ANIMATE-X: UNIVERSAL CHARACTER IMAGE ANIMATION WITH ENHANCED MOTION REPRESENTATION

Anonymous authors
Paper under double-blind review

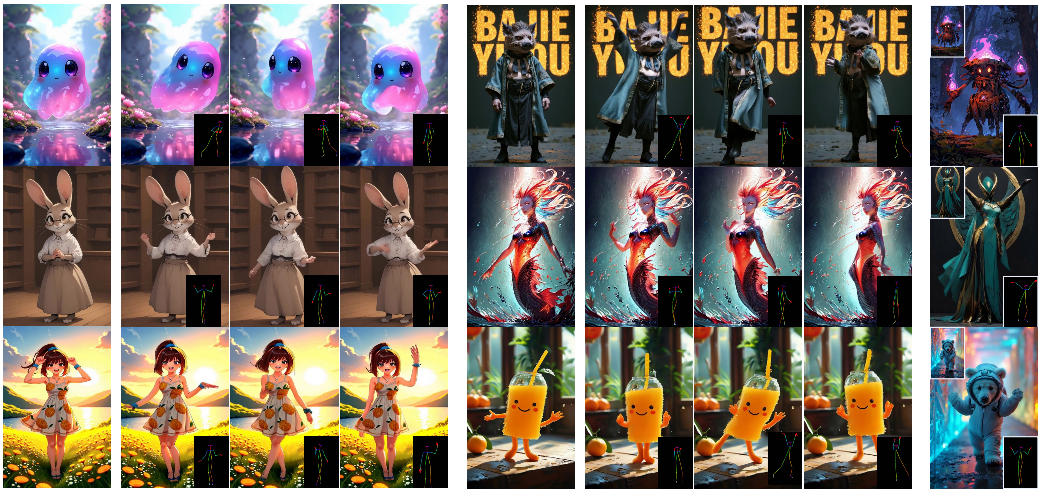


Figure 1: Illustrative animations produced by `Animate-X`. Given a character image, our method generates pose-controllable animation videos without requiring individual pose alignment. `Animate-X` demonstrates strong generalization capabilities, extending beyond human figures to anthropomorphic characters of various forms with dramatically different body structures, *e.g.*, even without limbs, from games, animations, and posters.

ABSTRACT

Character image animation, which generates high-quality videos from a reference image and target pose sequence, has seen significant progress in recent years. However, most existing methods only apply to human figures, which usually do not generalize well on anthropomorphic characters commonly used in industries like gaming and entertainment. Our in-depth analysis suggests to attribute this limitation to their insufficient modeling of motion, which is unable to comprehend the movement pattern of the driving video, thus imposing a pose sequence rigidly onto the target character. To this end, this paper proposes `Animate-X`, a universal animation framework based on LDM for various character types (collectively named `X`), including anthropomorphic characters. To enhance motion representation, we introduce the Pose Indicator, which captures comprehensive motion pattern from the driving video through both implicit and explicit manner. The former leverages CLIP visual features of a driving video to extract its gist of motion, like the overall movement pattern and temporal relations among motions, while the latter strengthens the generalization of LDM by simulating possible inputs in advance that may arise during inference. Moreover, we introduce a new Animated Anthropomorphic Benchmark (A^2`Bench`) to evaluate the performance of `Animate-X` on universal and widely applicable animation images. Extensive experiments demonstrate the superiority and effectiveness of `Animate-X` compared to state-of-the-art methods. Please use any web browser to open the *.html* file in the *Supplementary Materials* to view the generated videos.

1 Introduction

Character image animation Yang et al. (2018); Zablotskaia et al. (2019b) is a compelling and challenging task that aims to generate lifelike, high-quality videos from a reference image and a target pose sequence. A modern image animation method shall ideally *balance* the identity preservation and motion consistency, which contribute to the promise of broad utilization Hu et al. (2023); Xu et al. (2023a); Chang et al. (2023a); Jiang et al. (2022). The phenomenal successes of GAN Goodfellow et al. (2014); Yu et al. (2023); Zhang et al. (2022b) and generative diffusion models Ho et al. (2022; 2020); Guo et al. (2023) have reshaped the performance of character animation generation. Nevertheless, most existing methods only apply to the human-specific character domain. In practice, the concept of *"character"* encompasses a much broader concept than human, including anthropomorphic figures in cartoons and games, collectively referred to as X, which are often more desirable in gaming, film, short videos, etc. The difficulty in extending current models to these domains can be attributed to two main factors: (1) the predominantly human-centered nature of available datasets, and (2) the limited generalization capabilities of current motion representations.

The limitations are clearly evidenced for non-human characters in Fig. 5. To replicate the given poses, the diffusion models trained on human dance video datasets tend to introduce unrelated human characteristics which may not make sense to reference figures, resulting in abnormal distortions. In other words, these models treat identity preservation and motion consistency as *conflicting* goals and struggle to balance them, while motion control often prevails. This issue is particularly pronounced for non-human anthropomorphic characters, whose body structures often differ from human anatomy—such as disproportionately large heads or the absence of arms, as shown in Fig. 1. The primary cause is that the motion representations extracted merely from pose conditions are hard to generalize to a broad range of common cartoon characters with unique physical characteristics, leading to their excessive sacrifices in identity preservation in favor of strict pose consistency, which is an unsensible trade-off between these *conflicting* goals.

To address this issue, the natural approach is to enhance the flexibility of motion representations without discarding current pose condition, which can prevent the model from making unsensible trade-offs between overly precise poses and low fidelity to reference images. To this end, we identify two key limitations of existing methods. **First**, the simple 2D pose skeletons, constructed by connecting sparse keypoints, lack of image-level details and therefore cannot capture the essence of the reference video, such as motion-induced deformations (e.g., body part overlap and occlusion) and overall motion patterns. **Second**, the self-driven reconstruction strategy aligns reference and pose skeletons by body shape, simplifying animation but ignoring shape differences during inference. These inspire us to design the new Pose Indicator from both implicit and explicit perspectives.

In this paper, we propose `Animate-X` for animating any character X. Sparked by generative diffusion models Rombach et al. (2022), we employ a 3D-UNet Blattmann et al. (2023) as the denoising network and provide it with motion feature and figure identity as condition. To fully capture the gist of motion from the driving video, we introduce the Pose Indicator, which consists of the Implicit Pose Indicator (IPI) and the Explicit Pose Indicator (EPI). Specifically, IPI extracts implicit motion-related features with the assistance of CLIP image feature, isolating essential motion patterns and relations that cannot be directly represented by the pose skeletons from the driving video. Meanwhile, EPI enhances the representation and understanding of the pose encoder by simulating real-world misalignments between the reference image and driven poses during training, strengthening the ability to generate explicit pose features. With the combined power of implicit and explicit features, `Animate-X` demonstrates strong character generalization and pose robustness, enabling general X character animation even though it is trained solely on human datasets. Moreover, we introduce a new **A**nimated **A**nthropomorphic **Bench**mark (A^2`Bench`), which includes 500 anthropomorphic characters along with corresponding dance videos, to evaluate the performance of `Animate-X` on other types of characters. Extensive experiments on both public human animation datasets and A^2`Bench` demonstrate that `Animate-X` outperforms state-of-the-art methods in preserving identity and maintaining motion consistency in animating X. Main contributions summarized as follows:

- We present `Animate-X`, which facilitates image-conditioned pose-guided video generation with high generalizability, particularly for attractive anthropomorphic characters. To the best of our knowledge, this is the first work to animate generic cartoon images without the need for strict pose alignment.

- The rethinking about the motion inspire us to propose Pose Indicator, which extracts motion representation suitable for anthropomorphic characters in both implicit and explicit manner, enhancing the robustness of `Animate-X`.
- Since the popular datasets only contain human video with limited character diversity, we present a new A^2`Bench`, specifically for evaluating performance on anthropomorphic characters. Extensive experiments demonstrate that our `Animate-X` outperforms the competing methods quantitatively and qualitatively on both A^2`Bench` and current human animation benchmark.

2 Related Work

2.1 Diffusion models for image/video generation

In recent years, diffusion models Song et al. (2021); Ho et al. (2020) have demonstrated strong generative capabilities, pushing image generation technique towards a daily productivity tool Nichol et al. (2022); Ramesh et al. (2022); Mou et al. (2023); Huang et al. (2023); Zhang et al. (2023a); Liu et al. (2023). Pioneering works such as DALL-E 2 Ramesh et al. (2022) and Imagen Saharia et al. (2022) have showcased the extraordinary potential of diffusion models for high-quality image synthesis. Notable contributions, including Stable Diffusion Rombach et al. (2022), have well balanced scalability and efficiency, making diffusion-based image generation accessible and versatile across various applications. On the video generation front, diffusion models are making amazing progress Singer et al. (2023); Wang et al. (2023a; 2024c); Wu et al. (2023); Chai et al. (2023); Ceylan et al. (2023); Guo et al. (2023); Zhou et al. (2022); An et al. (2023); Xing et al. (2023); Qing et al. (2023); Yuan et al. (2023); Tan et al. (2024d); Gong et al. (2024). These methods joint spatio-temporal modeling to generate realistic motion dynamics and ensure temporal consistency, marking a substantial step forward in generative models for video content. In this work, we aim to tackle the character-centered image animation task, a dedicated of conditional video generation. Our approach enables the transformation of static images into dynamic animations by conditioning on desired motion. This innovation bridges the gap between image and video generation, highlights the versatility and adaptability of diffusion models in creating engaging visual narratives.

2.2 Pose-guided character motion transfer

Character image animation aims to transfer motion from the source character to the target identity Zhang et al. (2024); Chang et al. (2023b), which has experienced an impressive journey to improve animation quality and versatility. Early works Li et al. (2019); Siarohin et al. (2019b; 2021b); Zhao & Zhang (2022b); Tan et al. (2024a); Wang et al. (2022); Tan et al. (2024c;b; 2023) predominantly utilize Generative Adversarial Networks (GANs) to generate animated human images. However, these GAN-based models are often confronted by the emergence of various artifacts in the generated outputs. With the advent of diffusion models, researchers Shen et al. (2024); Zhu et al. (2024) explored how to go beyond GANs. One effort is Disco Wang et al. (2023b), which leverages ControlNet Zhang et al. (2023b) to facilitate human dance generation, demonstrating the potential of diffusion models in generating dynamic human poses. Following this, MagicAnimate Xu et al. (2023b) and Animate Anyone Hu et al. (2023) introduce transformer-based temporal attention modules Vaswani (2017), enhancing the temporal consistency of animations and resulting in more smooth movement transitions. Sparked by the linear time efficiency of Mamba Gu & Dao (2023); Gu et al. (2021) conceptually merges the merits of parallelism and non-locality, Unianimate Wang et al. (2024b) resorts to it resorts to Mamba for efficient temporal modeling.

While these approaches have improved the realism of the animations, a notable limitation remains: most current methods require strict alignment between a reference image and driving video. This restricts their applicability in the scenarios where poses cannot be easily extracted, such as anthropomorphic characters, often resulting in bizarre and unsatisfactory outputs. In contrast, our approach adopts a robust and flexible motion representation to mitigate the dependence on pose alignment. This enables the generation of high-quality animations even in cases where previous methods struggle with non-alignable poses. In this manner, our method enhances the versatility and applicability of character image animation across a broad range of contexts (`X` character).

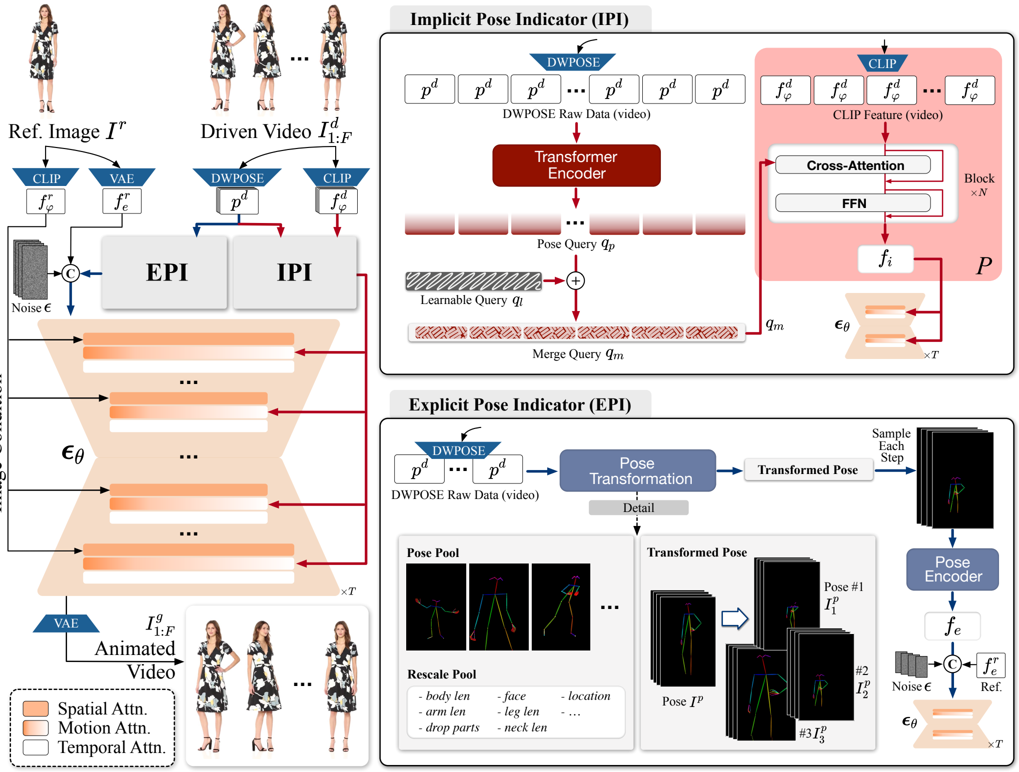


Figure 2: (a) The overview of our `Animate-X`. Given a reference image I^r, we first extract CLIP image feature f^r_φ and latent feature f^r_e via CLIP image encoder Φ and VAE encoder $\mathcal{E}$. The proposed Implicit Pose Indicator (**IPI**) and Explicit Pose Indicator (**EPI**) produce motion feature f_i and pose feature f_e, respectively. f_e is concatenated with the noised input ϵ along the channel dimension, then further concatenated with f^r_e along the temporal dimension. This serves as the input to the diffusion model ϵ_θ for progressive denoising. During the denoising process, f^r_φ and f_i provide appearance condition from I^r and motion condition from $I^d_{1:F}$. At last, a VAE decoder $\mathcal{D}$ is adopted to map the generated latent representation z_0 to the animation video. (b) The detailed structure of Implicit Pose Indicator. (c) The pipeline of pose transformation by Explicit Pose Indicator.

3 METHOD

In this work, we aim to generate an animated video that maintains consistency in identity with a reference image I^r and body movement with a driving video $I^d_{1:F}$. Different from previous works, our primary objective is to animate a general characters beyond human, particularly like anthropomorphic ones, which has broader applications in entertainment industry.

3.1 PRELIMINARIES OF LATENT DIFFUSION MODEL

A diffusion model (DM) operates by learning a probabilistic process that models data generation through noise. To mitigate the heavy computational load of traditional pixel-based diffusion models in high-dimensional RGB spaces, latent diffusion models (LDMs) Rombach et al. (2022) propose to shift the process into a lower-dimensional latent space using a pre-trained variational autoencoder (VAE) Kingma (2013). It encodes the input data into a compressed latent representation z_0. Gaussian noise is then incrementally added to this latent representation over several steps, reducing computational requirements while maintaining the generative capabilities of the model. The process can be formalized as:

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

where $\beta_t \in (0,1)$ represents the noise schedule. As $t \in 1, 2, ..., \mathcal{T}$ increases, the cumulative noise applied to the original $\mathbf{z}_0$ intensifies, causing $\mathbf{z}_t$ to progressively resemble random Gaussian noise.

Compared to the forward diffusion process, the reverse denoising process p_θ aims to reconstruct the clean sample $\mathbf{z}_0$ from the noisy input $\mathbf{z}_t$. We represent the denoising step $p(\mathbf{z}t-1|\mathbf{z}t)$ as follows:

$$p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{z}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{z}_t, t)), \tag{2}$$

in which $\boldsymbol{\mu}_\theta(\mathbf{z}_t, t)$ refers to the estimated target of the reverse diffusion process and the process typically is achieved by a diffusion model $\boldsymbol{\epsilon}_\theta$ with the parameters θ. To model the temporal dimension, the denoising model $\boldsymbol{\epsilon}_\theta$ is commonly built on a 3D-UNet architecture Blattmann et al. (2023) in video generation methods Hu et al. (2023); Wang et al. (2023c). Given the input conditional guidance c, they usually use an L2 loss to reduce the difference between the predicted noise and the ground-truth noise during the optimization process:

$$\mathcal{L} = \mathbb{E}_\theta\Big[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, c)\|^2\Big], \tag{3}$$

Once the reversed denoising stage is complete, the predicted clean latent is passed through the VAE decoder to reconstruct the predicted video in pixel space.

3.2 Pose Indicator

To extract motion representations, previous works typically detect the pose keypoints via DWPose Yang et al. (2023) from the driven video $I^d_{1:F}$ and further visualize them as pose image I^p, which are trained using self-driven reconstruction strategy. However, it brings several limitations as mentioned in Sec. 1: (1) The sole pose skeletons lack image-level details and are therefore unable to capture the essence of the reference video, such as motion-induced deformations and overall motion patterns. (2) The self-driven reconstruction training strategy naturally aligns the reference and pose images in terms of body shape, which simplifies the animation task by overlooking likely body shape differences between the reference image and the pose image during inference. Both limitations weaken the model to develop a deep, holistic motion understanding, leading to **inadequate** motion representation. To address these issues, we propose Pose Indicator, which consists of Implicit Pose Indicator (IPI) and Explicit Pose Indicator (EPI).

Implicit Pose Indicator (IPI). To extract unified motion representations from the driving video in the first limitation, we resort to the CLIP image feature $f^d_\varphi = \Phi(I^d_{1:F})$ extracted by a CLIP Image Encoder. CLIP utilizes contrastive learning to align the embeddings of related images and texts, which may include descriptions of appearance, movement, spatial relationships and etc. Therefore, the CLIP image feature is actually a highly entangled representation, containing motion patterns and relations helpful to animation generation. As presented in Fig. 2 (a), we introduce a lightweight extractor P which is composed of N stacked layers of cross-attention and feed-forward networks (FFN). In cross attention layer, we employ f^d_φ as the keys (K) and values (V). Consequently, the challenge becomes designing an appropriate query (Q), which should act as a guidance for motion extraction. Considering that the keypoints p^d extracted by DWPose provide a direct description of the motion, we design a transformer-based encoder to obtain the embedding q_p, which is regarded as an ideal candidate for Q. Nevertheless, motion modeling using sole sparse keypoints is overly simplistic, resulting in the loss of underlying motion patterns. To this end, we draw inspiration from query transformer architecture Awadalla et al. (2023); Jaegle et al. (2021) and initialize a learnable query vector q_l to complement sparse keypoints. Subsequently, we feed the merged query $q_m = q_p + q_l$ and f^d_φ into P and get the implicit pose indicator f_i, which contains the essential representation of motion that cannot be represented by the simple 2D pose skeletons.

Explicit Pose Indicator (EPI). To deal with the second limitation in the training strategy, we propose EPI, designed to train the model to handle misaligned input pairs during inference. The *key insight* lies in simulating misalignments between reference image and pose images during training while ensuring the motion remains consistent with the given driving video $I^d_{1:F}$. Therefore, we explore two pose transformation schemes: Pose Realignment and Pose Rescale. As shown in Fig. 2 (b), in the pose realignment scheme, we first establish a pose pool containing pose images from the training set. In each training step, we first sample the reference image I^r and the driving pose I^p following previous works. Additionally, we randomly select an align anchor pose I^p_{anchor} from the pose pool. This anchor serves as a reference for aligning the driving pose, producing the aligned pose $I^p_{realign}$. However, since the characters we aim to animate are often anthropomorphic characters, whose shapes can significantly differ from human, such as varying head-to-shoulder ratios, extremely short legs, or even the absence of arms (as shown in Fig. 1 and Fig. 5), relying solely on pose realignment is insufficient to capture these variations for simulation. Therefore, we further introduce Pose Rescale. Specifically, we define a set of keypoint rescaling operations, including modifying the length of the body, legs, arms, neck, and shoulders, altering face size, even adding or removing specific body parts and etc. These transformations are stored in a rescale pool. After obtaining the realigned poses $I^p_{realign}$, we apply a random selection of transformations from this pool with a certain probability on them, generating the final transformed poses I^p_n (additional examples of transformations are provided in the Appendix A). Note that we set the probability of $\lambda \in [0, 1]$ to apply the pose transformation, and with a probability of $1 - \lambda$, the pose image remains unchanged. Subsequently, I^p_n is encoded to the explicit feature f_e via a Pose Encoder.

3.3 Framework and Implement Details

In light of the success of previous works Hu et al. (2023); Zhang et al. (2024), `Animate-X` follows the main framework, which consists of several encoders for feature extraction and a 3D-UNet Wang et al. (2023a;c); Blattmann et al. (2023) for video generation. As shown in Fig. 2, given a reference image I^r, we employ the pretrained CLIP Image Encoder Φ Radford et al. (2021) to extract appearance feature f^r_φ from I^r. To reduce the parameters of the framework and facilitate appearance alignment, we exclude the Reference Net presented in most of the previous works Hu et al. (2023); Zhang et al. (2024); Zhu et al. (2024). Instead, a VAE encoder $\mathcal{E}$ is utilized to extract the latent representation f^r_e from I^r, which is then directly used as part of the input for the denoising network ϵ_θ following Wang et al. (2024b). For the driven video $I^d_{1:F}$, we detect the pose keypoints p^d and CLIP feature I^d via a DWPose Yang et al. (2023) and CLIP Image Encoder Φ. Subsequently, IPI and EPI introduced in Sec. 3.2 extract the implicit latent f_i and explicit latent f_e, respectively. The explicit f_e is first concatenated with the noised latent ϵ to obtain the fused features along the channel dimension, which is further stacked with f^r_e along the temporal dimension, resulting in combined features f_{merge}. Then, the combined features are fed into the video diffusion model ϵ_θ for jointly appearance alignment and motion modeling. The diffusion model ϵ_θ comprises multiple stacked layers of Spatial Attention, Motion Attention and Temporal Attention. The Spatial Attention receives inputs from f_{merge} and f^r_i and fuses the identity condition from I^r with the motion condition from I^d through cross-attention (CA), producing an intermediate representation x. To further enhance motion consistency, the implicit representation f_i is fed into the Motion Attention module, along with x in the form of a residual connection, resulting in the representation $x' = x + \text{CA}(x, f_i)$. Inpsired by the linear time efficiency of Mamba Gu & Dao (2023) in long sequence processing, we employ it as Temporal Attention module to maintain the temporal consistency.

Training and Inference. To improve the model's robustness against pose and reference image misalignments, we adopt two key training schemes. First, we set a high transformation probability λ (over 98%) in the EPI, enabling the model to handle a wide range of misalignment scenarios. Second, we apply random dropout to the input conditions at a predefined rate Wang et al. (2024b). After that, while the reference image and driven video are from the same human dancing video during training, in the inference phase (Fig. 9 (b)), `Animate-X` can handle an arbitrary reference image and driven video, which may differ in appearance.

3.4 A^2Bench

The main task of our `Animate-X` is to animate an anthropomorphic character with vivid and smooth motions. However, current publicly available datasets Jafarian & Park (2021); Zablotskaia et al. (2019a) primarily focus on human animation and fall short in capturing a broad range of anthropomorphic characters and corresponding dancing videos. This gap makes these datasets and benchmarks unsuitable for quantitatively evaluating different methods in anthropomorphic character animation.

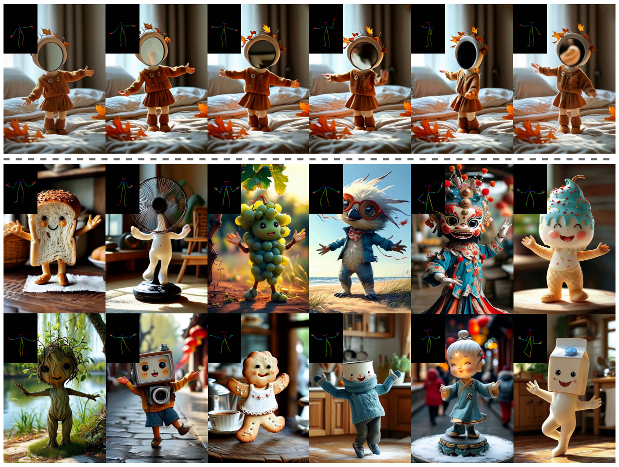


Figure 3: Examples from our A^2`Bench`.

Method	PSNR* ↑	SSIM ↑	L1 ↑	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
Moore-AnimateAnyone Corporation (2024)	9.86	0.299	1.58E-04	0.626	50.97	75.11	1367.84
MimicMotion Zhang et al. (2024) (ArXiv24)	10.18	0.318	1.51E-04	0.622	122.92	129.40	2250.13
ControlNeXt Peng et al. (2024) (ArXiv24)	10.88	0.379	1.38E-04	0.572	68.15	81.05	1652.09
MusePose Tong et al. (2024) (ArXiv24)	11.05	0.397	1.27E-04	0.549	100.91	114.15	1760.46
Unianimate Wang et al. (2024b) (ArXiv24)	11.82	0.398	1.24E-04	0.532	48.47	61.03	1156.36
Animate-X	**13.60**	**0.452**	**1.02E-04**	**0.430**	**26.11**	**32.23**	**703.87**

Table 1: Quantitative comparisons with SOTAs on A^2Bench with the rescaled pose setting. "PSNR*" means using the modified metric Wang et al. (2024a) to avoid numerical overflow.

Method	PSNR* ↑	SSIM ↑	L1 ↑	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
FOMM Siarohin et al. (2019a) (NeurIPS19)	10.49	0.363	1.47E-04	0.613	183.18	147.82	2535.12
MRAA Siarohin et al. (2021a) (CVPR21)	12.62	0.420	1.09E-04	0.556	161.57	196.87	3094.68
LIA Wang et al. (2022) (ICLR22)	13.78	0.445	9.70E-05	0.497	105.13	78.51	1813.28
DreamPose Karras et al. (2023) (ICCV23)	7.76	0.305	2.28E-04	0.534	277.64	315.58	4324.42
MagicAnimate Xu et al. (2023a) (CVPR24)	11.90	0.396	1.17E-04	0.523	117.09	117.54	2021.93
Moore-AnimateAnyone Corporation (2024) (CVPR24)	11.56	0.360	1.27E-04	0.532	37.82	59.80	1117.29
MimicMotion Zhang et al. (2024) (ArXiv24)	12.66	0.407	1.07E-04	0.497	96.46	61.77	1368.83
ControlNeXt Peng et al. (2024) (ArXiv24)	12.82	0.421	1.02E-04	0.472	46.66	59.41	1152.96
MusePose Tong et al. (2024) (ArXiv24)	12.92	0.438	9.90E-05	0.470	80.22	87.97	1401.96
Animate-X	**14.10**	**0.463**	**8.92E-05**	**0.425**	**31.58**	**33.15**	**849.19**

Table 2: Quantitative comparisons with existing methods on A^2Bench in the self-driven setting. Underline means the second best result.

To bridge this gap, we propose the **A**nimated **A**nthropomorphic character **Bench**mark (A^2Bench) to comprehensively evaluate the performance of different methods. Specifically, we first provide a prompt template to GPT-4 OpenAI (2024) and leverage it to generate 500 prompts, each of which contains a textual description of an anthropomorphic character. Please refer to Appendix B.2 for details. Inspired by the powerful image generation capability of KLing AI Technology (2024), we feed the produced prompts into its Text-To-Image module, which synthesizes the corresponding anthropomorphic character images according to the given text prompts. Subsequently, the Image-To-Video module is employed to further make the characters in the images dance vividly. For each prompt, we repeat the process for 4 times and filter the most satisfactory image-video pairs as the output corresponding to this prompt. In this manner, we collect 500 anthropomorphic characters and the corresponding dance videos, as shown in Fig. 3. Please refer to Appendix B for details.

4 Experiments

4.1 Experimental Settings

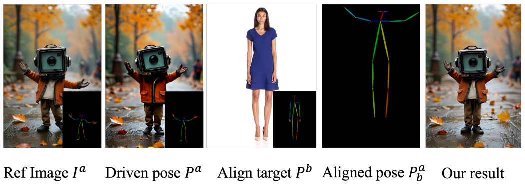


Figure 4: The illustration of comparison settings.

Dataset. We collect approximately 9,000 human videos from the internet and supplement this with TikTok dataset Jafarian & Park (2021) and Fashion dataset Zablotskaia et al. (2019a) for training. Following previous works Hu et al. (2023); Zablotskaia et al. (2019a); Jafarian & Park (2021), we use 10 and 100 videos for both qualitative and quantitative comparisons from TikTok and Fashion dataset, respectively. We additionally experimented on 100 image-video pairs selected from the newly proposed A^2Bench introduced in Sec 3.4. Please note that, to ensure a fair comparison, the data in the A^2Bench are **not** included in the training set to train our model. The data are only used to evaluate the quantitative results and provide interesting reference image cases.

Evaluation Metrics. We assess the results using evaluation metrics in Appendix B.1, including PSNR Hore & Ziou (2010), SSIM Wang et al. (2004), L1, LPIPS Zhang et al. (2018), which are widely-used image metrics for measuring the visual quality of the generated results. In addition, we introduce FID Heusel et al. (2017), FID-VID Balaji et al. (2019) and FVD Unterthiner et al. (2018) to quantify the discrepancy between the generated video distribution and the real video distribution.

4.2 Experimental Results

Quantitative Results. Since our Animate-X primarily focuses on animating the anthropomorphic characters, very few of which, if not none, can be extracted the pose skeleton accurately by

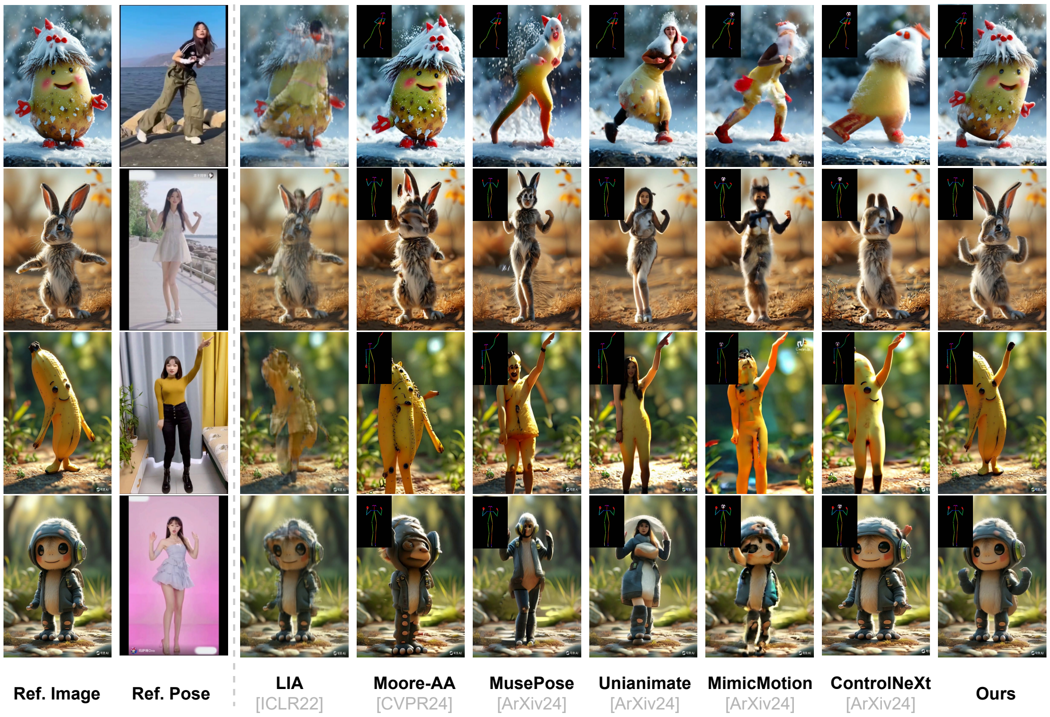


Figure 5: Qualitative comparisons with state-of-the-art methods.

DWPose Yang et al. (2023). It naturally leads to a misalignment of the input reference image with the driving pose images. To compute quantitative results in this case, we set up a new comparison setting. For each case in A^2Bench (*i.e.*, a reference image I^a and a pose P^a, as shown in Fig. 4), we randomly select one human's pose image P^b and align the anthropomorphic character's pose P^a to it, such that the aligned pose p^a_b retains the movements of P^a but has the same body shape (fat/thin, tall/short, *etc.*) as p^b. Ultimately, we take the anthropomorphic character I_a and the aligned driving pose image p^a_b as inputs to the model, generating results that allow it to calculate quantitative metrics with the original anthropomorphic character dancing video in A^2Bench. In this setting, we compare our method with Animate Anyone Hu et al. (2023), Unianimate Wang et al. (2024b), MimicMotion Zhang et al. (2024), ControlNeXt Peng et al. (2024) and MusePose Tong et al. (2024), which also use pose images (*e.g.,* P^b in Fig. 4) as input. The results of Animate Anyone Hu et al. (2023) are obtained by leveraging the publicly available reproduced code Corporation (2024). Tab. 1 presents the quantitative results, where Animate-X markedly surpasses all comparative methods in terms of all metrics. It is worth noting that, we do not use A^2Bench as training data to avoid overfitting and ensure fair comparisons, in line with other comparative methods.

Following previous works which evaluate quantitative results in self-driven and reconstruction manner, we additionally compare our method with (a) GAN-based image animate works: FOMM Siarohin et al. (2019a), MRAA Siarohin et al. (2021a), LIA Wang et al. (2022). (b) Diffusion model-based image animate works: DreamPose Karras et al. (2023), MagicAnimate Xu et al. (2023a) and present the results in Tab. 2, which indicates that our method achieves the best performance across all the metrics. Moreover, we provide the quantitative results on the human dataset (TikTok and Fashion) in Tab. 7 and Tab. 8, respectively. Please refer to Appendix D.2 for details. Animate-X reaches the comparable score to Unianimate and exceeds other SOTA methods, which demonstrates the superiority of Animate-X on **both** anthropomorphic and human benchmarks.

Qualitative Results. Qualitative comparisons of anthropomorphic animation are shown in Fig. 5. We observe that GAN-based LIA Wang et al. (2022) does not generalize well, which can only work on a specific dataset like Siarohin et al. (2019b). Benefiting from the powerful generative capabilities of the diffusion model, Animate Anyone Hu et al. (2023) renders a higher resolution image, but the identity of the image changes and do not generate an accurate reference pose motion. Although MusePose Tong et al. (2024), Unianimate Wang et al. (2024b) and MimicMotion Zhang et al. (2024) improve the accuracy of the motion transfer, these methods generate a unseen person, which is not

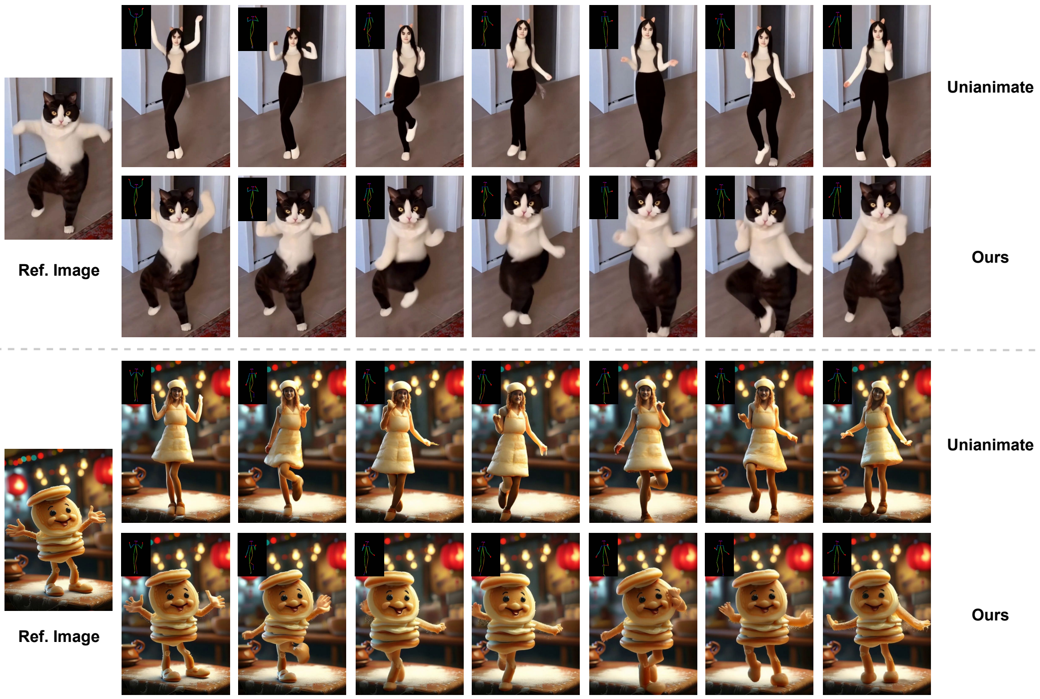


Figure 6: Qualitative comparisons with Unianimate in terms of long video generation.

Method	Moore-AA	MimicMotion	ControlNeXt	MusePose	Unianimate	**Animate-X**
Identity preservation ↑	60.4%	14.8%	52.0%	31.3%	43.0%	**98.5%**
Temporal consistency ↑	19.8%	24.9%	36.9%	43.9%	81.1%	**93.4%**
Visual quality ↑	27.0%	17.2%	40.4%	40.3%	79.3%	**95.8%**

Table 3: User study results.

the desired result. ControlNeXt combines the advantages of the above two types of methods, so maintains the consistency of identity and motion transfer to some extent, yet the results are somewhat unnatural and unsatisfactory, *e.g.*, the ears of the rabbit and the legs of the banana in Fig. 5. In contrast, Animate-X ensures both identity and consistency with the reference image while generating expressive and exaggerated figure motion, rather than simply adopting quasi-static motion of the target character. Further, we present some long video comparisons in Fig. 6. Unianimate generates a woman out of thin air who dances according to the given pose images. Animate-X animates the reference image in a cute way while preserving appearance and temporal continuity, and it does not generate parts that do not originally exist. In summary, Animate-X excels in maintaining appearance and producing precise, vivid animations with a high temporal consistency. Please refer to Appendix D.1 for details.

User Study. To estimate the quality of our method and SOTAs from human perspectives, we conduct a blind user study with 10 participants. Specifically, we randomly select 10 characters from A^2Bench and collect 10 driving video from the website. For each of 6 methods tested, 10 animation clips are generated, resulting in a total of 60 clips. Each participant is presented two results generated by different methods for the same set of inputs and asked to choose which one is better in terms of *visual quality*, *identity preservation*, and *temporal consistency*. This process is repeated C_2^6 times. The results are summarized in Tab. 3, where our method noticeably outperforms other methods in all aspects, demonstrating its superiority and effectiveness. Details in Appendix C.

4.3 Ablation Study

Ablation on Implicit Pose Indicator. To analyze the contributions of Implicit Pose Indicator, we remove it from Animate-X as w/o IPI and compare it with Baseline and Animate-X. From the first row of Fig. 7, we observe that Baseline generates a person whose appearance is appreciably distinct from the reference image. With the help of EPI, this problem is mildly mitigated. However, due to the absence of IPI, compared to Ours, there are still strange things and human-like hands

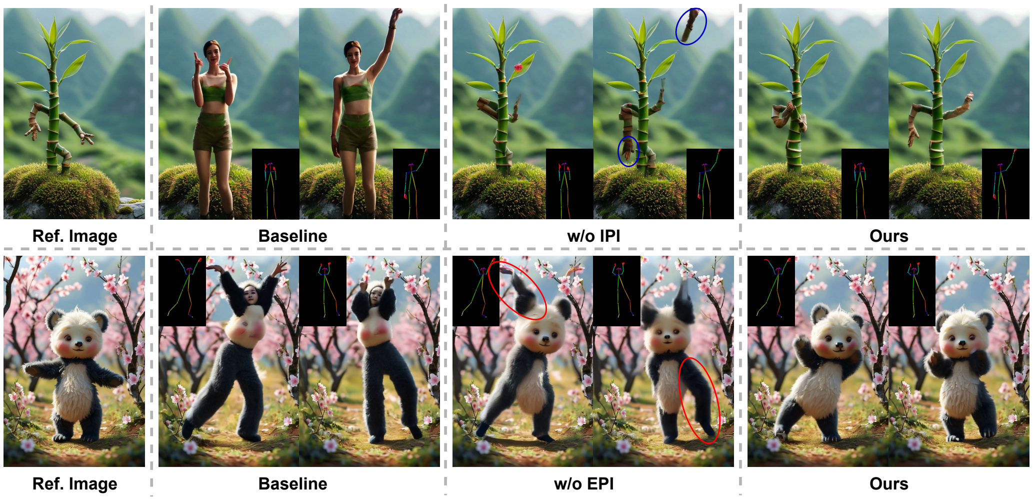


Figure 7: Visualization of ablation study on IPI and EPI.

appearing, as indicated by the blue circle. For more detailed analysis about the structure of IPI, we set up several variants: (1) remove IPI: w/o IPI. (2) remove learnable query: w/o LQ. (3) remove DWPose query: w/o DQ. The quantitative results are shown in Tab. 4. It can be seen that removing the entire IPI presents the worst performance. By modifying the IPI module, although it improves on the w/o IPI, it still falls short of the final result of Animate-X, which suggests that our current IPI structure is the most reasonable and achieves the best performance.

Ablation on Explicit Pose Indicator. We demonstrate the visual results of ablating EPI setting in the second row of Fig. 7 by removing EPI. Without EPI, although the appearance of the panda is preserved thanks to IPI, the model incorrectly treats the panda's ears as arms and forcibly stretches the legs to match the length of the legs in the pose image indicated by red circles. In contrast, these issues are completely resolved by the assistance of EPI. We further conduct more detailed ablation experiments for different pairs of pose transformations by (1) removing the entire EPI: w/o EPI. (3) remove Pose Realignment: w/o Realignment. (2) removing Pose Rescale: w/o Rescale; From the results displayed in Tab. 4, we found that Pose Realignment contributes the most. It suggests that simulating misalignment case in inference is the the key factor.

Method	PSNR* ↑	SSIM ↑	L1 ↑	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
w/o IPI	13.30	0.433	1.35E-04	0.454	32.56	64.31	893.31
w/o LQ	13.48	0.445	1.76E-04	0.454	28.24	42.74	754.37
w/o DQ	13.39	0.445	**1.01E-04**	0.456	30.33	62.34	913.33
w/o EPI	12.63	0.403	1.80E-04	0.509	42.17	58.17	948.25
w/o Realign	12.27	0.433	1.17E-04	0.434	34.60	49.33	860.25
w/o Rescale	13.23	0.438	1.21E-04	0.464	27.64	35.95	721.11
Animate-X	**13.60**	**0.452**	1.02E-04	**0.430**	**26.11**	**32.23**	**703.87**

Table 4: Quantitative results of ablation study.

In summary, we can draw conclusions: (1) IPI facilitates the preservation of appearance and prevents the generation of content that does not exist in the reference image like human arms. (2) EPI prevents the forced alignment of a pose image that is not naturally aligned with the reference image during animation, thus avoiding the unintended animation of parts that should remain static like the panda's ears shown in Fig. 7. Please refer to Appendix D.4 for details.

5 Conclusions

In this study, we present Animate-X, a novel approach to character animation capable of generalizing across different types of characters named X. To address the imbalance between identity preservation and movement consistency caused by the insufficient motion representation, we introduce the Pose Indicator, which leverages both implicit and explicit features to enhance the motion understanding of the model. In this way, Animate-X demonstrates strong generalization and robustness, achieving general X character animation. The proposed framework showcases significant improvements over state-of-the-art methods in terms of identity preservation and motion consistency, as evidenced by experiments on both public datasets and the newly introduced A^2Bench, which features anthropomorphic characters. Limitation and ethical considerations see Appendix E.

REFERENCES

Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation. arXiv preprint arXiv:2304.08477, 2023.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390, 2023.

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, volume 1, pp. 2, 2019.

Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In CVPR, pp. 5968–5976, 2023.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In CVPR, pp. 22563–22575, 2023.

Zinelabidine Boulkenafet, Jukka Komulainen, and Abdenour Hadid. Face anti-spoofing based on color texture analysis. In 2015 IEEE international conference on image processing (ICIP), pp. 2636–2640. IEEE, 2015.

Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In ICCV, pp. 23206–23217, 2023.

Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In ICCV, pp. 23040–23050, 2023.

Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Xiao Yang, and Mohammad Soleymani. Magicdance: Realistic human dance video generation with motions & facial expressions transfer. arXiv preprint arXiv:2311.12052, 2023a.

Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In Forty-first International Conference on Machine Learning, 2023b.

Moore Threads Corporation. Moore-AnimateAnyone. 2024. URL https://github.com/MooreThreads/Moore-AnimateAnyone.

Biao Gong, Siteng Huang, Yutong Feng, Shiwei Zhang, Yuyuan Li, and Yu Liu. Check locate rectify: A training-free layout calibration system for text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6624–6634, 2024.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. NeurIPS, 27, 2014.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396, 2021.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725, 2023.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. NeurIPS, 33: 6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303, 2022.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In 2010 20th international conference on pattern recognition, pp. 2366–2369. IEEE, 2010.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. arXiv preprint arXiv:2311.17117, 2023.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. ICML, 2023.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In International conference on machine learning, pp. 4651–4664. PMLR, 2021.

Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In CVPR, pp. 12753–12762, 2021.

Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. ACM Transactions on Graphics, 41(4):1–11, 2022.

Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. In ICCV, pp. 22680–22690, 2023.

Diederik P Kingma. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3693–3702, 2019.

Ming Liu, Yuxiang Wei, Xiaohe Wu, Wangmeng Zuo, and Lei Zhang. Survey on leveraging pre-trained generative adversarial networks for image editing and restoration. Science China Information Sciences, 66(5):151101, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453, 2023.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In ICML, pp. 16784–16804, 2022.

OpenAI. Chatgpt-4o. 2024. URL https://chat.openai.com/chat.

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.

Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. arXiv preprint arXiv:2312.04483, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In ICML, pp. 8748–8763, 2021.

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.

Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In CVPR, pp. 13535–13544, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, pp. 10684–10695, 2022.

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. NeurIPS, 35:36479–36494, 2022.

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, 2024. URL https://openreview.net/forum?id=rHzapPnCgT.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. NeurIPS, 32, 2019a.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. Advances in neural information processing systems, 32, 2019b.

Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In CVPR, pp. 13653–13662, 2021a.

Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13653–13662, 2021b.

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. ICLR, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In ICLR, 2021.

Shuai Tan, Bin Ji, and Ye Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22146–22156, 2023.

Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. arXiv preprint arXiv:2404.01647, 2024a.

Shuai Tan, Bin Ji, Yu Ding, and Ye Pan. Say anything with any style. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 5088–5096, 2024b.

Shuai Tan, Bin Ji, and Ye Pan. Flowvqtalker: High-quality emotional talking face generation through normalizing flow and quantization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26317–26327, 2024c.

Shuai Tan, Bin Ji, and Ye Pan. Style2talker: High-resolution talking head generation with emotion style and art style. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pp. 5079–5087, 2024d.

Kuaishou Technology. Kling ai. 2024. URL https://klingai.kuaishou.com.

Zhengyan Tong, Chao Li, Zhaokang Chen, Bin Wu, and Wenjiang Zhou. Musepose: a pose-driven image-to-video framework for virtual human generation. arxiv, 2024.

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.

A Vaswani. Attention is all you need. Advances in Neural Information Processing Systems, 2017.

Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. arXiv preprint arXiv:2308.06571, 2023a.

Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. arXiv e-prints, pp. arXiv–2307, 2023b.

Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. In ICLR, 2024a.

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. NeurIPS, 2023c.

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, 2024b.

Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. In CVPR, 2024c.

Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. arXiv preprint arXiv:2203.09043, 2022.

Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5042–5051, 2020.

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

Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In ICCV, pp. 7623–7633, 2023.

Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. arXiv preprint arXiv:2308.09710, 2023.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. arXiv preprint arXiv:2311.16498, 2023a.

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 arXiv, 2023b.

Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In ECCV, pp. 201–216, 2018.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In ICCV, pp. 4210–4220, 2023.

Wing-Yin Yu, Lai-Man Po, Ray CC Cheung, Yuzhi Zhao, Yu Xue, and Kun Li. Bidirectionally deformable motion modulation for video-based human pose transfer. In ICCV, pp. 7502–7512, 2023.

Zitong Yu, Chenxu Zhao, Zezheng Wang, Yunxiao Qin, Zhuo Su, Xiaobai Li, Feng Zhou, and Guoying Zhao. Searching central difference convolutional networks for face anti-spoofing. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 5295–5305, 2020.

Hangjie Yuan, Shiwei Zhang, Xiang Wang, Yujie Wei, Tao Feng, Yining Pan, Yingya Zhang, Ziwei Liu, Samuel Albanie, and Dong Ni. Instructvideo: Instructing video diffusion models with human feedback. arXiv preprint arXiv:2312.12490, 2023.

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, 2019a.

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, 2019b.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In ICCV, pp. 3836–3847, 2023a.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023b.

Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In CVPR, pp. 7713–7722, 2022a.

Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In CVPR, pp. 7713–7722, 2022b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In CVPR, pp. 586–595, 2018.

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

Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In CVPR, pp. 3657–3666, 2022a.

Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3657–3666, 2022b.

Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. Magicvideo: Efficient video generation with latent diffusion models. arXiv preprint arXiv:2211.11018, 2022.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In European Conference on Computer Vision (ECCV), 2024.

APPENDICES

A NETWORK DETAILS

Due to space constraints in the main paper, we only present a brief overview of the EPI process. Here, in Fig. 8, we provide a more detailed explanation of the pose transformation in EPI, along with additional case examples. First, we sample a driving pose I^p and then randomly select an anchor pose I^p_{anchor} from the pose pool (two examples are shown in Fig. 8). The driving pose I^p is aligned to the anchor pose I^p_{anchor}, resulting in the aligned pose $I^p_{realign}$. Next, we apply several rescaling operations randomly chosen from the rescale pool to further modify the aligned pose $I^p_{realign}$. By combining different rescaling options, we can obtain multiple transformed poses I^p_n. However, it is important to note that in each training step, only one anchor pose I^p_{anchor} and one rescaling combination are selected, so only one transformed pose I^p_n is used for training. As shown in the Fig. 8, the transformed pose I^p_n retains the same motion as the sampled pose I^p but has a body shape similar to the anchor pose I^p_{anchor}. This simulates scenarios during inference where there are body shape differences between the reference image and the driving pose, enabling the model to generalize to such cases.

In the experiments, we use the visual encoder of the multi-modal CLIP-Huge model Radford et al. (2021) in Stable Diffusion v2.1 Rombach et al. (2022) to encode the CLIP embedding of the reference image and driving videos. The pose encoder, composed of several convolutional layers, follows a similar structure to the STC-encoder in VideoComposer Wang et al. (2023c). For model initialization, we employ a pre-trained video generation model Wang et al. (2024c), as done in previous approaches Xu et al. (2023a); Hu et al. (2023); Zhu et al. (2024); Wang et al. (2024b). The experiments are carried out using 8 NVIDIA A100 GPUs. During training, videos are resized to a spatial resolution of 768×512 pixels, and we feed the model with uniformly sampled video segments of 32 frames to ensure temporal consistency. We use the AdamW optimizer Loshchilov & Hutter (2017) with learning rates of 5e-7 for the implicit pose indicator and 5e-5 for other modules. For noise sampling, DDPM Ho et al. (2020) with 1000 steps is applied during training. In the inference phase, we adjust the length of the driving pose to align roughly with the reference pose and used the DDIM sampler Song et al. (2021) with 50 steps for faster sampling.

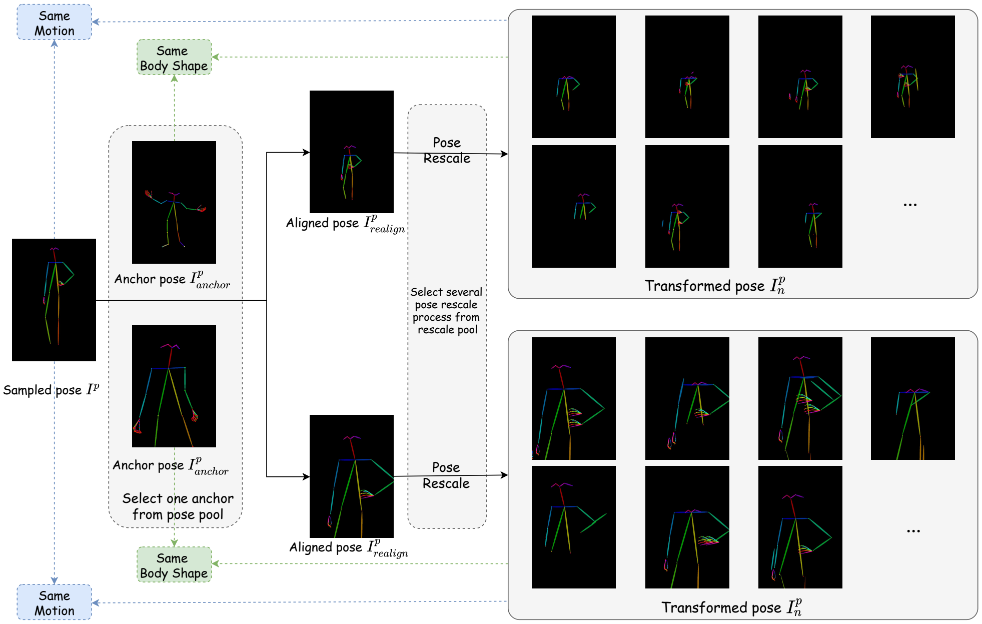

Same Motion
Same Body Shape
Aligned pose $I^p_{realign}$
Pose Rescale
Anchor pose I^p_{anchor}
Sampled pose I^p
Select several pose rescale process from rescale pool
Transformed pose I^p_n
...
Anchor pose I^p_{anchor}
Select one anchor from pose pool
Pose Rescale
Aligned pose $I^p_{realign}$
Same Body Shape
Same Motion
...
Transformed pose I^p_n

Figure 8: More example for EPI.

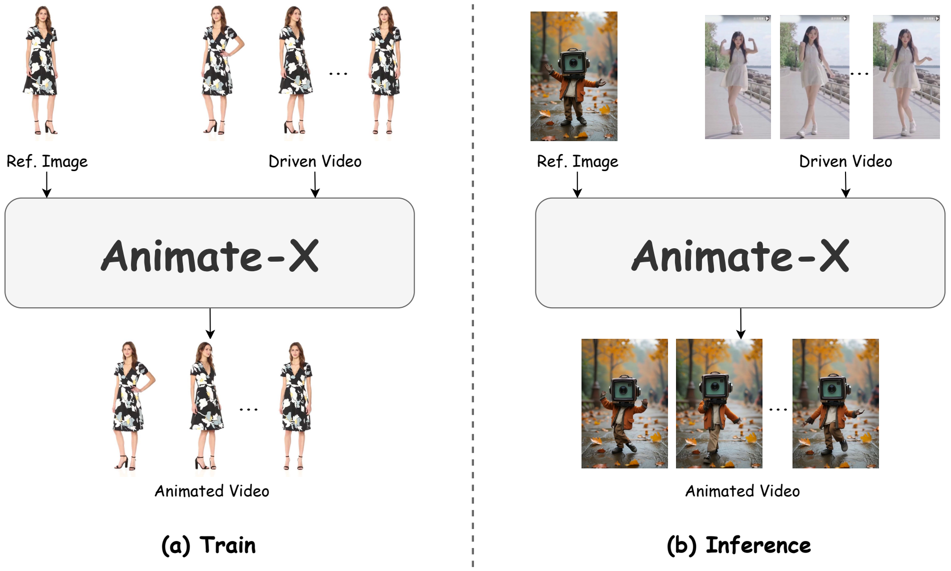


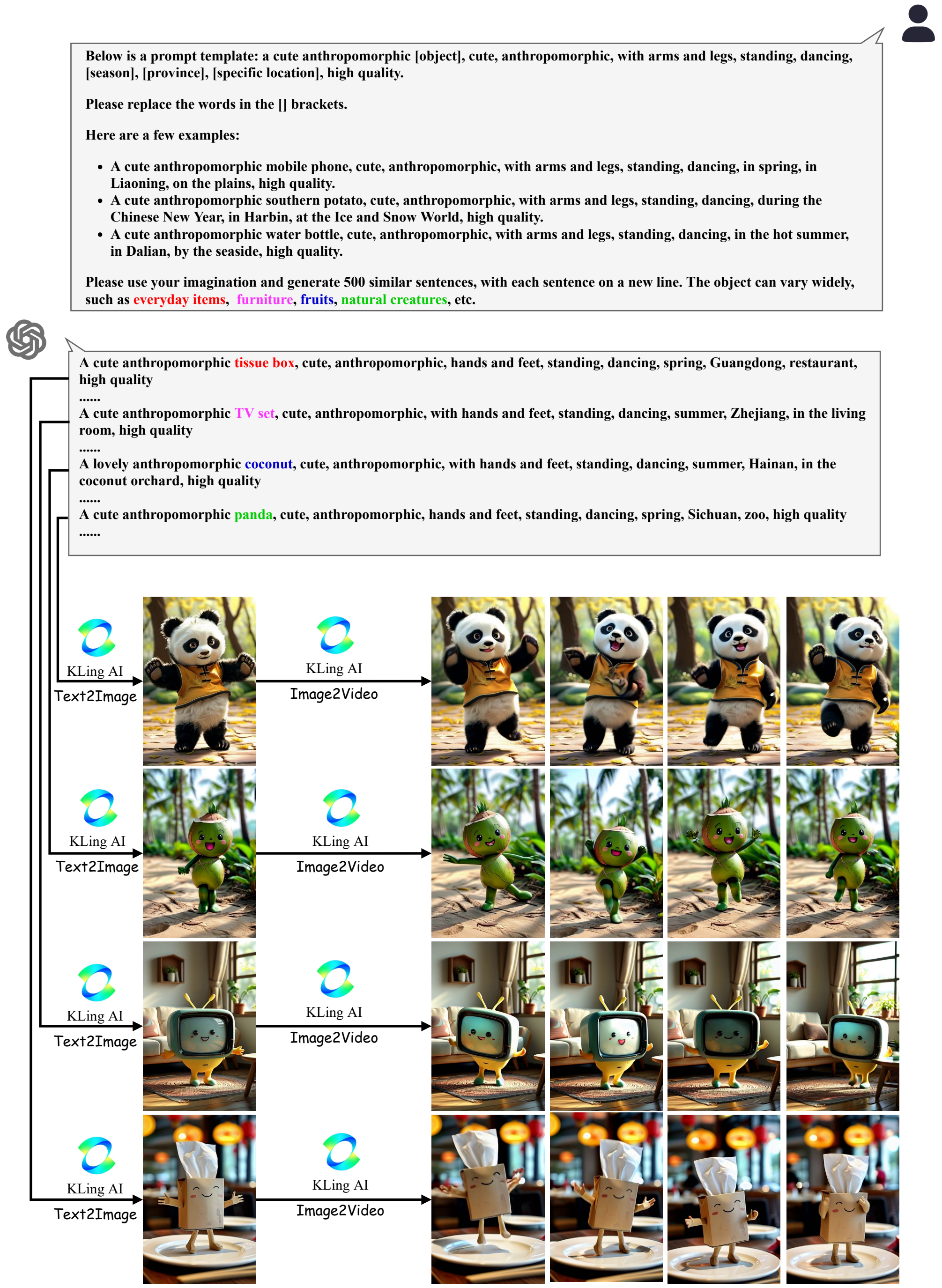
Figure 9: The difference of training and inference pipeline. During training, the reference image and the driven video come from the same video, while in the inference pipeline, the reference image and the driven video can be from any sources and appreciably different.

B Benchmark Details

B.1 Evaluation Metric

We employ several evaluation metrics to quantitatively assess our results, including PSNR, SSIM, L1, LPIPS, FID, FID-VID and FVD. The detailed metrics are introduced as follows:

- PSNR is a measure used to evaluate the quality of reconstructed images compared to the original ones. It is expressed in decibels (dB) and higher values indicate better quality. PSNR is commonly used in image compression and restoration fields.
- SSIM assesses the similarity between two images based on their luminance, contrast, and structural information. It considers perceptual phenomena affecting human vision and thus provides a better correlation with perceived image quality than PSNR.
- The L1 metric refers to the mean absolute difference between the corresponding pixel values of two images. It quantifies the average magnitude of errors in predictions without considering their direction, making it useful for measuring the extent of differences.
- LPIPS is a perceptual distance metric based on deep learning. It evaluates the similarity between images by analyzing the feature representations of image patches and tends to align well with human visual perception, making it suitable for tasks like image generation.
- FID is used to assess the quality of images generated by generative models (like GANs) by comparing the distribution of generated images to that of real images in feature space (extracted by a pretrained CNN). Lower FID values suggest that the generated images are more similar to real images.
- FID-VID extends the FID metric to video data. It measures the quality of generated videos by comparing the distribution of generated video features to real video features, providing insights into the temporal aspects of video generation.
- FVD is another metric for evaluating video generation, similar to FID. It measures the distance between the feature distributions of real and generated videos, taking both spatial and temporal dimensions into account. Lower FVD indicates that generated videos are closer to real ones regarding visual quality and dynamics.



Below is a prompt template: a cute anthropomorphic [object], cute, anthropomorphic, with arms and legs, standing, dancing, [season], [province], [specific location], high quality.

Please replace the words in the [] brackets.

Here are a few examples:

- A cute anthropomorphic mobile phone, cute, anthropomorphic, with arms and legs, standing, dancing, in spring, in Liaoning, on the plains, high quality.
- A cute anthropomorphic southern potato, cute, anthropomorphic, with arms and legs, standing, dancing, during the Chinese New Year, in Harbin, at the Ice and Snow World, high quality.
- A cute anthropomorphic water bottle, cute, anthropomorphic, with arms and legs, standing, dancing, in the hot summer, in Dalian, by the seaside, high quality.

Please use your imagination and generate 500 similar sentences, with each sentence on a new line. The object can vary widely, such as everyday items, furniture, fruits, natural creatures, etc.

A cute anthropomorphic tissue box, cute, anthropomorphic, hands and feet, standing, dancing, spring, Guangdong, restaurant, high quality
......
A cute anthropomorphic TV set, cute, anthropomorphic, with hands and feet, standing, dancing, summer, Zhejiang, in the living room, high quality
......
A lovely anthropomorphic coconut, cute, anthropomorphic, with hands and feet, standing, dancing, summer, Hainan, in the coconut orchard, high quality
......
A cute anthropomorphic panda, cute, anthropomorphic, hands and feet, standing, dancing, spring, Sichuan, zoo, high quality
......

Figure 10: Detailed pipeline for building A^2Bench based on large-scale pretrained models, including Open-ChatGPT 4o and KLing AI.

B.2 Data Details

The detailed process for constructing A^2Bench is outlined in Fig. 10. We initially provide GPT-4o with a template that clearly specifies the demand to generate ‘anthropomorphized’ images. The images were required to be cute, with arms and legs, standing, dancing, and of high quality. To allow for a variety of image outputs, we left the fields for ‘object’, ‘season’, ‘province’, and ‘specific location' empty. For the key factor influencing diversity and relevance, i.e., 'object', we provide a selectable range, such as everyday items, furniture, fruits, and natural creatures. To help GPT-4o better understand our intent, we additionally provide two examples, where the prompts had already been proven to generate satisfactory images by text-to-image module of KLing AI. Thanks to the text understanding and generation capabilities of GPT-4o, we collect 500 prompts for image generation. We then fed these 500 prompts into the text-to-image module of Keling AI, obtaining corresponding anthropomorphic characters images. Based on these images, we further generate videos of them dancing using the image-to-video module of Keling AI. In this way, we collect 500 pairs of images and videos of anthropomorphic characters, forming our $A^2\texttt{Bench}$.

Since most current animation methods Wang et al. (2024b); Hu et al. (2023); Zhang et al. (2024) take a pose image sequence as motion source, we also provide our $A^2\texttt{Bench}$ with additional pose images. To achieve this, we employ DWPose Yang et al. (2023) to extract pose sequences from the videos. However, since DWPose is trained on human data, it does not accurately extract every pose in the dancing video of the anthropomorphic character, so after extraction, we manually screen 100 videos with accurate poses, and view them as test videos for calculating quantitative metrics. Fig. 3 displays several examples, which include anthropomorphic characters of plants, animals, food, furniture, etc. For images and videos where pose extraction is not feasible, we take them as key sources of reference images in our qualitative demonstrations. This will inspire the community to animate a wider range of interesting cases. We also anticipate that these data could serve as an important resource for future pose extraction algorithms tailored to anthropomorphic datasets, making them accessible for broader use.

C User Study

In Fig. 11, we present examples shown to participants for evaluation in our user study. To obtain genuine feedback reflective of practical applications, the ten participants in our user study experiment come from diverse academic backgrounds. Since many of them do not major in computer vision, we provide detailed explanations for each question to assist their judgments.

- Identity Preservation: By comparing the reference image with the two generated videos by different methods, determine which video's character more closely resembles the character in the image.
- Temporal Consistency: Evaluate the motion changes of the character within the video and compare which video exhibits more coherent movement.
- Visual Quality: Compared to the previous two questions, this one involves more subjective judgment. Participants should assess the videos comprehensively based on visual content (e.g., flashes, distortions, afterimages), motion effects (e.g., smoothness, physical logic), and overall plausibility.

D Additional Experimental Results

D.1 More qualitative results

In the main paper, we present qualitative comparison results between our method and the state-of-the-art (SOTA) methods under a cross-driven setting on a human-like character, where our approach demonstrates outstanding performance. Considering that the other methods are primarily self-driven and trained on human characters, making them more suitable for inference in such settings, we additionally provide comparison results under a self-reconstruction setting on Tiktok and Abench. As shown in Fig. 14, when there is a appreciably difference between the reference pose and the reference image, the GAN-based LIA Wang et al. (2022) produces noticeable artifacts. Thanks to the powerful generative capabilities of diffusion models, diffusion-based models generate higher-quality results. However, MusePose Tong et al. (2024) and MimicMotion Zhang et al. (2024) generate awkward arms and blurry hands, respectively, while ControlNeXt Peng et al. (2024) synthesizes incorrect movements. Only Unianimate Wang et al. (2024b) can obtain results comparable to ours. Yet, when the reference image is a non-human character, even in a self-driven setting with the same training strategy as Unianimate, their results still show distorted heads. Fig. 15 provides results of

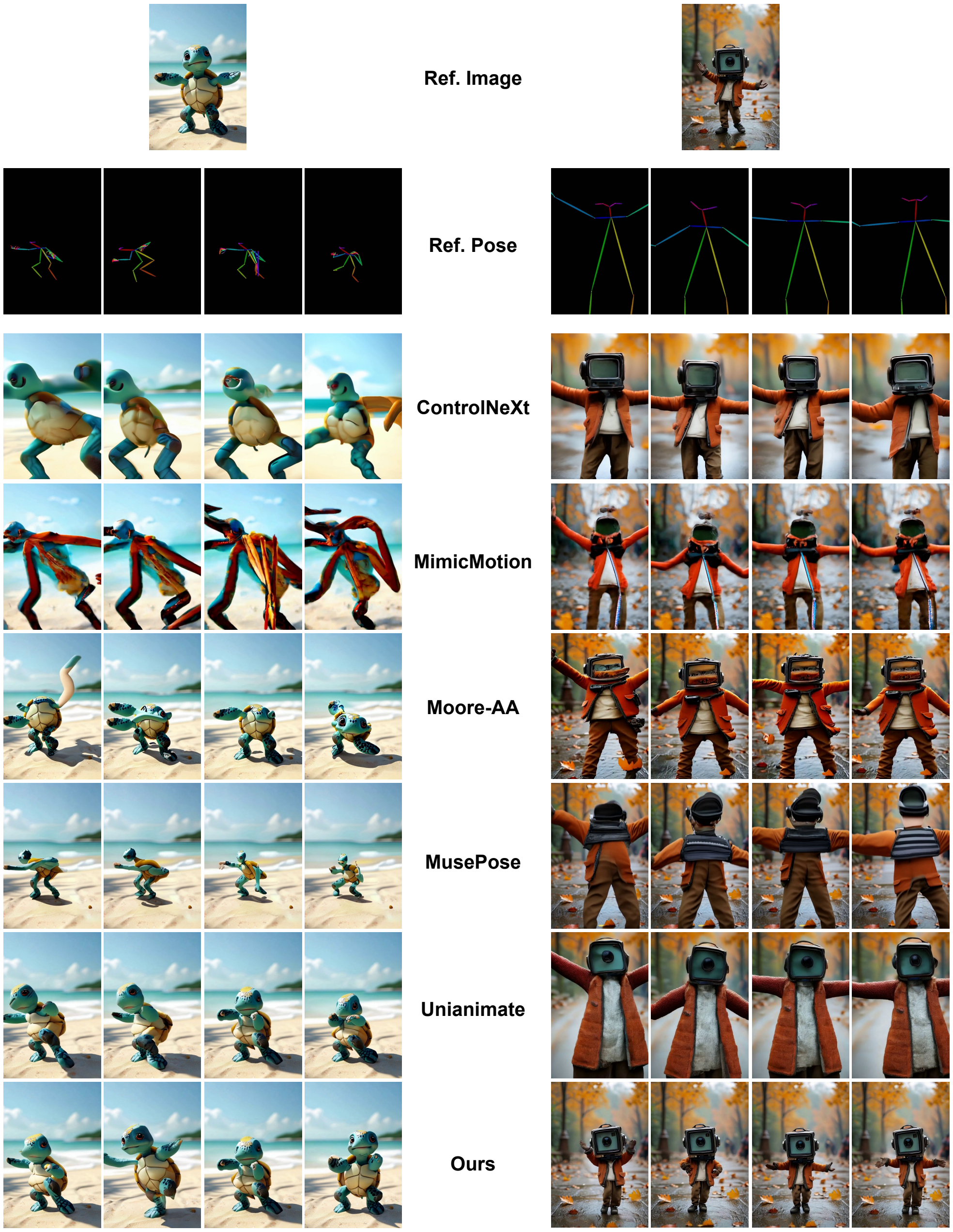


Figure 11: Visualization of cases in the user study

more comparison results, including MRAA Siarohin et al. (2021a), MagicAnimate Xu et al. (2023a) and Moore-AnimateAnyone Corporation (2024). In contrast, our method consistently generates satisfactory results for both human and anthropomorphic characters, demonstrating its ability to drive X character and highlighting its strong generalization and robustness.

D.2 More quantitative results

Tab. 7 and Tab. 8 presents the quantitative results on TikTok Jafarian & Park (2021) and Fashion Zablotskaia et al. (2019a) dataset, which suggests the superiority of methods over the comparison SOTA methods. Only Unianimate achieves comparable performance; however, our method is applicable to a wider range of characters and various unaligned pose inputs, as demonstrated in

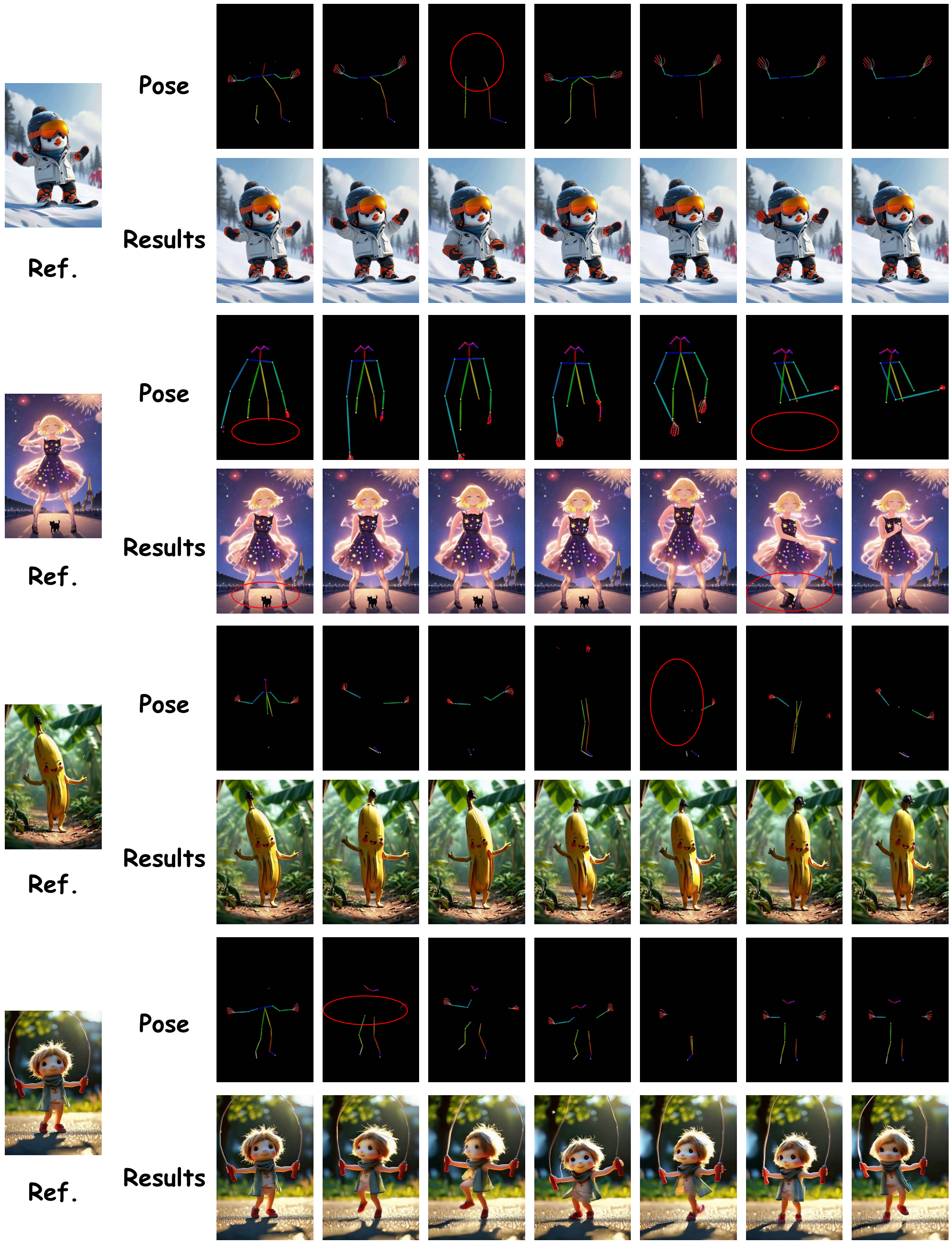


Figure 12: Visualization of the robustness of Animate-X.

Tab. 1. This addresses the main issue that this paper aims to solve: developing a universal character image animation model.

D.3 Robustness

Our method demonstrates robustness to both input X character and pose variations. On the one hand, as shown in Fig. 1, our approach successfully handles inputs from diverse subjects, including characters vastly different from humans, such as those without limbs, as well as game characters or those generated by other models. Despite these variations, our method consistently produces satisfactory results without crashing, showcasing its robustness to the input reference images. On

Method	PSNR* ↑	SSIM ↑	L1 ↑	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
w/o IPI	13.30	0.433	1.35E-04	0.454	32.56	64.31	893.31
w/o LQ	13.48	0.445	1.76E-04	0.454	28.24	42.74	754.37
w/o DQ	13.39	0.445	**1.01E-04**	0.456	30.33	62.34	913.33
PA	13.25	0.436	1.11E-04	0.464	27.63	46.54	785.36
KV_Q	13.34	0.443	1.17E-04	0.459	26.75	42.14	785.69
w/o EPI	12.63	0.403	1.80E-04	0.509	42.17	58.17	948.25
w/o Add	13.28	0.442	1.56E-04	0.459	34.24	52.94	804.37
w/o Drop	13.36	0.441	1.94E-04	0.458	26.65	44.55	764.52
w/o BS	13.27	0.443	1.08E-04	0.461	29.60	56.56	850.17
w/o NF	13.41	0.446	1.82E-04	0.455	29.21	56.48	878.11
w/o AL	13.04	0.429	1.04E-04	0.474	27.17	33.97	765.69
w/o Rescalings	13.23	0.438	1.21E-04	0.464	27.64	35.95	721.11
w/o Realign	12.27	0.433	1.17E-04	0.434	34.60	49.33	860.25
Animate-X	**13.60**	**0.452**	1.02E-04	**0.430**	**26.11**	**32.23**	**703.87**

Table 5: Quantitative results of ablation study.

the other hand, as illustrated in Fig. 12, even when the pose images exhibit body part omissions (highlighted by the red circles), our method correctly interprets the intended motion and generates coherent results for the reference images. This highlights the robustness of our approach to different pose images.

D.4 More ablation study

In the main paper, we present the results of the primary ablation experiments for IPI and EPI. In this section, we supplement those results with additional ablation experiments to further demonstrate the contribution of each individual module.

Ablation on Implicit Pose Indicator. For more detailed analysis about the structure of IPI, we set up several variants: (1) remove IPI: w/o IPI. (2) remove learnable query: w/o LQ. (3) remove DWPose query: w/o DQ. (4) set IPI and spatial Attention to Parallel: PA. (5) set CLIP features as Q and DWPose as K,V in IPI: KV_Q. The quantitative results are shown in Tab. 5. It can be seen that removing the entire IPI presents the worst performance. By modifying the IPI module, although it improves on the w/o IPI, it still falls short of the final result of Animate-X, which suggests that our current IPI structure is the most reasonable and achieves the best performance.

Since IPI is embedded in Animate-X in the form of residual connection, i.e., $x = x + \alpha IPI(x)$, we also explore the impact of the weight α of IPI on performance as illustrated in Fig. 13, as α increases from 0 to 1, all metrics show a stable improvement despite some fluctuations. The best performance is achieved when α is set to 1, so we empirically set α to 1 in the final configuration.

Ablation on Explicit Pose Indicator. We conduct more detailed ablation experiments for different pairs of pose transformations by (1) removing the entire EPI: w/o EPI; (2)&(3) removing adding and dropping parts; canceling the change of the length of (4) body and should: w/o BS; (5) neck and face: w/o NF; (6) arm and leg: w/o AL; (7) removing all rescaling process: w/o Rescalings; (8) remove another person pose alignment: w/o Realign. From the results displayed in Tab. 5, we found that each pose transformation contributes compared to w/o EPI, with aligned transformations with another person's pose contributing the most. It suggests that maintaining the overall integrity of the pose while allowing for some variations is the most important factor, and EPI also learns the overall integrity of the pose. The final result indicates that all the transformations together achieve the best performance.

To explore the effect of different probabilities λ of using pose transformation for EPI on the model performance, we set λ as 100%, 98%, 95%, 90% and 80% for the ablation experiments on two datasets. The results presented in Tab. 6 suggest that a high λ performs better on A^2Bench, i.e., it performs better when the reference image and pose image are not aligned, but harms performance on the TikTok dataset, i.e., when the reference image and pose image are strictly aligned. In contrast,

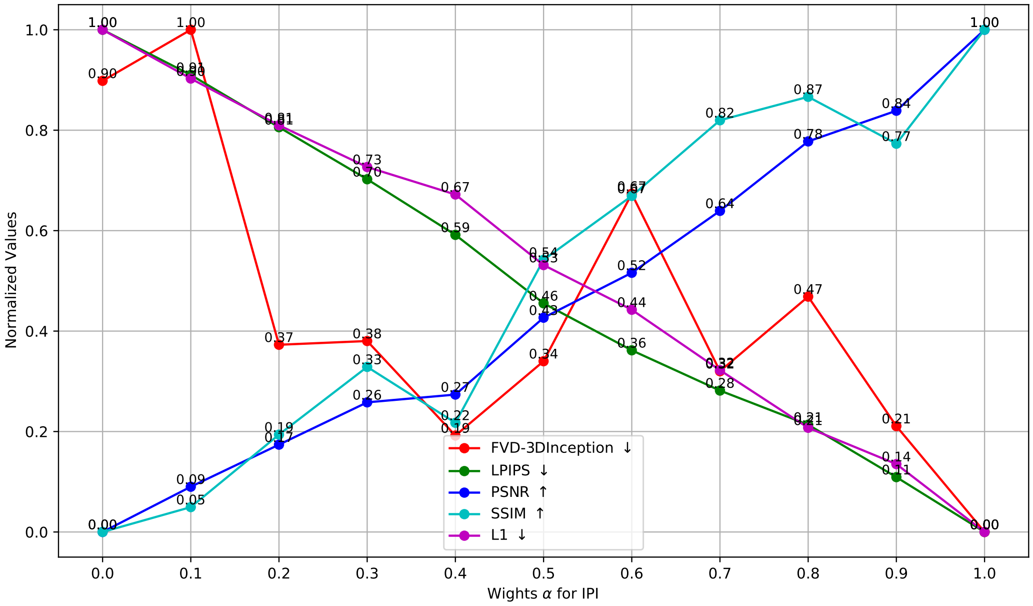


Figure 13: Ablation study on the weight α of Implicit Pose Indicator. To better visualize the impact of α on performance, we normalize all the values to the range of 0 to 1.

Method	A^2Bench				TikTok Jafarian & Park (2021)			
	SSIM↑	FID↓	FID-VID↓	FVD↓	SSIM↑	FID↓	FID-VID↓	FVD↓
100%	**0.452**	**26.11**	**32.23**	**703.87**	0.802	55.26	17.47	138.36
98%	<u>0.448</u>	<u>26.93</u>	<u>37.67</u>	<u>775.24</u>	0.797	55.81	16.28	<u>129.48</u>
95%	0.447	27.46	39.21	785.55	<u>0.804</u>	**52.72**	<u>14.61</u>	**124.92**
90%	0.444	27.15	38.03	775.38	**0.806**	<u>52.81</u>	14.82	139.01
80%	0.442	29.13	47.93	803.97	0.802	54.51	**14.42**	133.78

Table 6: Quantitative results for different probabilities of using pose transformation.

Method	L1 ↓	PSNR ↑	PSNR* ↑	SSIM ↑	LPIPS ↓	FVD ↓
FOMM Siarohin et al. (2019a) (NeurIPS19)	3.61E-04	-	17.26	0.648	0.335	405.22
MRAA Siarohin et al. (2021a) (CVPR21)	3.21E-04	-	18.14	0.672	0.296	284.82
TPS Zhao & Zhang (2022a) (CVPR22)	3.23E-04	-	18.32	0.673	0.299	306.17
DreamPose Karras et al. (2023) (ICCV23)	6.88E-04	28.11	12.82	0.511	0.442	551.02
DisCo Wang et al. (2024a) (CVPR24)	3.78E-04	29.03	16.55	0.668	0.292	292.80
MagicAnimate Xu et al. (2023a) (CVPR24)	3.13E-04	29.16	-	0.714	0.239	179.07
Animate Anyone Hu et al. (2023) (CVPR24)	-	29.56	-	0.718	0.285	171.90
Champ Zhu et al. (2024) (ECCV24)	2.94E-04	29.91	-	0.802	0.234	160.82
Unianimate Wang et al. (2024b) (ArXiv24)	**2.66E-04**	<u>30.77</u>	<u>20.58</u>	**0.811**	**0.231**	<u>148.06</u>
MusePose Tong et al. (2024) (ArXiv24)	3.86E-04	-	17.67	0.744	0.297	215.72
MimicMotion Zhang et al. (2024) (ArXiv24)	5.85E-04	-	14.44	0.601	0.414	232.95
ControlNeXt Peng et al. (2024) (ArXiv24)	6.20E-04	-	13.83	0.615	0.416	326.57
Animate-X	<u>2.70E-04</u>	**30.78**	**20.77**	<u>0.806</u>	<u>0.232</u>	**139.01**

Table 7: Quantitative comparisons with existing methods on TikTok dataset.

a relatively low λ, e.g., 90%, would be in this case perform better. It is reasonable that in the case of strict alignment, we expect the pose to provide a strictly accurate motion source, and thus need to reduce the percentage λ of pose transformation. However, in the non-strictly aligned case, we expect the pose image to provide an approximate motion trend, so we need to increase λ.

Method	PSNR ↑	PSNR* ↑	SSIM ↑	LPIPS ↓	FVD ↓
MRAA Siarohin et al. (2021a) (CVPR21)	-	-	0.749	0.212	253.6
TPS Zhao & Zhang (2022a) (CVPR22)	-	-	0.746	0.213	247.5
DPTN Zhang et al. (2022a) (CVPR22)	-	24.00	0.907	0.060	215.1
NTED Ren et al. (2022) (CVPR22)	-	22.03	0.890	0.073	278.9
PIDM Bhunia et al. (2023) (CVPR23)	-	-	0.713	0.288	1197.4
DBMM Yu et al. (2023) (ICCV23)	-	24.07	0.918	0.048	168.3
DreamPose Karras et al. (2023) (ICCV23)	-	-	0.885	0.068	238.7
DreamPose w/o Finetune Karras et al. (2023) (ICCV23)	34.75	-	0.879	0.111	279.6
Animate Anyone Hu et al. (2023) (CVPR24)	**38.49**	-	0.931	0.044	81.6
Unianimate Wang et al. (2024b) (ArXiv24)	37.92	27.56	**0.940**	**0.031**	**68.1**
MimicMotion Zhang et al. (2024) (ArXiv24)	-	27.06	0.928	0.036	118.48
`Animate-X`	36.73	**27.78**	**0.940**	**0.030**	79.4

Table 8: Quantitative comparisons with existing methods on the Fashion dataset. "w/o Finetune" represents the method without additional finetuning on the fashion dataset.

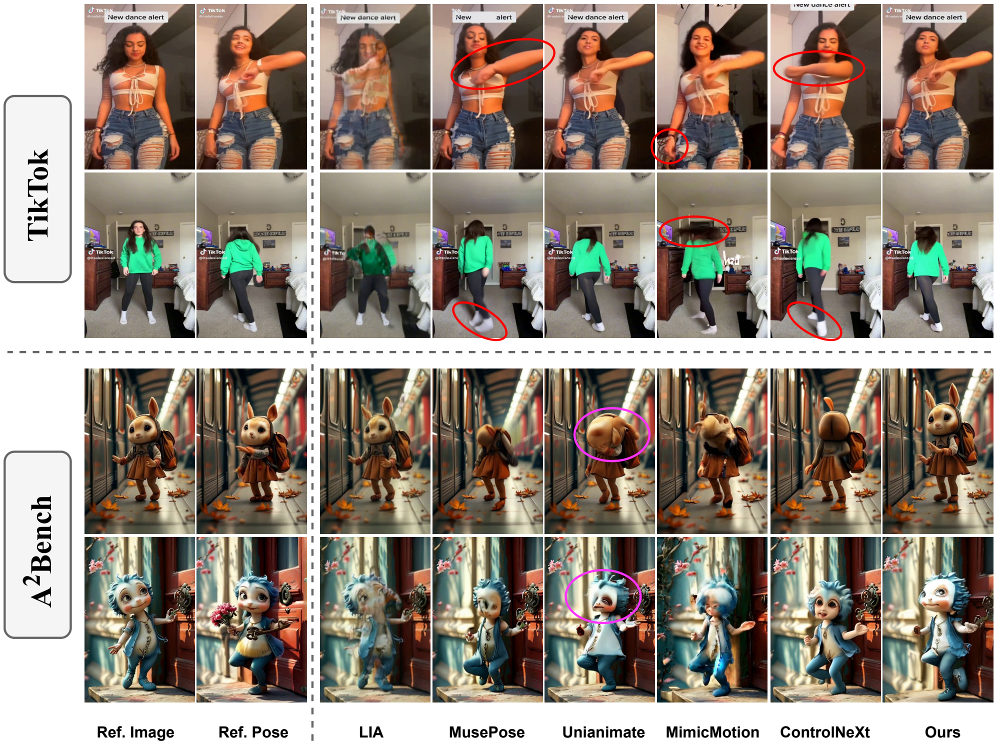


Figure 14: Visualization comparison on TikTok dataset and A^2Bench.

E Discussion

E.1 Limitation and Future Work

Although our method has made remarkable progress, it still has certain limitations. Firstly, its ability to model hands and faces remains insufficient, a limitation commonly faced by most current generative models. While our **IPI** leverages CLIP features to extract implicit information such as motion patterns from the driving video, mitigating the reliance on potentially inaccurate hand and face detection by DWPose, there is still a gap between our results and the desired realism. Secondly, due to the multiple denoising steps in the diffusion process, even though we replace the transformer with a more efficient Mamba model for temporal modeling, `Animate-X` still cannot achieve real-

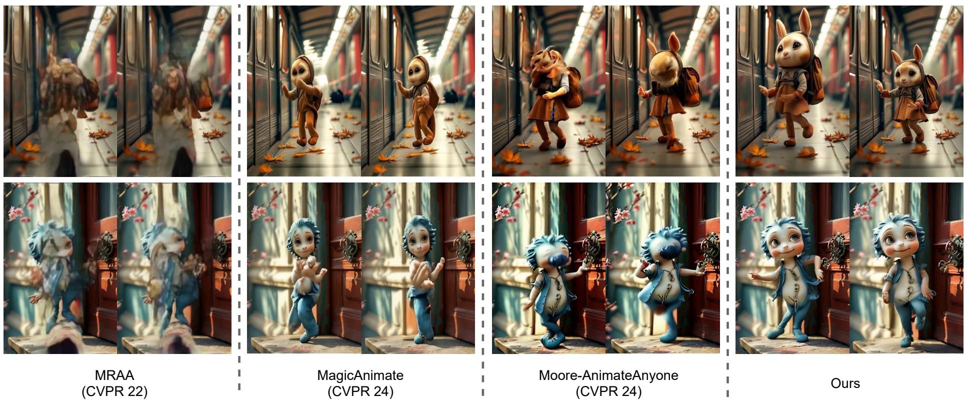


Figure 15: Comparison with more SOTAs on A^2Bench.

time animation. In future work, we aim to address these two limitations. Additionally, we will focus on studying interactions between the character and the surrounding environment, such as the background, as a key task to resolve.

E.2 Ethical Considerations

Our approach focuses on generating high-quality character animation videos, which can be applied in diverse fields such as gaming, virtual reality, and cinematic production. By providing body movement, our method enables animators to create more lifelike and dynamic characters. However, the potential misuse of this technology, particularly in creating misleading or harmful content on digital platforms, is a concern. While greatly progress has been made in detecting manipulated animations Boulkenafet et al. (2015); Wang et al. (2020); Yu et al. (2020), challenges remain in accurately identifying increasingly sophisticated forgeries. We believe that our animation results can contribute to the development of better detection techniques, ensuring the responsible use of animation technology across different domains.

Rebuttal for Animate-X

ICLR 2025,

Manuscript ID: 37

Reviewer: #1 aHUH

We sincerely thank **Reviewer #1 aHUH** for acknowledging the *"notable improvements of Animate-X"* and the *"comprehensive experiments and ablation studies presented in our work"*. Below, we have addressed each questions in detail and hope to clarify any concerns.

> **Comment #1**
>
> *"No video samples from A^2Bench are provided; only selected frames are shown in the paper. Given that the generated videos still struggle with maintaining strict logic and good spatial and temporal consistency, I question the rationale for using T2I + I2V to generate benchmark videos."*

Response: Thanks. We have provided video samples of A^2Bench in the **updated *Supplementary Materials*** (.zip/for_reviewer_aHUH/xxx.mp4). We kindly invite the reviewer to check these videos. Below, we address the reviewer's concerns regarding *"strict logic"* and *"good spatial and temporal consistency"* using T2I + I2V:

1. **Strict logic:** The choice to use T2I models stems from a clear need: current T2V models often struggle with imaginative and logically complex inputs, such as "*personified refrigerators*" or "*human-like bees*". T2I models offer strict logic and imagination in these scenarios, allowing to generate reasonable cartoon characters as the ground-truth. To prove this point, as shown in Table I, we assessed the semantic accuracy of A^2Bench using CLIP scores, which are commonly used to evaluate whether the semantic logic of images and text is strictly aligned (*i.e.*, Does the generated "*human-like bee*" maintain the visual essence of a bee while seamlessly incorporating human-like features, such as hands and feet?). For comparison, we also evaluate the publicly available TikTok and Fashion datasets using the same metric. These experimental results demonstrate that A^2Bench achieves the highest level of strict logical alignment. **Furthermore**, we input the images from A^2Bench into a multimodal large language model (MLLM) with logical reasoning, such as QWen Bai et al. (2023), to conduct a logical analysis of the visual outputs generated by the T2I model. The results, shown in Figure 1, reveal that the image descriptions answered by the MLLM closely aligns with our input prompts, which verifies again that the data in A^2Bench maintains strict logic.

2. **Good spatial and temporal consistency:** We have added several metrics from VBench Huang et al. (2024), such as *Background Consistency*, *Motion Smoothness*, *Aesthetic Quality* and *Image Quality*, to assess the spatial and temporal consistency of the videos in A^2Bench. As shown in Table I, A^2Bench outperforms TikTok dataset in all aspects and achieve comparable scores to Fashion dataset, where both TikTok and Fashion are collected from real-world scenarios. It demonstrates that the video generated by our method has the same level of spatial and temporal consistency as the real videos.

Table I: Quantitative results of different benchmarks. The best and second results for each column are **bold** and underlined, respectively.

Benchmark	CLIP Score	Background Consistency	Motion Smoothness	Aesthetic Quality	Image Quality
TikTok	26.92	94.10 %	99.05 %	55.14 %	62.54 %
Fashion	20.18	**98.25** %	**99.45** %	49.62 %	49.96 %
A^{2}Bench	**33.24**	96.66 %	99.39 %	**69.86** %	**69.32** %

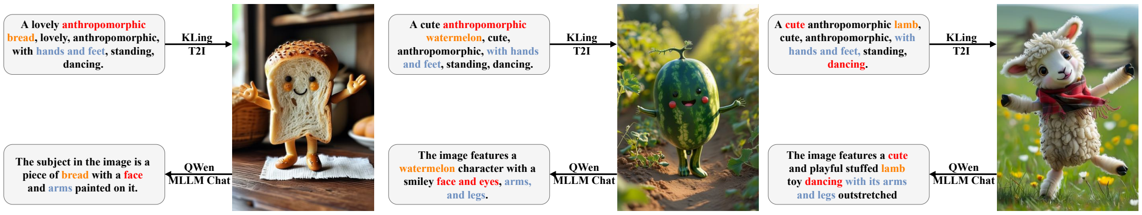

A lovely anthropomorphic bread, lovely, anthropomorphic, with hands and feet, standing, dancing.
KLing T2I
The subject in the image is a piece of bread with a face and arms painted on it.
QWen MLLM Chat
A cute anthropomorphic watermelon, cute, anthropomorphic, with hands and feet, standing, dancing.
KLing T2I
The image features a watermelon character with a smiley face and eyes, arms, and legs.
QWen MLLM Chat
A cute anthropomorphic lamb, cute, anthropomorphic, with hands and feet, standing, dancing.
KLing T2I
The image features a cute and playful stuffed lamb toy dancing with its arms and legs outstretched
QWen MLLM Chat

Figure 1: Prompts, generated images by T2I in A^{2}Bench, and logical answers from QWen.

In summary, to our best knowledge, T2I+I2V is the reasonable and effective solution currently available for automating the production of videos with anthropomorphic cartoon characters. Specifically, the T2I model can understand the prompt and generate well-aligned high-quality images with strict logic, while the I2V model can preserve the identity of the characters in the image and generate videos with good spatial and temporal consistency. Moreover, the T2I step allows human artists to check and make manual modification to the cartoon characters if necessary before generating the videos.

Comment #2

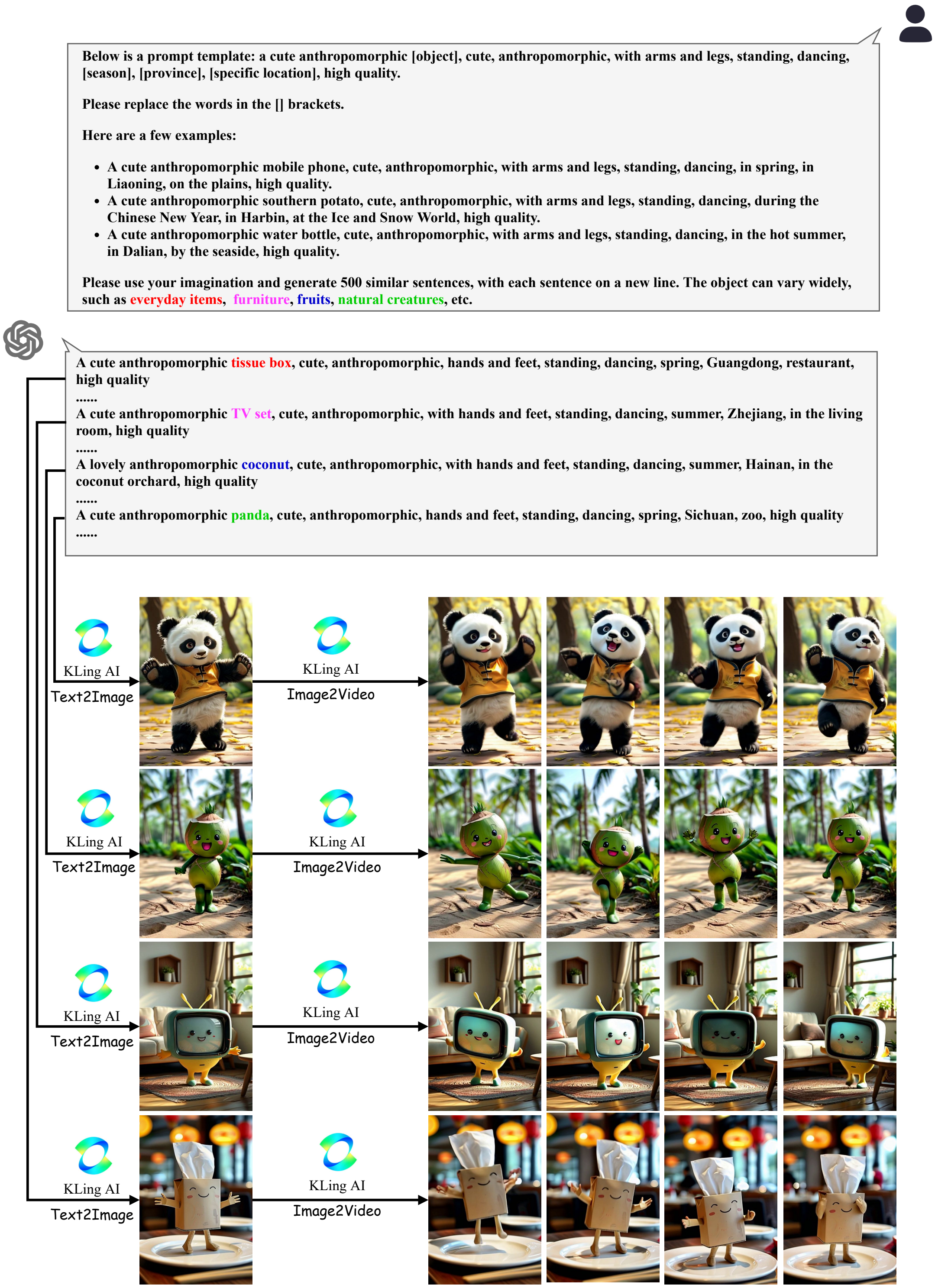
> *"Additionally, the benchmark lacks detailed information, such as video length and frame rate (Answer 2.1). Were any additional motion prompts used to generate videos from images (Answer 2.2)? If so, what is their diversity and complexity (Answer 2.3)?"*

Response:

Answer 2.1. Each video in A^{2}Bench is 5 seconds long, with a frame rate of 30-FPS and a resolution of 832×1216.

Answer 2.2. When generating videos from images, we supplement the prompt in Figure 2 (*i.e.*, Figure 10 in the original submission) regarding spatial relationships, physical logic, and temporal consistency, such as: "*reasonable movement*", "*varied dance*", and "*continuous dance*", , which further ensure strict logic and good spatial and temporal consistency.

Answer 2.3. To guarantee diversity and complexity, for each prompt, we first generate 4 images using 4 different random seeds. Then, for each image, we generate 4 videos. Thus, we ensure both diversity and complexity in the final results. Moreover, as suggested by **Reviewer #3 feUz**, we add style trigger words such as "*Watercolor Painting*", "*Cyberpunk Style*", "*Van Gogh*", "*Ukiyo-E*", "*Pixel Art*" and so on. The results are presented in Figure 3, which further enhances the diversity and complexity of A^{2}Bench.



Below is a prompt template: a cute anthropomorphic [object], cute, anthropomorphic, with arms and legs, standing, dancing, [season], [province], [specific location], high quality.

Please replace the words in the [] brackets.

Here are a few examples:

- A cute anthropomorphic mobile phone, cute, anthropomorphic, with arms and legs, standing, dancing, in spring, in Liaoning, on the plains, high quality.
- A cute anthropomorphic southern potato, cute, anthropomorphic, with arms and legs, standing, dancing, during the Chinese New Year, in Harbin, at the Ice and Snow World, high quality.
- A cute anthropomorphic water bottle, cute, anthropomorphic, with arms and legs, standing, dancing, in the hot summer, in Dalian, by the seaside, high quality.

Please use your imagination and generate 500 similar sentences, with each sentence on a new line. The object can vary widely, such as everyday items, furniture, fruits, natural creatures, etc.

A cute anthropomorphic tissue box, cute, anthropomorphic, hands and feet, standing, dancing, spring, Guangdong, restaurant, high quality
......
A cute anthropomorphic TV set, cute, anthropomorphic, with hands and feet, standing, dancing, summer, Zhejiang, in the living room, high quality
......
A lovely anthropomorphic coconut, cute, anthropomorphic, with hands and feet, standing, dancing, summer, Hainan, in the coconut orchard, high quality
......
A cute anthropomorphic panda, cute, anthropomorphic, hands and feet, standing, dancing, spring, Sichuan, zoo, high quality
......

Figure 2: Detailed pipeline for building A^{2}Bench based on large-scale pretrained models, including OpenChatGPT 4o and KLing AI.

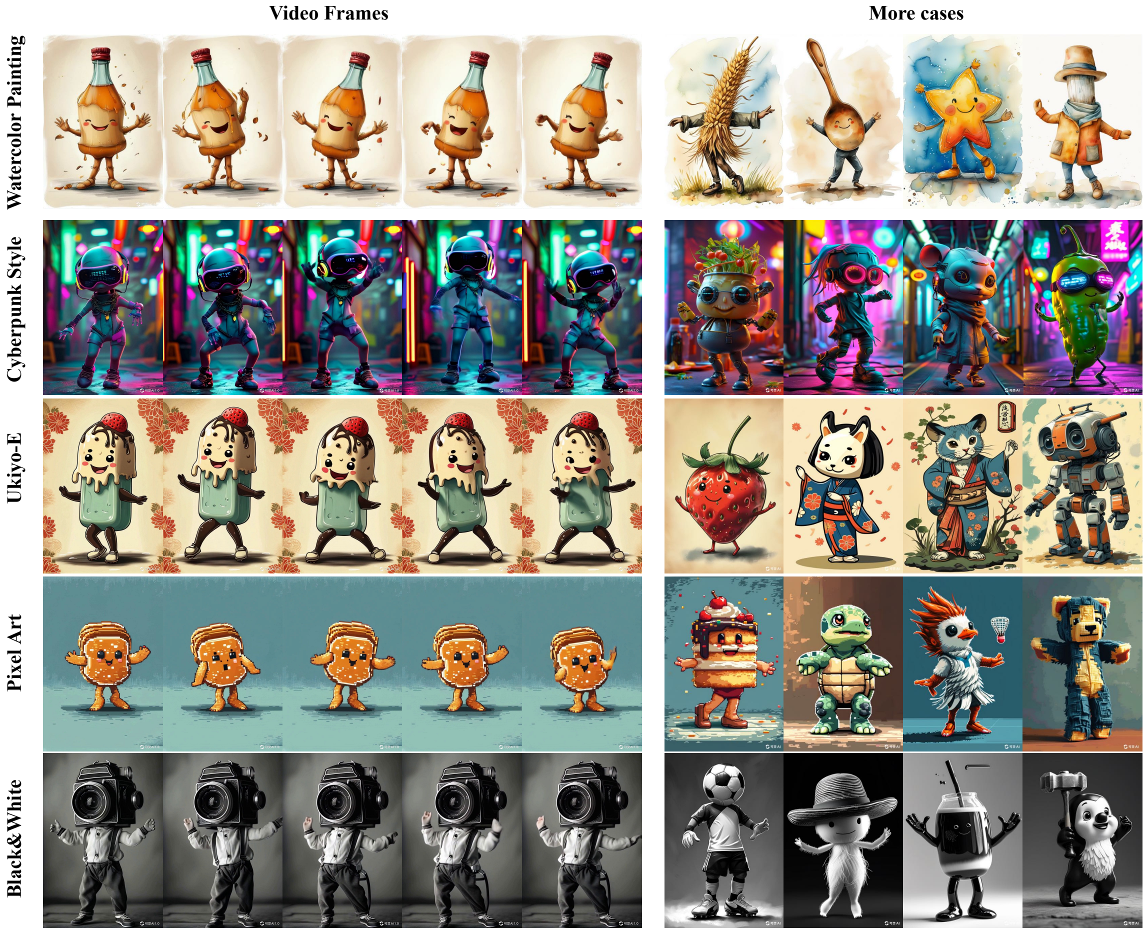


Figure 3: More styles in A^2Bench.

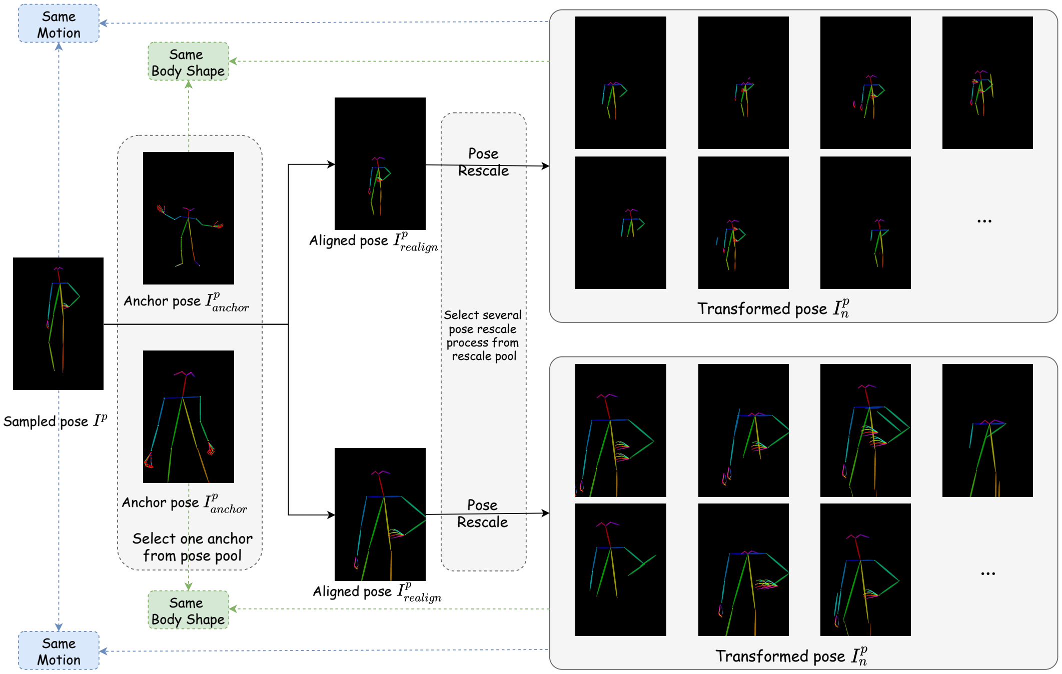
Comment #3

> *"The necessity of a pose pool and the selection of an anchor pose image need clarification (Answer 3.3). What operations are involved in the "align" process (Answer 3.1), specifically regarding translation and rescaling (Answer 3.2)? Why not use random translation and rescaling instead of relying on an anchor pose image (Answer 3.3)?"*

Response:

Answer 3.1. As shown in the left half of Figure 4 (*i.e.*, Figure 8 in the original submission), the operations in the "align" process are as follws:

- **Step1:** Given a driving pose I^p, we randomly select an anchor pose I^p_{anchor} from the pose pool (two examples are shown in Figure 4).
- **Step2:** We then calculate the proportion of each body part between these two poses. For example, the shoulder length of I^p_{anchor} divided by the shoulder length of I^p might be 0.45, and the leg length of I^p_{anchor} divided by the leg length of I^p might be 0.53, and so on.
- **Step3:** We multiply each body part of the driven pose (*i.e.*, I^p) by the corresponding ratio (*e.g.*, 0.45, 0.53, *etc.*) to obtain the aligned pose (*i.e.*, I^p_n), as shown in Figure 4.



Same Motion
Same Body Shape
Anchor pose I^p_{anchor}
Aligned pose $I^p_{realign}$
Pose Rescale
Transformed pose I^p_n
...
Sampled pose I^p
Select several pose rescale process from rescale pool
Anchor pose I^p_{anchor}
Select one anchor from pose pool
Aligned pose $I^p_{realign}$
Pose Rescale
...
Same Body Shape
Same Motion
Transformed pose I^p_n

Figure 4: The pipeline of EPI.

Answer 3.2. As shown in the right half of Figure 4:

- **Step4:** ("*rescaling*") Then we define a set of keypoint rescaling operations, including modifying the length of the body, legs, arms, neck, and shoulders, altering face size, adding or removing specific body parts, *etc*. These operations are stored in a rescale pool.
- **Step5:** ("*translation*") We apply the selected rescaling operations on the aligned pose $I^p_{realign}$ to obtain the final transformed poses I^p_n.

Answer 3.3. As shown in Figure 5, the reason of "*not using random translation and rescaling instead of relying on an anchor pose image*" is that random translation and rescaling disrupt the motion guidance originally conveyed by the driven pose image. This issue makes the animation model miss the accurate driving guidance, which diminishes its ability to generate proper animations. In contrast, using anchor pose images maintain harmonious proportions for each body part and preserve the consistency of all motion details.

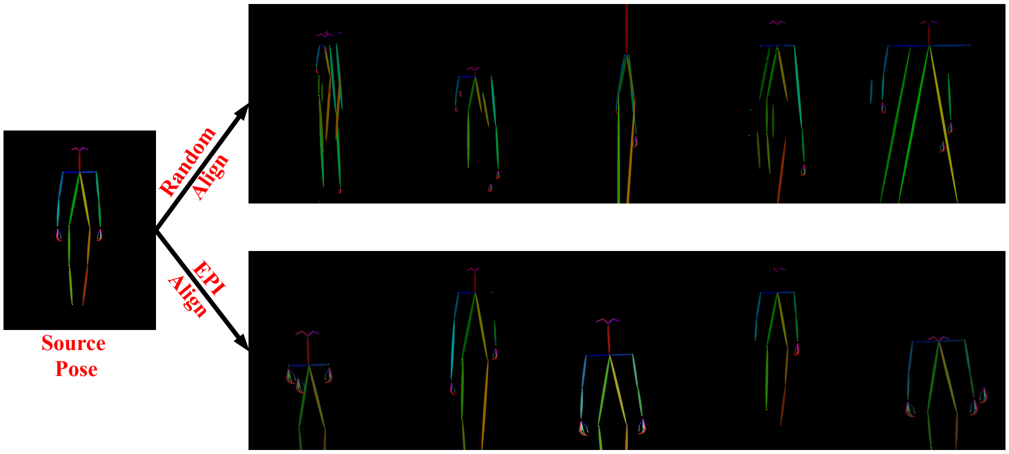


Figure 5: The different results between random alignment and EPI alignment.

To prove this point, we **have re-trained** our model using pose images which obtained by **random** translation and rescaling. Results are presented in Figure 6. The results indicate that the baseline achieves only a marginal improvement (*i.e.*, the content of the reference image only appears in the initial frames, while illogical human characteristics persist throughout), while our approach delivers satisfactory performance (*i.e.*, it perfectly preserves the cartoon ID of the reference image while adding dynamic motion).

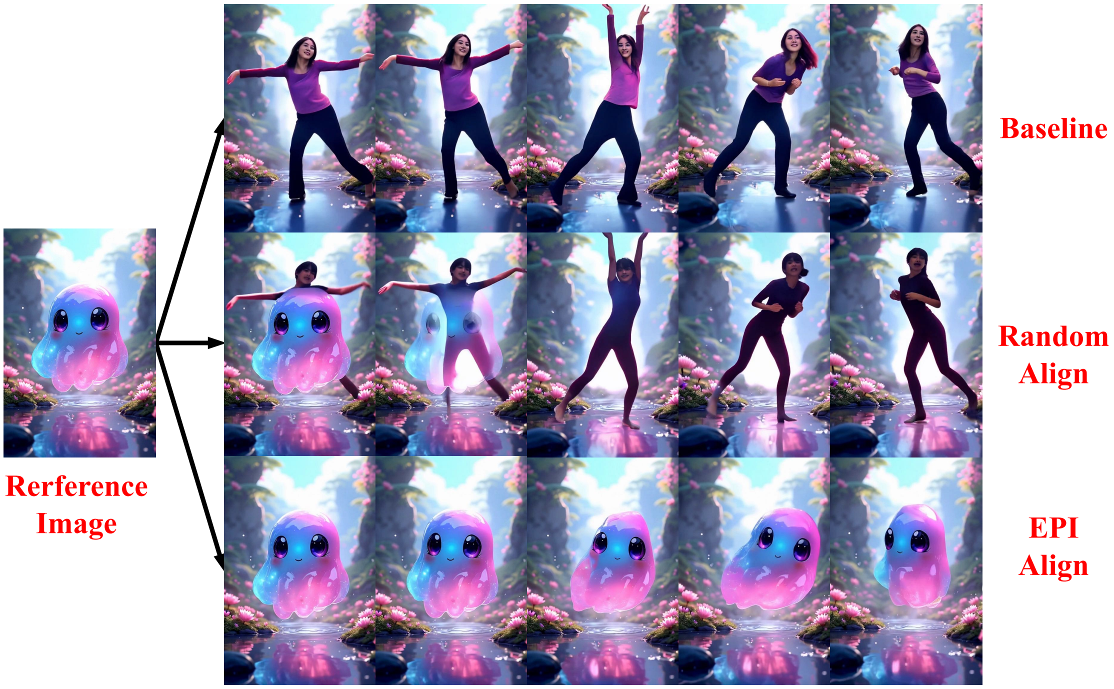


Figure 6: The results of different alignments.

Finally, as shown in Table II, quantitative results of ablation study indicate that the "realign" operation plays a crucial role in improving performance, which justifies both the pose pool and the selection of an anchor pose for EPI alignment.

Table II: Quantitative results of ablation study.

Method	PSNR* ↑	SSIM ↑	L1 ↓	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
w/o Add in EPI	13.28	0.442	1.56E-04	0.459	34.24	52.94	804.37
w/o Drop in EPI	13.36	0.441	1.94E-04	0.458	26.65	44.55	764.52
w/o BS in EPI	13.27	0.443	1.08E-04	0.461	29.60	56.56	850.17
w/o NF in EPI	13.41	0.446	1.82E-04	0.455	29.21	56.48	878.11
w/o AL in EPI	13.04	0.429	1.04E-04	0.474	27.17	33.97	765.69
w/o Rescalings in EPI	13.23	0.438	1.21E-04	0.464	27.64	35.95	721.11
w/o Realign in EPI	12.27	0.433	1.17E-04	0.434	34.60	49.33	860.25
with complete EPI	**13.60**	**0.452**	**1.02E-04**	**0.430**	**26.11**	**32.23**	**703.87**

Comment #4

"The effectiveness of the Implicit Pose Indicator (IPI) is also in question. The motivation for the IPI is that sparse keypoints lack image-level details, while IPI aims to retrieve richer information. However, Tables 7 and 8 indicate that Animate-X achieves comparable performance to Animate-Anyone and UniAnimate on human videos. This suggests that the IPI does not provide any benefits for human animation."

Response: The effectiveness of the Implicit Pose Indicator (IPI) have been demonstrated through the quantitative results in Table III (*i.e.*, Figure 4 in the original submission) and the qualitative analysis in Figure 7 in the original submission.

Table III: Quantitative results of ablation study on IPI.

Method	PSNR* ↑	SSIM ↑	L1 ↓	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
w/o IPI	13.30	0.433	1.35E-04	0.454	32.56	64.31	893.31
w/o LQ	13.48	0.445	1.76E-04	0.454	28.24	42.74	754.37
w/o DQ	13.39	0.445	**1.01E-04**	0.456	30.33	62.34	913.33
Animate-X	**13.60**	**0.452**	1.02E-04	**0.430**	**26.11**	**32.23**	**703.87**

1) The primary purpose of `Animate-X` is to animate universal characters, especially anthropomorphic figures in cartoons and games. Human animation is **NOT** the primary focus of this work as it is a small subset of 'X'. Table 7&8 verify that even for human figures, `Animate-X`'s performance is on par to latest works focusing on animating human figures, which actually well indicates the generalization capability of `Animate-X`;

2) IPI does retrieve richer information from driven video that is critical to some hard cases that lack of enough details in anthropomorphic figures, e.g., . It is reasonable that its contribution is marginal for those simple human-driven animations that the details are already sufficient to capture human motion, which are not the cases that IPI is designed to address. Therefore, for datasets like TiTok with exclusive human data only, we just want to show II also improves a bit and `Animate-X` is well backward compatible for human figures;

3) Anthropomorphic characters are arguably more desirable in gaming film and short videos. Therefore we introduce a novel benchmark beyond human, as detailed in Section 3.4. We kindly suggest the reviewer to watch the MP4 videos in the updated supplementary materials.

Reviewer: #2 mbHE

We sincerely thank **Reviewer #2 mbHE** for acknowledging the *"introduced* `A²Bench`*, the qualitative and quantitative experiments presented in our work"*. Below, we have addressed each question in detail and hope to clarify any concerns.

Comment #1

> *"Some parts of the writing can be quite confusing, words and sentences are bad orgnized. For example, in P5 L260, what exactly is in the pose pool (Answer 1.1)? And how is it aligned with the reference? (Answer 1.2)"*

Response:

Answer 1.1. The pose pool mentioned in P5 L260 consists of all the unenhanced pose images extracted from our training dataset. Specifically, we use DWPose as the pose extractor to obtain skeleton images with a black background from the training videos.

Answer 1.2. We have provided a detailed explanation of the pose pool and alignment process in Appendix A and Figure 4. The alignment process can be organized into the following steps:

- **Step1:** Given a driving pose I^p, we randomly select an anchor pose I^p_{anchor} from the pose pool (two examples are shown in Figure 4).
- **Step2:** We then calculate the proportion of each body part between these two poses. For example, the shoulder length of I^p_{anchor} divided by the shoulder length of I^p might be 0.45, and the leg length of I^p_{anchor} divided by the leg length of I^p might be 0.53, and so on.
- **Step3:** We multiply each body part of the driven pose (*i.e.*, I^p) by the corresponding ratio (*e.g.*, 0.45, 0.53, *etc.*) to obtain the aligned pose (*i.e.*, I^p_n), as shown in Figure 4.

Comment #2

> *" The dataset includes 9,000 independently collected videos. Could you analyze these videos (Answer 2.1), and did other baselines use the same data for training (Answer 2.2)? If not, could this lead to an unfair comparison (Answer 2.3)?"*

Response:

Thanks for your valuable comments. First, we would like to clarify that we have demonstrated the improvements in our approach stem from the IPI and EPI modules through the extensive and fair ablation experiments. Next, we will address each question in detail.

Answer 2.1. Following the commonly used public human animation TikTok datasets which consists of videos downloaded from TikTok, we additionally collect 9,000 TikTok-like videos. The distribution of the additional data is similar to the TikTok dataset, primarily consisting of human dance videos.

Answer 2.2. We notice that other baselines have also used their own collected data for model training. For example, UniAnimate Wang et al. (2024) uses 10,000 internal videos. Despite using more data than we did, `Animate-X` still improves the performance substantially, suggesting that these gains stem from the design of our modules rather than the data.

Answer 2.3. Data is also the essential contribution of each respective work. The use of independently collected videos, including in our work, is transparently explained in the papers and has become a well-established convention in prior researches. **To address potential concerns**, we have trained our `Animate-X` solely on the public TikTok and Fashion benchmarks, **without incorporating any extra videos**. We have conducted the same experiments as presented in Table 1, and reported results marked by # in Table IV. As shown in Table IV, our method still outperforms other approaches, which further demonstrates that the improvements in `Animate-X` are driven by the IPI and EPI modules, rather than the use of additional training data.

Table IV: Quantitative comparisons with SOTAs on A^2Bench.

Method	PSNR* ↑	SSIM ↑	L1 ↑	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
Moore-AnimateAnyone	9.86	0.299	1.58E-04	0.626	50.97	75.11	1367.84
MimicMotion (ArXiv24)	10.18	0.318	1.51E-04	0.622	122.92	129.40	2250.13
ControlNeXt (ArXiv24)	10.88	0.379	1.38E-04	0.572	68.15	81.05	1652.09
MusePose (ArXiv24)	11.05	0.397	1.27E-04	0.549	100.91	114.15	1760.46
Unianimate (ArXiv24)	11.82	0.398	1.24E-04	0.532	48.47	61.03	1156.36
Animate-X#	13.46	0.441	1.19E-04	0.468	37.76	40.19	933.43
Animate-X	**13.60**	**0.452**	**1.02E-04**	**0.430**	**26.11**	**32.23**	**703.87**

Comment #3

> *"The authors first identify the weaknesses of previous methods as a conflict between identity preservation and pose control. They further expand on this point by highlighting two specific limitations: the lack of image-level details in sole pose skeletons and pose alignment within the self-driven reconstruction training strategy. However, while the authors clearly state that differences in appearance between characters and humans can negatively impact animation, learning image-level details seems to contradict their viewpoint "sole pose skeletons lack image-level details", making this contribution appear more like a forced addition."*

Response: We disagree with this comment. "*sole pose skeletons lack image-level details*" and "*learning image-level details*" are not contradictory but rather represent a cause-and-effect relationship. As shown in Figure 7, previous methods extract only pose skeletons from original driving videos. The process can be represented as

video → pose skeletons → results.

These pose skeletons lack image-level motion-related details, *i.e.*, motion-induced deformations (*e.g.*, body part overlap and occlusion). These details play a crucial role in enhancing character animation, since personification cartoon characters have more unpredictable movement patterns compared to humans. Therefore, we design the IPI module specifically to extract these image-level motion-related details. The process can be represented as:

- **Step1:** (*as same as the previous method*) video → pose images;
- **Step2:** video → IPI → image-level motion-related features;
- **Step3:** pose images + image-level motion-related features → results.

Moreover, the introduction of our IPI module is a core contribution of this paper which is not "*a forced addition*". In previous approaches, temporal information in driven videos was derived solely from multi-frame pose skeletons, often set against pure black backgrounds. The original RGB videos were discarded during the training process. While this method works well for human animation, where carefully designed pose skeletons align perfectly with human joints, it falls short for anthropomorphic characters whose skeletons differ significantly from humans. Thus, pose skeletons alone can NOT provide sufficient driving guidance, as they lack the motion-related details found only in the original driving video. This is where our IPI module makes a difference, extracting these richer details from the original video to improve the generalization of motion representation modeling.

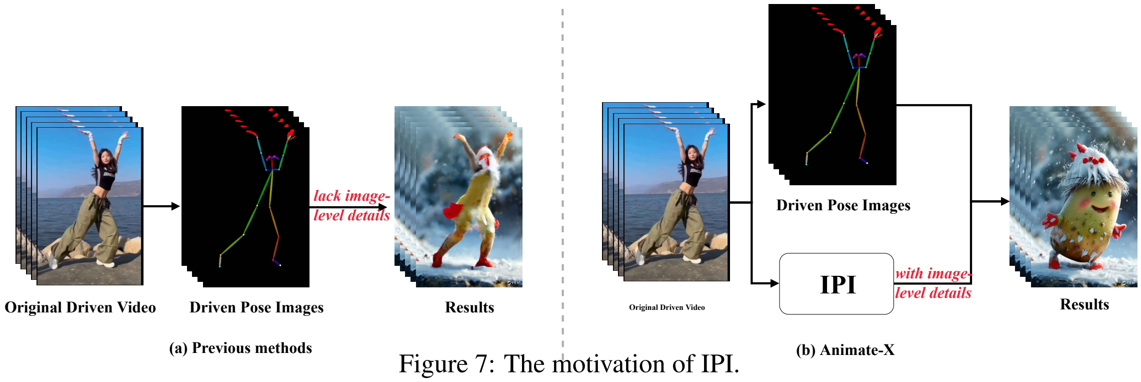


Figure 7: The motivation of IPI.

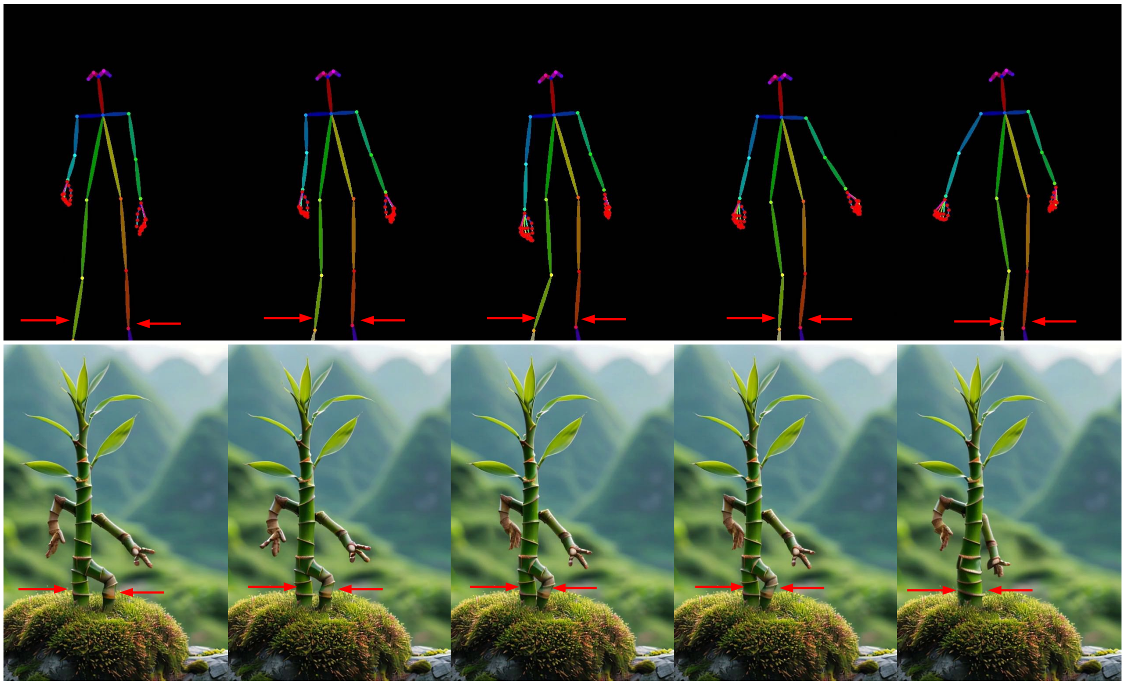


Figure 8: More frames in Figure 7 of the original submission.

Comment #4

> *"Additionally, the visualization in Figure 7 provided by the authors also supports w3. The inclusion or exclusion of the IPI appears to have minimal impact on the motion of the Ref image, and with IPI, part of the foot in the Ref image is even missing. This raises doubts about the effectiveness of the IPI module and seems inconsistent with the authors' stated motivation. "*

Response: Thanks for the comment. The "*missing foot*" is caused by the video not being fully displayed in our submission, rather than an issue caused by our IPI module. We have added more frames of the video in Figure 8. Please see (.zip/for_reviewer_mbHE/full_frame_of_figure7.mp4) for video result. As shown Figure 8, in the initial frames, the foot is present and highly consistent with the reference image. Subsequently, the driven pose image begins to perform a leg-merging motion, with the distance between the legs gradually decreasing. To allow the anthropomorphic bamboo character to follow this motion, it also gradually merges its legs, which gives the appearance of the "*missing foot*".

Comment #5

> *“Pose augmentation has already been widely explored in existing methods, such as MimicMotion, which makes the innovation in this paper insufficient.”*

Response: The primary contribution of our work is to animate anthropomorphic figures by two new IPI and EPI modules which are not limited to the “*pose augmentation*”. Pose augmentation is a training strategy and is not exclusive to any specific method. By itself, it cannot solve the animation issue in our work. The IPI and EPI modules designed to handle figures beyond human and human pose are novel to address the specific challenges in animating anthropomorphic figures. We then provide a detailed explanation of the concept beyond “*Pose Augmentation*”. Please refer to Figure 4 or Figure 8 in appendix for an illustration of the following process:

- **Step1:** We first construct the pose pool using the DWPose extractor. The pose pool is composed of pose skeletons (*i.e.*, pose images);
- **Step2:** Given a driving pose I^p, we randomly select an anchor pose I^p_{anchor} from the pose pool.
- **Step3:** We then calculate the proportion of each body part between these two poses. For example, the shoulder length of I^p_{anchor} divided by the shoulder length of I^p might be 0.45, and the leg length of I^p_{anchor} divided by the leg length of I^p might be 0.53, and so on.
- **Step4:** We multiply each body part of the driven pose (*i.e.*, I^p) by the corresponding ratio (*e.g.*, 0.45, 0.53, *etc.*) to obtain the aligned pose (*i.e.*, I^p_n).
- **Step5:** Then we define a set of keypoint rescaling operations, including modifying the length of the body, legs, arms, neck, and shoulders, altering face size, adding or removing specific body parts, *etc*. These transformations are stored in a rescale pool.
- **Step6:** We apply the selected transformations on the aligned pose $I^p_{realign}$ to obtain the final transformed poses I^p_n.

Comment #6

> *“This paper lacks comparisons with similar methods, such as MimicMotion, which makes the experimental results less convincing. [1]MimicMotion: High-Quality Human Motion Video Generation with Confidence-aware Pose Guidance.”*

Response: We have already conducted: **(1) quantitative comparisons** with MimicMotion Zhang et al. (2024) in Tables 1, 2, 7, and 8 in the original submission; **(2) qualitative comparisons** with MimicMotion in Figure 5 and the videos in the original *Supplementary Materials*; **(3) the user study comparison** with MimicMotionin Table 3 in the original submission. For your convenience, we highlight and summary these results below.

Table V: Quantitative comparisons with MimicMotion on A^2Bench with the rescaled pose setting.

Method	PSNR* ↑	SSIM ↑	L1 ↓	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
MimicMotion (ArXiv24)	10.18	0.318	1.51E-04	0.622	122.92	129.40	2250.13
Animate-X	**13.60**	**0.452**	**1.02E-04**	**0.430**	**26.11**	**32.23**	**703.87**

Table VI: Quantitative comparisons with MimicMotion on A^2Bench in the self-driven setting.

Method	PSNR* ↑	SSIM ↑	L1 ↓	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
MimicMotion (ArXiv24)	12.66	0.407	1.07E-04	0.497	96.46	61.77	1368.83
Animate-X	**14.10**	**0.463**	**8.92E-05**	**0.425**	**31.58**	**33.15**	**849.19**

Table VII: User study results.

Method	Moore-AA	MimicMotion	ControlNeXt	MusePose	Unianimate	**Animate-X**
Identity preservation ↑	60.4%	14.8%	52.0%	31.3%	43.0%	**98.5%**
Temporal consistency ↑	19.8%	24.9%	36.9%	43.9%	81.1%	**93.4%**
Visual quality ↑	27.0%	17.2%	40.4%	40.3%	79.3%	**95.8%**

Table VIII: Quantitative comparisons with existing methods on TikTok dataset.

Method	L1 ↓	PSNR* ↑	SSIM ↑	LPIPS ↓	FVD ↓
MimicMotion (ArXiv24)	5.85E-04	14.44	0.601	0.414	232.95
Animate-X	**2.70E-04**	**20.77**	**0.806**	**0.232**	**139.01**

Table IX: Quantitative comparisons with existing methods on the Fashion dataset.

Method	PSNR* ↑	SSIM ↑	LPIPS ↓	FVD ↓
MimicMotion (ArXiv24)	27.06	0.928	0.036	118.48
Animate-X	**27.78**	**0.940**	**0.030**	**79.4**

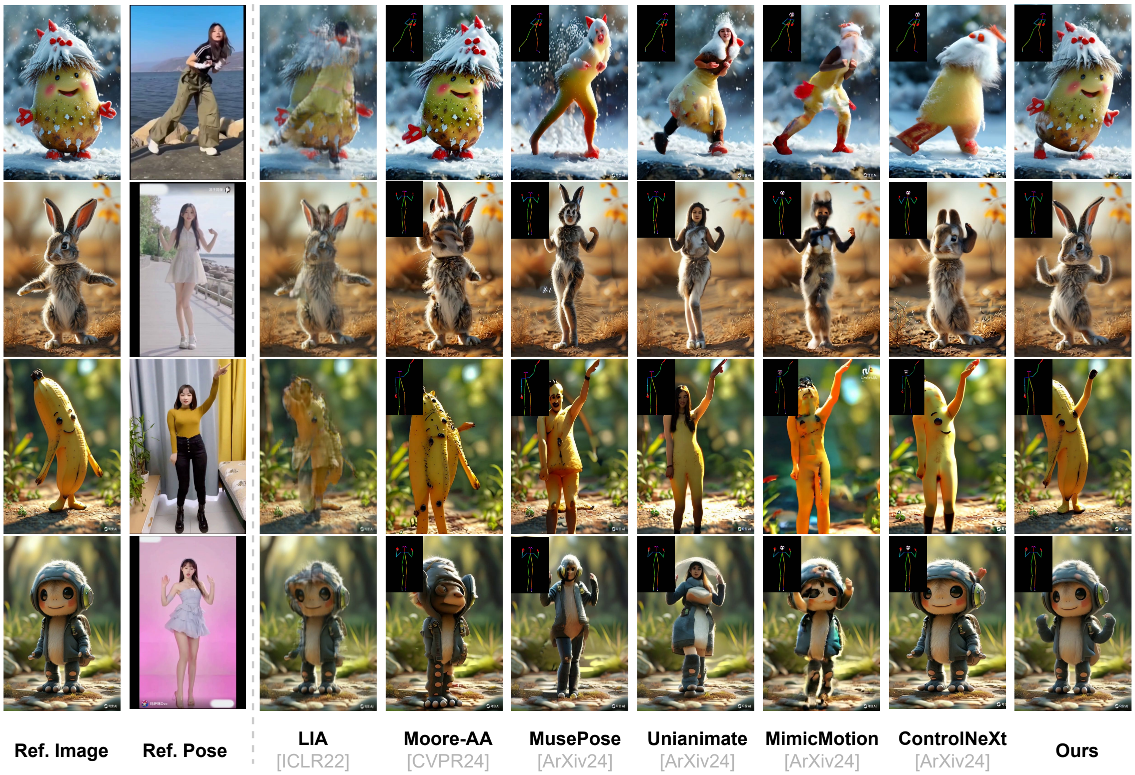


Figure 9: Qualitative comparisons with state-of-the-art methods.

Reviewer: #3 feUz

We sincerely thank **Reviewer #3 feUz** for acknowledging "*the new method, benchmark and animation results presented in our work*". Below, we have addressed each question in detail and hope to clarify any concerns.

Comment #1

"The paper lacks a detailed analysis of the construction of the augmentation pool, making it difficult to reproduce the method. Could the authors provide more details on the construction of the pose pool and alignment pool, such as the pool sizes and how poses are selected from the training set?"

Response: Thanks for your feedback. Yes, we present the detailed analysis of the construction of the augmentation pool. Please refer to Figure 4 or Figure 8 in appendix for an illustration of the following process:

- **Step1:** We first construct the pose pool using the DWPose extractor. The pose pool is composed of pose skeletons (*i.e.*, pose images);
- **Step2:** Given a driving pose I^p, we randomly select an anchor pose I^p_{anchor} from the pose pool.
- **Step3:** We then calculate the proportion of each body part between these two poses. For example, the shoulder length of I^p_{anchor} divided by the shoulder length of I^p might be 0.45, and the leg length of I^p_{anchor} divided by the leg length of I^p might be 0.53, and so on.
- **Step4:** We multiply each body part of the driven pose (*i.e.*, I^p) by the corresponding ratio (*e.g.*, 0.45, 0.53, *etc.*) to obtain the aligned pose (*i.e.*, I^p_n).
- **Step5:** Then we define a set of keypoint rescaling operations, including modifying the length of the body, legs, arms, neck, and shoulders, altering face size, adding or removing specific body parts, *etc*. These transformations are stored in a rescale pool.
- **Step6:** We apply the selected transformations on the aligned pose $I^p_{realign}$ to obtain the final transformed poses I^p_n.

Comment #2

"here is insufficient in-depth analysis of the model design, such as why the Implicit Pose Indicator (IPI) outperforms the reference network, which has more learnable parameters. Comparing the results in Table 4 and Table 1, Animate-X outperforms the baselines even without pose augmentation (EPI). Could the authors provide a deeper analysis of why the Implicit Pose Indicator (IPI), with fewer parameters, outperforms the reference network?"

Response: Sure, IPI outperforms the reference network because the latter focuses on extracting content features from reference images, while IPI focuses on motion, aiming to capture a universal motion representation. The reference network intends to capture all appearance details of the reference image. In contrast, IPI only models the motion-related image-level detais, so IPI can employ a smaller network to do the job. We provide a detailed explanation of how IPI improves the performance as follows:

1. **Reference network:** From the results using current methods using the reference network, *e.g.*, MimicMotion Zhang et al. (2024), we observe an inherent trade-off between overly precise poses and low fidelity to reference images. While the reference network attempts to address this by extracting additional appearance information from the reference image to improve fidelity through the denoising model, Figure 9 illustrates that the reference network based approach remains insufficient, as precise human poses still dominate.

2. **IPI:** To address the observed limitations, we shifted our focus from appearance information to motion as the critical factor in our work. Simple 2D pose skeletons, constructed by connecting sparse keypoints, lack the image-level details needed to capture the essence of the reference video, such as motion-induced deformations (*e.g.*, body part overlap and occlusion). This absence of image-level details causes previous methods, even those using a reference network, to produce results with consistent poses but compromised identity fidelity. To overcome this issue, we introduced the IPI module to recover these missing **motion-related** image-level details. Specifically, IPI employs a pretrained CLIP encoder to extract features from the driving image, followed by a lightweight extractor (P) to isolate the motion-related details. This approach enables IPI to outperform the reference network, which, despite having more learnable parameters, unable to capture these essential motion-related features.

As shown in Figure 9, methods utilizing reference networks, such as AnimateAnyone Hu et al. (2023), primarily focus on preserving colors from the reference image, as demonstrated by the white hat and yellow body of the potato in the first row. However, these methods cannot maintain the identity of the reference image, often generating videos that deviate from the original image, such as forcefully inserting human limbs onto potatoes. It highlights the limitation of reference networks, which prioritize color consistency over identity preservation, leading to weaker performance on quantitative metrics like SSIM, L1, and FID.

In contrast, as shown in Figure 7 in submission, even without the EPI module, `Animate-X` successfully generates a panda that retains the identity of the reference image. This leads to substantial improvements in SSIM, L1, and FID compared to baselines that rely on reference networks, even without the EPI module.

Comment #3

"What happens if the reference pose differs significantly from the candidates in the pose pool and alignment pool? The authors should provide a robustness analysis for this scenario."

Response: Thanks. We are a bit unsure whether the reviewer's question refers to the training process or the inference process, so we have analyzed both situations. We hope it helps clarify any confusion.

1. **During training:** Significant differences between the reference pose and the candidates in the pose and alignment pools can actually benefit training by enhancing the model's robustness. Different poses enable the model to understand the difference between complex reference image inputs and driven pose video inputs. For example, in the first row of Figure 1 (*i.e.*, the teaser), we use a human skeleton to drive a limb-less character. To achieve such capability, we need to simulate extreme scenarios during training. Therefore, when the reference pose differs significantly from the candidates in the pose pool and alignment pool during training, it enhances the robustness of the model.

2. **During inference:** Even when the reference pose differs significantly from the candidates in the pose and alignment pools, our model is still able to produce reasonable results, which is one of the core challenges addressed in this paper. Our pose pool and alignment pool are designed to encompass a wide range of local deformations, while the IPI module focuses on implicit motion modeling. This combination allows the model to learn generalized motion patterns from videos, rather than being constrained to specific actions. Thus, regardless of the input driver video or its corresponding pose, `Animate-X` ensures stable and reliable generation without excessive collapse.

Comment #4

"Could aligning the driving pose to a "standard" one in the pose pool further improve generation quality?"

Response: Yes. Aligning the driving pose to a "*standard*" one can further improve generation quality. This is because the "*aligning*" operation simplifies the complexity of the animation process, making it easier for the model to generate accurate results.

Comment #5

"Consider adding a difficulty level split for A2Bench."

Response: Thanks for your valuable suggestion. We have added the difficulty level split for `Animate-X`. As shown in Figure 10, we categorized the videos in A2Bench into three difficulty levels: Level 1, Level 2, and Level 3. The classification is based on their appearance characteristics. **First**, we classify characters that have body shapes and other appearance features similar to humans, as shown in the first row of Figure 10, into the easiest, Level 1 category. These characters are generally simpler to drive, produce fewer artifacts, and have better motion consistency. **In contrast**, characters that maintain more distinct structural features from humans, such as dragons and ducks in the third row of Figure 10, are classified into the most difficult Level 3 category. These characters often preserve their original structures (*e.g.*, a duck's webbed feet and wings), which makes balancing identity preservation and motion consistency more challenging. To ensure identity preservation, the consistency of motion may be compromised, and vice versa. Additionally, images involving interactions between characters, objects, environments, and backgrounds are also placed in Level 3, as they increase the difficulty for the model to distinguish the parts that need to be driven from those that do not. **Videos in between these two categories**, like those in the second row of Figure 10, are classified as Level 2. These characters often strike a good balance between anthropomorphism and their original form, making them easier to animate with better motion consistency than Level 3 characters and more interesting results than Level 1 characters.

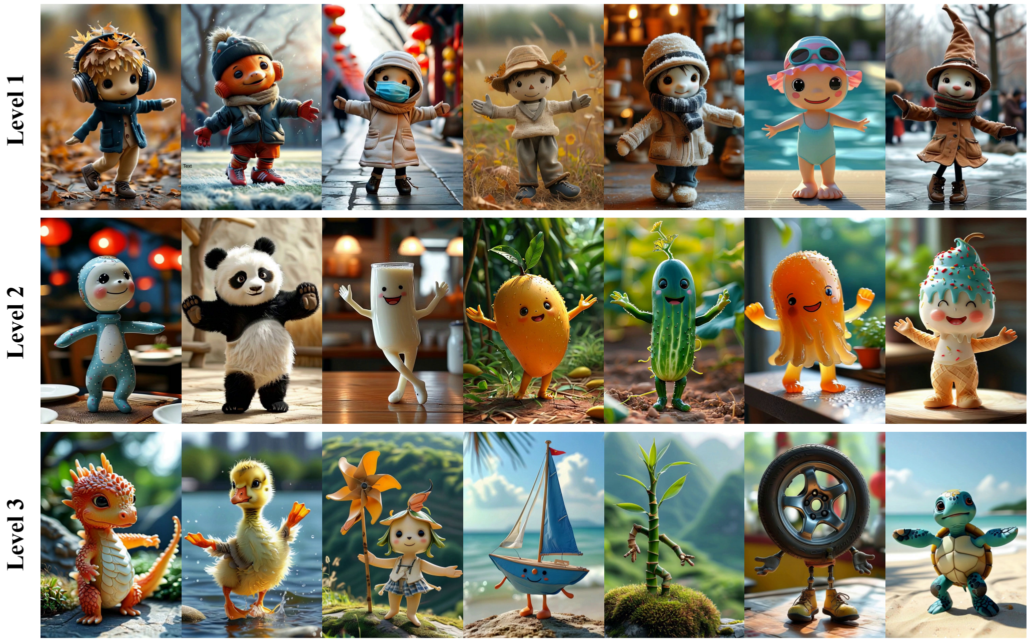


Figure 10: Difficulty levels in $\texttt{A}^2\texttt{Bench}$.

Comment #6

“Most styles in the A2Bench benchmark are "3D render style"; the benchmark should include a wider variety of visual styles. In the supplementary materials, the authors show results in various styles, yet most styles in A2Bench are in "3D render style." Would it be possible to add a "style trigger word" in the prompt template to diversify and strengthen the benchmark?”

Response: Following your suggestions, we have added style trigger words such as “*Watercolor Painting*”, “*Cyberpunk Style*”, “*Van Gogh*”, “*Ukiyo-E*”, “*Pixel Art*” and so on. Some results are shown in Figure 11, which indeed enriches the benchmark and strengthens its diversity. Please see (.zip/for_reviewer_feUz/more_style/xxx.mp4) for video results. Thanks for your valuable suggestions.

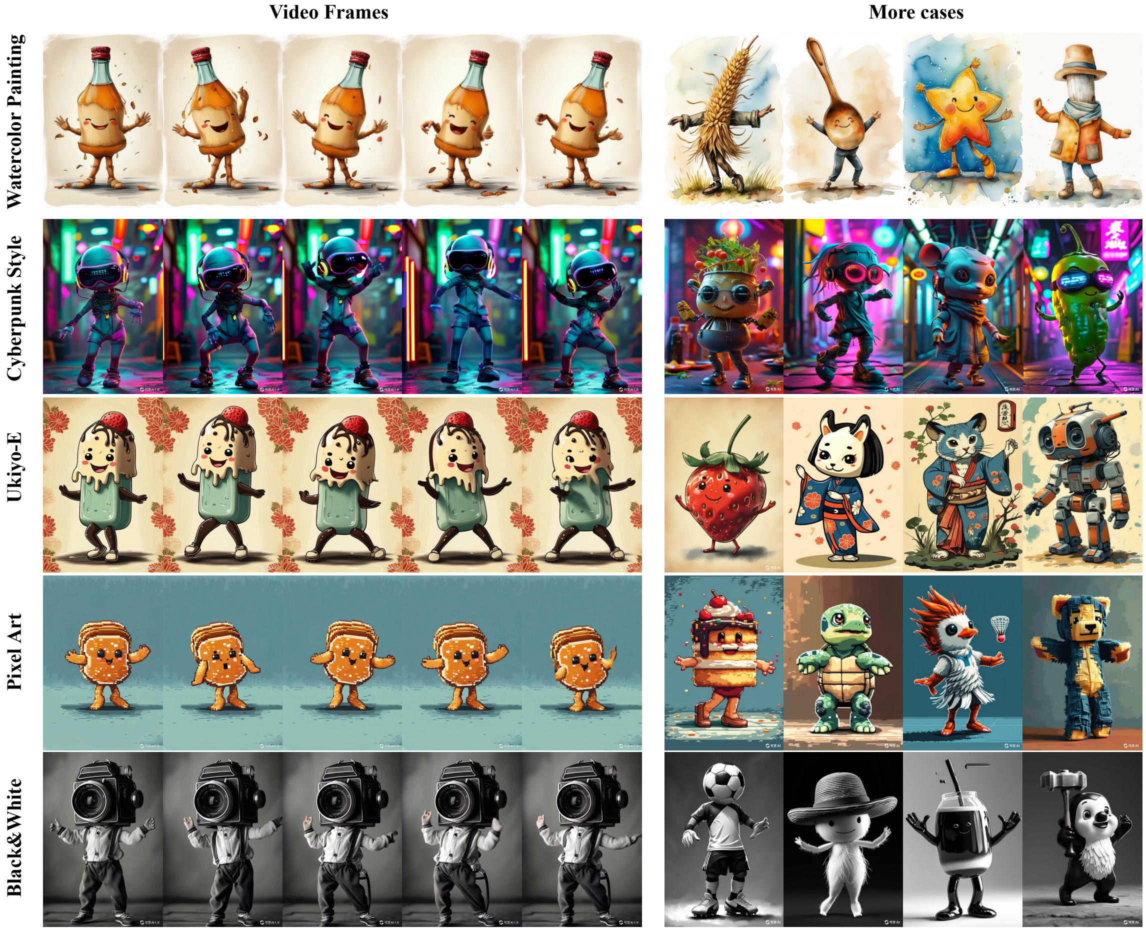


Figure 11: More styles in A^{2}Bench.

Reviewer: #4 ESK8

We sincerely thank **Reviewer #4 ESK8** for acknowledging "*the clear motivation, video results and benchmark presented in our work*". Below, we have addressed each question in detail and hope to clarify any concerns.

Comment #1

> *"The backbone of this work remains unchanged, it is quite similar to the prior works like your reference AnimateAnyone and MagicAnimate, which makes this work a straightforward extension of existing works and thus reduces the contribution of this paper."*

Response: Thanks for the comments. First of all, the primary contribution of this work is the introduction of the **universal** character image animation. We proposed `Animate-X` to addresses challenges by leveraging our proposed IPI and EPI modules to implicitly and explicitly model the universal pose indicator.

Using the same backbone as AnimateAnyone Hu et al. (2023) and MagicAnimate Xu et al. (2023), which have pioneered in latent diffusion models for human animation, allows us to have a fair comparison with these works and demonstrate the contribution of IPI and EPI to animate anthropomorphic figures.

Comment #2

> *"Leveraging driving videos to boost the animation performance has been already explored in a few prior works like [1]. The implicit pose indicator is also a similar design which aims to extract comprehensive motion patterns to improve the animation performance. [1] X-portrait: Expressive portrait animation with hierarchical motion attention."*

Response: Thanks for the comments and for introducing X-Portrait. We will cite it and discuss the diffeernce between X-Portrait and ours:

1. **Use of the Driven Video:** In `Animate-X`, we extract pose images from the driven video to serve as the primary source of motion. Given that a single pose image cannot provide image-level motion-related details (such as motion-induced deformations like body part overlap, occlusion, and overall motion patterns). In contrast, X-Portrait directly inputs the driven video into the model without any processing, which is following most of GAN-based animation methods.
2. **Different Technical Approaches:** X-Portrait follows the approach of ControlNet Zhang et al. (2023), where the driving video is fed into an SD U-Net, and then a zero-conv layer is inserted into the main branch of the U-Net. In comparison, our IPI module first uses a pre-trained CLIP encoder to extract features from the driven video and then decouples image-level motion-related features for motion modeling.

2. **Task Scope:** X-Portrait focuses on facial animation, but `Animate-X` handles full-body animation for universal characters, which includes anthropomorphic figures in cartoons and games.

In summary, `Animate-X` is different from X-Portrait in *Use of the Driven Video*, *Technical Approaches*, and *Task Scope*.

Comment #3

“The explicit pose indicator is a little bit confusing because I think this module is an augmentation of the driving pose sequences. Therefore, the novelty of the proposed method is not very significant. It is reasonable that the augmentation can break the strong correspondence between the driving video and motion representation. What is the advantage of this training time rescale augmentation and over the test time pose alignment (Answer 3.1)? Are there any ablation studies about this? (Answer 3.2)”

Response:

Answer 3.1: The advantages of training time rescale augmentation over the test time alignment are as follows:

1. **Generalization for Characters Without Extractable Poses:** For reference images with structures significantly different from human skeletons, such as the limb-less fairy shown in Figure 1, pose extraction using DWPose is not feasible, which is because DWPose is specifically designed for processing human poses. Consequently, pose alignment at test time cannot be performed, making the diffusion model challenging to generate reasonable videos. In contrast, training time rescale augmentation enables the diffusion model to learn how to handle misaligned reference and driven poses, enhancing its robustness and generalization. In this way, `Animate-X` can handle scenarios where poses cannot be extracted from the reference image, as it eliminates the need for pose alignment between the reference and driven pose images during inference.

2. **Reduced Dependency on Strict Pose Alignment:** Even when pose alignment is available at test time, the results often rely heavily on precise alignment. For example, if the aligned pose differs in arm length from the reference image (*e.g.*, a longer arm), the generated result will reflect this discrepancy, compromising identity preservation. In contrast, rescale augmentation during training reduces the model’s dependence on strict pose alignment, ensuring that even with imperfect or absent alignment, the generated results can still effectively preserve identity information.

3. **Simpler Test-Time Workflow and Faster Inference:** For example, animating 100,000,000 reference images with a single driven pose using previous methods would require extracting the pose for each of the 100,000,000 reference images, followed by an equal number of strict pose alignment operations. In contrast, our method removes the need for these alignment operations, significantly reducing inference time and simplifying the test-time process.

Answer 3.2: We have conducted extensive ablation experiments for different pairs of pose transformations in EPI, as detailed in Appendix D.4 and Table X. The results show that each pose transformation improves performance compared to the scenarios without augmentation, confirming the effectiveness of the augmentation operation in enhancing the model’s performance.

Table X: Quantitative results of ablation study.

Method	PSNR* ↑	SSIM ↑	L1 ↓	LPIPS ↓	FID ↓	FID-VID ↓	FVD ↓
w/o Add in EPI	13.28	0.442	1.56E-04	0.459	34.24	52.94	804.37
w/o Drop in EPI	13.36	0.441	1.94E-04	0.458	26.65	44.55	764.52
w/o BS in EPI	13.27	0.443	1.08E-04	0.461	29.60	56.56	850.17
w/o NF in EPI	13.41	0.446	1.82E-04	0.455	29.21	56.48	878.11
w/o AL in EPI	13.04	0.429	1.04E-04	0.474	27.17	33.97	765.69
w/o Rescalings in EPI	13.23	0.438	1.21E-04	0.464	27.64	35.95	721.11
w/o Realign in EPI	12.27	0.433	1.17E-04	0.434	34.60	49.33	860.25
w/o EPI	12.63	0.403	1.80E-04	0.509	42.17	58.17	948.25
`Animate-X`	**13.60**	**0.452**	**1.02E-04**	**0.430**	**26.11**	**32.23**	**703.87**

Comment #4

"From the results of the animation of anthropomorphic characters, the example of a banana shows that although the animation result looks like a banana, the motion precision is decreased. Therefore, I think the implicit pose indicator could harm the motion precision (Answer 4.1). The authors could conduct more experiments to study this issue (Answer 4.2)."

Response:

Answer 4.1: Thanks for pointing out. First of all, we need to clarify that the implicit pose indicator does not harm motion precision. We have demonstrated that adding the IPI module to the baseline results in improvements across all quantitative metrics, highlighting its contributions to every aspect of animation through extensive ablation experiments (i.e., Table III).

Answer 4.2: As shown in Figure 12, we have conducted additional experiments on the banana case and provided a detailed discussion. Specifically, we input the banana image and the driven poses into the model without the IPI module to generate the results. As shown in Figure 12, we observe that without the IPI module, the model generates the human-like arms, which was not the intended outcome. In contrast, `Animate-X` (with IPI) prioritized preserving the banana's identity and avoiding obvious artifacts. We believe this trade-off is reasonable and aligns with the limitation discussed in our paper: the excessive sacrifices in identity preservation in favor of strict pose consistency.

To balance pose consistency and identity preservation, we assigned an appropriate weight to the IPI module. In this way, we generated the preferrable result, as shown in the last row of Figure 12. To allow users to control the trade-off, we made this weight an adjustable parameter. Additionally, we conducted detailed experiments and analysis of this weight, as presented in Figure 12 in submission.

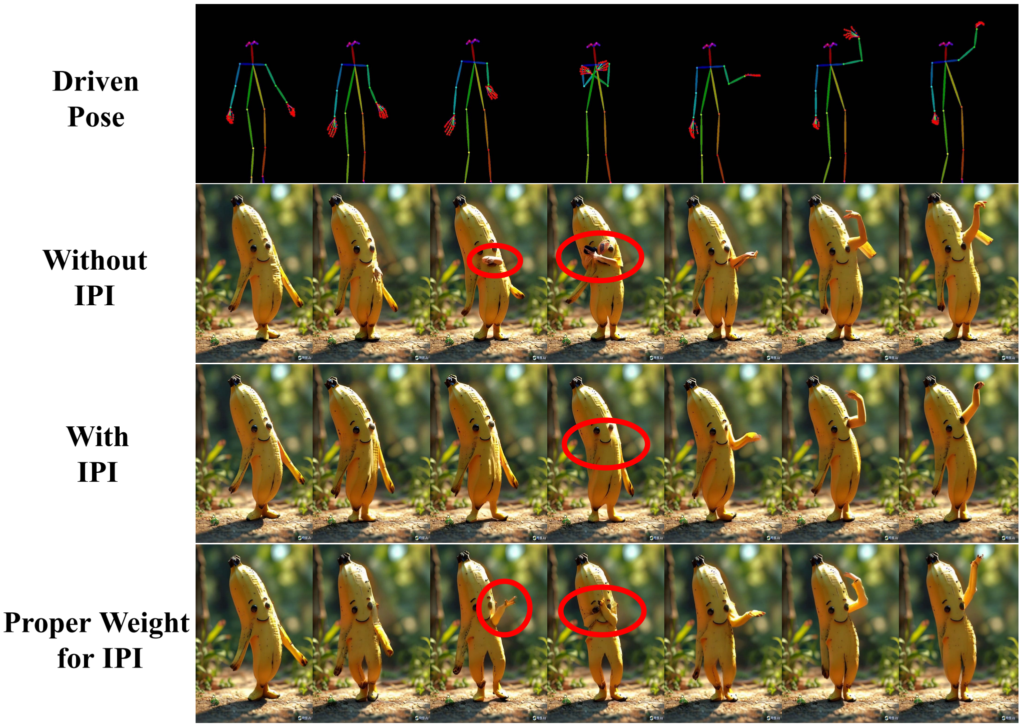


Figure 12: Ablation results of IPI on banana case.

Comment #5

"Does this model still use any input videos in the inference stage (Answer 5.1)? I am asking this question because there are no input driving videos in the "Animating anthropomorphic characters" section of the supplementary materials. Could the author explain the inference setting (Answer 5.2)? If there is a corresponding driving video, it is better to also include them into the results (Answer 5.3)."

Response:

Answer 5.1: Yes, this model can still use any input videos during the inference stage.

Answer 5.2: Yes. As shown in Figure 13, during inference, our method takes a reference image and a driven video as input and outputs an animated video that maintains the same identity as the reference image and the same motion as the driven video.

Answer 5.3: Thanks. Following your suggestions, we have included the corresponding driving video in the results. Please see the videos in (.zip/for_reviewer_ESK8/for_comment_5/xxx.mp4).

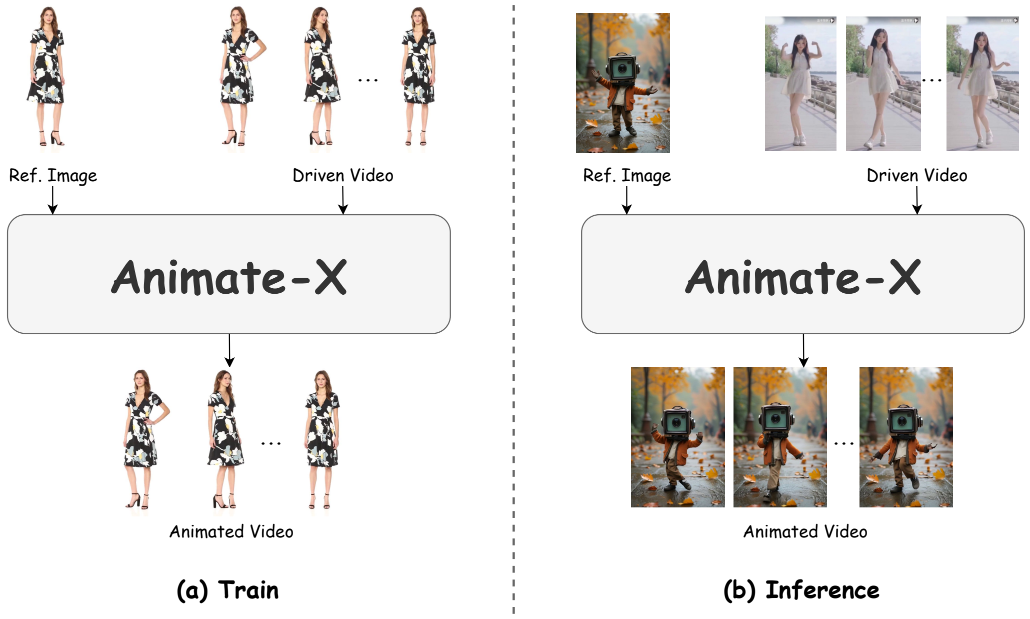


Figure 13: The difference of training and inference pipeline. During training, the reference image and the driven video come from the same video, while in the inference pipeline, the reference image and the driven video can be from any sources and appreciably different.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

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.

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. *arXiv preprint arXiv:2311.16498*, 2023.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pp. 3836–3847, 2023.

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.