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

Shuai Tan[1*], **Biao Gong**[1†], **Xiang Wang**[2], **Shiwei Zhang**[2],
Dandan Zheng[1], **Ruobing Zheng**[1], **Kecheng Zheng**[1], **Jingdong Chen**[1], **Ming Yang**[1]
[1]Ant Group [2]Alibaba Group
{tanshuai2001,a.biao.gong}@gmail.com,
{xiaolao.wx,zhangjin.zsw}@alibaba-inc.com,{yuandan.zdd,
zhengruobing.zrb,zhengkecheng.zkc,jingdongchen.cjd,m.yang}@antgroup.com

Project Page: https://lucaria-academy.github.io/Animate-X/

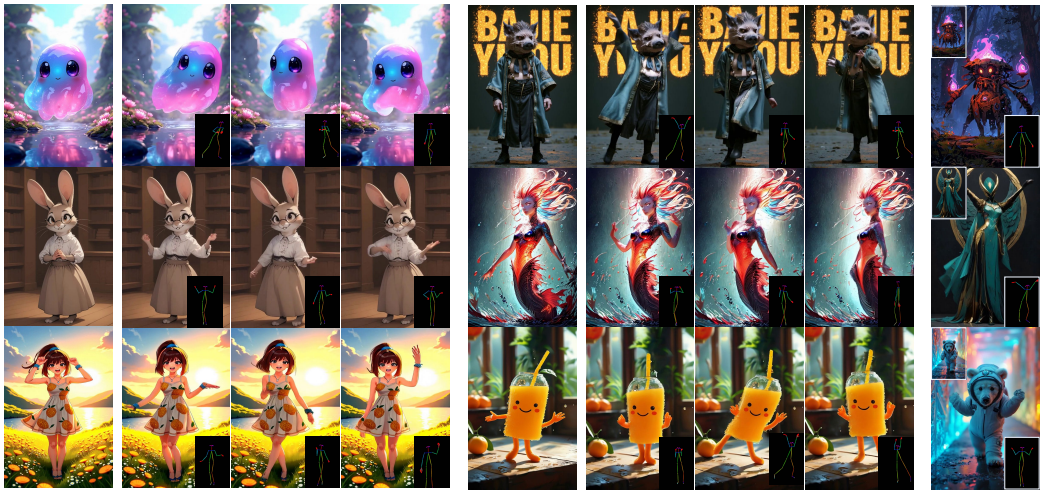

Figure 1: Animations produced by Animate-X which extends beyond human to anthropomorphic characters with various body structures, *e.g.*, without limbs, from games, animations, and posters.

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^2Bench) 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.

∗ Work done during internship at Ant Group.
† Project lead and corresponding author.

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 contributes 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.