PoseAnimate: Zero-shot high fidelity pose controllable character animation

Paper ID 2920

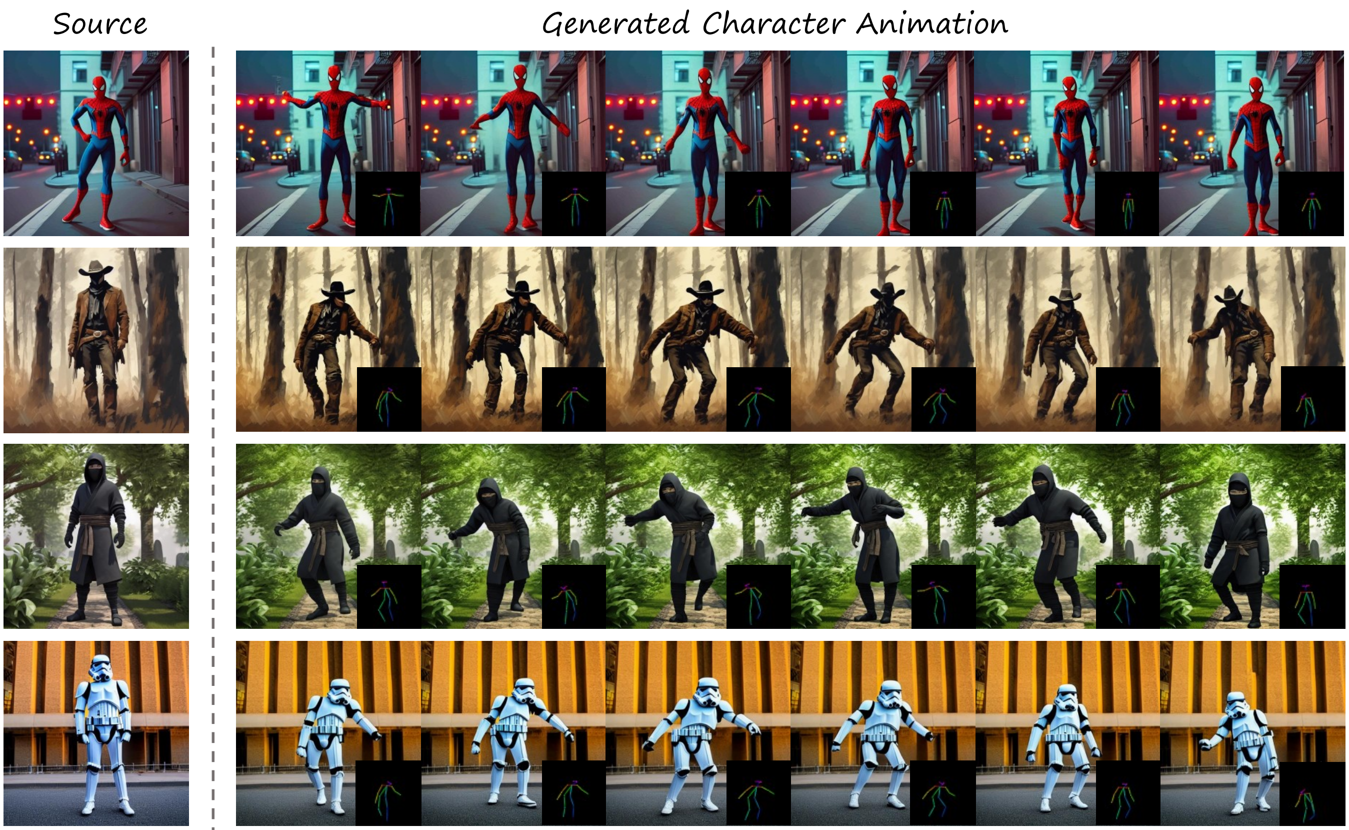

Figure 1: Our PoseAnimate framework is capable of generating smooth and high-quality character animations for character images across various pose sequences.

Abstract

Image-to-video(I2V) generation aims to create a video sequence from a single image, which requires high temporal coherence and visual fidelity with the source image. However, existing approaches suffer from character appearance inconsistency and poor preservation of fine details. Moreover, they require a large amount of video data for training, which can be computationally demanding. To address these limitations, we propose PoseAnimate, a novel zero-shot I2V framework for character animation. PoseAnimate contains three key components: 1) Pose-Aware Control Module (PACM) incorporates diverse pose signals into conditional embeddings, to preserve character-independent content and maintain precise alignment of actions. 2) Dual Consistency Attention Module (DCAM) enhances temporal consistency, and retains character identity and intricate background details. 3) Mask-Guided Decoupling Module (MGDM) refines distinct feature perception, improving animation fi-

delity by decoupling the character and background. We also propose a Pose Alignment Transition Algorithm (PATA) to ensure smooth action transition. Extensive experiment results demonstrate that our approach outperforms the state-of-the-art training-based methods in terms of character consistency and detail fidelity. Moreover, it maintains a high level of temporal coherence throughout the generated animations.

1 Introduction

Image animation [Siarohin *et al.*, 2019b; Siarohin *et al.*, 2019a; Siarohin *et al.*, 2021; Wang *et al.*, 2022; Zhao and Zhang, 2022] is a task that brings life into static images by seamlessly transforming them into dynamic and realistic videos. It involves the transformation of still images into a sequence of frames that exhibit smooth and coherent motions. In this task, character animation has gained significant attention due to its valuable applications in various scenarios, such as television production, game development, online retail and artistic creation, etc. However, minor motion variations hardly meet with the requirements. The goal of character animation is to make the character in image perform target pose sequences, while maintaining identity consistency and visual coherence. In early works, most of character animation were driven by traditional animation techniques, which involves meticulous frame-by-frame drawing or manipulation. In the subsequent era of deep learning, the advent of generative models [Goodfellow *et al.*, 2014; Zhu *et al.*, 2017; Karras *et al.*, 2019] drove the shift towards data-driven and automated approaches [Ren *et al.*, 2020; Chan *et al.*, 2019; Zhang *et al.*, 2022]. However, there are still ongoing challenges in achieving highly realistic and visually consistent animations, especially when dealing with complex motions, fine-grained details, and long-term temporal coherence.

Recently, diffusion models [Ho *et al.*, 2020] have demonstrated groundbreaking generative capabilities. Driven by the open source text-to-image diffusion model Stable Diffusion [Rombach *et al.*, 2022], the realm of video generation has achieved unprecedented progress in terms of visual quality and content richness. Hence, several endeavors [Wang *et al.*, 2023a; Xu *et al.*, 2023; Hu *et al.*, 2023] have sought to extrapolate the text-to-video(T2V) methods to image-to-video(I2V) by training additional image feature preserving networks and adapt them to the task of character animation. Nevertheless, these training-based methods do not possess accurate feature preservation capabilities for arbitrary open-domain images, and suffer from notable deficiencies in appearance control and loss of fine details. Furthermore, they require additional training data and computational overhead.

To this end, we contemplate employing a more refined and efficient resolution, image reconstruction for feature preservation, to tackle this problem. We propose PoseAnimate, depicted in Fig. 2, a zero-shot reconstruction-based I2V framework for pose controllable character animation video generation. PoseAnimate introduces a pose-aware control module(PACM), shown in Fig. 3 which optimizes the text embedding twice based on the original and target pose conditions respectively finally resulting a unique pose-aware embedding for each generated frame. This optimization strategy allows for the generated actions aligned to the target pose while contributing to keep the character-independent scene consistent. However, the introduction of a new target pose in the second optimization, which differs from the original pose, inevitably undermines the reconstruction of the character's identity and background. Thus, we further devise a dual consistency attention module(DCAM), as dedicated in the right part of Fig. 2, to address the disruption, in addition to maintain a smooth temporal progression. Since directly employing the entire attention map or key for attention fusion may result in the loss of fine-grained detail perception. We propose a mask-guided decoupling module(MGDM) to enable independent and focused spatial attention fusion for both the character and background. As such, our framework promises to capture the intricate character and background details, thereby effectively enhancing the fidelity of the animation. Besides, for the sake of adaptation to various scales and positions of target pose sequences, a pose alignment transition algorithm(PATA) is designed to ensure pose alignment and smooth transitions. Through combination of these novel modules, PoseAnimate achieves promising character animation results, as shown in Fig. 1, in a more efficient manner with lower computational overhead.

To summarize, our contributions are as follows: 1) We pioneer a reconstruction-based approach to handle the task of character animation and propose PoseAnimate, a novel zero-shot framework, which generates coherent high-quality videos for arbitrary character images under various pose sequences, without any training of the network. To the best of our knowledge, we are the first to explore a training-free approach to character animation. 2) We propose a pose-aware control module that enables precise alignment of actions while maintaining consistency across character-independent scenes. 3) We decouple the character and the background regions, performing independent inter-frame attention fusion for them, which significantly enhances visual fidelity. 4) Experiment results demonstrate the superiority of PoseAnimate compared with the state-of-the-art training-based methods in terms of character consistency and image fidelity.

2 Related work

2.1 Diffusion Models for Video Generation

Image generation has made significant progress due to the advancement of Diffusion Models(DMs) [Ho *et al.*, 2020]. Motivated by DM-based image generation [Rombach *et al.*, 2022], some works [Yang *et al.*, 2023; Ho *et al.*, 2022; Nikankin *et al.*, 2022; Esser *et al.*, 2023; Blattmann *et al.*, 2023b] explore DMs for video generation. Most video generation methods incorporate temporal modules to pretrained image diffusion models, extending 2D U-Net to 3D U-Net. Recent works control the generation of videos with multiple conditions. For text-guided video generation, these works [He *et al.*, 2022; Ge *et al.*, 2023; Gu *et al.*, 2023] usually tokenize text prompts with a pretrained image-language model, such as CLIP [Radford *et al.*, 2021], to control video generation through cross-attention. Due to the imperfect alignment between language and visual modalities in existing image-language models, text-guided video generation can't achieve high textual alignment. Alternative methods [Wang *et al.*, 2023b; Chen *et al.*, 2023; Blattmann *et al.*, 2023a] employ images as guidance for video generation. These works encode reference images to text token space, which benefits capturing visual semantic information. VideoComposer[Wang *et al.*, 2023b] combines textual conditions, spatial conditions(e.g., depth, sketch, reference image) and temporal conditions(e.g., motion vector) through Spatio-Temporal Condition encoders. VideoCrafter1[Chen *et al.*, 2023] introduces a text-aligned rich image embedding to capture details both from text prompts and reference images. Stable Video Diffusion [Blattmann *et al.*, 2023a] is a latent diffusion model for high-resolution T2V and I2V generation, which sets three different stages for training: text-to-image pretraining, video pretraining, and high-quality video finetuning.

2.2 Video Generation with Human Pose

Generating videos with human pose is currently a popular task. Compared to other conditions, human pose can better guide the synthesis of motions in videos, which ensures good temporal consistency. Follow your pose[Ma *et al.*, 2023] introduces a two-stage method to generate pose-controllable character videos. Many studies [Wang *et al.*, 2023a; Karras *et al.*, 2023; Xu *et al.*, 2023; Hu *et al.*, 2023] try to generate character videos from still images via pose sequence, which needs to preserve consistency of appearance from source images as well. Inspired by ControlNet[Zhang *et al.*, 2023], DisCo[Wang *et al.*, 2023a] realizes disentangled control of human foreground, background and pose, which enables faithful human video generation. To increase fidelity to the reference human images, DreamPose[Karras *et al.*, 2023] proposes an adapter to models CLIP and VAE image embeddings. MagicAnimate[Xu *et al.*, 2023] adopts ControlNet[Zhang *et al.*, 2023] to extract motion conditions. It also introduces a appearance encoder to model reference images embedding. Animate Anyone[Hu *et al.*, 2023] designs a ReferenceNet to extract detail features from reference images, combined with a pose guider to guarantee motion generation.

3 Method

Given a source character image I_s, and a desired pose sequence $P = \{p_i\}_{i=1}^{M}$, where M is the length of sequence. In the generated animation, we adopt a progressive approach to transition the character seamlessly from the original pose p_s to the desired pose sequence $P = \{p_i\}_{i=1}^{M}$. We first facilitate the Pose Alignment Transition Algorithm(PATA), derailed in supplementary material, to smoothly interpolate t intermediate frames between the source pose p_s and the target pose sequence $P = \{p_i\}_{i=1}^{M}$. Simultaneously, it aligns each target pose p_i with the source pose p_s to compensate for their discrepancies in terms of position and scale. As a result, the final target pose sequence is $P = \{p_i\}_{i=0}^{N}$, where $N = M + t$. It is worth noting that the first frame x_0 in our generated animation $X = \{x_i\}_{i=0}^{N}$ is identical to the source image I_s. Secondly, we propose a pose-aware control module(PACM) that optimizes a unique pose-aware embedding for each generated frame. This module can eliminate perturbation of original character posture, thereby ensuring the generated actions aligned with the target pose. Furthermore, it also maintains consistency of content irrelevant to characters. Thirdly, a dual consistency attention module(DCAM) is developed to ensure consistency of the character identity and improve temporal consistency. In addition, we design a mask-guided decoupling module(MGDM) to further enhance perception of characters and backgrounds details. The overview of our PoseAnimate is shown in Fig. 2.

In this section, we first give an introduction of Stable Diffusion in Sec 3.1. Subsequently, Sec 3.2 introduces the incorporation of motion awareness into pose-aware embedding. The proposed dual consistency control is elaborated in Sec 3.3, followed by mask-guided decoupling module in Sec 3.4.

3.1 Preliminaries on Stable Diffusion

Stable Diffusion [Rombach *et al.*, 2022] has demonstrated strong text-to-image generation ability through a diffusion model in a latent space constructed by a pair of image encoder $\mathcal{E}$ and decoder $\mathcal{D}$. For an input image $\mathcal{I}$, the encoder $\mathcal{E}$ first maps it to a lower dimensional latent code $z_0 = \mathcal{E}(\mathcal{I})$, then Gaussian noise is gradually added to z_0 through the diffusion forward process:

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

where $t = 1, ..., T$, denotes the timesteps, $\beta_t \in (0, 1)$ is a predefined noise schedule. Through a parameterization trick, we can directly sample z_t from z_0:

$$q(\mathbf{z}_t|\mathbf{z}_0) = \mathcal{N}(\mathbf{z}_t; \sqrt{\bar{\alpha}_t}\mathbf{z}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\alpha_t = 1 - \beta_t$. Diffusion model uses a neural network ϵ_θ to learn to predict the added noise ϵ by minimizing the mean square error of the predicted noise:

$$\min_\theta \mathbb{E}_{z,\epsilon\sim\mathcal{N}(0,I),t}[\|\epsilon - \epsilon_\theta(z_t, t, \mathbf{c})\|_2^2], \tag{3}$$

where $\mathbf{c}$ is embedding of textual prompt. And we can adopt a deterministic sampling process [Song *et al.*, 2020] to iteratively recover $z_0 \sim \mathcal{P}_{data}(z)$ from random noise z_T:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \underbrace{\hat{z}_{t\to 0}}_{\text{predicted } z_0} + \underbrace{\sqrt{1-\bar{\alpha}_{t-1}}\epsilon_\theta(z_t, t, \mathbf{c})}_{\text{direction pointing to } z_{t-1}}, \tag{4}$$

where $\hat{z}_{t\to 0}$ is the predicted z_0 at timestep t,

$$\hat{z}_{t\to 0} = \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(z_t, t, \mathbf{c})}{\sqrt{\bar{\alpha}_t}}. \tag{5}$$

3.2 Pose-Aware Control Module

For generating a high fidelity character animation from a source image, two tasks need to be accomplished. Firstly, it is critical to preserve the consistency of original character and background in generated animation. In contrast to other approaches [Karras *et al.*, 2023; Xu *et al.*, 2023; Hu *et al.*, 2023] that rely on training additional spatial preservation networks for consistency identity, we achieve

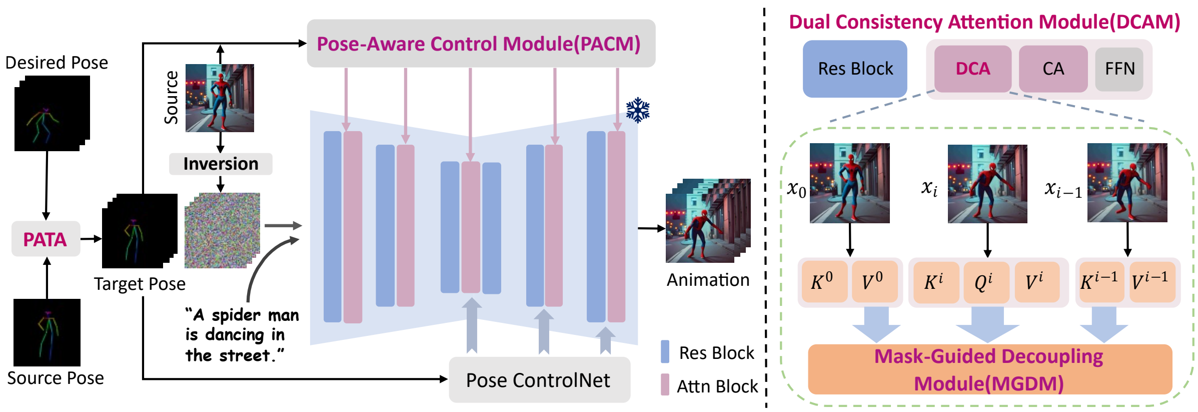

Figure 2: Overview of PoseAnimate. The pipeline is on the left, we first utilize the Pose Alignment Transition Algorithm(PATA) to align the desired pose with a smooth transition to the target pose. We utilize the inversion noise of the source image as the starting point for generation. The optimized pose-aware embedding of PACM, in Sec. 3.2, serves as the unconditional embedding for input. The right side is the illustration of DCAM in Sec. 3.3. The attention block in this module consists of Dual Consistency Attention(DCA), Cross Attention (CA), and Feed-Forward Networks (FFN). Within DCA, we integrate MGDM to independently perform inter-frame attention fusion for the character and background, which further enhance the fidelity of fine-grained details.

it through a computationally efficient reconstruction-based method. Secondly, the actions in generated frames needs to align with the target poses. Although the pre-trained OpenPose ControlNet [Zhang *et al.*, 2023] has great spatial control capabilities in controllable condition synthesis, our purpose is to discard the original pose and generate new continuous motion. Therefore, directly introducing pose signals through ControlNet may result in conflicts with the original pose, resulting in severe ghosting and blurring in motion areas.

In light of this, we propose the pose-aware control module, as illustrated in the Fig. 3. Inspired by the idea of inversion in image editing [Mokady *et al.*, 2023], we achieve the perception of pose signals by optimizing the text embedding $\varnothing_{text}$ twice based on the original pose p_s and target pose p_i respectively. In the first optimization, i.e. pose-aware inversion, we iteratively refine the original text embedding $\varnothing_{text}$ to accurately reconstruct the intricate details of the source image I_s under the original pose p_s. Building upon the optimized source embeddings $\{\varnothing_{s,t}\}_{t=1}^{T}$ obtained from this process, we then proceed with the second optimization, i.e. pose-aware embedding optimization, where we inject the target pose signals $P = \{p_i\}_{i=1}^{N}$ into the optimized pose-aware embeddings $\{\{\widetilde{\varnothing}_{x_i,t}\}_{t=1}^{T}\}_{i=1}^{N}$, as detailed in Alg. 1. Perceiving the target pose signals, these optimized pose-aware embeddings $\{\{\widetilde{\varnothing}_{x_i,t}\}_{t=1}^{T}\}_{i=1}^{N}$ ensure a flawless alignment between the generated character actions and the target poses, while upholding the consistency of character-independent content.

Specifically, to incorporate the pose signals, we integrate ControlNet into all processes of the module. Diverging from null-text inversion [Mokady *et al.*, 2023] that achieves image reconstruction by optimizing unconditional embeddings [Ho and Salimans, 2022], our pose-aware inversion optimizes the conditional embedding $\varnothing_{text}$ of text prompt C during the reconstruction process. The motivation stems from the observation that conditional embedding contains more abundant and robust semantic information, which endows it with a heightened potential for encoding pose signals.

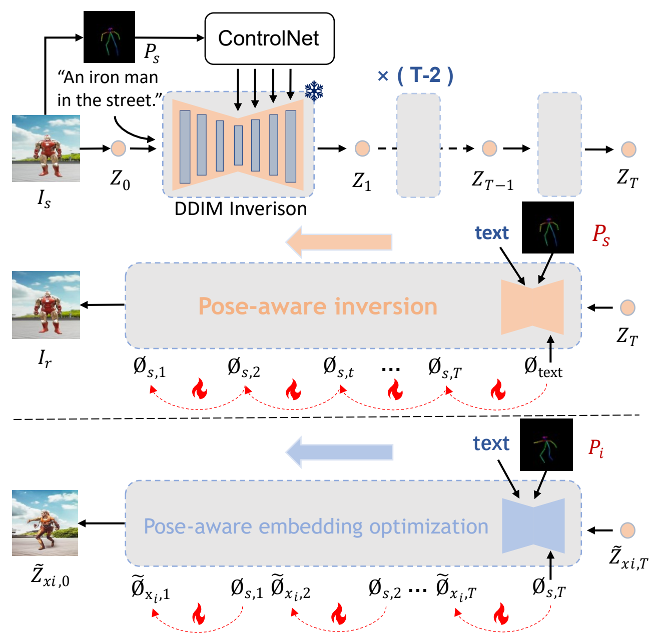

Figure 3: Illustration of pose-aware control module. We optimize the text embedding twice to inject motion awareness into pose-aware embedding.

3.3 Dual Consistency Attention Module

Although the pose-aware control module accurately captures and injects body poses, it may unintentionally alter the identity of the character and the background details due to the introduction of different pose signals, as demonstrated by the example $\widetilde{Z}_{x_i,0}$ in Fig. 3, which is undesirable. Since self-attention layers in the U-Net [Ronneberger *et al.*, 2015] play a crucial role in controlling appearance, shape, and fine-grained details, existing attention fusion paradigms commonly employ cross-frame attention mechanism [Ni *et al.*, 2022], to

Algorithm 1 Pose-aware embedding optimization.

Input: Source character image I_s, source character pose p_s, text prompt C, and target pose sequence $P = \{p_i\}_{i=1}^{N}$, number of frames N, timestep T.

Output: Optimized source embeddings $\{\varnothing_{s,t}\}_{t=1}^{T}$, Optimized pose-aware embeddings $\{\{\widetilde{\varnothing}_{x_i,t}\}_{t=1}^{T}\}_{i=1}^{N}$, and latent code Z_T .

Set guidance scale = 1.0. Calculate DDIM inversion latent code $Z_0, ..., Z_T$ corresponding to input image I_s.
Set guidance scale = 7.5. Obtain optimized source embeddings $\{\varnothing_{s,t}\}_{t=1}^{T}$ through pose-aware inversion (Fig. 3).
for $i = 1, 2, ..., N$ **do**
  Initialize $\widetilde{Z}_{x_i,T} = Z_T$, $\{\widetilde{\varnothing}_{x_i,t}\}_{t=1}^{T} = \{\varnothing_{s,t}\}_{t=1}^{T}$;
  for $t = T, T-1, ..., 1$ **do**
    $\widetilde{Z}_{x_i,t-1} \leftarrow \text{Sample}(\widetilde{Z}_{x_i,t}, \epsilon_\theta(\widetilde{Z}_{x_i,t}, \widetilde{\varnothing}_{x_i,t}, p_i, C, t))$;
    $\widetilde{\varnothing}_{x_i,t} \leftarrow \widetilde{\varnothing}_{x_i,t} - \eta \nabla_{\widetilde{\varnothing}} \text{MSE}(Z_{t-1}, \widetilde{Z}_{x_i,t-1})$;
  end for
end for
Return $Z_T, \{\varnothing_{s,t}\}_{t=1}^{T}, \{\{\widetilde{\varnothing}_{x_i,t}\}_{t=1}^{T}\}_{i=1}^{N}$

facilitate spatial information interaction across frames:

$$\text{Attention}(Q^i, K^j, V^j) = \text{softmax}\left(\frac{Q^i (K^j)^\top}{\sqrt{d}}\right) V^j, \quad (6)$$

where Q^i is the query feature of frame x_i, and K^j, V^j correspond to the key feature and value feature of frame x_j. As pose p_1 is identical to the original pose p_s, the reconstruction of frame x_0 remains undisturbed, allowing for a perfect restoration of the source image I_s. Hence, we can compute the cross-frame attention between each subsequent frame $\{x_i\}_{i=1}^{N}$ with the frame x_0 to ensure the preservation of identity and intricate details. However, solely involving frame x_0 in the attention fusion would bias the generated actions towards the original action, resulting in ghosting artifacts and flickering. Consequently, we develop the Dual Consistency Attention Module(DCAM) by replacing self-attention layers with our dual consistency attention(DC Attention) to address the issue of appearance inconsistency and improve temporal consistency. The DC Attention mechanism operates for each subsequent frame x_i as follows:

$$\begin{gathered} \text{CFA}_{i,j} = \text{Attention}(Q^i, K^j, V^j), \\ \text{Dual Consistency Attention}(x_i) := \text{DCA}_i = \\ \lambda_1 * \text{CFA}_{i,0} + \lambda_2 * \text{CFA}_{i,i-1} + \lambda_3 * \text{CFA}_{i,i}, \end{gathered} \quad (7)$$

where $\lambda_1, \lambda_2, \lambda_3 \in (0, 1)$ are hyper-parameters, and $\lambda_1 + \lambda_2 + \lambda_3 = 1$. $\text{CFA}_{i,j}$ refers to cross-frame attention between frames x_i and x_j.They jointly control the participation of the initial frame x_0, the current frame x_i and the preceding frame x_{i-1} in the DC Attention calculation. In the experiment, we set $\lambda_1 = 0.7$ and $\lambda_2 = \lambda_3 = 0.15$ to enable the frame x_0 to be more involved in the spatial correlation control of the current frame for the sake of better appearance preservation. Apart from this, retaining a relatively small portion of feature interaction for the current frame and the preceding frame simultaneously is promised to enhance motion stability and improve temporal coherence of the generated animation.

Furthermore, it is vital to note that we do not replace all the U-Net transformer blocks with DCAM. We find that incorporating the DC Attention only in the upsampling blocks of the U-Net architecture while leaving the remaining unchanged allows us to maintain consistency with the identity and background details of the source, without compromising the current frame's pose and layout.

3.4 Mask-Guided Decoupling Module

Directly utilizing the entire image features for attention fusion can lead to substantial loss of fine-grained details. To address this problem, we propose the mask-guided decoupling module, which decouples the character and background and enables individual inter-frame interaction to further refine spatial feature perception.

For the source image I_s, we obtain a precise body mask M_s (i.e. M_{x_0}) that separates the character from the background by an off-the-shelf segmentation model [Liu *et al.*, 2023a]. The target pose prior is insufficient to derive body mask for each generated frame of the character. Considering the strong semantic alignment capability of cross attention layers mentioned in Prompt-to-prompt [Hertz *et al.*, 2022], we extract the corresponding body mask M_{x_i} for each frame from the cross attention maps. With M_s and M_{x_i}, only attentions of character and background within corresponding region are calculated, according to the mask-guided decoupling module as follows:

$$\begin{aligned} \text{K}_j^c &= M_{x_j} \odot \text{K}_j, \text{K}_j^b = (1 - M_{x_j}) \odot \text{K}_j \\ \text{V}_j^c &= M_{x_j} \odot \text{V}_j, \text{V}_j^b = (1 - M_{x_j}) \odot \text{V}_j \\ \text{CFA}_{i,j}^c &= \text{Attention}(Q^i, K_j^c, V_j^c), \\ \text{CFA}_{i,j}^b &= \text{Attention}(Q^i, K_j^b, V_j^b), \end{aligned} \quad (8)$$

where $\text{CFA}_{i,j}^c$ is the attention output in character between frame x_i and x_j, and $\text{CFA}_{i,j}^b$ is for the background. Then we can get the final DC Attention output:

$$\begin{aligned} \text{DCA}_i^c &= \lambda_1 * \text{CFA}_{i,0}^c + \lambda_2 * \text{CFA}_{i,i-1}^c + \lambda_3 * \text{CFA}_{i,i}^c \\ \text{DCA}_i^b &= \lambda_1 * \text{CFA}_{i,0}^b + \lambda_2 * \text{CFA}_{i,i-1}^b + \lambda_3 * \text{CFA}_{i,i}^b \\ \text{DCA}_i &= M_{x_i} \odot \text{DCA}_i^c + (1 - M_{x_i}) \odot \text{DCA}_i^b, \end{aligned} \quad (9)$$

for $i = 1, ..., N$. The proposed decoupling module introduces explicit learning boundary between the character and background, allowing the network to focus on their respective content independently rather than blending features. Consequently, the intricate details of both the character and background are preserved, leading to a substantial improvement in the fidelity of the animation.

4 Experiment

4.1 Experiment Settings

We implement PoseAnimate based on the pre-trained weights of ControlNet [Zhang *et al.*, 2023] and Stable Diffu-

Figure 4: Qualitative comparison between our PoseAnimate and other training-based state-of-the-art character animation methods. We overlay the corresponding DensePose on the bottom right corner of the MagicAnimate(Densepose) synthesized frames. Previous methods suffer from inconsistent character appearance and details lost. Source prompt: "A firefighters in the smoke."(left)"A boy in the street."(right).

sion [Rombach *et al.*, 2022] v1.5. For each generated character animation, we generate $N = 16$ frames with a unified 512×512 resolution. All experiments are performed on a single NVIDIA A100 GPU.

4.2 Comparison Result

We compare our PoseAnimate with several state-of-the-art methods for character animation: MagicAnimate [Xu *et al.*, 2023] and Disco [Wang *et al.*, 2023a]. For MagicAnimate, both densepose [Güler *et al.*, 2018] and openpose signals of the same motion are applied to evaluate performances. We leverage the official open source code of disco to test its effectiveness. Additionally, we construct a competitive character animation baseline by IP-Adapter [Ye *et al.*, 2023] with ControlNet [Zhang *et al.*, 2023] and spatio-temporal attention [Wu *et al.*, 2023], which is termed as IP+CtrlN. It is worth noting that these methods are all training based, while ours does not require training.

Qualitative Results. We set up two different levels of pose for experiments to fully demonstrate the superiority of our method. The visual comparison results are shown in Fig. 4, with the left side displaying simple actions and the right side complex actions. Although IP+CtrlN has good performance on identity preservation, it fails to maintain details and inter-frame consistency. Disco loses the character appearance completely, and severe frame jitter leads to ghosting shadows and visual collapse for complex actions. MagicAnimate performs better than the other two methods, but it still encounters inconsistencies in character appearance at a more fine-grinded level guided by Densepose. It is also unable to preserve background and character details accurately, e.g. vehicle textures and mask of firefighter and the boy in Fig. 4. MagicAnimate under OpenPose signal conditions has worse performances

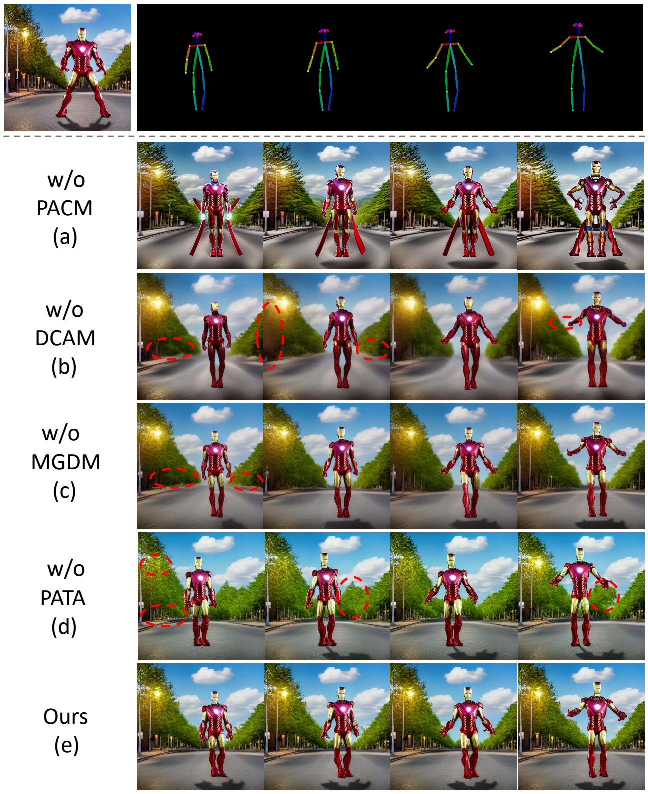

Figure 5: Ablation study. Source prompt: "An iron man on the road."

Method	LPIPS ↓	CLIP-I ↑	FC ↑	WE ↓
IP+CtrlN	0.466	0.937	94.88	0.1323
Disco	0.278	0.811	92.23	0.0434
MA(DensePose)	0.273	0.870	**97.87**	**0.0193**
MA(OpenPose)	0.411	0.867	97.63	0.0261
Ours	**0.247**	**0.948**	97.33	0.0384

Table 1: Quantitative comparison between our PoseAnimate and other training-based state-of-the-art methods. The best average performance is in bold. ↑ indicates higher metric value and represents better performance and vice versa.

than that under DensePose. While our method exhibits the best performance on image fidelity to the source image, and effectively preserves complex fine-grained appearance details and temporal consistency.

Quantitative Results. For quantitative analysis, we randomly sample 50 in-the-wild image-text pairs and 10 different disered pose sequences to conduct evaluations. We adopt four evaluation metrics: (1) LPIPS [Zhang *et al.*, 2018] measures the fidelity between generated frames and source image. (2) CLIP-I [Ye *et al.*, 2023] represents the similarity of CLIP [Radford *et al.*, 2021] image embedding between generated frames and the source image. (3) Frame Consistency(FC) [Esser *et al.*, 2023] evaluates video continuity by computing the average CLIP cosine similarity of two consecutive frames. (4) Warping Error(WE) [Liu *et al.*, 2023b] evaluates the temporal consistency of the generated animation through the Optical Flow algorithm [Teed and Deng, 2020].

Quantitative results are provided in Table. 1. Our method achieves the best scores on LPIPS and CLIP-I and greatly surpasses other comparison methods in terms of fidelity to the source image, demonstrating outstanding detail preservation capability. In addition, PoseAnimate outperforms the other two training-based methods in terms of inter-frame consistency. A good Warping Error score is also achieved, illustrating that our method is able to maintain good temporal coherence without additional training.

4.3 Ablation Study

We conduct ablation study to verify effectiveness of each component of our framework and present results in Fig. 5. The leftmost one in the first row is the source image, and the others are the target pose sequences. The following rows are generation results without certain components: (a) Pose-Aware Control Module (PACM) that effectively removes the interference of character original pose and maintains consistency of content unrelated to character; (b) Dual Consistency Attention Module (DCAM) that maintains image fidelity to the source image and improves temporal consistency; (c) Masked-Guided Decoupling Module (MGDM) that preserves image details; and (d) Pose Alignment Transition Algorithm (PATA) that tackles the issue of misalignments.

PACM. Fig. 5(a) illustrates the significant interference of original pose on the generated actions. Due to the substantial difference between the posture of Iron Man's legs in the source and in the target, there is a severe breakdown in the leg area of the generated frame, undermining the generation of a reasonable target action. Moreover, the character-irrelevant scenes also have noticeable distortion.

DCAM. From Fig. 5(b) we can find that it fails to maintain content consistency without Dual Consistency Attention Module. And the missing pole and Iron Man's hand in the red box reveal inter-frame inconsistency, indicating that both spatial and temporal content cannot be effectively maintained.

MGDM. Compared with our results in Fig. 5(e), we can observe that small signs are missing without MGDM. It proves that Masked-Guided Decoupling Module can effectively enhance the fine-grained feature perception and image fidelity.

PATA. Fig. 5(d) verifies the proposed Pose Alignment Transition Algorithm. The red circles in the first frame indicate the spatial content misalignment. When Iron Man in the original image does not match with the input pose position, an extra tree appears in the original position of Iron Man. And such misalignment can also leads to disappearance of background details, e.g., streetlights and distant signage.

5 Conclusion

This paper proposes a novel zero-shot approach PoseAnimate to tackle the task of character animation for the first time. PoseAnimate can generate temproal coherent and high-fidelity animations for arbitrary images under various pose sequences. Extensive experiment results demonstrate that PoseAnimate outperforms the state-of-the-art training based methods in terms of character consistency and detail fidelity.

References

[Blattmann *et al.*, 2023a] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[Blattmann *et al.*, 2023b] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[Chan *et al.*, 2019] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019.

[Chen *et al.*, 2023] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[Esser *et al.*, 2023] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.

[Ge *et al.*, 2023] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22930–22941, 2023.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Gu *et al.*, 2023] Jiaxi Gu, Shicong Wang, Haoyu Zhao, Tianyi Lu, Xing Zhang, Zuxuan Wu, Songcen Xu, Wei Zhang, Yu-Gang Jiang, and Hang Xu. Reuse and diffuse: Iterative denoising for text-to-video generation. *arXiv preprint arXiv:2309.03549*, 2023.

[Güler *et al.*, 2018] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018.

[He *et al.*, 2022] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity video generation with arbitrary lengths. *arXiv preprint arXiv:2211.13221*, 2022.

[Hertz *et al.*, 2022] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations*, 2022.

[Ho and Salimans, 2022] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

[Ho *et al.*, 2022] 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.

[Hu *et al.*, 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.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[Karras *et al.*, 2023] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023.

[Liu *et al.*, 2023a] Peng Liu, Fanyi Wang, Jingwen Su, Yanhao Zhang, and Guojun Qi. Lightweight high-resolution subject matting in the real world. *arXiv preprint arXiv:2312.07100*, 2023.

[Liu *et al.*, 2023b] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. *arXiv preprint arXiv:2310.11440*, 2023.

[Ma *et al.*, 2023] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023.

[Mokady *et al.*, 2023] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[Ni *et al.*, 2022] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022.

[Nikankin *et al.*, 2022] Yaniv Nikankin, Niv Haim, and Michal Irani. Sinfusion: Training diffusion models on a single image or video. *arXiv preprint arXiv:2211.11743*, 2022.

[Radford *et al.*, 2021] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Ren *et al.*, 2020] Yurui Ren, Ge Li, Shan Liu, and Thomas H Li. Deep spatial transformation for pose-guided person image generation and animation. *IEEE Transactions on Image Processing*, 29:8622–8635, 2020.

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

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[Siarohin *et al.*, 2019a] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. Animating arbitrary objects via deep motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2377–2386, 2019.

[Siarohin *et al.*, 2019b] 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, 2019.

[Siarohin *et al.*, 2021] 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*, pages 13653–13662, 2021.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[Teed and Deng, 2020] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.

[Wang *et al.*, 2022] 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.

[Wang *et al.*, 2023a] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. *arXiv preprint arXiv:2307.00040*, 2023.

[Wang *et al.*, 2023b] 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. *arXiv preprint arXiv:2306.02018*, 2023.

[Wu *et al.*, 2023] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[Xu *et al.*, 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*, 2023.

[Yang *et al.*, 2023] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *Entropy*, 25(10):1469, 2023.

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

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

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

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhao and Zhang, 2022] 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*, pages 3657–3666, 2022.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.