are pretrained on mixed video-image datasets, which also gives them strong image generation capability. This makes them well-suited for both static image synthesis and dynamic video generation tasks, such as the pose-driven personalized generation discussed below.

2.2. Pose-Driven Personalized Generation

Pose-driven personalized generation aims to transfer appearance from a reference image to a specified pose. Based on the spatial alignment between two inputs, existing tasks can be divided into two settings: (*i*) unaligned image-to-image generation, where the reference image and target pose differ in position, scale, and viewpoint; (*ii*) aligned image-to-video generation, where the driving video poses are spatially consistent with the reference frame. This division stems from **data availability**: image-to-image tasks benefit from abundant training data with diverse pose variations [29], while video generation typically relies on self-reconstruction data, naturally resulting in spatial alignment.

In the image domain, image pose transfer [26, 31, 37, 38, 41, 43, 60] represents the unaligned setting. Early GAN-based methods [37, 43] struggle with distorted textures under large pose disparities. Recent diffusion-based approaches [26, 31, 41] leverage stronger generative priors for better appearance preservation, but remain constrained to low-resolution outputs with limited facial quality. In the video domain, character animation methods [4, 6, 17, 47, 52, 62] perform aligned image-to-video generation by animating a reference character according to a driving video. These methods ensure temporal consistency through temporal modules [10] or video foundation models [2, 49]. However, they heavily rely on spatial alignment between the reference and driving poses, experiencing significant performance degradation when this alignment is disrupted. Despite progress in both directions, existing methods remain task-specific. In contrast, our method unifies unaligned image-to-image, aligned image-to-video, and unaligned image-to-video generation in a single framework.

3. Method

Given a reference image and a driving video, our goal is to generate an animated video that preserves the identity from the reference while following the motion from the driving video. Unlike existing methods, which rely on assumptions such as similar camera distances and well-aligned skeletons, we explicitly tackle the challenging real-world scenarios where the reference image and driving video exhibit large scale variations, diverse spatial framings, and mismatched skeletal layouts. We address these challenges through both data and model design. On the data side, we introduce a self-supervised training scheme that synthesizes spatially mismatched reference-driving pairs (Sec. 3.1). On the model side, we propose Reference Extractor to handle such occluded reference (Sec. 3.2). Additionally, we introduce Identity-Robust Pose Control (Sec. 3.3) to reduce pose overfitting, and TokenReplace (Sec. 3.4) to enable long-video generation.

3.1. One-to-All Animation

Inference pipeline. An overview of the proposed **One-to-All** framework is presented in Fig. 3. At the inference stage, since the reference image $\mathbf{I}^r$ could have significant scale variations relative to the driving sequence $\mathbf{P}^{1:N}$, we first perform pose-guided translation between the reference image and the driving video. Specifically, we identify an anchor frame from the driving sequence with the most similar pose orientation to the reference. We estimate their scale ratio using visible body parts present in both (e.g., inter-shoulder or inter-ear distance), then resize the reference accordingly and zero-pad it to the driving video resolution, producing a spatially adjusted input $\tilde{\mathbf{I}}^r$ and a mask $\mathbf{M}^r$ indicating the padded areas. The model takes the triplet $(\tilde{\mathbf{I}}^r, \mathbf{M}^r, \mathbf{P}^{1:N})$ as input and generates a video that follows the driving motion while synthesizing the padded areas to complete the reference appearance.

Self-supervised training via outpainting. We follow the self-reconstruction training setup, but with a key modification: instead of treating all frames as naturally aligned, we simulate spatial mismatches through outpainting preprocessing. During training, we apply face-centered random masking to the reference image to simulate various scale conditions. We generate a binary outpainting mask that indicates the missing regions. The triplet of masked reference, mask and pose sequence is fed as input, with the original complete video as supervision. This training setup produces inputs identical to those generated by pose-guided translation at inference, forcing the network to learn to hallucinate occluded regions and generate coherent motions.

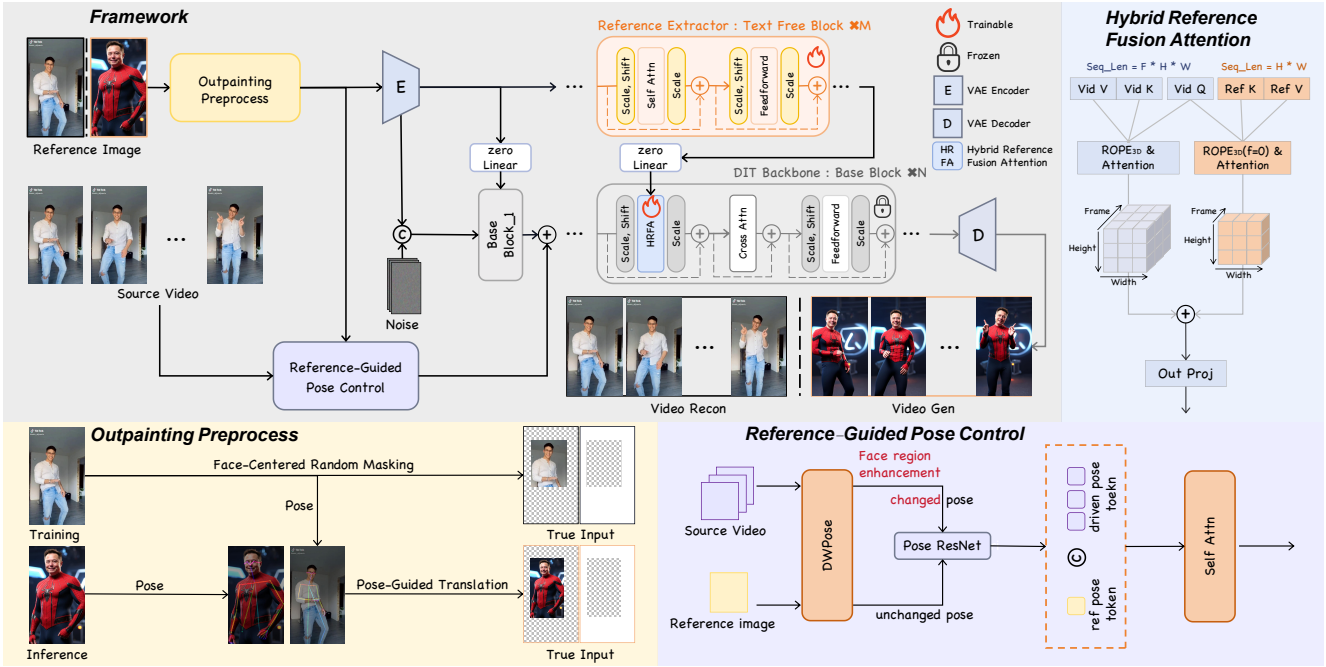


Figure 3. Overview of the proposed framework. We introduce outpainting preprocess to handle diverse body proportions through face-centered random masking during training and pose-guided translation at inference. The driving poses are encoded and refined via reference-guided pose control to preserve facial identity despite skeletal mismatch. Reference features are progressively injected through hybrid reference fusion attention, supporting variable resolutions and dynamic sequence lengths.

3.2. Reference Extractor.

Motivation. The outpainting-based training presents a unique challenge: extracting reliable appearance features from severely occluded reference. Existing methods are not designed for this scenario. Previous works typically use CLIP encoders for semantic feature extraction or directly embed reference frames into I2V backbones. However, CLIP encoders focus on global representation and lack fine-grained identity details. I2V backbones are limited by the strict first-frame "copy-paste" [27, 35] nature, which cannot effectively complete large occluded regions. In contrast, we design a dedicated Reference Extractor that extracts multi-level appearance features from occluded references.

Architecture details. As shown in Fig. 3, the extractor operates in parallel with the main denoising DiT backbone, producing features in the same latent space. Given reference $\tilde{\mathbf{I}}^r$ and mask $\mathbf{M}^r$, both are first encoded into latent representations z^r and z^m via the 3D VAE, where $\mathbf{M}^r$ is repeated along the channel dimension to match the encoder input format. These latents are concatenated along the channel dimension and converted to patch tokens:

$$r^0 = \text{patchify}\left([z^r,\, z^m]_{\text{channel}}\right), \tag{1}$$

forming the initial reference feature r^0. This feature is then refined through M text-free blocks. Each block is initialized from the DiT backbone but excluding the text cross-attention. The M block outputs, together with the initial r^0, form $M+1$ reference features. These are injected into the N-block denoising backbone ($M < N$) through zero-initialized linear projections. Each reference feature is shared by n consecutive denoising blocks, where $n = \frac{N}{M+1}$.

Hybrid Reference Fusion Attention. The key design of our Reference Extractor is the Hybrid Reference Fusion Attention (**HRFA**), which enables robust identity preservation across variable resolutions and dynamic sequence lengths. Specifically, we add a new cross-attention layer on the self-attention layer in the DiT block. Given video latent $h \in \mathbb{R}^{F\times H\times W\times C}$, the self-attention with 3D Rotary Position Encoding (RoPE) is formulated as:

$$\text{Attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \text{softmax}\left(\frac{\mathrm{Q}(\mathrm{K})^\top}{\sqrt{d}}\right)\mathrm{V}, \tag{2}$$

$$\mathrm{Q} = \text{RoPE}_{3D}(hW_q), \tag{3}$$

$$\mathrm{K} = \text{RoPE}_{3D}(hW_k), \tag{4}$$

$$\mathrm{V} = hW_v, \tag{5}$$

where W_q, W_k, W_v are learnable projection matrices. For reference feature $r \in \mathbb{R}^{1\times H\times W\times C}$, the output of the new cross-attention layer is computed as:

$$\text{Attention}(\mathrm{Q}',\mathrm{K}',\mathrm{V}') = \text{softmax}\left(\frac{\mathrm{Q}'(\mathrm{K}')^\top}{\sqrt{d}}\right)\mathrm{V}', \tag{6}$$

$$\mathrm{Q}' = \text{RoPE}_{3D}(hW_q, f=0), \tag{7}$$

$$\mathrm{K}' = \text{RoPE}_{3D}(rW'_k, f=0), \tag{8}$$

$$\mathrm{V}' = rW'_v, \tag{9}$$

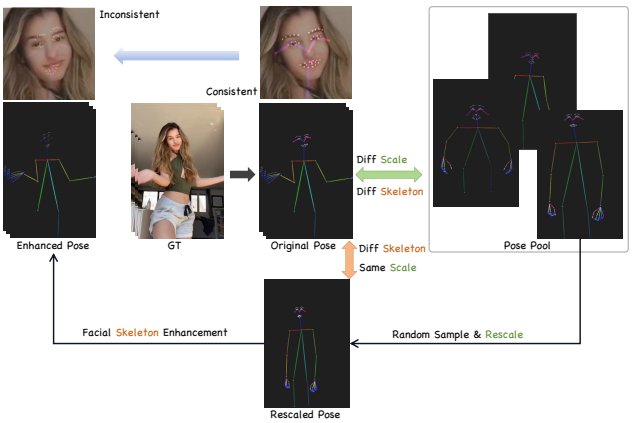


Figure 4. Face region enhancement. We sample a random pose, rescale it to match the driving pose, and apply their skeletal difference to create intentional facial skeleton inconsistency. A color-lighten signal indicates augmented poses.

where W'_k and W'_v are two newly introduced projection matrices for the reference features.

Notably, although h is a video latent with frame dimension F, we apply RoPE_{3D} with $f = 0$ to hW_q when computing Q'. This design prevents the cross-attention from learning absolute frame position dependencies between the reference and video frames. As a result, the model preserves its temporal extrapolation capability and can be seamlessly applied to both image and video generation tasks. The fused output of the self-attention and cross-attention layer is given by:

$$\mathbf{z}'_{\text{fusion}} = \mathrm{Attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) + \mathrm{Attention}(\mathrm{Q}', \mathrm{K}', \mathrm{V}'). \tag{10}$$

3.3. Identity-Robust Pose Control

Motivation. Following most mainstream approaches, we represent the driving sequence using 2D pose keypoints. The keypoints are skeleton-matched to the generated content and directly added to the initial noisy latent. However, this creates a coupling problem: the generated face identity is influenced by both reference appearance and driving pose skeleton. Existing methods mitigate this via precise pose retargeting and removing specific facial keypoints, but fail when retargeting is inaccurate, causing results to overfit the driving skeleton. Instead of constraining inference data preparation, we address this from training. We propose face region enhancement to decouple facial identity from skeletal structure, enabling the model to learn identity solely from the reference image.

Face region enhancement. An overview of the proposed face region enhancement is shown in Fig. 4. During training, we first sample a random pose from the training set and rescale it to match the scale of the current driving pose. Then we compute the structural difference between the sampled pose and driving pose by comparing the relative positions and bone-length ratios of their facial landmarks and skeletal keypoints. This difference is applied to the driving pose, intentionally creating facial skeleton inconsistency with the ground-truth. Under such inconsistency, the model is forced to learn facial appearance from the reference image rather than the driving pose, thereby enhancing identity robustness and reduces overfitting problem. We apply face region enhancement to 70% training samples, and use a color-lighten signal to indicate whether a pose is original or changed. At inference time, we retain the landmark's facial control. When retargeting is unreliable, the model can accept lightened-color pose signals to preserve identity without strict skeletal matching.

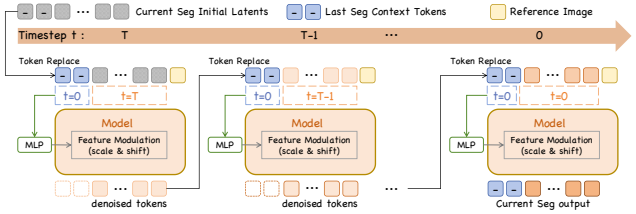


Figure 5. At each denoising timestep, context tokens from the last segment (blue) are replaced into the current segment's initial latents, modulated with $t = 0$ to serve as clean signals. This process ensures smooth transitions across segment boundaries.

Reference-guided pose control. While face region enhancement improves robustness to skeletal misalignment, it breaks spatial correspondence and introduces training instability. To address this, we propose reference-guided pose control that leverages the reference image to refine the driving sequence. Specifically, the reference image $\tilde{\mathbf{I}}^r$ is encoded by VAE to obtain its latent feature $\mathbf{z}^r$, which is concatenated with the video latents $\mathbf{z}^{1:n}$ along frame dimension:

$$\tilde{\mathbf{z}}^{1:(n+1)} = [\,\mathbf{z}^r,\, \mathbf{z}^{1:n}\,]_{\text{frame}}, \tag{11}$$

and fed into the DiT backbone. In parallel, we extract the pose $\tilde{\mathbf{P}}^r$ of the reference image $\tilde{\mathbf{I}}^r$. Importantly, reference pose does not undergo enhancement and remains aligned with the reference image. A Pose ResNet encodes both reference pose $\tilde{\mathbf{P}}^r$ and driving pose sequence $\mathbf{P}^{1:n}$ into features $\mathbf{p}^r$ and $\mathbf{p}^{1:n}$. which are also concatenated along frame dimension and processed by a self-attention (SA) block to capture intra-sequence dependencies:

$$\tilde{\mathbf{p}}^{1:(n+1)} = \mathrm{SA}([\,\mathbf{p}^r,\, \mathbf{p}^{1:n}\,]_{\text{frame}}). \tag{12}$$

This process propagates consistent appearance cues across frames. The refined pose representation $\tilde{\mathbf{p}}^{1:(n+1)}$ is then element-wise added to the output of the first DiT block.

3.4. Token Replace for Long-Video Generation.

For exceedingly long video generation, the pose sequence is divided into multiple segments and generated sequentially. To ensure smooth transitions across segment boundaries,

Table 1. Quantitative comparisons on TikTok and Cartoon datasets. In the table, a / b denotes results on TikTok / Cartoon.

Model	PNSR↑	SSIM↑	LPIPS↓	FID↓	FID-VID↓	FVD↓
~1.3B Models						
MimicMotion [62]	15.43/15.09	0.721 / 0.647	0.315 / 0.368	67.09 / 87.06	18.56 / 58.35	412.50 / 943.38
StableAnimator [47]	14.92/15.16	0.737 / 0.638	0.315 / 0.333	76.50 / 70.92	21.57 / 31.30	477.28 / 720.48
Animate-X [46]	15.22/15.63	0.741 / 0.659	0.329 / 0.330	**52.96 / 59.68**	**17.81** / 33.47	375.56 / 723.25
One-to-All-1.3B	**17.75/16.24**	**0.788 / 0.677**	**0.269 / 0.289**	74.96 / 67.82	21.37 / **29.11**	**361.85 / 549.27**
~14B Models						
UniAnimate-DiT [52]	**19.07**/17.03	**0.816** / 0.699	0.265 / 0.269	55.32 / 55.29	17.42 / 24.26	358.42 / 510.61
Wan-Animate [6]	17.57/16.43	0.763 / 0.659	0.306 / 0.318	66.98 / 58.94	16.79 / 27.74	**282.86** / 485.92
One-to-All-14B	18.07/**17.10**	0.812 / **0.701**	**0.254 / 0.259**	**50.49 / 50.07**	**13.93 / 15.07**	297.94 / **403.47**

we adopt a token replace strategy. As illustrated in Fig. 5, the last five frames of the preceding segment are encoded by VAE into two latent frames, serving as context tokens $\mathbf{z}_{\text{ctx}}$. At each denoising timestep t, the first two latent frames in the noisy latent $\mathbf{z}_t^{1:n}$ are replaced by these context tokens:

$$\tilde{\mathbf{z}}_t^{1:n} = [\,\mathbf{z}_{\text{ctx}},\, \mathbf{z}_t^{3:n}\,].$$

During training, context tokens are excluded from the reconstruction loss and act solely as temporal guidance; at inference, they are treated as clean signals with $t = 0$ in the feature modulation module throughout the denoising process. After denoising, the first two tokens are retained as context, and the VAE decodes the complete latent sequence to obtain the final video frames.

3.5. Training and Inference

Training setup. We train the model following Rectified Flow (RF) formulation [23, 28]. In the forward process, Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I)$ is added to a clean latent sample $\mathbf{x}_0$ to produce:

$$\mathbf{x}_t = (1-t)\,\mathbf{x}_0 + t\,\boldsymbol{\varepsilon}, \tag{13}$$

where the time step t is obtained by sampling an integer $\tau \in \{0, \ldots, T\}$ (with $T = 1000$) and normalising it to $[0, 1]$. The network G_θ is trained to predict the target velocity:

$$u_t = \frac{\partial \mathbf{x}_t}{\partial t} = \boldsymbol{\varepsilon} - \mathbf{x}_0, \tag{14}$$

using the regression loss:

$$\mathcal{L}_{\text{RF}} = \|v_t - u_t\|^2, \tag{15}$$

where $v_t = G_\theta(\mathbf{x}_t,\, t,\, C)$ and C denotes the conditioning inputs (pose sequence, reference image, and mask).

Three-stage training. Our training process is divided into three stages. In the first stage, we train the reference extractor and the W_k' / W_v' components of the HRFA solely using appearance information as condition. In the second stage, we introduce pose condition and train the reference-guided pose control jointly with all components of the HRFA. In the third stage, we include the token replace mechanism. Throughout all stages, the text prompt is fixed as an empty string. We perform joint image-video training with mixed resolutions (512px and 768px) and various aspect ratios. The ratio of video to image data per epoch is set to 6:1. To reduce computational cost, each video sample is limited to 29 frames. Further details are provided in the Supplementary Material.

Inference setting. At inference, we employ the Euler method for sampling over 30 steps. We adopt cumulative classifier-free guidance [3] to strengthen both the reference appearance and pose guidance. Specifically, the denoised output at each step $t-1$ is computed as:

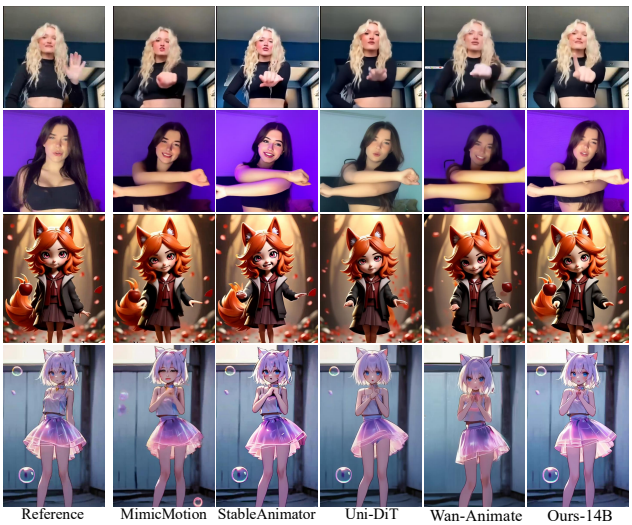
$$x_{t-1} = x_{\varnothing}^{t-1} + \lambda_P \left(x_P^{t-1} - x_{\varnothing}^{t-1}\right) + \lambda_I \left(x_{RP}^{t-1} - x_P^{t-1}\right), \tag{16}$$

where $x_{\varnothing}^{t-1}$ is the unconditional prediction, x_P^{t-1} is the pose-conditioned prediction, and x_{RP}^{t-1} is the joint reference-plus-pose prediction. The guidance scales λ_P and λ_I are both set to 1.5 in most cases. For long video generation, the sequence is segmented into clips of 65–81 frames and processed sequentially. Starting from the second segment, the last five frames of the preceding segment are encoded via VAE into two latent frames for token replace (Sec. 3.4), , thereby ensuring seamless transitions across segments.

4. Experiments

4.1. Implementations

Our model is built upon the open-source text-to-video backbone Wan2.1 [49]. We train two model variants with 1.3B and 14B parameters. Experiments are conducted on 8 NVIDIA H20 GPUs. We collect approximately 7,000 human videos from the internet and expand the training corpus by incorporating additional samples from the TikTok [18], Champ [64], and UBC [58] datasets. To enhance cross-style generalization, we curate 200 cartoon character images and synthesize animation clips using Seedance [9] as a supplemental training subset. We also include the DeepFashion



(a) Results on TikTok (top two) and Cartoon testset (bottom two).

(b) Results on DeepFashion benchmark.

Figure 6. Qualitative comparisons with state-of-the-art methods.

dataset [29], which contains 52,712 high-resolution fashion model images, to facilitate mixed image–video training.

4.2. Comparisons

Metrics. We follow the previous evaluation metrics for pose-driven personlizaed generation. Specifically, for single-frame quality assessment, we employ PSNR [16], SSIM [53], LPIPS [61] and FID [14]. For video fidelity, we utilize FID-VID [1] and FVD [48].

Evaluation on Video Dataset. For character animation, we follow previous works [50, 54] by evaluating on TikTok benchmark. In addition, we evaluate on 12 cartoon-style image–video pairs outside the cartoon training set for cross-domain evaluation. To ensure fairness, we adopt the same reference images for all methods and compare models within similar parameter scales. Specifically, we benchmark both ∼1.3B and ∼14B variants of our model against state-of-the-art competitors on the two datasets. As shown in Table 1, One-to-All-1.3B achieves superior results on most metrics among small-scale backbones, while One-to-All-14B consistently outperforms large-scale counterparts

Table 2. Quantitative comparison on DeepFashion dataset.

Method	FID↓	LPIPS↓	PSNR↑	SSIM↑
Evaluate on 512×352 ***resolution***				
CFLD [30] (CVPR24)	7.11	0.279	**17.13**	**0.753**
MCLD [25] (CVPR25)	7.07	0.275	16.51	0.736
One-to-All-1.3B	**6.85**	**0.249**	16.84	0.742
Evaluate on 944×624 ***resolution***				
CFLD [30] (CVPR24)	8.38	0.314	**17.38**	0.758
MCLD [25] (CVPR25)	8.96	0.322	16.33	**0.761**
One-to-All-1.3B	**6.92**	**0.285**	16.24	0.754

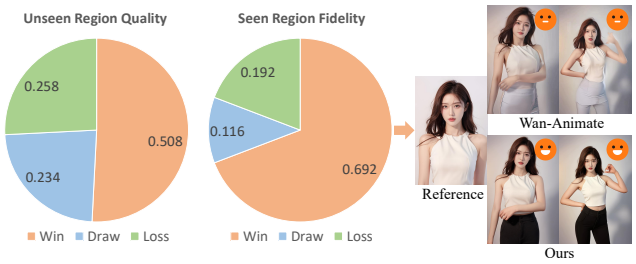


Figure 7. Human evaluation with current SOTA.

across both datasets, indicating strong scalability and generalization. Qualitative results are shown in Fig. 6a.

Evaluation on Image Dataset. For image pose transfer, following prior works [25, 30], we evaluate on 8,570 test pairs from the DeepFashion dataset. Previous methods typically perform inference at a low resolution of 512×352, which often leads to loss of fine details. To explore high-resolution generation, we additionally conduct inference at 768 px while maintaining the same aspect ratio (944×624). We compare our One-to-All-1.3B model with CFLD [30] and MCLD [25] under both resolution settings. As shown in Table 2, our method achieves the lowest FID and LPIPS scores at both resolutions, indicating superior perceptual quality. Although PSNR and SSIM scores are slightly lower than some competing methods, Fig. 6b demonstrates that our approach produces significantly clearer facial details and better visual quality.

User Study. To evaluate misaligned animation quality, we conduct a user study comparing with the current SOTA Wan-Animate [6]. We randomly select 5 characters generated by Flux [21] and collect 6 driving videos, resulting in 30 misaligned test cases. Four participants evaluated anonymized video pairs based on: (1) Unseen region quality: visual quality and plausibility of newly generated regions; and (2) Seen region fidelity: preservation of reference appearance and background. The results in Figure 7 show that our method achieves superior performance (50.8% vs 25.8% for unseen region quality, and 69.2% vs 19.2% for seen region fidelity), especially in identity preservation, where it maintains consistent facial features and overall appearance better than Wan-Animate.

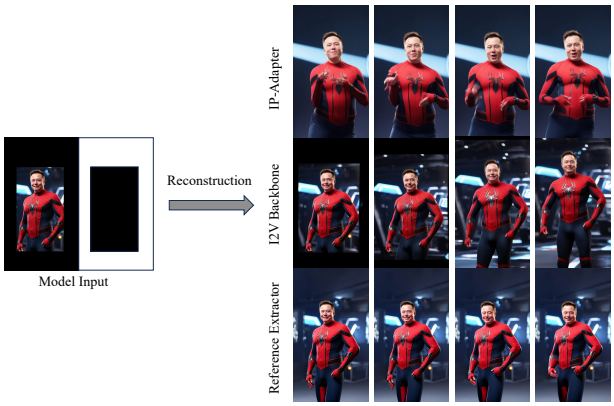


Figure 8. Qualitative comparison of different reference feature extraction methods in the first training stage.

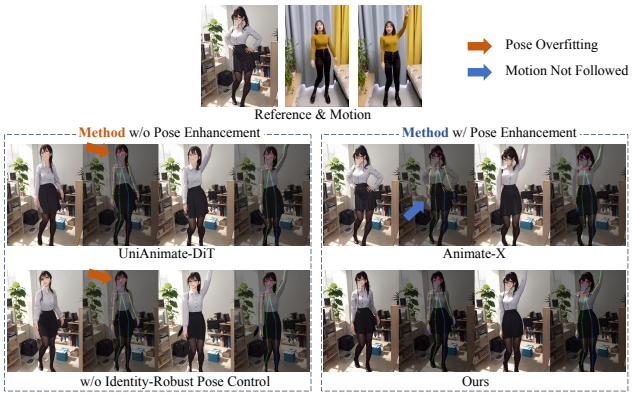


Figure 9. Qualitative ablation of Identity-Robust Pose Control.

4.3. Ablation Study

Reference Extractor. To demonstrate the effectiveness of our proposed Reference Extractor, we compare against two alternatives during the first training stage: (1) IP-Adapter [57], which uses CLIP encoder to extract semantic features, and (2) I2V Backbone, which relies on the model's first-frame prior for reference feature extraction. Reconstruction results are presented in Fig. 8. IP-Adapter exhibits poor consistency in both background and foreground details across frames. The I2V Backbone is constrained by its "copy-paste" nature and can only gradually fill masked regions from the original input, which often leads to incomplete recovery in cases of large occlusion. In contrast, our Reference Extractor is able to reconstruct occluded regions while preserving fine-grained identity features.

Identity-Robust Pose Control. We evaluate the effectiveness of Identity-Robust Pose Control. As shown in Fig. 9, when pose retargeting is inaccurate, both traditional methods (e.g., UniAnimate-DiT) and our baseline without pose enhancement suffer from pose overfitting (orange arrows), where the generated identity is dictated by the driving skeleton instead of the reference image. In contrast, our method successfully preserves identity under such conditions.

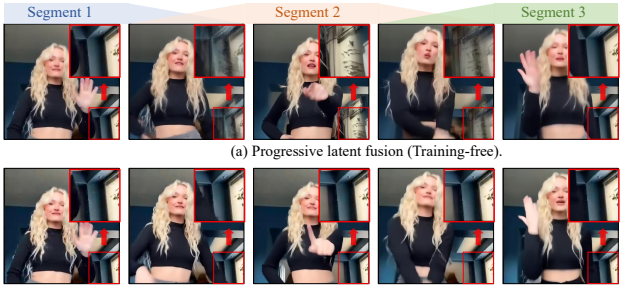
1) Face region enhancement. While we are not the first to employ pose augmentation during training, our approach differs notably from prior work. We compare against Animate-X [46], which rescales the entire pose skeleton. However, such aggressive whole-body augmentation disrupts spatial correspondence across all body parts, causing motion deviation (blue arrows in Fig. 9). In contrast, our face region enhancement selectively modifies only facial part, thereby preserving body motion accuracy while improving identity robustness.

2) Reference-guided pose control. Table 3 reveals a critical insight: adding face region enhancement alone actually degrades performance. The reason is straightforward: while we intentionally introduce skeletal inconsistencies to prevent pose overfitting, the driving pose is still directly added to the noisy latent. This mismatch between enhanced pose and ground-truth causes training instability. Our reference-guided pose control addresses this through relationship modeling. By concatenating the unchanged reference pose with the driving sequence, the model learns to capture their structural dependencies and adaptively refine the poses accordingly. This stabilizes training and yields substantial improvements.

Token Replace. We conduct a comparison with the default long video generation strategy: progressive latent fusion. This training-free approach generates video segments separately and fuses their overlapping frames at each denoising step [62]. However, as shown in Fig. 10a, the overlap regions fail to maintain consistency with both adjacent segments. This exposes the inherent limitation of training-free fusion approaches. In contrast, our trainable token replace strategy leverages frames from overlap regions as context token, ensuring strong temporal consistency across segments. Quantitative results in Table 3 demonstrate that incorporating token replace achieves the best performance across all metrics.

Table 3. Ablation study on model components. Experiments are conducted on the TikTok benchmark using 14B model.

Components	SSIM↑	LPIPS↓	FVD↓
Base Ref. Extractor	0.773	0.280	355.2
+ Face region enhancement	0.748	0.335	412.8
+ Reference-guided pose control	0.795	0.275	325.7
Full + Token Replace	**0.812**	**0.254**	**297.9**



(a) Progressive latent fusion (Training-free).

(b) Token replace (Trainable).

Figure 10. Comparison of long video generation strategies.

5. Conclusion

In this paper, we present *One-to-All Animation*, a unified framework for pose-driven personalized generation under diverse misalignment scenarios. We cast training as a self-supervised outpainting task, enabling the model to handle arbitrary spatial layouts. We design a Reference Extractor with hybrid fusion attention to preserve fine-grained identity features across variable resolutions and frames. We propose Identity-Robust Pose Control to decouple facial appearance from skeletal structure and prevent pose overfitting. Moreover, we introduce Token Replace to ensure temporal consistency in long video generation. Extensive experiments and ablation studies validate our method's robustness across diverse layouts and identity preservation, demonstrating practical flexibility for real-world applications.

References

[1] Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. Conditional gan with discriminative filter generation for text-to-video synthesis. In *IJCAI*, page 2, 2019. 7

[2] 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. *CoRR*, 2023. 3

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 6

[4] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. *arXiv preprint arXiv:2311.12052*, 2023. 2, 3

[5] Chaofeng Chen. Iqa-pytorch: Pytorch toolbox for image quality assessment. https://github.com/chaofengc/IQA-PyTorch, 2022. 2

[6] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, et al. Wan-animate: Unified character animation and replacement with holistic replication. *arXiv preprint arXiv:2509.14055*, 2025. 2, 3, 6, 7

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3

[8] Qijun Gan, Yi Ren, Chen Zhang, Zhenhui Ye, Pan Xie, Xiang Yin, Zehuan Yuan, Bingyue Peng, and Jianke Zhu. Humandit: Pose-guided diffusion transformer for long-form human motion video generation. *arXiv preprint arXiv:2502.04847*, 2025. 2

[9] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025. 6, 1

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *12th International Conference on Learning Representations, ICLR 2024*, 2024. 3

[11] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 3

[12] Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

[13] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022. 3

[14] 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. *Advances in neural information processing systems*, 30, 2017. 7

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[16] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 7

[17] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 2, 3

[18] Yasamin Jafarian and Hyun Soo Park. 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*, pages 12753–12762, 2021. 6

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[20] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3

[21] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 3, 7

[22] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, Chao Liang, Yuan Zhang, and Jingtuo Liu. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human an-

imation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13847–13858, 2025. 3

[23] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 6

[24] Dongyang Liu, Shicheng Li, Yutong Liu, Zhen Li, Kai Wang, Xinyue Li, Qi Qin, Yufei Liu, Yi Xin, Zhongyu Li, et al. Lumina-video: Efficient and flexible video generation with multi-scale next-dit. *arXiv preprint arXiv:2502.06782*, 2025. 3

[25] Jiaqi Liu, Jichao Zhang, Paolo Rota, and Nicu Sebe. Multi-focal conditioned latent diffusion for person image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16019–16028, 2025. 7

[26] Jiaqi Liu, Jichao Zhang, Paolo Rota, and Nicu Sebe. Multi-focal conditioned latent diffusion for person image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16019–16028, 2025. 3

[27] Lijie Liu, Tianxiang Ma, Bingchuan Li, Zhuowei Chen, Jiawei Liu, Gen Li, Siyu Zhou, Qian He, and Xinglong Wu. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*, 2025. 4

[28] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 6

[29] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 3, 7

[30] Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2024. 7

[31] Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6420–6429, 2024. 3

[32] PaddlePaddle. Paddleocr: Awesome multilingual ocr toolkits. https://github.com/PaddlePaddle/PaddleOCR, 2021. 1

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 2, 3

[34] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 3

[35] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 4

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 3

[37] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7690–7699, 2020. 3

[38] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13535–13544, 2022. 3

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[40] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[41] Fei Shen, Hu Ye, Jun Zhang, Cong Wang, Xiao Han, and Yang Wei. Advancing pose-guided image synthesis with progressive conditional diffusion models. In *The Twelfth International Conference on Learning Representations*. 3

[42] Shijun Shi, Jing Xu, Lijing Lu, Zhihang Li, and Kai Hu. Self-supervised controlnet with spatio-temporal mamba for real-world video super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7385–7395, 2025. 3

[43] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3408–3416, 2018. 3

[44] 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. 3

[45] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*. 3

[46] Shuai Tan, Biao Gong, Xiang Wang, Shiwei Zhang, Dandan Zheng, Ruobing Zheng, Kecheng Zheng, Jingdong Chen, and Ming Yang. Animate-x: Universal character image animation with enhanced motion representation. *arXiv preprint arXiv:2410.10306*, 2024. 2, 6, 8

[47] Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21096–21106, 2025. 2, 3, 6

[48] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 7

[49] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 3, 6

[50] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9336, 2024. 7, 2

[51] 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. *arXiv preprint arXiv:2406.01188*, 2024. 2

[52] Xiang Wang, Shiwei Zhang, Longxiang Tang, Yingya Zhang, Changxin Gao, Yuehuan Wang, and Nong Sang. Unianimate-dit: Human image animation with large-scale video diffusion transformer. *arXiv preprint arXiv:2504.11289*, 2025. 2, 3, 6

[53] 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. 7

[54] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 7

[55] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 1

[56] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 3

[57] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 8

[58] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *arXiv preprint arXiv:1910.09139*, 2019. 6

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 3

[60] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 7713–7722, 2022. 3

[61] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[62] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 2, 3, 6, 8

[63] Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025. 3

[64] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 6

One-to-All Animation: Alignment-Free Character Animation and Image Pose Transfer

Supplementary Material

A. Details of Cartoon Dataset

To enhance cross-style generalization, we construct a specialized cartoon and anime dataset. We collect 1000 images from the AI art platform Shakker*, which provides diverse artistic styles and character designs. To ensure reliable pose conditioning, we manually filter out images where DWPose fails to detect clear body keypoints, resulting in 212 high-quality samples. We split these into 200 for training and 12 for evaluation. The overall curation process and representative examples are shown in Fig. 11. For each cartoon image, we generate a corresponding video clip using Seedance [9] with the text prompt "a character is dancing". Since Seedance outputs contain watermarks, we detect these regions and assign them zero loss weight during training, implementation details in Sec. B.1.

(a) Dataset curation pipeline.

(b) Examples from Cartoon dataset.

Figure 11. Cartoon Dataset construction: (a) pose filtering and video generation pipeline, (b) representative samples.

B. Implementation Details

B.1. Training Details

Multi-resolution training. We adopt a bucket-based sampler to support varying resolutions and aspect ratios training. Each bucket is defined by a target resolution (e.g., 512px or 768px) and a set of aspect ratios. During sampling, we first select a resolution according to predefined probabilities, then match the sample to the closest aspect ratio within that bucket. Frames are center-cropped and resized to the target dimensions. For example, a 720×1300 video assigned to the 512px bucket would be resized to 384×672 (closest to 9:16 aspect ratio). Images are treated as single-frame videos and processed identically. To maintain a specific video-to-image data ratio (6:1 in our experiments) per epoch, we partition images into $K/6$ groups by aspect ratio, where K is the number of video samples. When sampling images, we first select one group and then randomly pick an image from it.

Region-weighted loss. To improve generation quality, we apply spatially adaptive loss weights during training. We partition each frame into regions of interest (ROI) and non-ROI. ROI includes head regions and high-confidence hand regions, while non-ROI includes text regions. We extend the base rectified flow loss $\mathcal{L}_{\text{RF}}$ (Eq. 15) with spatial weighting:

$$\mathcal{L}_{\text{weighted}} = \mathcal{L}_{\text{RF}} \cdot \mathcal{M}^{\text{text}} \cdot \left((w_{\text{roi}} - 1) \cdot \mathcal{M}^{\text{roi}} + 1 \right), \quad (17)$$

where $\mathcal{M}^{\text{text}} \in \{0, 1\}$ is the inverse text mask (0 for text regions, 1 otherwise), $\mathcal{M}^{\text{roi}} \in \{0, 1\}$ denotes the binary ROI mask, and w_{roi} is the ROI weight usually set 3.0.

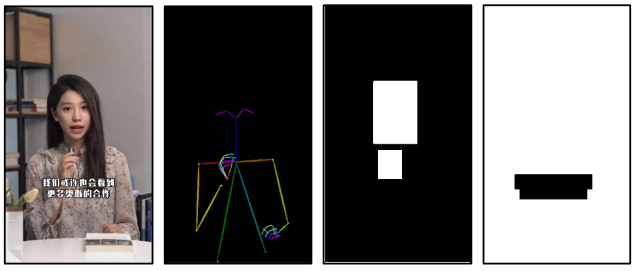


Figure 12. Visualization of region-weighted loss components.

In implementation, we use DWPose [55] for body detection and PaddleOCR [32] for text detection. Hand regions

*https://www.shakker.ai/

Table 4. Model architecture, training hyperparameters, and inference cost.

Config	**One-to-All-1.3B**	**One-to-All-14B**
Hidden Dim	1536	5120
Num Layers – Extractor	14	7
Num Layers – DiT	30	40
Pre-trained Model	Wan2.1-T2V-1.3B	Wan2.1-T2V-14B
Training Parameters		
Batch Size per GPU	1	
Optimizer	AdamW	
Learning Rate	1×10^{-5}	
Weight Decay	1×10^{-4}	
LR Schedule	constant with warmup	
Training Steps (Stage 1/2/3)	40k / 40k / 20k	30k / 30k / 20k
Inference Cost ($81 \times 576 \times 1024$, 30steps, bfloat16, H100)		
Memory (GB)	29	65
Time (mm:ss, $\lambda_P = \lambda_I = 1.0$)	0:59	3:41
Time (mm:ss, $\lambda_P = \lambda_I > 1.0$)	1:58	7:43
Time (mm:ss, $\lambda_P \neq \lambda_I > 1.0$)	2:57	11:12

PoseResNet	
Conv3d	3→16, kernel=3, stride=1
ResNet Block 1	–
Conv3d	16→16, kernel=3, stride=(1,2,2)
ResNet Block 2	–
Conv3d	16→16, kernel=3, stride=(2,2,2)
ResNet Block 3	–
Conv3d	16→16, kernel=3, stride=(2,2,2)
ResNet Block 4	–
Conv3d	16→16, kernel=3, stride=(1,2,2)
Conv3d	16→1536/5120, kernel=3, stride=1
ResNet Block	
GroupNorm	4 groups, eps=1e-6
SiLU	–
Conv3d	16→16, kernel=3, stride=1
GroupNorm	4 groups, eps=1e-6
SiLU	–
Conv3d	16→16, kernel=3, stride=1

are treated as ROI only when all corresponding keypoint confidences exceed 0.8. The resulting masks are spatially downsampled by $16\times$ and temporally merged over every 4 consecutive frames to match the latent's feature dimension. Fig. 12 visualizes example region masks.

B.2. Evaluation Details

Metrics. We quantitatively evaluate our results using several metrics, including PSNR, SSIM, LPIPS, FID, FID-VID, and FVD. The detailed metrics are introduced as follows:

- **PSNR** quantifies the pixel-level reconstruction accuracy by measuring the ratio between maximum signal power and noise power, expressed in decibels (dB). Higher PSNR values indicate better reconstruction quality.
- **SSIM** evaluates image similarity by comparing luminance, contrast, and structural patterns. Unlike PSNR, SSIM better reflects human perception of image quality.
- **LPIPS** measures perceptual distance using deep features extracted from pretrained networks. It correlates more closely with human judgment than traditional metrics.
- **FID** computes the distance between feature distributions of real and generated images using Inception network embeddings. Lower FID indicates generated images are closer to real data in feature space.
- **FID-VID** extends FID to videos by computing distribution distance on frame-level features, capturing temporal consistency.
- **FVD** measures video quality by comparing feature distributions that encode both spatial appearance and temporal dynamics. Lower values indicate better video fidelity.

Implementation. For video datasets, we follow the evaluation codebase from DisCo [50]. For image datasets, we use the PyIQA [5] library for metric computation. Specifically, we use LPIPS-VGG for all LPIPS measurements. All evaluations are performed on a single NVIDIA H100 GPU. We will release the full inference and evaluation code to facilitate reproduction.

B.3. Hyperparameters

Table 4 summarizes the model architecture, training hyperparameters, and inference cost. The table also details the PoseResNet architecture, which applies the same spatial-temporal compression as the VAE encoder. Its final layer projecting features to match the backbone dimension for element-wise addition. For training, we use the AdamW optimizer with a learning rate of 1×10^{-5} and weight decay of 1×10^{-4} following a constant schedule with warmup. The three-stage training runs for 40k / 40k / 20k steps for the 1.3B model and 30k / 30k / 20k steps for the 14B model. For inference, we benchmark both models on a single H100 GPU to generate 81 frames at 576×1024 resolution. Under our default guidance setting ($\lambda_P = \lambda_I = 1.5$), the 1.3B model requires 29 GB of memory with a runtime of 1 min 58 s, while the 14B model requires 65 GB with a runtime of 7 min 43 s. These numbers are reported without any compression or acceleration techniques and can be further optimized to support lower-memory devices.

C. Additional Results

C.1. More Ablation Studys

Ablation on RoPE Design in HRFA. A key design of HRFA is applying 3D RoPE with $f = 0$ to the cross-attention query. Here we provide additional ablation results to justify this design. We train a variant that retains full 3D RoPE encoding for Q′ (computed as $\mathrm{RoPE}_{3D}(hW_q)$ without setting $f = 0$) on 29-frame video clips. During

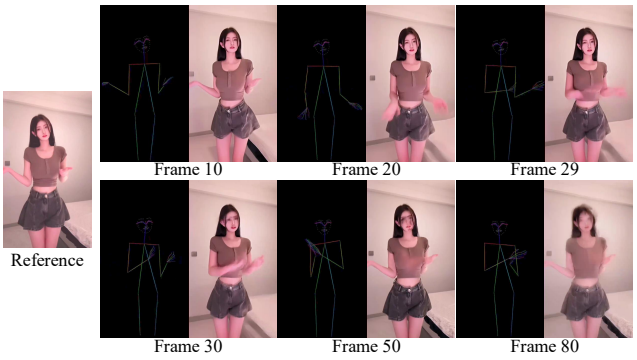


Figure 13. Failure case of retaining full 3D RoPE.

Table 5. Ablation on training stage design. All methods evaluated on TikTok benchmark using 14B model.

Training Strategy	SSIM↑	LPIPS↓	FVD↓
Pose → Ref+Pose	0.732	0.332	408.7
Joint (Ref+Pose)	0.729	0.343	415.3
Ref → Ref+Pose (Ours)	**0.773**	**0.280**	**355.2**

inference with 81-frame sequences, this variant maintains reasonable quality for the first 29 frames but exhibits severe collapse afterward, as shown in Fig. 13. This results indicates that the model overfits to absolute frame positions and cannot extrapolate beyond the training sequence length. In contrast, our design with $f = 0$ preserves the model's temporal extrapolation capability and avoids such degradation.

Ablation on Training Stage. As described in Sec. 3.5, we employ a two-stage training before token replace: stage 1 trains the reference extractor with reference condition only, and stage 2 adds pose conditioning. To validate this design, we compare three training strategies: (1) Ref → Ref+Pose (Ours), (2) Pose → Ref+Pose, and (3) Joint training from scratch. All variants exclude face region enhancement and reference-guided pose control for fair comparison. Table 5 shows our approach outperforms both alternatives across all metrics. We observe that introducing pose conditioning too early causes the model to over-rely on pose signals for both structure and appearance, making it difficult to learn robust identity features. This consistent with findings from OmniHuman [22], which also found that introducing stronger conditions (pose) too early can hinder the learning of weaker conditions (appearance). Our staged design addresses this by training appearance first, then layering motion control on top.

Ablation on outpainting strategy. To validate our self-supervised outpainting strategy, we compare it with a baseline that applies random crop-and-resize augmentation in training stage 2. As shown in Fig. 14, the baseline produces severe artifacts due to spatial misalignment, which we attribute to conflicting training goals. In stage 1, the model learns self-reconstruction with only the reference image as

Figure 14. Failure cases of crop-and-resize baseline.

condition. Without pose guidance, pixel-level alignment between the reference and output is essential for effective learning. Introducing crop-and-resize augmentation at such stage would break the alignment. Applying it only in stage 2 creates a distribution shift: the reconstruction prior learned in stage 1 becomes invalid, forcing the model to relearn spatial mappings from scratch and resulting in training instability and blur.

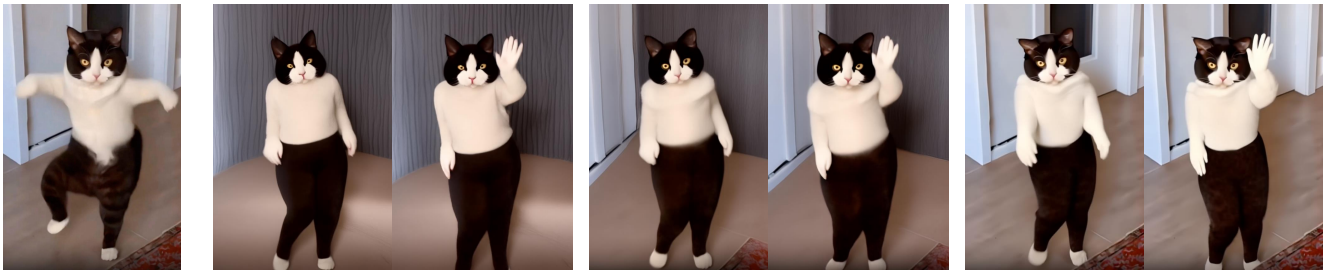
In contrast, our masking-based approach preserves the spatial alignment of visible regions and enables outpainting training from the beginning, allowing seamless transfer of the reconstruction prior to pose-guided generation.

D. More Results

We provide additional visualizations in Fig. 15, Fig. 16 and Fig. 17. Fig. 15 visualizes the performance of the 1.3B model under different guidance settings. We find that increasing the image guidance scale λ_I progressively improves character and background consistency, resulting in more robust performance on out-of-domain inputs. Fig. 16 shows that our model supports text edits of unseen elements, such as pants or environment, while maintaining identity and motion. Fig. 17 presents additional qualitative results on Cartoon benchmark.

E. Limitations and Future Work

Despite strong performance, our method has several limitations. First, we have not fully explored the optimal ratio between image and video training data. Our 1.3B model uses different checkpoints for image and video benchmarks, with the former fine-tuning on a higher image sampling ratio. Second, high-resolution or long-duration generation with the 14B model remains computationally expensive, requiring 65GB memory and over 7 minutes per inference. Future work could explore more efficient architectures or quantization methods to reduce this cost. Finally, our method currently relies solely on pose pose sequences, which only implicitly capture camera motion through character positioning. This limits precise control over camera trajectories independent of character motion. Future work could introduce explicit camera parameters or motion cues to enable independent control over camera movement.



Reference $\lambda_P = 1.5, \lambda_I = 1.5$ $\lambda_P = 1.5, \lambda_I = 2.5$ $\lambda_P = 1.5, \lambda_I = 3.0$

Figure 15. Results from the 1.3B model showing that higher λ_I improves character and background consistency.

a clown in a red suit dancing joyfully on a street.

Reference

a clown in a red suit and blue pants dancing joyfully on a street.

a clown in a red suit dancing joyfully on a street, with a McDonald's sign visible in the background.

Figure 16. Our model enables prompt-based editing while maintaining identity and motion.

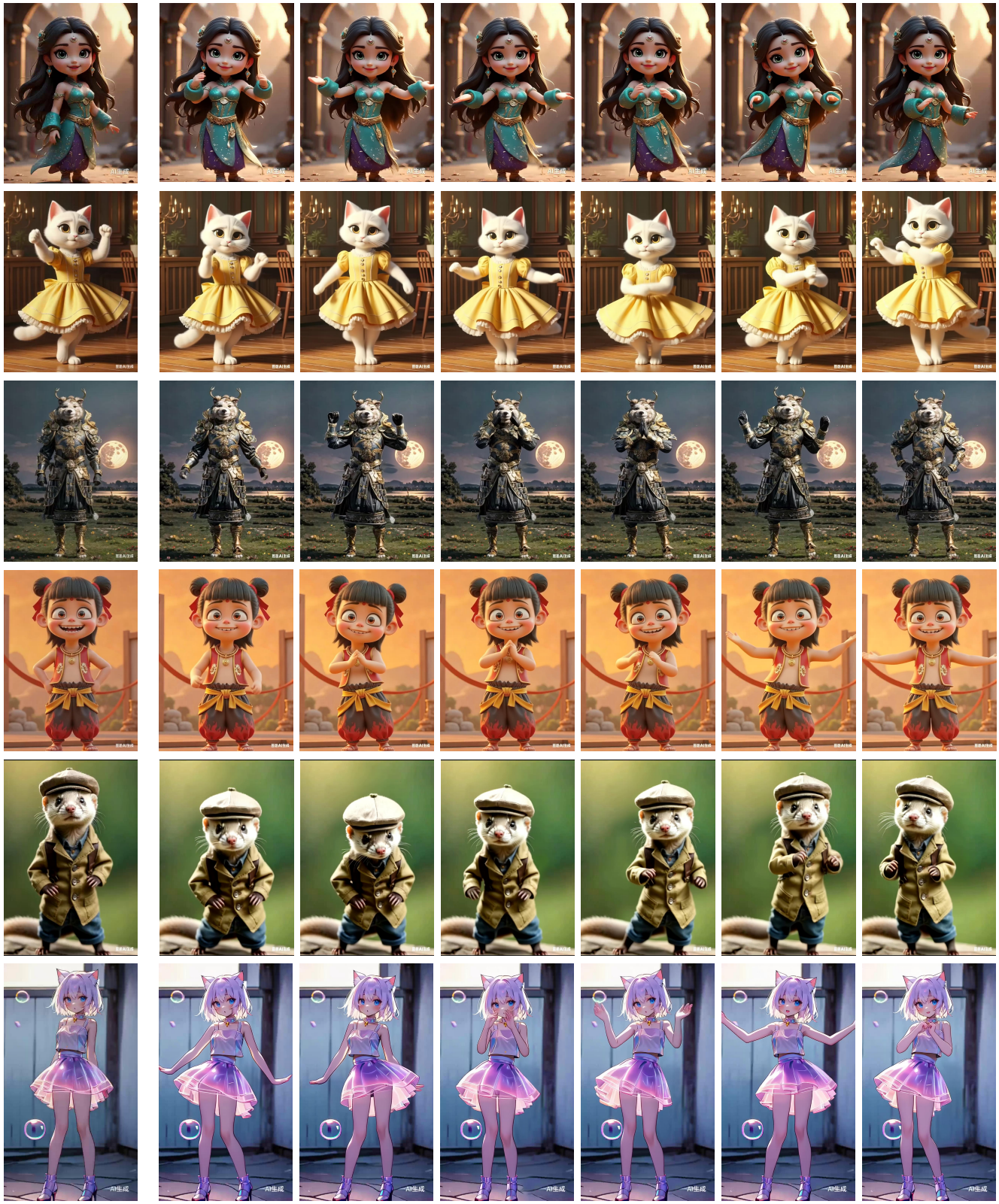


Figure 17. Additional Cartoon benchmark results.