MagicPose: Realistic Human Poses and Facial Expressions Retargeting with Identity-aware Diffusion

Di Chang [1 2]   **Yichun Shi** [2]   **Quankai Gao** [1]   **Hongyi Xu** [2]   **Jessica Fu** [1]
Guoxian Song [2]   **Qing Yan** [2]   **Yizhe Zhu** [2]   **Xiao Yang** [2]   **Mohammad Soleymani** [1]

Abstract

In this work, we propose MagicPose, a diffusion-based model for 2D human pose and facial expression retargeting. Specifically, given a reference image, we aim to generate a person's new images by controlling the poses and facial expressions while keeping the identity unchanged. To this end, we propose a two-stage training strategy to disentangle human motions and appearance (e.g., facial expressions, skin tone and dressing), consisting of (1) the pre-training of an appearance-control block and (2) learning appearance-disentangled pose control. Our novel design enables robust appearance control over generated human images, including body, facial attributes, and even background. By leveraging the prior knowledge of image diffusion models, MagicPose generalizes well to unseen human identities and complex poses without the need for additional fine-tuning. Moreover, the proposed model is easy to use and can be considered as a plug-in module/extension to Stable Diffusion. The project website is https://boese0601.github.io/magicdance/. The code is available at https://github.com/Boese0601/MagicDance.

1. Introduction

Human motion transfer is a challenging task in computer vision. This problem involves retargeting body and facial motions, from one source image to a target image. Such methods can be used for image stylization, editing, digital human synthesis, and possibly data generation for training perception models.

[1]University of Southern California [2]ByteDance Inc. Correspondence to: Di Chang <dichang@usc.edu>.
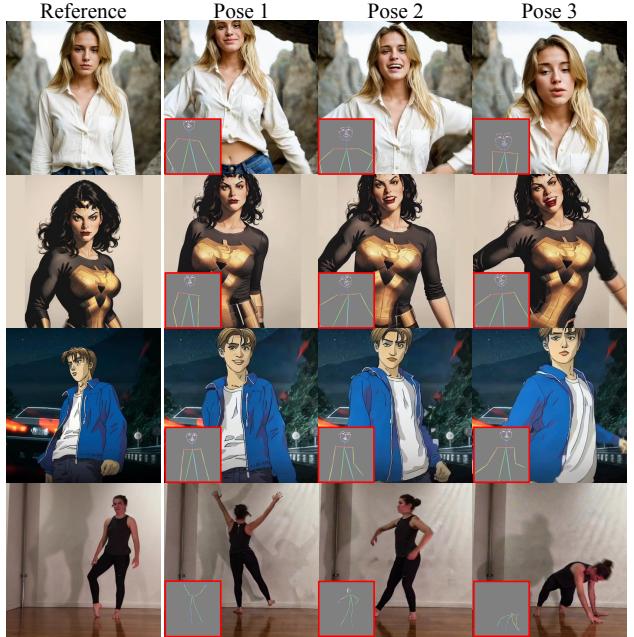
Figure 1. MagicPose can provide zero-shot and realistic human poses and facial expressions retargeting for human images of different styles and poses. A shared model is used here for in-the-wild generalization without any fine-tuning on target domains. Our proposed modules can be treated as an extension/plug-in to the original text-to-image model without modifying its pre-trained weight.

Traditionally, human motion transfer is achieved by training a task-specific generative model, such as generative adversarial networks (GANs) on specific datasets, e.g., (Siarohin et al., 2018; 2019b; Liu et al., 2019; Wei et al., 2020; Sun et al., 2022) for body pose and (Wu et al., 2020; Qiao et al., 2018; Hong et al., 2022) for facial expressions. Such methods commonly suffer from two issues: (1) they are typically dependent on an image warping module (Siarohin et al., 2018; 2019b) and hence struggle to interpolate the body parts that are invisible in the reference image due to perspective change or self-occlusion, and (2) they can hardly generalize to images that are different from the training data, greatly limiting their application scope.

Recently, diffusion models (Ho et al., 2020; Song et al., 2020; Rombach et al., 2021; Zhang et al., 2023) have exhibited impressive ability on image generation (Bertalmio et al.,

2000; Yeh et al., 2017; Lugmayr et al., 2022). By learning from web-scale image datasets, these models present powerful visual priors for different downstream tasks, such as image inpainting (Lugmayr et al., 2022; Saharia et al., 2022a; Jam et al., 2021), video generation (Ho et al., 2022; Wu et al., 2023; Singer et al., 2022), 3D generation (Poole et al., 2022; Raj et al., 2023; Shi et al., 2023) and even image segmentations (Amit et al., 2021; Baranchuk et al., 2021; Wolleb et al., 2022). Thus, such diffusion priors are great candidates for human motion transfer. Two recent studies, DreamPose (Karras et al., 2023) and DisCo (Wang et al., 2023), have attempted to adapt diffusion models for human body re-posing. However, we found that they are still limited in either generation quality, identity preservation (as discussed in Section. 5.3), or temporal consistency due to the limits in model design and training strategy. Moreover, there is no clear advantage of these methods over GAN-based methods in generalizability. For example, Disco (Wang et al., 2023) still needs to be fine-tuned to adapt to images of out-of-domain styles.

In this work, we propose MagicPose to fully exploit the potential of image diffusion priors for human pose retargeting, demonstrating superior visual quality, identity preservation ability, and domain generalizability, as illustrated in Figure. 1. Our key idea is to decompose the problem into two tasks: (1) identity/appearance control and (2) pose/motion control, which we consider useful capabilities required by image diffusion priors to achieve accurate motion transfer. Correspondingly, as shown in Figure. 2, MagicPose has two sub-modules besides the Stable Diffusion (SD) (Rombach et al., 2021): 1) Appearance Control Model that provides appearance guidance from a reference image to the SD via **Multi-Source Attention Module**, and 2) Pose ControlNet, which provides pose/expression guidance from a condition image. A multi-stage training strategy is also proposed to effectively learn these sub-modules to disentangle the appearance and pose control. Extensive experiments demonstrate the effectiveness of MagicPose which can retain well the key features of the reference identities, including skin tone and clothing, while following the pose skeleton and facial landmark inputs. Moreover, MagicPose can generalize well to unseen identities and motions without any fine-tuning. The main contributions of this work are as follows:

- An effective method (MagicPose) for human pose and expression retargeting as a plug-in for Stable Diffusion.

- Multi-Source Attention Module that offers detailed appearance guidance.

- A two stage training strategy that enables appearance-pose-disentangled generation.

- Experiment on out-of-domain data demonstrating strong generalizability of our model to diverse image styles and human poses.

- Comprehensive experiments conducted on TikTok dataset showing model's superior performance in pose retargeting.
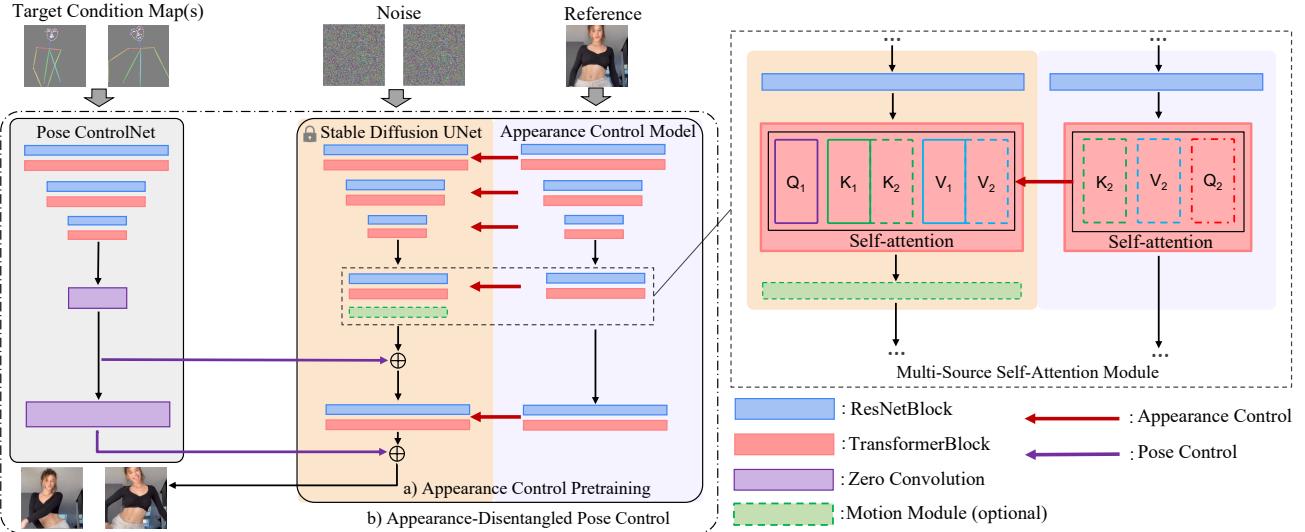
2. Related Work

2.1. Human Motion/Expression Transfer

Early work in human motion transfer primarily involved manipulation of given image sequence segments to create a desired action (Bregler et al., 1997; Efros et al., 2003; Beier & Neely, 1992). Subsequent solutions shifted their focus towards generating three-dimensional (3D) representations of human subjects and performing motion transfer within 3D environments (Cheung et al., 2004; Xu et al., 2011). However, these approaches were characterized by significant time and labor requirements. In contrast, recent advancements leverage deep learning to learn detailed representations of the input (Tulyakov et al., 2018; Kim et al., 2018; Chan et al., 2019a). This shift has facilitated motion transfer with heightened realism and increased automation. Generative Adversarial Networks (GANs) have been a clear deep learning approach to motion transfer tasks (AlBahar et al., 2021; Bregler et al., 1997; Efros et al., 2003), providing realistic image generation and Conditional GANs adding further conditioning (Mirza & Osindero, 2014). Kim et al. (Kim et al., 2018) took synthetic renderings, interior face model, and gaze map to transfer head position and facial expression from one human subject to another, presenting the results as detailed portrait videos. MoCoGAN (Tulyakov et al., 2018) also implements unsupervised adversarial training to perform motion and facial expression transfer onto novel subjects. Chan et al. (Chan et al., 2019a) further advanced this approach to full-body human motion synthesis by utilizing a video-to-video approach, taking in 2D video subjects and 2D pose stick figures to produce transferred dance sequences on new human subjects. In the sub-domain of fashion video synthesis, DreamPose (Karras et al., 2023) used SD with human image input and pose sequence input to generate videos featuring human subjects executing pose sequences with intricate fabric motion. DisCo (Wang et al., 2023), another SD-based model, contributed to the use-case of human dance generation, enabling controllable human reference, background reference, and pose maps to produce arbitrary compositions that maintain faithfulness and generalizability to unseen subjects.

2.2. Image/Video Diffusion Models

Previous research has demonstrated the effectiveness of diffusion probabilistic models (Song et al., 2021a;b) for image generation (Ramesh et al., 2022; Saharia et al., 2022b; Nichol et al., 2021). Latent diffusion models (Ho et al., 2020) have further advanced this domain by reducing computational costs by executing the diffusion step in a lower-dimensional latent space rather than pixel space. With

Figure 2. Overview of the proposed MagicPose pipeline for controllable human poses and facial expressions retargeting with motions & facial expressions transfer. The Appearance Control Model is a copy of the entire Stable-Diffusion UNet, initialized with the same weight. The Stable-Diffusion UNet is frozen throughout the training. During a) Appearance Control Pretraining, we train the appearance control model and its Multi-Source Self-Attention Module. During b) Appearance-disentangled Pose Control, we jointly fine-tune the Appearance Control Model, initialized with weights from a), and the Pose ControlNet. After these steps, an optional motion module can be integrated into the pipeline and fine-tuned for better sequential output generation quality.

customization and specification being important aspects of content generation, the text-to-image approach has gained popularity as a means of achieving controllable image generation, with notable examples such as Imagen (Saharia et al., 2022b) and SD (Rombach et al., 2021). The introduction of ControlNet (Zhang et al., 2023) extended the approach to controllable generation by introducing additional conditioning to SD models, enabling input sources such as segmentation maps, pose key points, and more. Additional condition inputs has enabled a higher degree of customization and task-specificity in the generated outputs, providing a contextual foundation for conditional image generation. With the advancement of conditional image generation, there is a natural extension towards the synthesis of dynamic visual content. Blattmann et al. (Blattmann et al., 2023) showed the use-case of latent diffusion models for video generation by integrating a temporal dimension to the latent diffusion model and further fine-tuning the model on encoded image sequences. Similar to image generation, video generation has seen both text-based as well as condition-based approaches to control the synthesized output.

3. Preliminary

Latent Diffusion Models (Rombach et al., 2021) (LDM) (Rombach et al., 2021), represent those diffusion models uniquely designed to operate within the latent space facilitated by an autoencoder, specifically $\mathcal{D}(\mathcal{E}(\cdot))$. A notable instance of such models is the Stable Diffusion (SD) (Rombach et al., 2022), which integrates a Vec-

tor Quantized-Variational AutoEncoder (VQ-VAE) (Van Den Oord et al., 2017) and a U-Net structure (Ronneberger et al., 2015). SD employs a CLIP-based transformer architecture as a text encoder (Radford et al., 2021) to convert text inputs into embeddings, denoted by c_{text}. The training regime of SD entails presenting the model with an image I and a text condition c_{text}. This process involves encoding the image to a latent representation $z_0 = \mathcal{E}(I)$ and subjecting it to a predefined sequence of T diffusion steps governed by a Gaussian process. This sequence yields a noisy latent representation z_T, which approximates a standard normal distribution $\mathcal{N}(0, 1)$. SD's learning objective is iteratively denoising z_T back into the latent representation z_0, formulated as follows:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(I), c_{\text{text}}, \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, c_{\text{text}})\|_2^2 \right] \quad (1)$$

where ϵ_θ is the UNet with learnable parameters θ and $t = 1, ..., T$ denotes the time-step embedding in denoising. These modules employ convolutional layers, specifically Residual Blocks (ResNetBlock), and incorporate both self- and cross-attention mechanisms through Transformer Blocks (TransformerBlock).

ControlNet is an extension of SD that is able to control the generated image layout of SD without modifying the original SD's parameters. It achieves this by replicating the encoder of SD to learn feature residuals for the latent feature maps in SD. It has been successfully applied to different controlled image generation tasks including pose-conditioned human image generation (Zhang et al., 2023).

4. MagicPose

Given a image I_R with a person in it, the objective of MagicPose to re-pose the person in the given image to the target pose $\{P, F\}$, where P is the human pose skeleton and F is the facial landmarks. Such a pipeline can be decomposed into two sub-tasks: (1) keeping and transferring the appearance of the human individual and background from reference image and (2) controlling generated images with the pose and expression defined by $\{P, F\}$. To ensure the generazability of the model, MagicPose is designed to inherit the structures and parameters as much as possible from pre-trained stable diffusion models. To this end, we propose an attention-based appearance controller by replicating the structures of the original UNet. An additional ControlNet is then trained jointly to control the pose and expression of the person. We train MagicPose on human video datasets where image pairs of the same person but different poses are available. Then during testing, the reference I_R and poses $\{P, F\}$ could come from different sources for pose transfer. The overview of the proposed method (MagicPose) is illustrated in Figure. 2. We first presents our preliminary experiments in terms of appearance control in Sec. 4.1, which motivates us to propose the Appearance Control Module as elaborated in Sec. 4.2. Then, Sec. 4.3 presents the fine-tuning of the Appearance-disentangled Pose Control.

4.1. Exploration of Appearance Control Mechanism

We first evaluated vanilla ControlNet for appearance control. As shown in Figure 3, we found that ControlNet is not able to maintain the appearance when generating human images of different poses, making it unsuitable for the re-targeting task. On the other side, recent studies (Cao et al., 2023; Lin et al., 2023b; Zhang) have found that self-attention layers in the diffusion models is highly relevant to the appearance of the generated images. Inspired by them, we conduct an experiment on self-attention for zero-shot appearance control, where the reference image and the noisy image are both forwarded through the diffusion UNet with their self-attention layers connected. A critical observation is that is such an architecture can naturally lead to an appearance resemblance between the two images, even without any fine-tuning (Figure 3 connected attention). One plausible explanation is that self-attention layers in the UNet plays an important role to transmit the appearance information spatially and hence it could serve as a deformation module to generate similar images with different geometric structures. From another perspective, such an forward process mimics the generation of two image as a single one, and thus, their appearance tend to be similar. However, the problem with such a zero-shot approach is that the generation results are not stable.

Figure 3. Identity and appearance control ability comparison between different architectural designs.

4.2. Appearance Control Pretraining

Given the above observations, we introduce our Appearance Control Model, which inherits the structure and capability of the zero-shot attention-based control but further extends its stability by introducing task-specific parameters. In particular, it is designed as an auxiliary UNet branch to provide layer-by-layer attention guidance. As shown in Figure. 2, our Appearance Control Model consists of another trainable copy of the original SD-UNet, which connects to the Appearance Control Model by sharing the key and value through the Multi-Source Self Attention Module.

Formally, the calculation of self-attention in TransformerBlocks of SD-UNet can be written as:

$$Self\_Attn = softmax(\frac{Q \cdot K^T}{\sqrt{d}}) \cdot V \qquad (2)$$

where Q, K, V are query, key, and value. d denotes the dimension of the key and query. In our Multi-Source Self Attention Module, we concatenate the key-value pairs from the Appearance Control Model with SD-UNet together as new key-value pairs and calculate the attention similar to Eq. 2:

$$Our\_Attn = softmax(\frac{Q_1 \cdot (K_1 \oplus K_2)^T}{\sqrt{d}}) \cdot (V_1 \oplus V_2) \qquad (3)$$

where Q_1, K_1, V_1 are query, key, and value from self-attention layers in the TransformerBlocks of SD-UNet and K_2, V_2 are from the Appearance Control Model. $\oplus$ refers to vector concatenation. In essence, the only modification for the SD-UNet is to change the calculation of self-attention from Eq. 2 to Eq. 3.

In order to maintain the generalizability of the SD, in the first training stage (**Appearance Control Pre-training**), we fix the original UNet and only train the Appearance Control module. The pose ControlNet is not included in this stage. The objective of Appearance Control Pretraining is:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(I), A_\theta(I_R), \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \epsilon_\theta(z_t, t, A_\theta(I_R))\|_2^2 \right] \quad (4)$$

where A_θ is the Appearance Control Model taking reference image I_R as input. ϵ_θ is the SD-UNet, which takes the noisy latent z_t, denoising step t and Our_Attn as inputs.

4.3. Appearance-disentangled Pose Control

To control the pose in the generated images, a naive solution directly integrates the pre-trained OpenPose ControlNet model (Zhang et al., 2023) with our pre-trained

Appearance Control Model without fine-tuning. However, our experiments indicate that such a combination struggles with appearance-independent pose control, leading to severe errors between the generated poses and the input poses. To address the issue, we reuse our pre-trained Appearance Control module to disentangle the pose ControlNet from appearance information. In particular, assuming the Appearance Controller already provides a complete guidance for the generated image's appearance, we fine-tune the Pose ControlNet jointly with our Appearance Control Model. As such, Pose ControlNet exclusively modulates the pose attributes of the human, while the Appearance Control Model focuses on appearance control. Specifically, we fine-tune MagicPose with an objective similar to latent diffusion training (Rombach et al., 2022):

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(I), A_\theta(I_R), P_\theta(I_C), \epsilon \sim \mathcal{N}(0,1), t} \left[\| \epsilon - \epsilon_\theta(z_t, t, A_\theta(I_R), P_\theta(I_C)) \|_2^2 \right] \quad (5)$$

where P_θ is the Pose ControlNet taking poses I_C as inputs.

4.4. Comparison with prior works

Comparison with ControlNet In terms of control objective, ControlNet was introduced to control the **geometrical** shape and **structural** information in the text-to-image model, while our proposed Appearance Control Model aims to provide **identity** and **appearance** information for the generated subject regardless of the given text. In term of structure design, ControlNet copies the encoder and middle blocks of SD-UNet, whose output feature maps are added to the decoder of SD-UNet to realize **pose** control. On the other side, the proposed Appearance Control Model replicates a whole UNet model to controls the generation process of pre-trained diffusion model via attention layers, enabling more flexible information interchange among distant pixels. And therefore it is more suited for the task of pose retargeting.

Comparison with MasaCtrl MasaCtrl (Cao et al., 2023) is an inference-only model and requires text as appearance guidance input. MagicPose is a pipeline that can be fine-tuned on customized data and provide consistent identity-preserving generation without any text prompt.

MasaCtrl also utilizes parallel UNet architecture, however, there are several major differences between MasaCtrl and MagicPose. 1) The self-attention key-value pairs from the reference branch in MasaCtrl **replace** those in SD-UNet, while MagicPose's **Multi-Source** Self-Attention Module concatenates the key-value pairs from both SD-UNet and appearance control model. 2) The replacement of key-value pairs only exists in **decoder** after specific denoising timestep S and after specific layer index L in MasaCtrl, while MagicPose manipulates all self-attention layers in the encoder, middle block, and the decoder. For both inference and train-

ing, the manipulation always exists regardless of timesteps. This ensures our model learns both **encoding** appearance from the reference image (encoder) and **generating** identity-preserving results(decoder) from customized data. 3) Text is required as input to generate an extraction **mask** for key-value pairs in MasaCtrl, while MagicPose doesn't need any additional text information so that appearance information **only** comes from the reference image. This further helps our model to strictly preserve the identity and make our approach suitable for the motion retargeting (usually there's no text prompt for this task, since pose map defines the human motion and reference image controls human appearance and background).

Comparison with Reference Only ControlNet Reference Only ControlNet (Zhang) is another inference-only model and does not have a parallel UNet architecture like the original ControlNet (Zhang et al., 2023). Reference Only ControlNet shares the same architecture and weight as SD-UNet and first takes a noisy reference image as the only input. During the denoising process of the reference image, the query key and value from the self-attention layers are saved temporarily in a *cache*. Then the text is fed as input to the SD-UNet again and the denoising process yields image generation output, while the self-attention layers take query key value from the *cache*. In contrast, MagicPose introduces a **trainable** parallel UNet architecture without text input and the appearance control model implicitly learns how to provide identity control for SD-UNet with Multi-Sourse Self-Attnetion Module during fine-tuning.

5. Experiments

5.1. Datasets

TikTok (Jafarian & Park, 2021) dataset consists of 350 single-person dance videos (with video length of 10-15 seconds). Most of these videos contain the face and ***upper-body*** of a human. For each video, we extract frames at 30fps and run OpenPose (Cao et al., 2019; Simon et al., 2017; Cao et al., 2017; Wei et al., 2016) on each frame to infer the human pose skeleton, facial landmarks, and hand poses. 335 videos are sampled as the training split. We follow (Wang et al., 2023) and use their 10 TikTok-style videos depicting different people from the web as the testing split.

Everybody Dance Now (Chan et al., 2019b) consists of **full-body** videos of five subjects. Experiments on this dataset aim to test our method's generalization ability to in-the-wild, full-body motions.

Self-collected Out-of-Domain Images come from online resources. We use them to test our method's generalization ability to in-the-wild appearance.
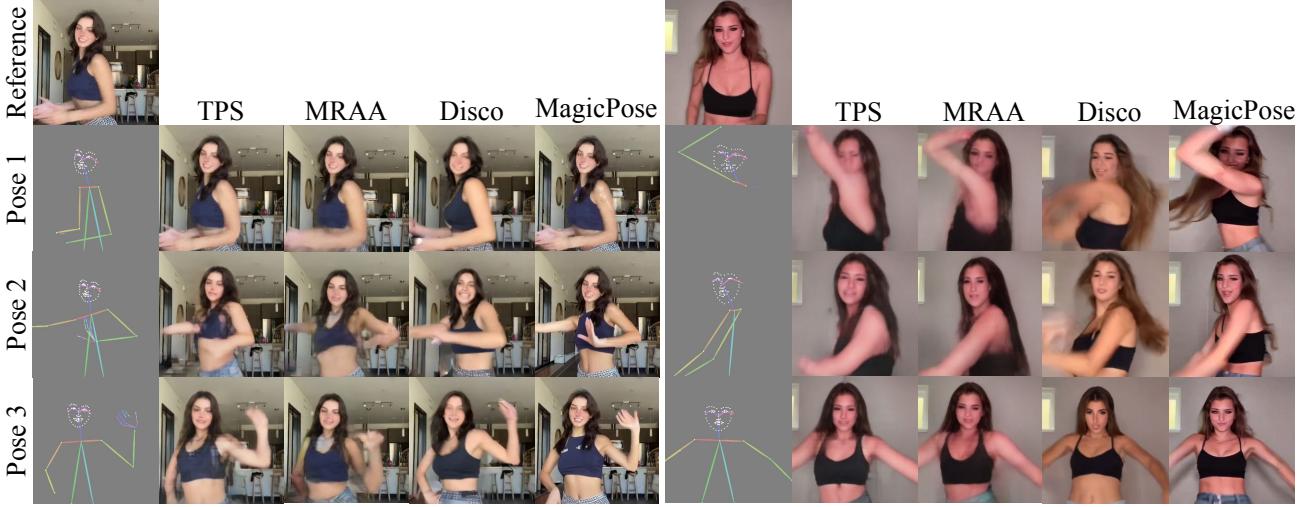
Figure 4. Qualitative comparison of human poses and facial expressions retargeting between TPS (Zhao & Zhang, 2022), MRAA (Siarohin et al., 2019a), Disco (Wang et al., 2023) and MagicPose. Previous methods suffer from inconsistent facial expressions and human pose identity.

Table 1. Quantitative comparisons of MagicPose with the recent SOTA methods DreamPose (Karras et al., 2023) and Disco (Wang et al., 2023). ↓ indicates that the lower the better, and vice versa. Methods with ∗ directly use the target image as the input, including more information compared to the OpenPose (Cao et al., 2019; Simon et al., 2017; Cao et al., 2017; Wei et al., 2016). † represents that Disco (Wang et al., 2023) is pre-trained on other datasets (Fu et al., 2022; Ge et al., 2019; Schuhmann et al., 2021; Lin et al., 2014) more than our proposed MagicPose, which uses only 335 video sequences in the TikTok (Jafarian & Park, 2021) dataset for pretraning and fine-tuning. **Face-Cos** represents the cosine similarity of the extracted feature by AdaFace (Kim et al., 2022) of face area between generation and ground truth image.

Method	Image						Video
	FID ↓	SSIM ↑	PSNR ↑	LPIPS ↓	L1 ↓	Face-Cos ↑	FID-VID ↓
FOMM∗ (Siarohin et al., 2019a)	85.03	0.648	29.01	0.335	3.61E-04	0.190	90.09
MRAA∗ (Siarohin et al., 2021)	54.47	0.672	29.39	0.296	3.21E-04	0.337	66.36
TPS∗ (Zhao & Zhang, 2022)	53.78	0.673	29.18	0.299	3.23E-04	0.280	72.55
DreamPose (Karras et al., 2023)	72.62	0.511	28.11	0.442	6.88E-04	0.085	78.77
DisCo (Wang et al., 2023)	50.68	0.648	28.81	0.309	4.27E-04	-	69.68
DisCo† (Wang et al., 2023)	30.75	0.668	29.03	**0.292**	3.78E-04	0.166	59.90
MagicPose	**25.50**	**0.752**	**29.53**	**0.292**	**0.81E-04**	**0.426**	**46.30**

5.2. Implementation Details

We first pre-train the appearance control model on 8 NVIDIA A100 GPUs with batch size 192 for 10k steps with image size 512×512 and learning rate 0.0001. We then jointly fine-tune the appearance and pose control model on 8 NVIDIA A100 GPUs with batch size 16 for 20K steps. The Stable-Diffusion UNet weight is frozen during all experiments. During training, we randomly sampled the two frames of the video as the reference and target. Both reference and target images are randomly cropped at the same position along the height dimension with the aspect ratio of 1 before resizing to 512×512. For evaluation, we apply center cropping instead of random cropping. We initialize the U-Net model with the pre-trained weights of Stable-Diffusion Image Variations (Justin & Lambda, 2022). The

Appearance Control Model branch is initialized with the same weight as the U-Net model. After Appearance Control pre-training, we initialize the U-Net and Appearance Control Model branch with the previous pre-trained weights and initialize the Pose ControlNet branch with the weight from (Zhang et al., 2023), for joint fine-tuning. After these steps, an optional motion module can be further fine-tuned.

5.3. Qualitative and Quantitative Comparison

We conduct a comprehensive evaluation of TikTok (Jafarian & Park, 2021) in comparison to established motion transfer methodologies, including FOMM (Siarohin et al., 2019a), MRAA (Siarohin et al., 2021), and TPS (Zhao & Zhang, 2022), as well as recent advancements in the field such as Disco (Wang et al., 2023). Disco (Wang et al., 2023) lever-

Table 2. User study of MagicPose. We collect the number of votes from 100 participants for eight subjects in the test set. The participants found that MagicPose preserves the best identity and appearance information in pose and facial expression retargeting.

Method	Average
MRAA (Siarohin et al., 2019a)	4%
FOMO (Siarohin et al., 2021)	3%
TPS (Zhao & Zhang, 2022)	4%
Disco (Wang et al., 2023)	16%
MagicPose	**73%**

ages a CLIP encoder to integrate appearance information from the reference image into the Transformer Blocks of the Stable-Diffusion UNet and Pose ControlNet while retaining OpenPose (Cao et al., 2019; Simon et al., 2017; Cao et al., 2017; Wei et al., 2016) as the pose condition. Though OpenPose has the limitation of incomplete detection of the human skeleton (More details in supplementary), we follow previous work and adopt OpenPose as the pose detector. For image quality evaluation, we adhere to the methodology outlined in Disco (Wang et al., 2023) and report metrics such as frame-wise FID (Heusel et al., 2017), SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), PSNR (Hore & Ziou, 2010), and L1. In addition to these established metrics, we introduce a novel image-wise metric called **Face-Cos**, which stands for Face Cosine Similarity. This metric is designed to gauge the model's capability to preserve the identity information of the reference image input. To compute this metric, we first align and crop the facial region in both the generated image and the ground truth. Subsequently, we calculate the cosine similarity between the extracted feature by AdaFace (Kim et al., 2022), frame by frame of the same subject in the test set, and report the average value as the **Face-Cos** score. The experimental results, in Table 1, unequivocally establish MagicPose as a front-runner, showcasing substantial improvements across all metrics compared to alternative approaches. Notably, MagicPose attains a Face-Cos score of ~ 0.426, representing a substantial **+0.260** enhancement over Disco. This performance shows MagicPose's robust capacity to preserve identity information. The substantial performance improvement over the state-of-the-art methods demonstrates the superiority of the MagicPose framework. We qualitatively compare MagicPose to previous methods (Zhao & Zhang, 2022; Siarohin et al., 2019a; Wang et al., 2023) in Figure 4. TPS (Zhao & Zhang, 2022), MRAA (Siarohin et al., 2019a), and Disco (Wang et al., 2023) suffer from inconsistent facial expressions and human appearances. Please check the supplementary materials to see more examples of real-human poses and facial expressions re-targeting.

User Study We provide a user study for comparison between MagicPose and previous works (Siarohin et al., 2019a; 2021; Zhao & Zhang, 2022; Wang et al., 2023).

We collect reference images, openpose conditions, and pose retargeting results from prior works and MagicPose of 8 subjects in the test set. For each subject, we visualize different human poses and facial expressions and ask 100 users to choose **only one** method, which preserves the best identity and appearance information for each subject. We present the averaged vote result in Table. 2. Visualization examples and detailed user studies can be found in supplementary material.

5.4. Ablation Analysis

In this section, a comprehensive ablation analysis of MagicPose on the TikTok (Jafarian & Park, 2021) dataset is presented. The impact of various training and inference configurations within MagicPose is systematically analyzed in Table 3. We examine the proposed Appearance Control Model and its Multi-Source Self-Attention Module, specifically assessing their contributions when omitted. The absence of Appearance Control Pretraining and Appearance-disentangled Pose Control reveals the significance of these components, which can be observed in Figure. 5 as well. Notably, the introduction of Appearance Control Pretraining markedly enhances generation quality, evidenced by a substantial increase of **+944.73%** in Face-Cos and **+149.82%** in SSIM. Additionally, the implementation of Appearance-disentangled Pose Control demonstrates its efficacy, yielding improvements of **+7.30%** in Face-Cos and **+3.43%** in SSIM. Furthermore, we highlight the necessity of incorporating the data augmentation technique of randomly masking facial landmarks and hand poses during training. This is particularly crucial due to the occasional limitations of OpenPose (Cao et al., 2019; Simon et al., 2017; Cao et al., 2017; Wei et al., 2016) in providing complete and accurate detection of hand pose skeletons and facial landmarks, which can result in artifacts in generated images. Therefore, to enhance the robustness of MagicPose against incomplete human pose estimations by OpenPose (Cao et al., 2019; Simon et al., 2017; Cao et al., 2017; Wei et al., 2016), this data augmentation strategy is proposed and leads to incremental improvements in Face-Cos and SSIM by **+2.20%** and **+0.13%**, respectively. Moreover, the application of classifier-free guidance (Image-CFG) in the training process, as discussed in prior work (Wang et al., 2023; Ho, 2022; Lin et al., 2023a; Balaji et al., 2022; Dao et al., 2022) on diffusion models, further augments the quality of generation. The implementation of Image-CFG enhances Face-Cos by **+56.62%** and SSIM by **+14.11%**, underscoring its value in the image generation context.

5.5. Generalization Ability

It is also worth highlighting that MagicPose can generalize to out-of-domain reference images of unseen styles and poses with surprisingly good appearance controllabil-

Figure 5. Ablation Analysis of MagicPose. The proposed Appearance Control Pretraining and Appearance-disentangled Pose Control provide better identity control and generation quality effectively.

Table 3. Ablation Analysis of MagicPose with different training and inference settings. **App-Pretrain** stands for Appearance Control Pretraining through Multi-Source Attention Module and **Disentangle** denotes Appearance-disentangled Pose Control. **Image-CFG** denotes classifier free guidance. **Data Aug** indicates the model is trained with data augmentation of random masking of facial landmarks and hand poses.

| App-Pretrain | Disentangle | Image CFG. | Data Aug. | Image | | | | | | Video |
				FID ↓	SSIM ↑	PSNR ↑	LPIPS ↓	L1 ↓	Face-Cos ↑	FID-VID ↓
✗	✗	✓	✓	288.64	0.291	27.85	0.695	2.69E-04	0.038	382.10
✓	✗	✓	✓	35.76	0.727	28.79	0.321	0.97E-04	0.397	65.72
✓	✓	✗	✓	61.33	0.659	28.31	0.357	1.29E-04	0.272	98.96
✓	✓	✓	✗	28.71	0.751	29.14	0.296	0.86E-04	0.417	47.26
✓	✓	✓	✓	**25.50**	**0.752**	**29.53**	**0.292**	**0.81E-04**	**0.426**	**46.30**

Table 4. Quantitative evaluation of generalization ability of MagicPose. MagicPose† denotes the pipeline is directly evaluated on test set of Everybody Dance Now (Chan et al., 2019b) after being trained on TikTok (Jafarian & Park, 2021), and MagicPose‡ represents the pipeline is further fine-tuned on Everybody Dance Now (Chan et al., 2019b) train set and evaluated on test set.

| Method | Subject1 | | Subject2 | | Subject3 | | Subject4 | | Subject5 | | Average | |
	FID ↓	PSNR ↑	FID ↓	PSNR ↑	FID ↓	PSNR ↑	FID ↓	PSNR ↑	FID ↓	PSNR ↑	FID ↓	PSNR ↑
MagicPose†	22.59	**30.67**	**22.21**	**30.13**	35.43	**29.35**	31.72	29.53	31.24	28.48	28.64	29.63
MagicPose‡	**22.50**	**30.67**	22.61	28.40	**27.38**	29.10	36.73	**33.95**	**21.99**	**30.94**	**26.24**	**30.61**

ity, even without any further fine-tuning on the target domain. Figure. 6 compares the zero-shot results of applying TPS (Zhao & Zhang, 2022), MRAA (Siarohin et al., 2019a), Disco (Wang et al., 2023) and MagicPose to out-of-domain images, whose visual style is distinct from corresponding training data of the real-human upper-body images. For real-human reference images, we observe that most of the human subjects from TikTok (Jafarian & Park, 2021) dataset and the self-collected test set of Disco (Wang et al., 2023) are young women. So we test our method on more in-the-wild real-human examples, e.g. elder people, in Figure 7. We also evaluate the in-the-wild motions generalization ability of MagicPose on Everybody Dance Now (Chan et al., 2019b), which is a full-body dataset, in contrast to the upper-body images used in the TikTok dataset. We directly apply MagicPose to such full-body reference images and visualize the qualitative results in Figure. 8 and provide a quantita-

tive evaluation in Table. 4. Experiments show that Magic-Pose generalizes surprisingly well to full body images even though it has never been trained on such data. Furthermore, better quality of generation can be achieved after fine-tuning on specific datasets as well. More visualizations of zero-shot Animation and results on Everybody Dance Now (Chan et al., 2019b) can be found in the supplementary materials.

6. Conclusion

In this work, we propose MagicPose, a novel approach in the realm of realistic human poses and facial expressions retargeting. By seamlessly incorporating motion and facial expression transfer and enabling the generation of consistent in-the-wild animations without any further fine-tuning, Mag-icPose shows a significant advancement over prior methods. Notably, our approach demonstrates a superior capacity to
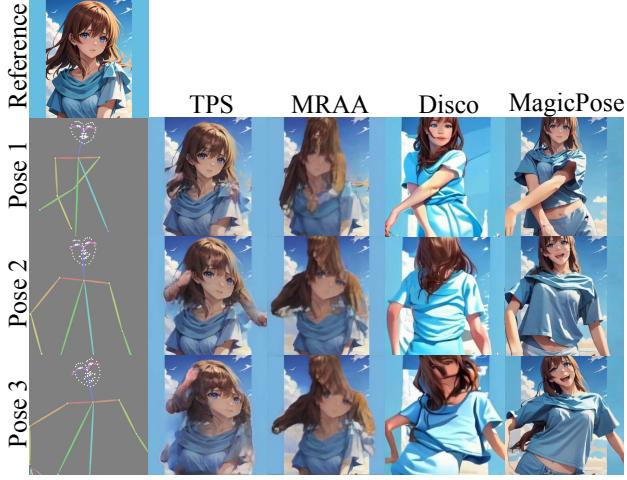
Figure 6. Comparison of zero-shot pose and facial expression retargeting on out-of-domain image.

Figure 7. Visualization of zero-shot pose and facial expression retargeting on in-the-wild real-human with different ethnicity and age from training data (Tiktok).

generalize over diverse human identities and complex motion sequences. Moreover, MagicPose boasts a practical implementation as a plug-in module or extension compatible with existing models such as Stable Diffusion (Rombach et al., 2022). This combination of innovation, efficiency, and adaptability establishes MagicPose as a promising tool in the field of poses and facial expressions retargeting.

Acknowledgements
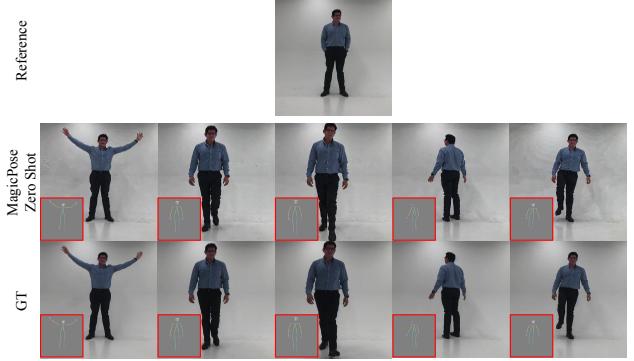
Figure 8. Visualization of zero-shot Human Motion and Facial Expression Transfer on EverybodyDancNow Dataset (Chan et al., 2019b). MagicPose can generalize perfectly to in-the-wild human motions.

Impact Statement

The proposed MagicPose for retargeting human poses and facial expressions has a wide range of applications. It can significantly improve communication in digital environments, enabling individuals to express themselves more effectively through avatars or digital characters, thereby enhancing interactions in virtual meetings, online classrooms, and social networking platforms. Additionally, MagicPose has the potential to revolutionize entertainment and media production, allowing for the creation of more lifelike and expressive characters in movies, video games, and animations, consequently fostering more immersive storytelling experiences and increased audience engagement. The experiment demonstrates our model can generalize across different real-human ethnicities and ages, and even to out-of-domain images, e.g., cartoon-style images, painting-style images, and AI-generated images.

Potential Negative Social Impact: The method can potentially be used in a malicious way, e.g., making fake animated videos of people, which could be used in fraud. To avoid the potential misuse of such technologies, It is essential to employ several solutions like digital watermarking and detection algorithms, enact and enforce stringent legal measures, enhance public awareness and education on media literacy, and establish ethical guidelines within the tech industry. This involves collaboration among tech companies, governments, educators, and the public to create a safer digital environment and mitigate the risks of fraudulent AI-generated content

References

AlBahar, B., Lu, J., Yang, J., Shu, Z., Shechtman, E., and Huang, J.-B. Pose with Style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics*, 2021.

Amit, T., Shaharbany, T., Nachmani, E., and Wolf, L. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., Karras, T., and Liu, M.-Y. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *ArXiv*, abs/2211.01324, 2022.

Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V., and Babenko, A. Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv:2112.03126*, 2021.

Beier, T. and Neely, S. Feature-based image metamorphosis. In *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '92, pp. 35–42, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 0897914791. doi: 10.1145/133994.134003. URL https://doi.org/10.1145/133994.134003.

Bertalmio, M., Sapiro, G., Caselles, V., and Ballester, C. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 417–424, 2000.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Bregler, C., Covell, M., and Slaney, M. Video rewrite: Driving visual speech with audio. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, pp. 353–360, USA, 1997. ACM Press/Addison-Wesley Publishing Co. ISBN 0897918967. doi: 10.1145/258734.258880. URL https://doi.org/10.1145/258734.258880.

Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., and Zheng, Y. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. *arXiv preprint arXiv:2304.08465*, 2023.

Cao, Z., Simon, T., Wei, S.-E., and Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., and Sheikh, Y. A. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

Chan, C., Ginosar, S., Zhou, T., and Efros, A. A. Everybody dance now. In *IEEE International Conference on Computer Vision (ICCV)*, 2019a.

Chan, C., Ginosar, S., Zhou, T., and Efros, A. A. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5933–5942, 2019b.

Cheung, G., Baker, S., Hodgins, J., and Kanade, T. Markerless human motion transfer. In *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004.*, pp. 373–378, 2004. doi: 10.1109/TDPVT.2004.1335262.

Dao, T., Fu, D. Y., Ermon, S., Rudra, A., and Ré, C. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022.

Efros, Berg, Mori, and Malik. Recognizing action at a distance. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 726–733 vol.2, 2003. doi: 10.1109/ICCV.2003.1238420.

Fu, J., Li, S., Jiang, Y., Lin, K.-Y., Qian, C., Loy, C. C., Wu, W., and Liu, Z. Stylegan-human: A data-centric odyssey of human generation. In *ECCV*, 2022.

Ge, Y., Zhang, R., Wang, X., Tang, X., and Luo, P. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *CVPR*, 2019.

Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.

Ho, J. Classifier-free diffusion guidance. *ArXiv*, abs/2207.12598, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020.

Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D. P., Poole, B., Norouzi, M., Fleet, D. J., et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Hong, F.-T., Zhang, L., Shen, L., and Xu, D. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3397–3406, 2022.

Hore, A. and Ziou, D. Image quality metrics: Psnr vs. ssim. In *ICPR*, 2010.

Jafarian, Y. and Park, H. S. Learning high fidelity depths of dressed humans by watching social media dance videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12753–12762, June 2021.

Jam, J., Kendrick, C., Walker, K., Drouard, V., Hsu, J. G.-S., and Yap, M. H. A comprehensive review of past and present image inpainting methods. *Computer vision and image understanding*, 203:103147, 2021.

Justin, P. and Lambda. Stable Diffusion Image Variations. https://huggingface.co/lambdalabs/sd-image-variations-diffusers, 2022.

Karras, J., Holynski, A., Wang, T.-C., and Kemelmacher-Shlizerman, I. Dreampose: Fashion image-to-video synthesis via stable diffusion. *arXiv preprint arXiv:2304.06025*, 2023.

Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollöfer, M., and Theobalt, C. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4): 163, 2018.

Kim, M., Jain, A. K., and Liu, X. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Lin, S., Liu, B., Li, J., and Yang, X. Common diffusion noise schedules and sample steps are flawed. 2023a.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *ECCV*, 2014.

Lin, Y., Han, H., Gong, C., Xu, Z., Zhang, Y., and Li, X. Consistent123: One image to highly consistent 3d asset using case-aware diffusion priors. *arXiv preprint arXiv:2309.17261*, 2023b.

Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., and Gao, S. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5904–5913, 2019.

Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., and Van Gool, L. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Mirza, M. and Osindero, S. Conditional generative adversarial nets, 2014.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, abs/2112.10741, 2021. URL https://arxiv.org/abs/2112.10741.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., and Wang, H. Geometry-contrastive gan for facial expression transfer. *arXiv preprint arXiv:1802.01822*, 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Raj, A., Kaza, S., Poole, B., Niemeyer, M., Ruiz, N., Mildenhall, B., Zada, S., Aberman, K., Rubinstein, M., Barron, J., et al. Dreambooth3d: Subject-driven text-to-3d generation. *arXiv preprint arXiv:2303.13508*, 2023.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., and Norouzi, M. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022a.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding, 2022b.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Shi, Y., Wang, P., Ye, J., Long, M., Li, K., and Yang, X. Mv-dream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

Siarohin, A., Sanguineto, E., Lathuilière, S., and Sebe, N. Deformable gans for pose-based human image generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., and Sebe, N. First order motion model for image animation. *NeurIPS*, 2019a.

Siarohin, A., Lathuilière, S., Sanguineto, E., and Sebe, N. Appearance and pose-conditioned human image generation using deformable gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019b.

Siarohin, A., Woodford, O. J., Ren, J., Chai, M., and Tulyakov, S. Motion representations for articulated animation. In *CVPR*, 2021.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

Sun, Y.-T., Fu, Q.-C., Jiang, Y.-R., Liu, Z., Lai, Y.-K., Fu, H., and Gao, L. Human motion transfer with 3d constraints and detail enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4682–4693, 2022.

Tulyakov, S., Liu, M.-Y., Yang, X., and Kautz, J. MoCoGAN: Decomposing motion and content for video generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1526–1535, 2018.

Van Den Oord, A., Vinyals, O., et al. Neural discrete representation learning. *NeurIPS*, 2017.

Wang, T., Li, L., Lin, K., Lin, C.-C., Yang, Z., Zhang, H., Liu, Z., and Wang, L. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Wei, D., Xu, X., Shen, H., and Huang, K. Gac-gan: A general method for appearance-controllable human video motion transfer. *IEEE Transactions on Multimedia*, 23:2457–2470, 2020.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. Convolutional pose machines. In *CVPR*, 2016.

Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., and Cattin, P. C. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, pp. 1336–1348. PMLR, 2022.

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X., and Shou, M. Z. 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*, pp. 7623–7633, 2023.

Wu, R., Zhang, G., Lu, S., and Chen, T. Cascade ef-gan: Progressive facial expression editing with local focuses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Xu, F., Liu, Y., Stoll, C., Tompkin, J., Bharaj, G., Dai, Q., Seidel, H.-P., Kautz, J., and Theobalt, C. Video-based characters: Creating new human performances from a multi-view video database. *ACM Trans. Graph.*, 30(4), jul 2011. ISSN 0730-0301. doi: 10.1145/2010324.1964927. URL https://doi.org/10.1145/2010324.1964927.

Yeh, R. A., Chen, C., Yian Lim, T., Schwing, A. G., Hasegawa-Johnson, M., and Do, M. N. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5485–5493, 2017.

Zhang, L. [major update] reference-only control · mikubill/sd-webui-controlnet · discussion #1236. URL https://github.com/Mikubill/sd-webui-controlnet/discussions/1236.

Zhang, L., Rao, A., and Agrawala, M. Adding conditional control to text-to-image diffusion models, 2023.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Zhao, J. and Zhang, H. Thin-plate spline motion model for image animation. In *CVPR*, 2022.

A. Detailed User Study

In this section, we provide a comprehensive user study for qualitative comparison between MagicPose and previous works (Siarohin et al., 2019a; 2021; Zhao & Zhang, 2022; Wang et al., 2023). As we mentioned in the experiment, we collect reference images, openpose conditions, and pose retargeting results from prior works and MagicPose of 8 subjects in the test set. For each subject, we visualize different human poses and facial expressions. Some examples are shown in Figure. 10 and Figure. 11. The methods are anonymized as A, B, C, D, E, and the order of the generated image from the corresponding method is randomized in each subject comparison. We ask 100 users to choose **only one** method which preserves the best identity and appearance information for each subject. We present the full result in Table. 5.

Participants We use **Prolific**, an online platform designed to connect researchers with participants for academic studies and market research for all our user studies. The participants are English-speaking random users verified by this platform without prior knowledge of computer vision.

Criteria for Judgment Since our paper focuses on the motion retargeting task, the criteria for evaluation are 1) The appearance (Face, Clothes on the body, Background) of the generation should strictly match the given reference image input. 2) The motions and facial expressions of the generation should strictly match the given pose condition map input. As mentioned in Section A of our Supplementary Material, We ask the participants to choose **the only one** method which preserves the best identity information for each video subject. In order to compare in a professional manner, all the methods are anonymized as A, B, C, D, E, and the order of the generated image from the corresponding method is randomized in each subject comparison, e.g. For comparison of video subject 1, A,B,C,D,E correspond to MRAA, FOMO, TPS, Disco, MagicPose. For comparison of video subject 2, A,B,C,D,E correspond to Disco, MRAA, MagicPose, FOMO, TPS.

Results In order to make the conclusion from our user study statistically significant, we recruited 100 participants from Prolific. The result is presented in Table. 5 and we can conclude that MagicPose preserves the identity the best compared to prior works.

Statistical Analysis We perform a chi-square test on our results. For each vote by an participant, we label the chosen method as 1 and the rest methods as 0. Results of such a test are shown in Table. 5, we compare five methods (A, B, C, D, E) on eight video subjects with the following steps:

1. State the *Null Hypothesis*: There is no association between the video subjects and the choice of method.

The distribution of votes for each method is the same across all groups, meaning any observed differences are due to chance.

2. Compute chi-square statistic and p-value: Chi-square statistic: 116.02. p-value: 1.11×10^{-12}. Degrees of freedom: 28.

3. Conclusion and Discussion: Given the extremely small p-value (much less than 0.05), we can reject the Null Hypothesis. This indicates that there is a statistically significant association between the video subjects and the choice of method. In simpler terms, the differences in vote distribution for the methods across the eight groups are unlikely to have occurred by chance, pointing towards a significant preference pattern among the groups. We conclude that the participants indeed prefer our proposed MagicPose more than other methods.

B. Additional Visulizations

B.1. TikTok

We provide more visualizations on the test set of the experiments on TikTok (Jafarian & Park, 2021) in Figure. 10, Figure. 11, Figure. 12, Figure. 13, and Figure. 14.

Figure 9. Visualization of generalization to unseen image styles that are different from our training set (Tiktok).

B.2. EverybodyDanceNow

We provide more visualizations of zero-shot generation on Everybody Dance Now dataset (Chan et al., 2019b) in Figure. 15 and Figure. 16.

Table 5. The user study with 100 participants. We collect the number of votes for eight video subjects from test set by five methods and report the percentage. Our MagicPose preserves the best identity information in pose and facial expression retargeting on all subjects.

Method	Subject1	Subject2	Subject3	Subject4	Subject5	Subject6	Subject7	Subject8	Average
MRAA (Siarohin et al., 2019a)	8%	6%	0%	5%	2%	2%	8%	4%	4%
FOMO (Siarohin et al., 2021)	3%	1%	3%	1%	1%	0%	5%	8%	3%
TPS (Zhao & Zhang, 2022)	4%	16%	0%	4%	2%	3%	4%	2%	4%
Disco (Wang et al., 2023)	12%	3%	9%	18%	5%	20%	33%	27%	16%
MagicPose	**73%**	**74%**	**88%**	**72%**	**90%**	**75%**	**50%**	**59%**	**73%**
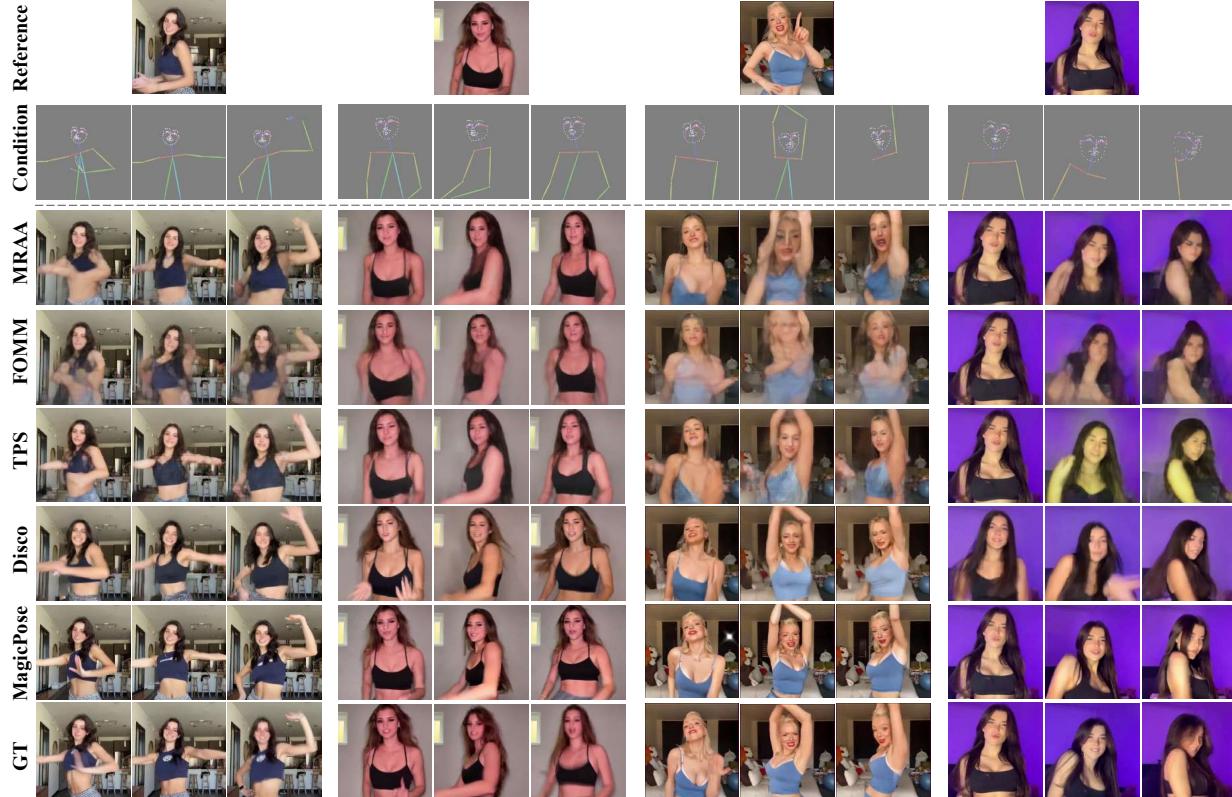
Figure 10. Visualization of human pose and facial expression retargeting (Subjects 1-4 in the user study): MagicPose preserves identity and appearance information from the reference image the best.

B.3. Zero-Shot Animation

B.3.1. OUT-OF-DOMAIN IMAGES

We provide more visualizations of zero-shot generation of out-of-domain images in Figure. 9, Figure. 17, Figure. 18, and Figure. 19.

B.3.2. COMBINE WITH T2I MODEL

A potential application of our proposed model is that it can be combined with the existing Text to Image (T2I) generation model (Zhang et al., 2023; Rombach et al., 2022) and used to edit the generation result. We visualized some samples in Figure. 20.

C. Sequence Generation with Motion Module

As mentioned in our main paper, the Appearance Control Model and Apperance-disentangled Pose ControlNet together already achieve accurate image-to-image motion transfer, but we can further integrate an optional motion module into the primary SD-UNet architecture to improve the temporal consistency. We initially employed the widely-used AnimateDiff (Guo et al., 2023), which provides an assortment of motion modules tailored to the stable diffusion model v1.5., but we found that AnimateDiff faces limitations in achieving seamless transition across frames, particularly with more complex movement patterns present in human dance, as opposed to more subdued video content.
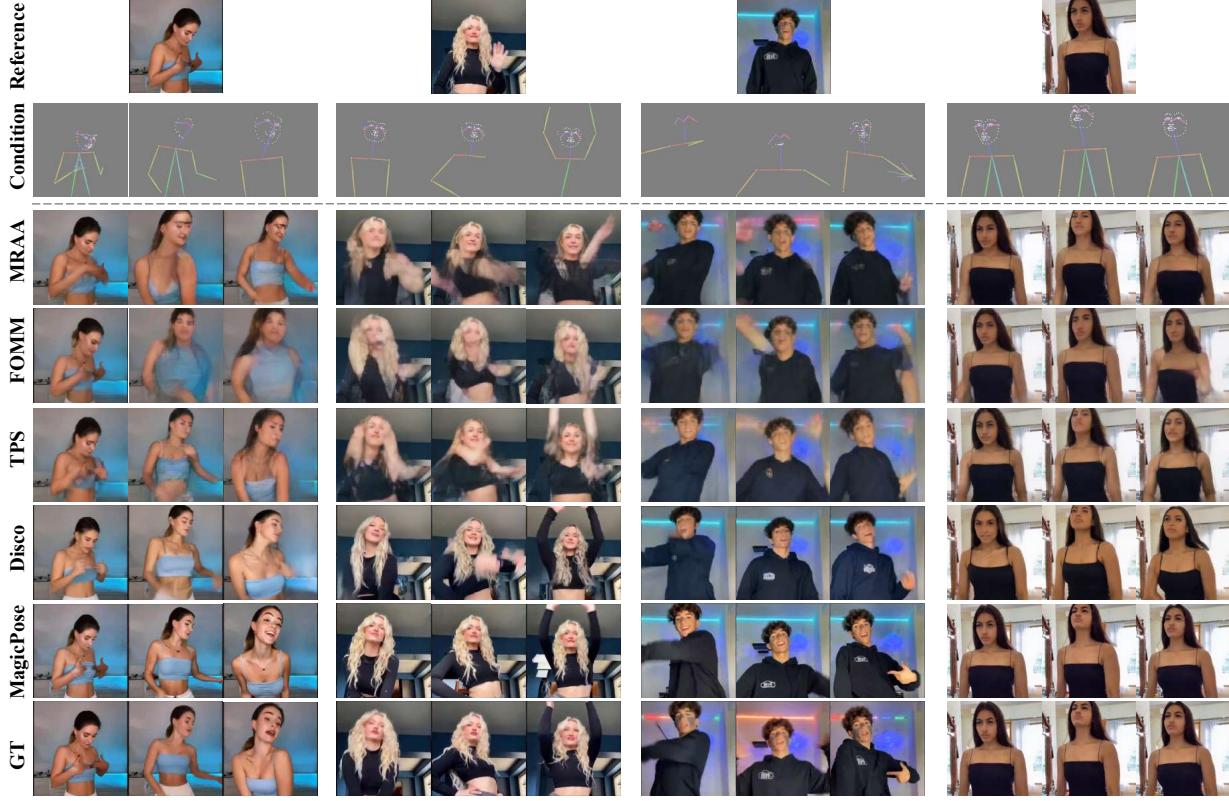
Figure 11. Visualization of human pose and facial expression retargeting (Subjects 5-8 in the user study): MagicPose preserves identity and appearance information from the reference image the best.

To solve this issue, we fine-tuned the AnimateDiff motion modules until satisfactory temporal coherence was observed during the evaluation. We freeze the weights of all parts in our Appearance Control Model and Apperance-disentangled Pose ControlNet, and fine-tune the motion module with pretrained weights from AnimateDiff (Guo et al., 2023) for 30k steps with a batch size of 8. Each batch contains 16 frames of a video sequence as the target output. For more smooth and consistent video generation quality, we also propose a special sampling strategy for DDIM (Song et al., 2021a) during inference. Figure. 12, Figure. 17, Figure. 18, and Figure. 19 are examples of sequential output from our model.

D. Limitations

In MagicPose, We follow previous work (Wang et al., 2023) and adopt OpenPose (Cao et al., 2019; Simon et al., 2017; Cao et al., 2017; Wei et al., 2016) as the human pose detector, which is crucial for pose control, significantly affecting the generated images' quality and temporal consistency. However, challenges arise in accurately detecting complete pose skeletons and facial landmarks, especially under rapid movement, occlusions, or partial visibility of subjects. As

illustrated in the second row of Figure 4, we can observe that the skeleton and hand pose are partially missing in the detection result, especially in the right half of the row. In future works, a more advanced pose detector can be adopted for better image editing quality.

E. Discussion on motivation and future works

In addition to the suggestion of replacing openpose with a more advanced pose detector, we also would like to discuss future works from our motivation. Our understanding of image generation is that it can be decomposed into two aspects: (1) identity control (appearance of human) and (2) shape/geometry control (pose and motion of human). MagicPose was introduced to maintain the appearance and identity information in generation from reference image input strictly while editing the geometry shape and structural information under the guidance of human pose skeleton. In this paper, we demonstrate the **identity-preserving** ability of the Appearance Control Model and its Multi-Source Attention Module by human pose and facial expression retargeting task. The design of this Multi-Source Attention Module can be further extended to other tasks as well, e.g. novel view synthesis of general objects under the shape
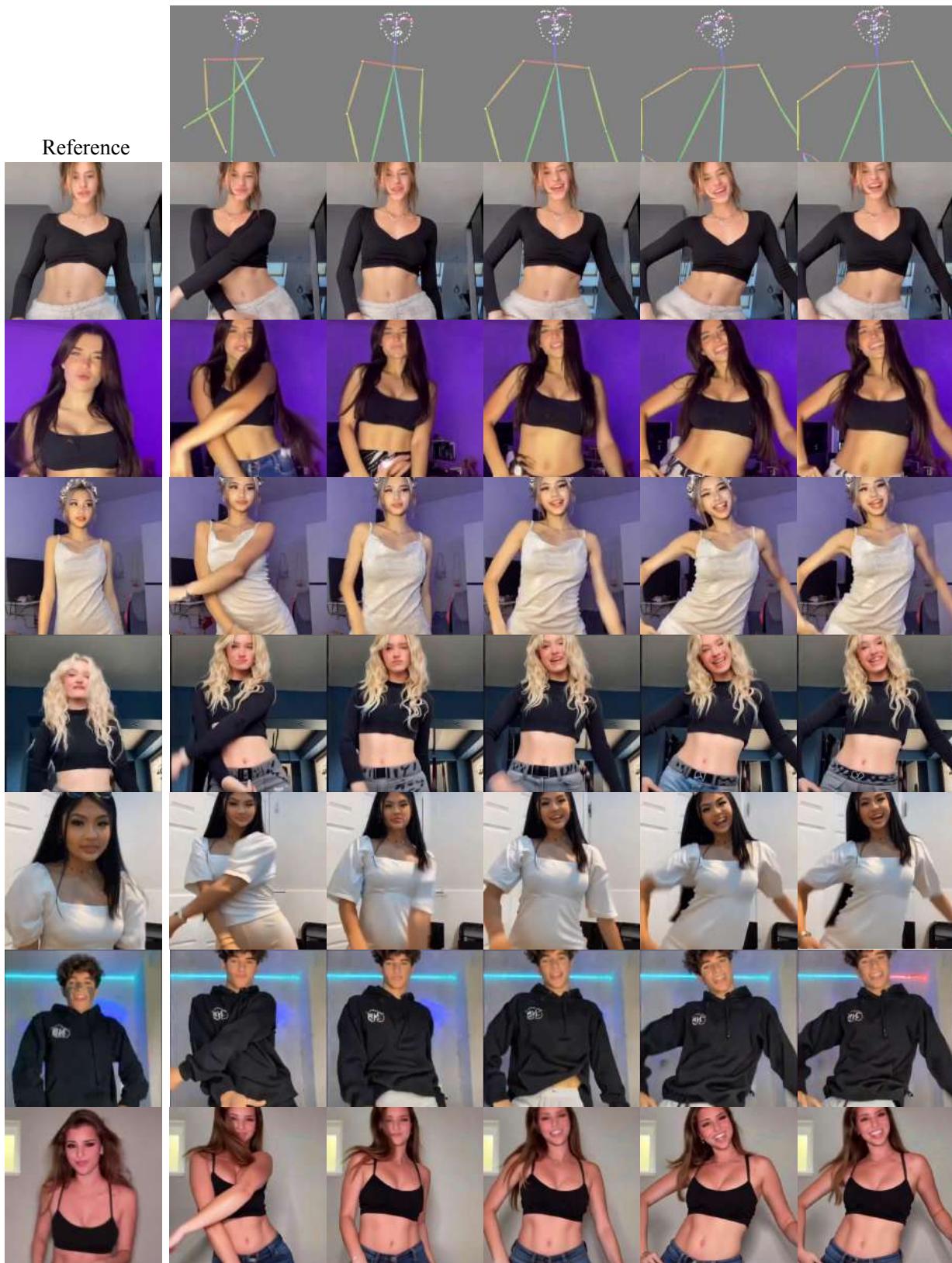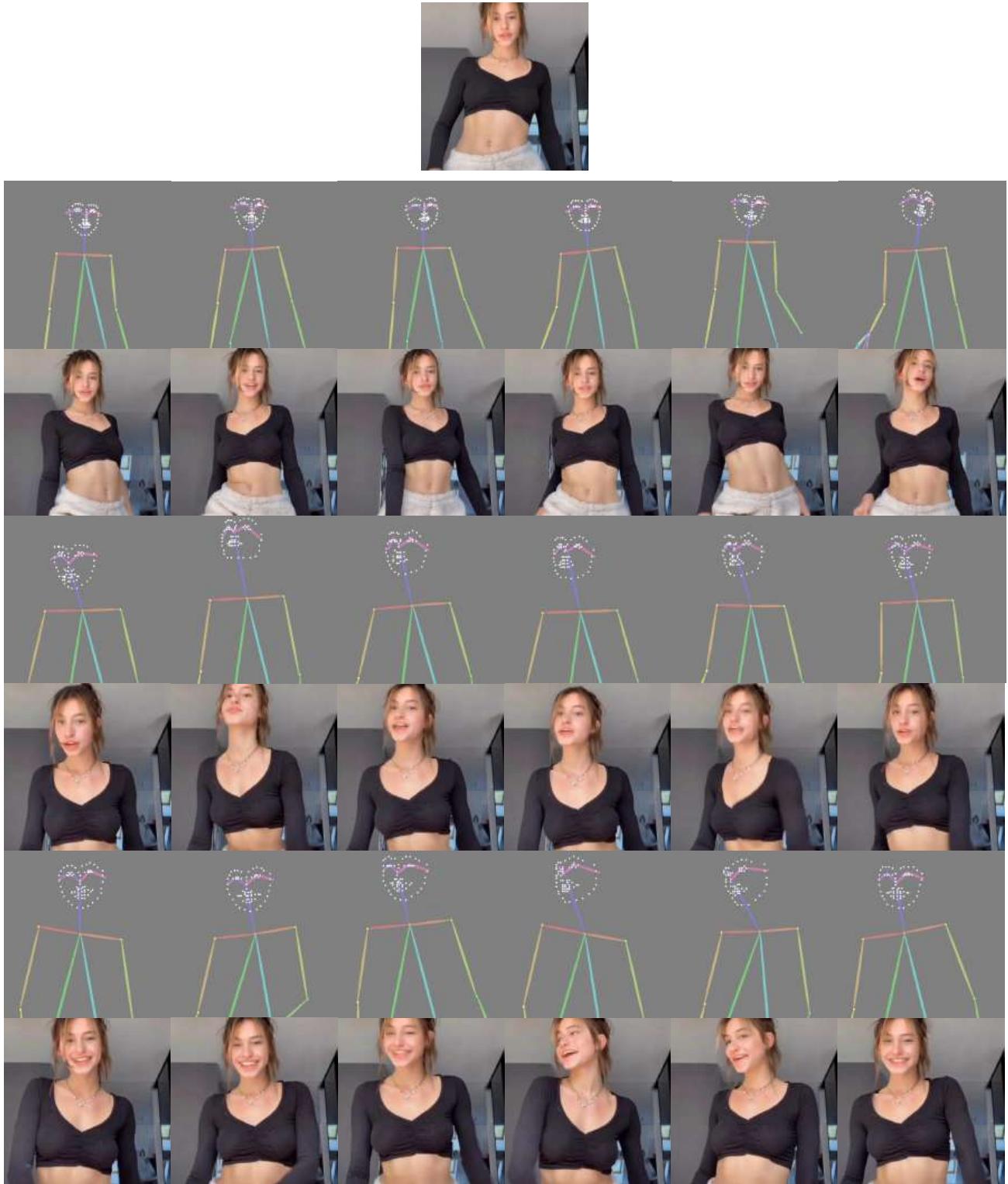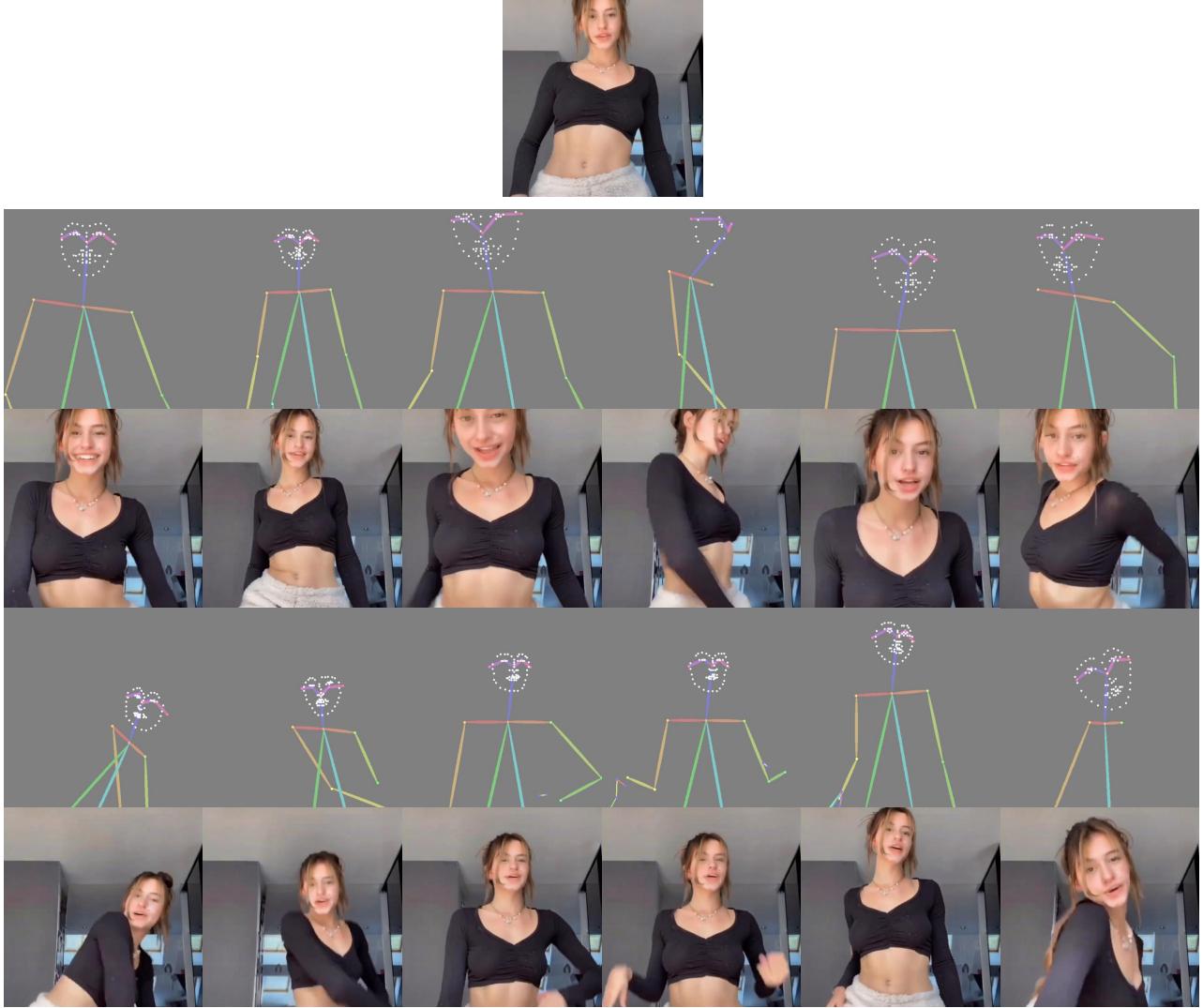
Figure 12. Visualization of Human Motion and Facial Expression Transfer on TikTok (Jafarian & Park, 2021).

Figure 13. Visualization of Human Motion and Facial Expression Transfer on TikTok (Jafarian & Park, 2021). MagicPose is able to generate vivid and realistic motion and expressions under the condition of diverse pose skeleton and face landmark input, while accurately maintaining identity information from the reference image input.

Figure 14. Visualization of Human Motion and Facial Expression Transfer on TikTok (Jafarian & Park, 2021).

condition of the camera, shape manipulation of the natural scenes under the geometry condition of depth/segmentation map, and motion transfer of animals under the animal pose condition of skeletons, etc.
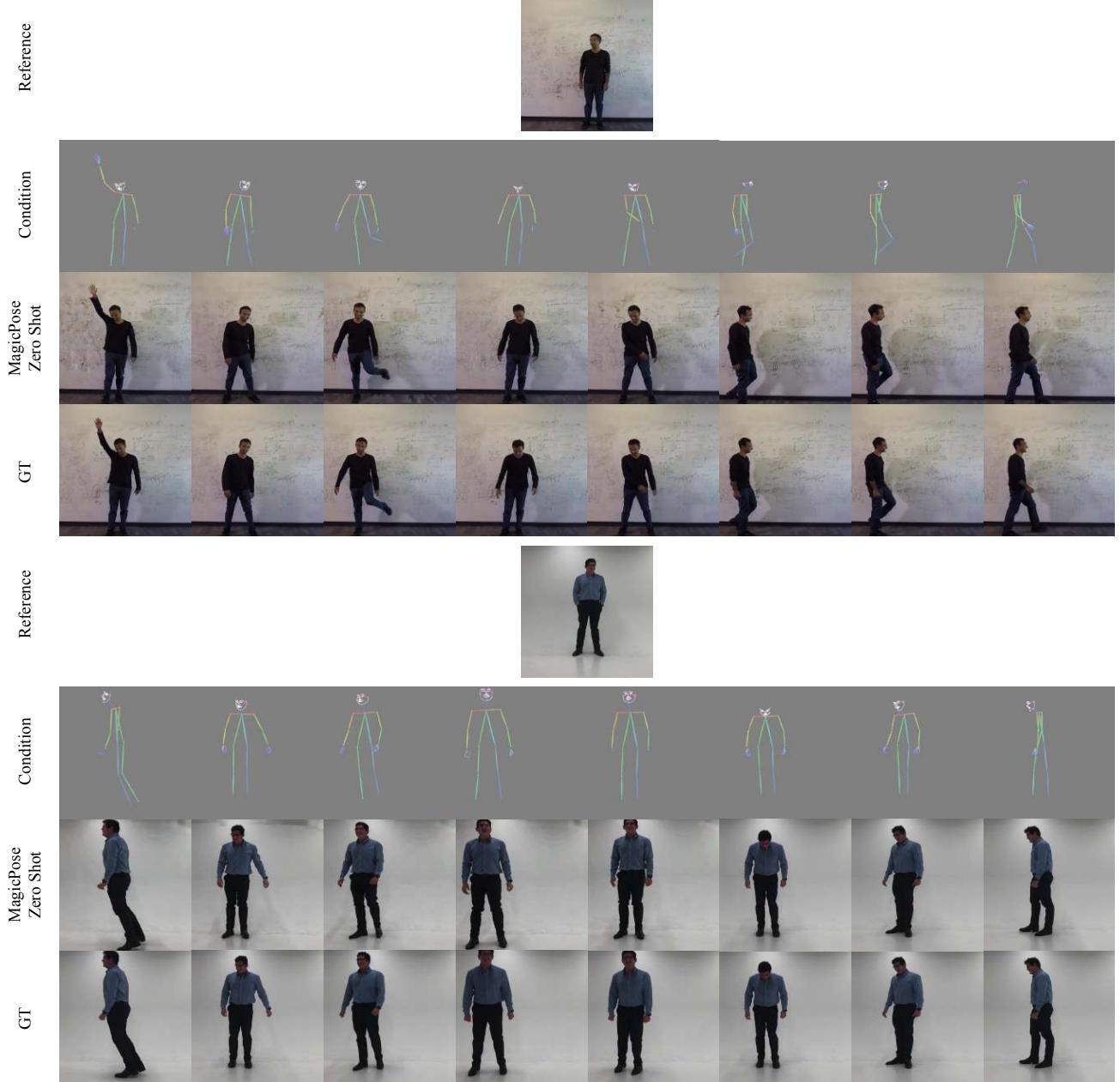
Figure 15. Visualization of Zero-Shot Human Motion and Facial Expression Transfer on Everybody Dance Now Dataset (Chan et al., 2019b).
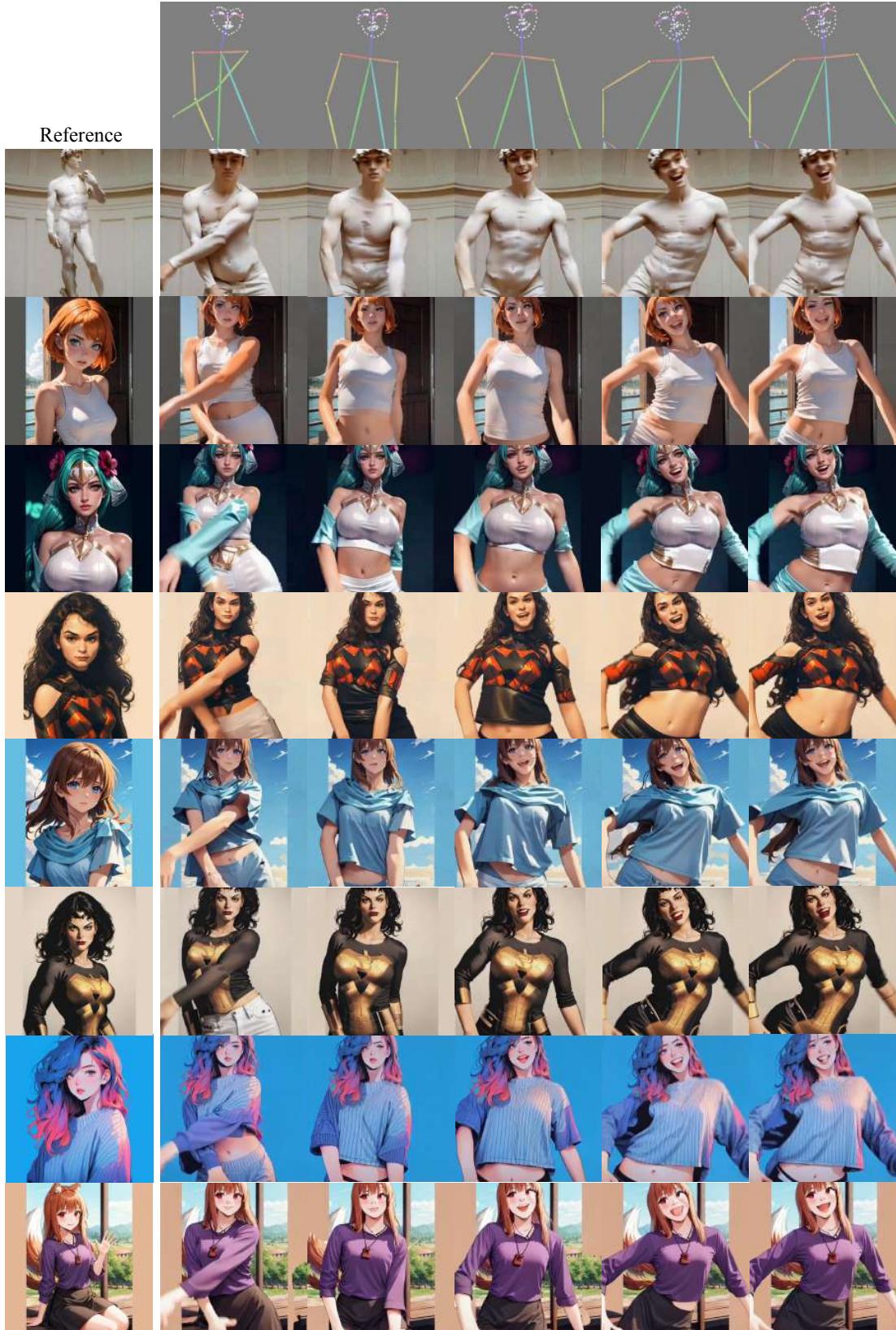
Figure 16. Visualization of Zero-Shot Human Motion and Facial Expression Transfer on Everybody Dance Now Dataset (Chan et al., 2019b).

Figure 17. Visualization of Zero-Shot Animation. MagicPose can provide a precise generation with identity information from out-of-domain images even without any further fine-tuning after being trained on real-human dance videos.
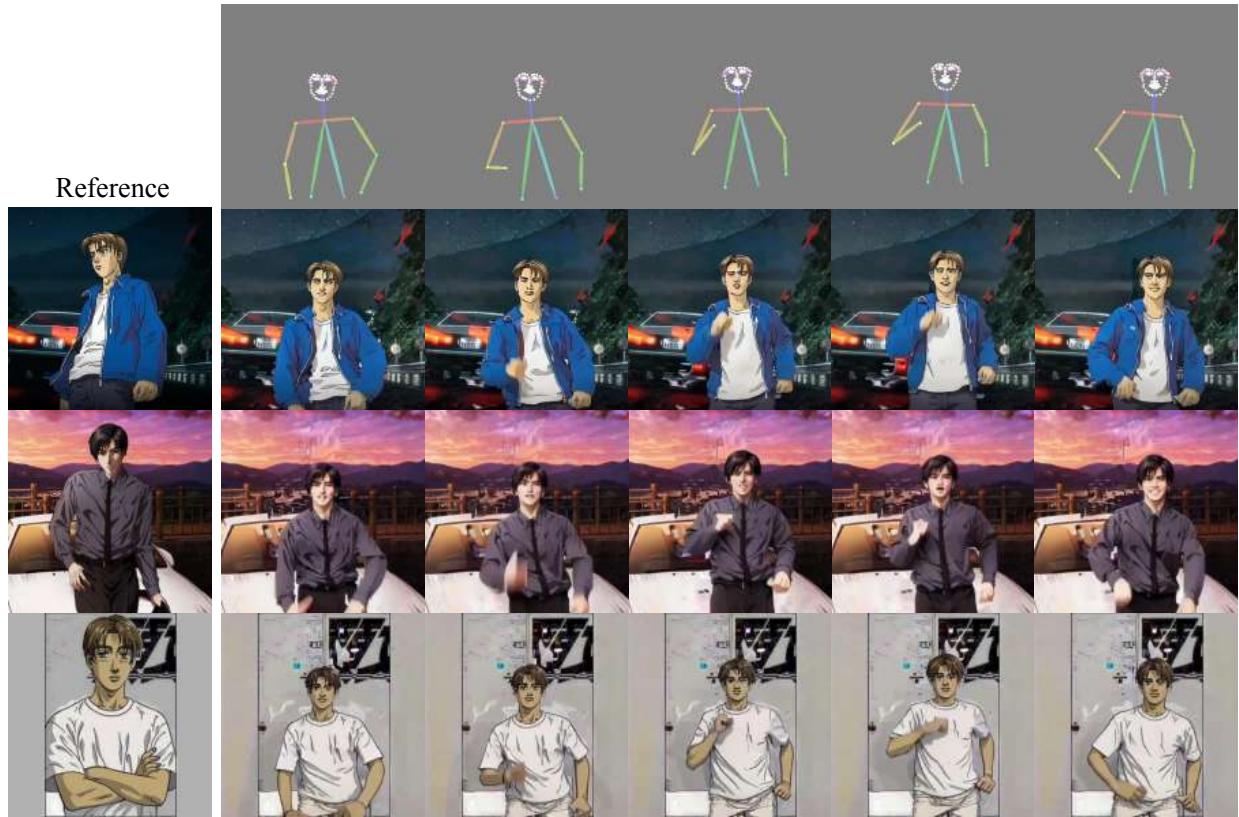
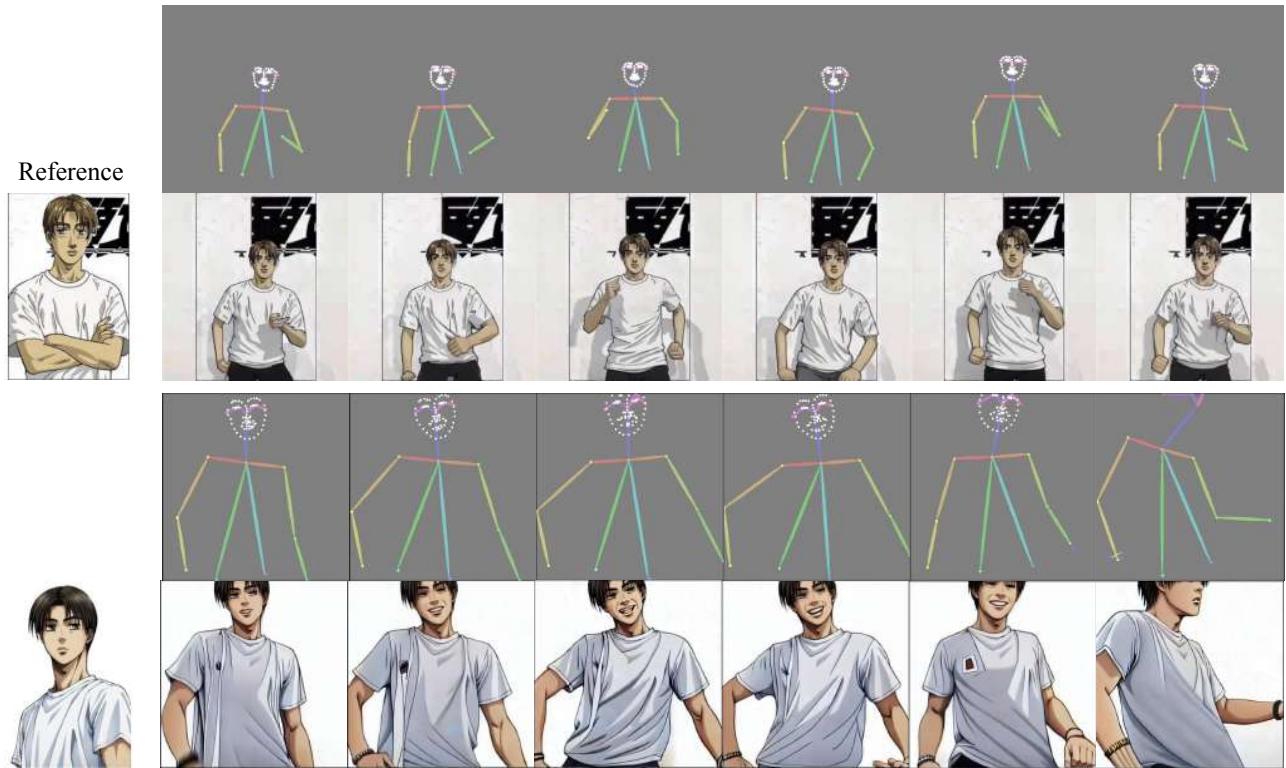*Figure 18.* Visualization of Zero-Shot 2D Cartoon Generation.

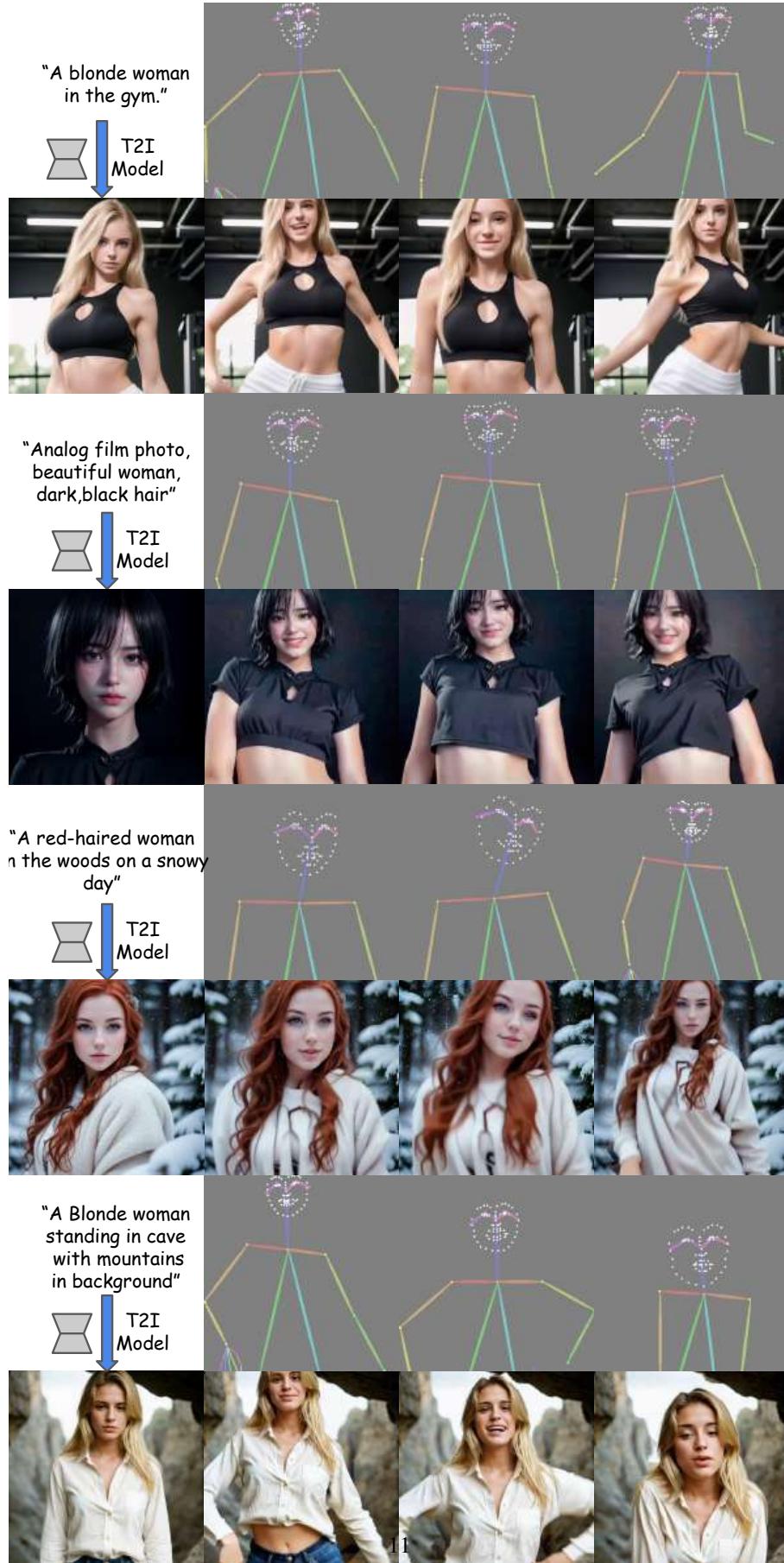*Figure 19.* Visualization of Zero-Shot 2D Cartoon Generation.

Figure 20. Usage of combining with T2I model. MagicPose can provide a precise generation with identity information from T2I-generated images even without further fine-tuning after training on real-human dance videos.