

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	<u>26.92</u>	94.10 %	99.05 %	<u>55.14 %</u>	<u>62.54 %</u>
Fashion	20.18	98.25 %	99.45 %	49.62 %	49.96 %
A ² Bench	33.24	<u>96.66 %</u>	<u>99.39 %</u>	69.86 %	69.32 %

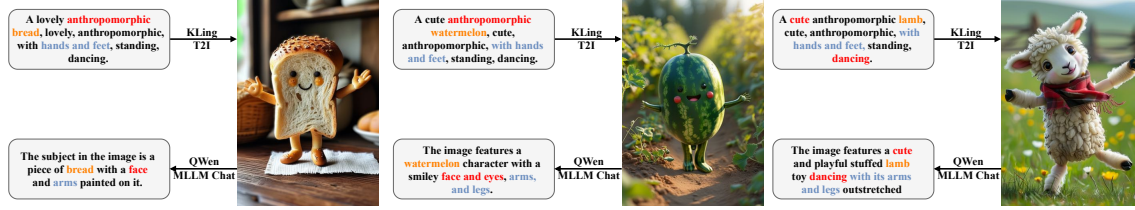


Figure 1: Prompts, generated images by T2I in A²Bench, and logical answers from QWen.

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 A²Bench 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 A²Bench.

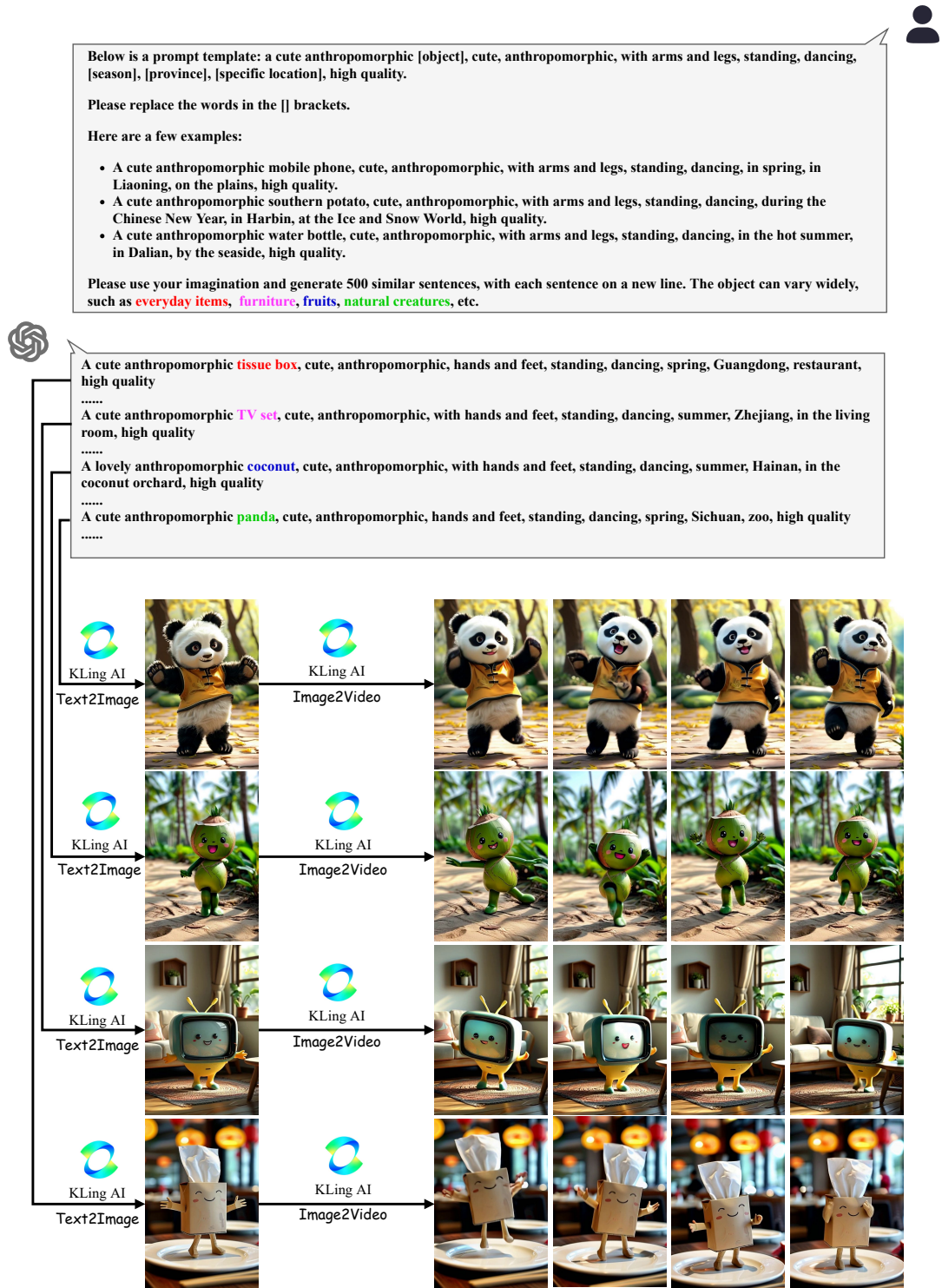


Figure 2: Detailed pipeline for building A²Bench based on large-scale pretrained models, including Open-ChatGPT 4o and KLing AI.

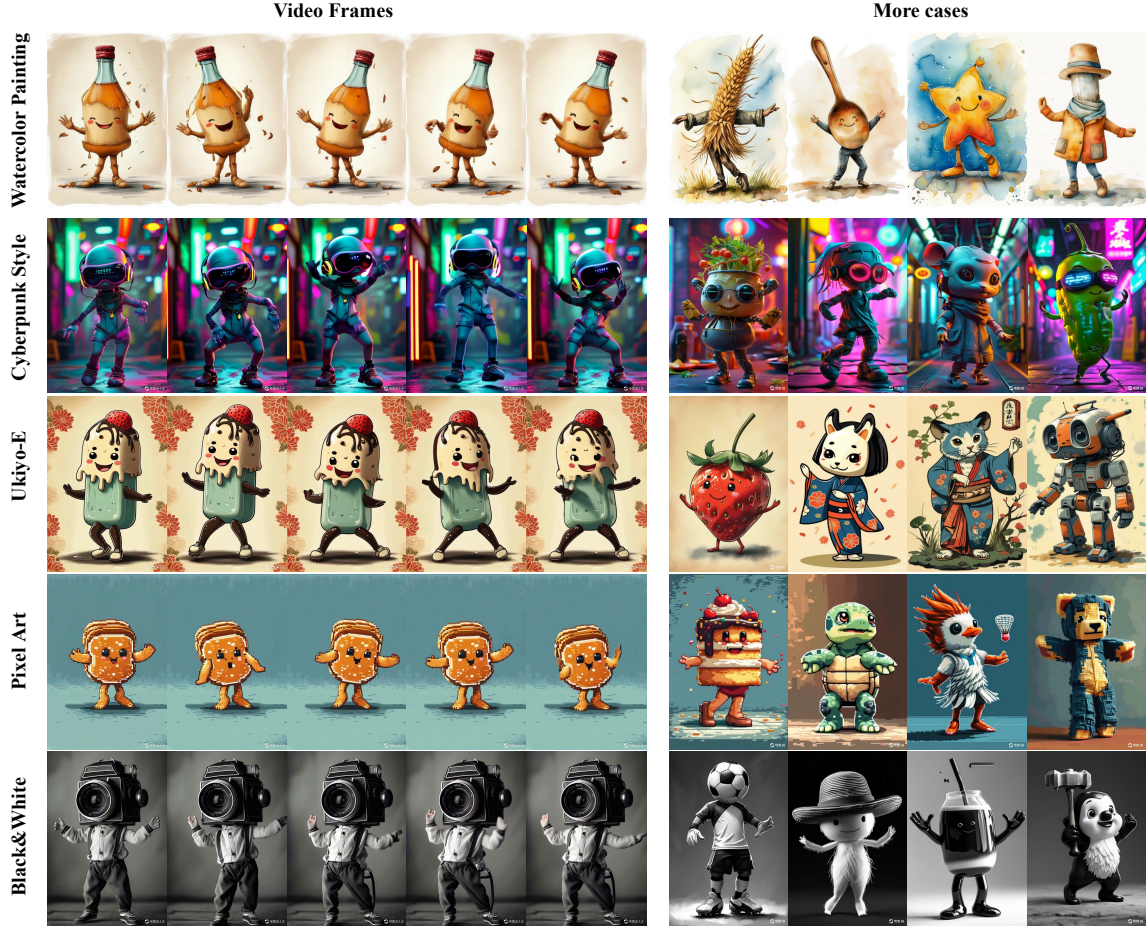


Figure 3: More styles in A^2Bench .

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 follows:

- **Step1:** Given a driving pose I^p , we randomly select an anchor pose I_{anchor}^p 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_{anchor}^p divided by the shoulder length of I^p might be 0.45, and the leg length of I_{anchor}^p 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_n^p), as shown in Figure 4.

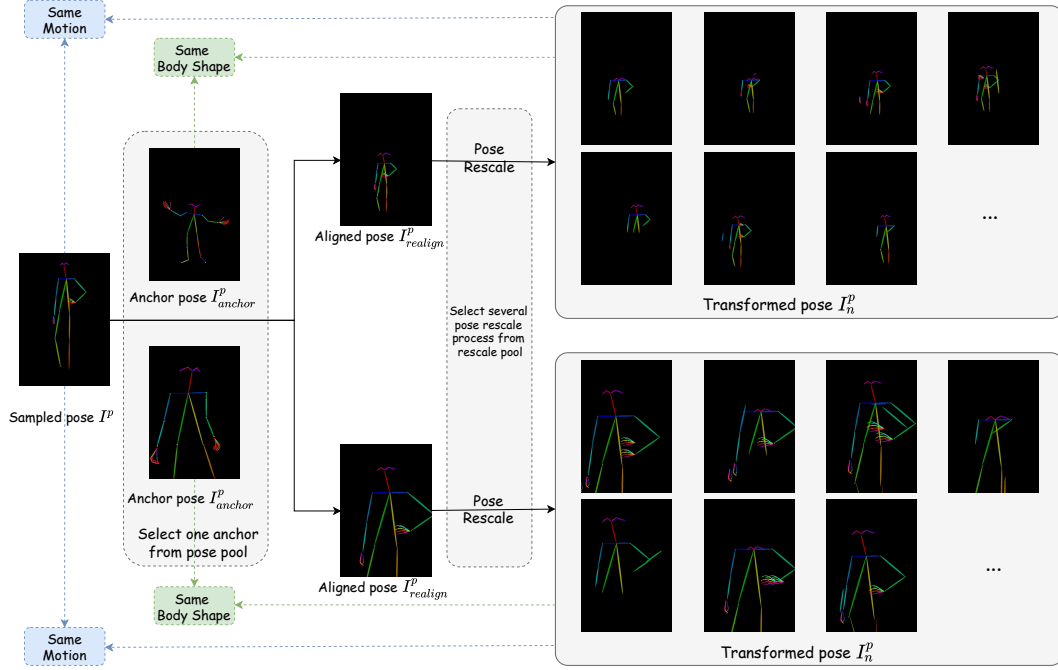


Figure 4: The pipeline of EPI.

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_n^p .

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.

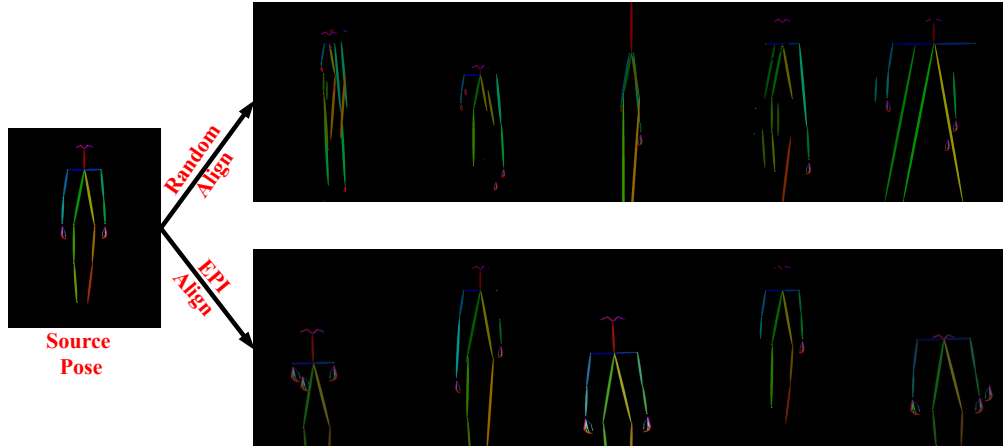


Figure 5: The different results between random alignment and EPI alignment.

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).

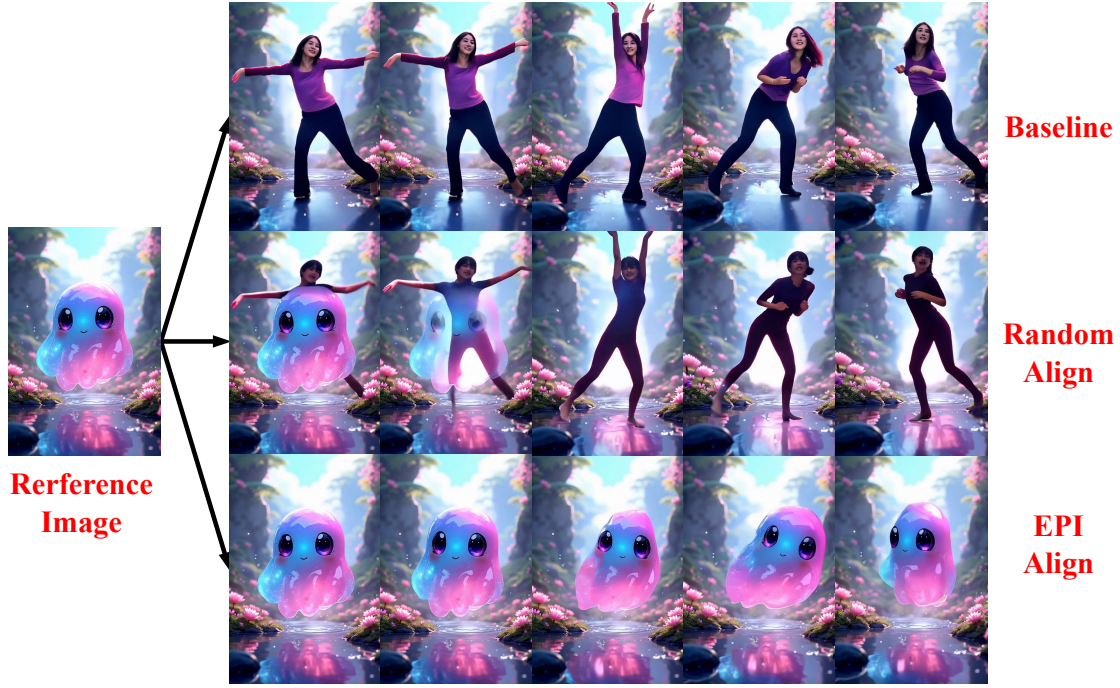


Figure 6: The results of different alignments.

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* \uparrow	SSIM \uparrow	L1 \downarrow	LPIPS \downarrow	FID \downarrow	FID-VID \downarrow	FVD \downarrow
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	<u>13.41</u>	<u>0.446</u>	1.82E-04	0.455	29.21	56.48	878.11
w/o AL in EPI	13.04	0.429	<u>1.04E-04</u>	0.474	27.17	<u>33.97</u>	765.69
w/o Rescalings in EPI	13.23	0.438	1.21E-04	0.464	27.64	35.95	<u>721.11</u>
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* \uparrow	SSIM \uparrow	L1 \downarrow	LPIPS \downarrow	FID \downarrow	FID-VID \downarrow	FVD \downarrow
w/o IPI	13.30	0.433	1.35E-04	0.454	32.56	64.31	893.31
w/o LQ	<u>13.48</u>	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	<u>1.02E-04</u>	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 A^2Bench , 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_{anchor}^p 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_{anchor}^p divided by the shoulder length of I^p might be 0.45, and the leg length of I_{anchor}^p 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_n^p), 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* \uparrow	SSIM \uparrow	L1 \uparrow	LPIPS \downarrow	FID \downarrow	FID-VID \downarrow	FVD \downarrow
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

$$\text{video} \rightarrow \text{pose skeletons} \rightarrow \text{results}.$$

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 \rightarrow pose images;
- **Step2:** video \rightarrow IPI \rightarrow image-level motion-related features;
- **Step3:** pose images + image-level motion-related features \rightarrow 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.

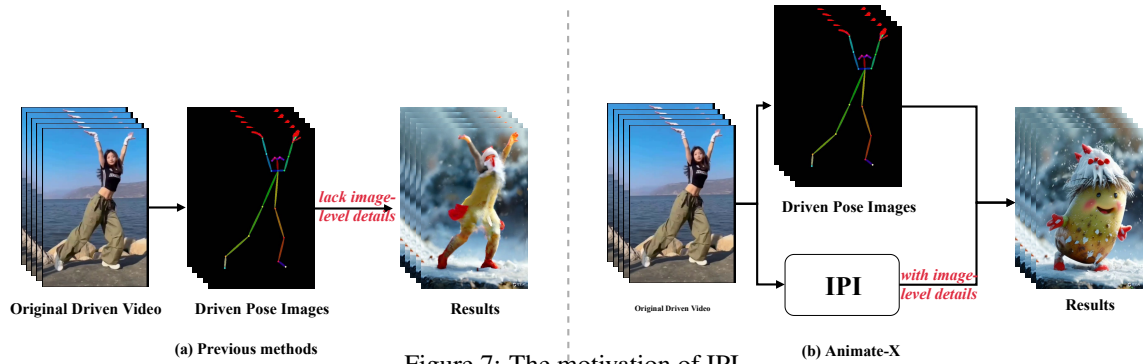


Figure 7: The motivation of IPI.

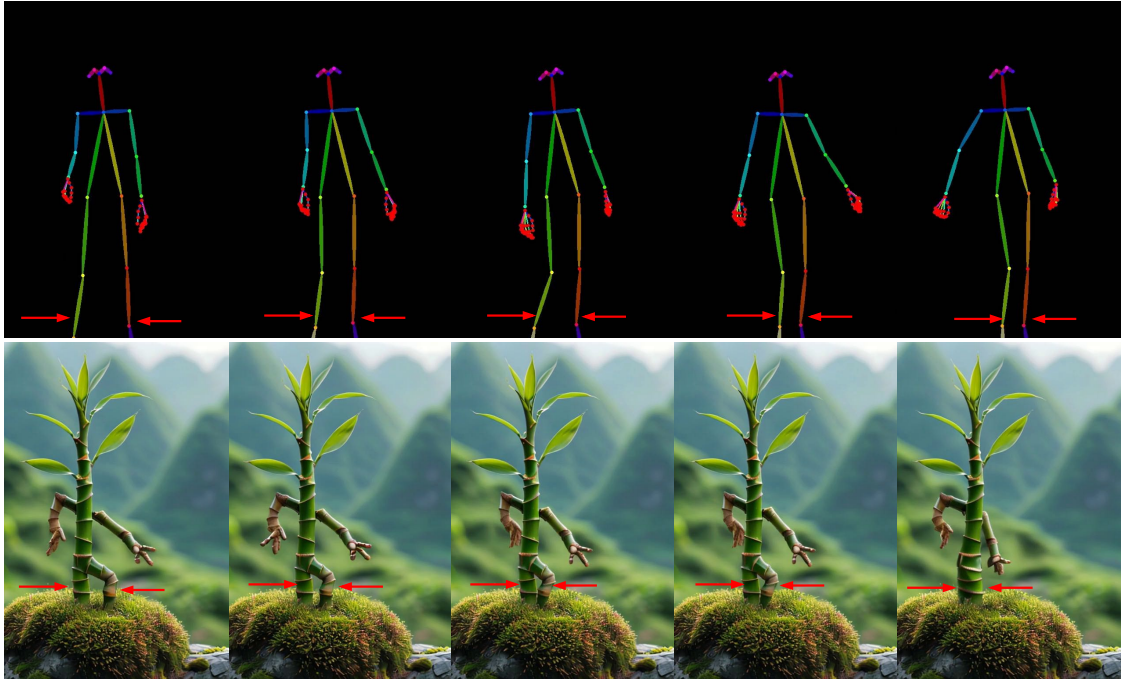


Figure 8: More frames in Figure 7 of the original submission.

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_{anchor}^p 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_{anchor}^p divided by the shoulder length of I^p might be 0.45, and the leg length of I_{anchor}^p 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_n^p).
- **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 $I_{realign}^p$ to obtain the final transformed poses I_n^p .

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 MimicMotion in 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* \uparrow	SSIM \uparrow	L1 \downarrow	LPIPS \downarrow	FID \downarrow	FID-VID \downarrow	FVD \downarrow
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^2 Bench in the self-driven setting.

Method	PSNR* \uparrow	SSIM \uparrow	L1 \downarrow	LPIPS \downarrow	FID \downarrow	FID-VID \downarrow	FVD \downarrow
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 \uparrow	60.4%	14.8%	52.0%	31.3%	43.0%	98.5%
Temporal consistency \uparrow	19.8%	24.9%	36.9%	43.9%	81.1%	93.4%
Visual quality \uparrow	27.0%	17.2%	40.4%	40.3%	79.3%	95.8%

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

Method	L1 \downarrow	PSNR* \uparrow	SSIM \uparrow	LPIPS \downarrow	FVD \downarrow
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* \uparrow	SSIM \uparrow	LPIPS \downarrow	FVD \downarrow
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_{anchor}^p 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_{anchor}^p divided by the shoulder length of I^p might be 0.45, and the leg length of I_{anchor}^p 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_n^p).
- **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 $I_{realign}^p$ to obtain the final transformed poses I_n^p .

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 details, 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 A²Bench.

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.

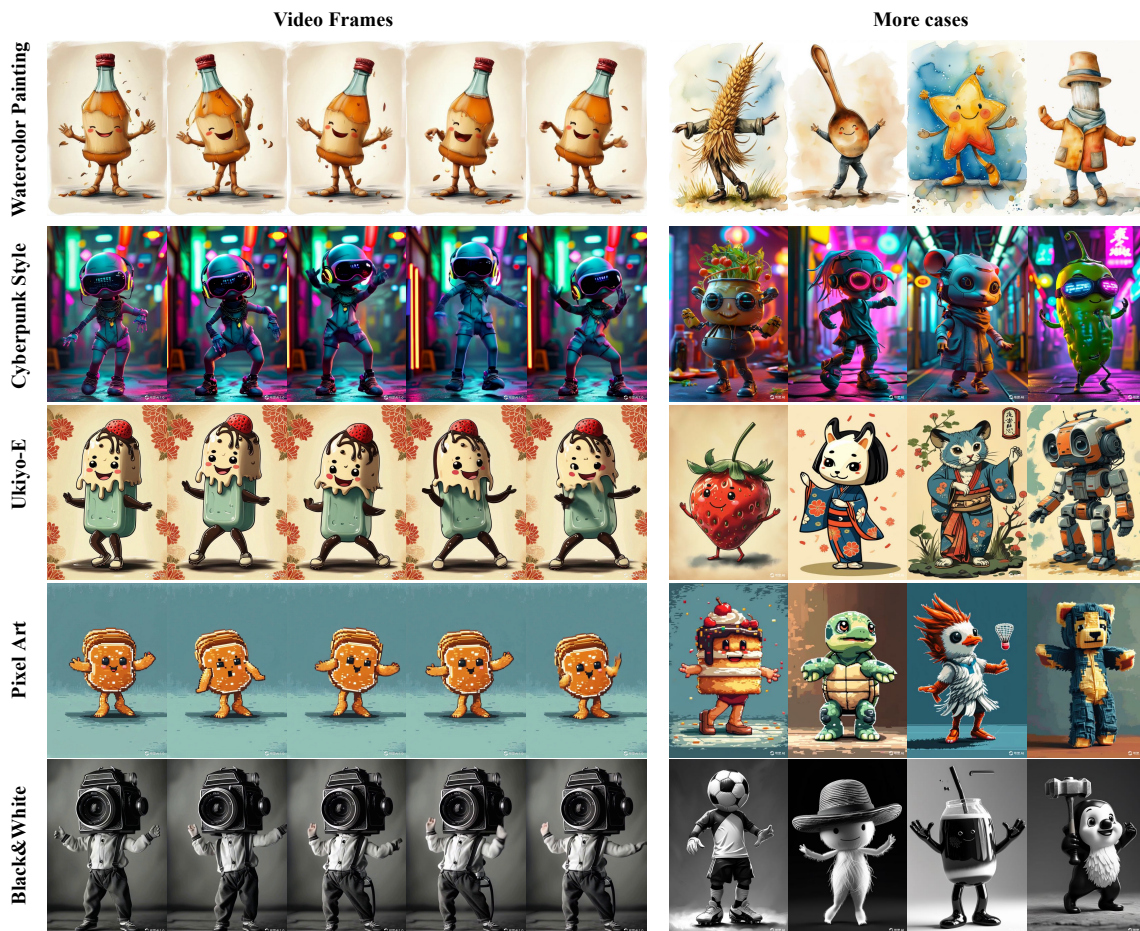


Figure 11: More styles in A²Bench.

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 difference 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* \uparrow	SSIM \uparrow	L1 \downarrow	LPIPS \downarrow	FID \downarrow	FID-VID \downarrow	FVD \downarrow
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	<u>26.65</u>	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	<u>13.41</u>	<u>0.446</u>	1.82E-04	0.455	29.21	56.48	878.11
w/o AL in EPI	13.04	0.429	<u>1.04E-04</u>	0.474	27.17	<u>33.97</u>	765.69
w/o Rescalings in EPI	13.23	0.438	1.21E-04	0.464	27.64	35.95	<u>721.11</u>
w/o Realign in EPI	12.27	0.433	1.17E-04	<u>0.434</u>	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 preferable 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.

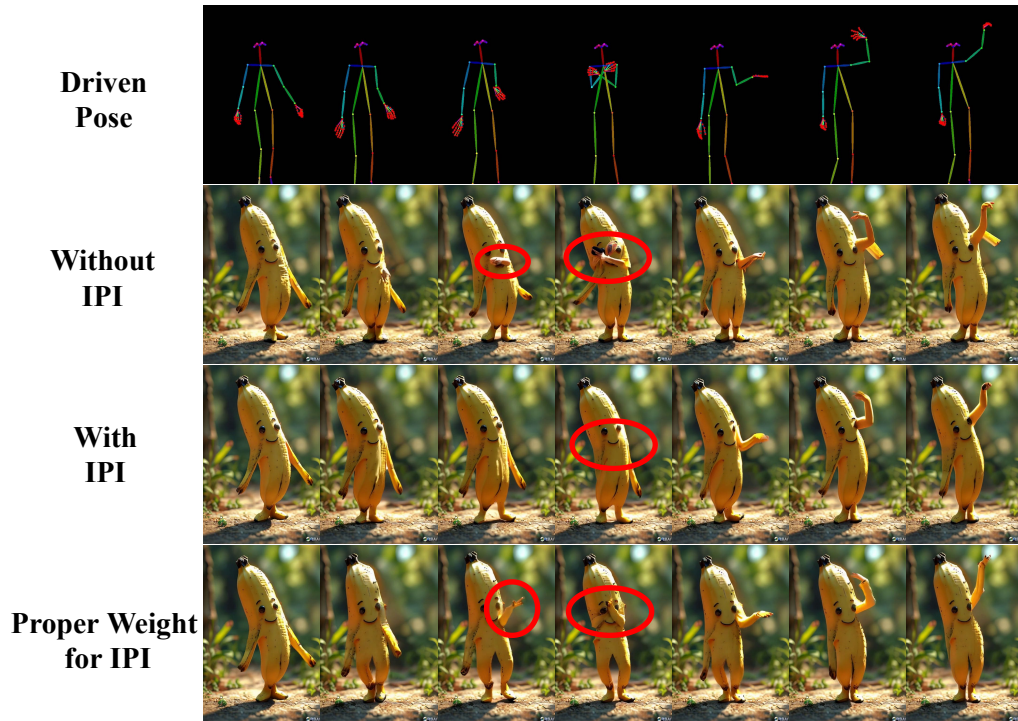


Figure 12: Ablation results of IPI on banana case.

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).

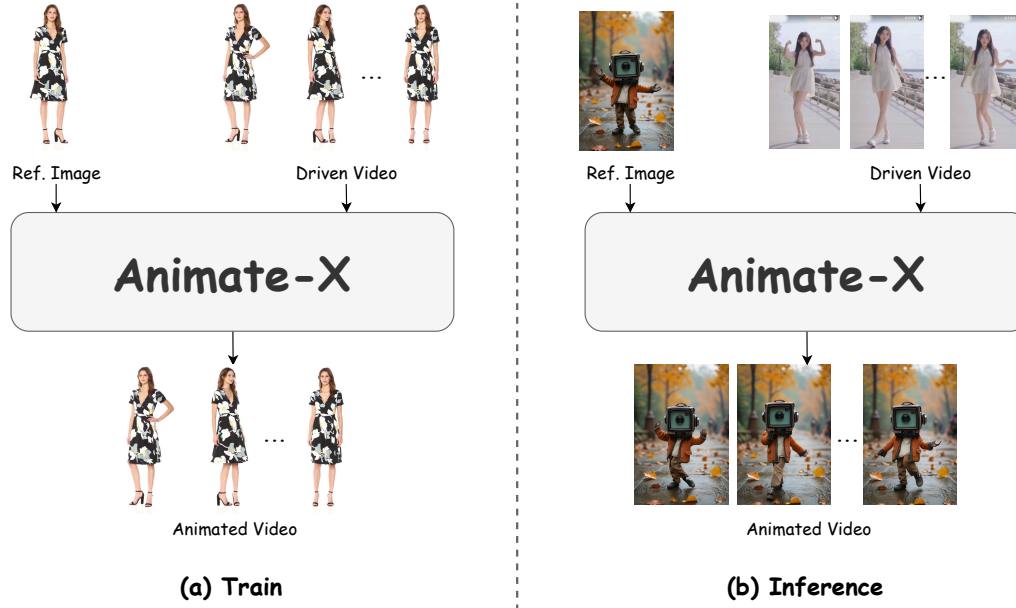


Figure 13: The difference of training and inference pipeline. During training, the reference image and the driven video come from the same video, while in the inference pipeline, the reference image and the driven video can be from any sources and appreciably different.

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.