Supplementary Material for StableAnimator: High-Quality Identity-Preserving Human Image Animation

Shuyuan Tu[1,2] Zhen Xing[1,2] Xintong Han[4] Zhi-Qi Cheng[5] Qi Dai[3] Chong Luo[3] Zuxuan Wu[1,2] *
[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
[3]Microsoft Research Asia [4]Huya Inc. [5]Carnegie Mellon University
https://francis-rings.github.io/StableAnimator

A. Evaluation Metrics

Following previous human image animation evaluation settings, we implement numerous quantitative evaluation metrics, including L1, PSNR, SSIM, LPIPS, FVD, and CSIM, to compare our StableAnimator with current state-of-the-art animation models. The details of the above metrics are described as follows:

(1) L1 refers to the average absolute difference between the corresponding pixel values of two images. It measures the typical magnitude of prediction errors without considering their direction, making it a valuable tool for quantifying the extent of discrepancies.

(2) PSNR measures the ratio between the maximum possible power of a signal (in this case, the original image) and the power of corrupting noise that affects the fidelity of its representation. PSNR is expressed in decibels (dB), with higher values indicating better quality.

(3) SSIM refers to the similarity between two images based on their luminance, contrast, and structural information.

(4) LPIPS measures the similarity between images by analyzing the feature representations of their patches, reflecting human visual perception effectively.

(5) FVD evaluates the disparity between the feature distributions of real and generated videos, considering both spatial and temporal dimensions. FVD is often used to measure the video fidelity.

(6) CSIM refers to the cosine similarity between the facial embeddings of two face images. The facial embeddings are extracted by ArcFace.

B. Preliminaries

The diffusion model includes a forward diffusion process and a reverse denoising process. In the forward process, the Gaussian noise is progressively added to the data sample $\boldsymbol{x}_0 \sim \boldsymbol{p}_{\text{data}}$ from the particular data distribution $\boldsymbol{p}_{\text{data}}$:

$$\boldsymbol{q}(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I}). \tag{1}$$

The data sample $\boldsymbol{x}_0$ is ultimately converted into Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(0,1)$ after $\boldsymbol{T}$ diffusion forward steps. α_t is a constant noise schedule. In the reverse process, the diffusion model $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)$ tends to recover $\boldsymbol{x}_0$ from $\boldsymbol{x}_T$ by predicting the noise $\boldsymbol{\varepsilon}$ based on the current sample $\boldsymbol{x}_t$ and time step $\boldsymbol{t}$. The MSE loss is applied to train $\boldsymbol{\varepsilon}(\cdot)$:

$$\mathcal{L} = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\varepsilon},t}(\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)\|^2). \tag{2}$$

Moreover, the denoising process can be regarded as a continuous process (reverse-SDE):

$$d\boldsymbol{X}_t = [f(\boldsymbol{X}_t, \boldsymbol{t}) - g^2(\boldsymbol{X}_t, \boldsymbol{t})\nabla \log p(\boldsymbol{X}_t, \boldsymbol{t})]d\boldsymbol{t} + g(\boldsymbol{X}_t, \boldsymbol{t})d\boldsymbol{W}_t, \tag{3}$$

where $\boldsymbol{W}_t$ and $\nabla \log p(\boldsymbol{X}_t, \boldsymbol{t})$ refer to the standard Brownian motion and score function. $f(\boldsymbol{X}_t, \boldsymbol{t})$ and $g(\boldsymbol{X}_t, \boldsymbol{t})$ are drift and volatility. The diffusion model $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)$ approximates $\nabla \log p(\boldsymbol{X}_t, \boldsymbol{t})$ during the continuous denoising process.

C. Details of Testing Dataset

We select 100 unseen videos (10-20 seconds long) from the internet to construct the testing dataset Unseen100. Some examples are shown in Fig. 1. The first row refers to five frames of a video, while the following rows represent individual frames of different videos. The sources of videos come from numerous social media platforms, including YouTube, TikTok, and BiliBili. These videos showcase individuals across ethnicities, genders, portrayed in full-body, half-body, and close-up shots against varied indoor and outdoor settings. In contrast to existing open-source testing datasets (TikTok dataset), our Unseen100 contains relatively complicated motion information and intricate protagonist appearances. Moreover, positions and facial expressions in some Unseen100 videos dynamically change,

*Corresponding authors.

Algorithm 1 HJB Equation-based Face Optimization ($\sigma(t) = t$ and $s(t) = 1$)

Input: A diffusion model $\mathtt{D}_\theta(\boldsymbol{x}; \boldsymbol{\sigma})$, Timesteps $t_{i\in\{0,\dots,N\}}$, Pre-defined factors $\boldsymbol{\gamma}_{i\in\{0,\dots,N-1\}}$, A reference image $\boldsymbol{y}$
Sample $\boldsymbol{x}_0 \sim \mathcal{N}(0, t_0^2 \boldsymbol{I})$
For $i \in \{0, \dots, N-1\}$ **do**
  $\boldsymbol{\gamma}_i = 0$
  if $t_i \in [\boldsymbol{S}_{t_{\text{min}}}, \boldsymbol{S}_{t_{\text{max}}}]$: ▷ $\boldsymbol{S}_{\text{noise}}$,$\boldsymbol{S}_{\text{churn}}$, $\boldsymbol{S}_{t_{\text{min}}}$, and $\boldsymbol{S}_{t_{\text{max}}}$ are the pre-defined values of EDM
    $\boldsymbol{\gamma}_i = \min\left(\frac{\boldsymbol{S}_{\text{churn}}}{N}, \sqrt{2} - 1\right)$
  Sample $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{S}_{\text{noise}}^2 \boldsymbol{I})$
  $\hat{t}_i = t_i + \boldsymbol{\gamma}_i t_i$ ▷ Select temporarily increased noise level $\hat{t}_i$
  $\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \sqrt{\hat{t}_i^2 - t_i^2}\boldsymbol{\epsilon}_i$ ▷ Add new noise to move from t_i to $\hat{t}_i$
  $\boldsymbol{x}_{\text{pred}} = \mathtt{D}_\theta(\hat{\boldsymbol{x}}_i; \hat{t}_i)$ ▷ The diffusion model predicts the denoised sample
  $\boldsymbol{x}_{\text{op}} = \boldsymbol{x}_{\text{pred}}.\mathtt{clone}().\mathtt{detach}()$ ▷ Starting optimization
  $\boldsymbol{op} = \mathtt{Adam}([\boldsymbol{x}_{\text{op}}], \boldsymbol{\eta})$ ▷ Adam and $\boldsymbol{\eta}$ are an Adam optimizer and a learning rate
  $\boldsymbol{x}_{\text{op}}$.requires_grad = True ▷ $\boldsymbol{x}_{\text{op}}$ is a HJB variable (trainable)
  For $k \in \{1, 2, \dots, 10\}$ **do** ▷ k is the optimization step
    $\boldsymbol{f}_{\text{pred}} = \mathtt{Decoder}(\boldsymbol{x}_{\text{op}})$ ▷ Decoder is a VAE decoder, which converts predicted sample to the pixel level
    $\boldsymbol{loss} = (1 - \mathtt{Cos}(\mathtt{Arc}(\boldsymbol{f}_{\text{pred}}), \mathtt{Arc}(\boldsymbol{y}))).\mathtt{abs}().\mathtt{mean}()$ ▷ Cos$(\cdot)$ computes the similarity between given embeddings
    $\boldsymbol{op}$.zero_grad() ▷ Arc is the Arcface model which extracts face embeddings
    $\boldsymbol{loss}$.backward(retain_graph=True) ▷ $\boldsymbol{x}_{\text{op}}$ is updated towards optimal face consistency by the gradient of the loss
    $\boldsymbol{op}$.step()
  $\boldsymbol{x}_{\text{pred}} = \boldsymbol{x}_{\text{op}}$ ▷ End of Optimization
  $\boldsymbol{d}_i = (\hat{\boldsymbol{x}}_i - \boldsymbol{x}_{\text{pred}})/\hat{t}_i$ ▷ Evaluate $d\boldsymbol{x}/dt$ at $\hat{t}_i$
  $\boldsymbol{x}_{i+1} = \hat{\boldsymbol{x}}_i + (t_{i+1} - \hat{t}_i)\boldsymbol{d}_i$ ▷ Take Euler step from $\hat{t}_i$ to t_{i+1}
  if $t_{i+1} \neq 0$:
    $\boldsymbol{d}'_i = (\boldsymbol{x}_{i+1} - \mathtt{D}_\theta(\boldsymbol{x}_{i+1}; t_{i+1}))/t_{i+1}$ ▷ Apply 2nd order correction
    $\boldsymbol{x}_{i+1} = \hat{\boldsymbol{x}}_i + (t_{i+1} - \hat{t}_i)\left(\frac{1}{2}\boldsymbol{d}_i + \frac{1}{2}\boldsymbol{d}'_i\right)$
return $\boldsymbol{x}_N$

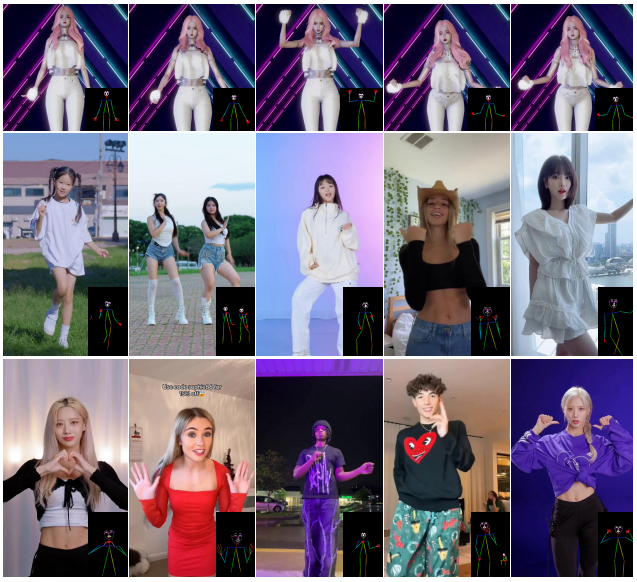

Figure 1. Examples from Unseen100.

such as shaking heads, making it more challenging to maintain identity consistency.

D. Long Animation

We conduct several comparison experiments of our StableAnimator and SOTA human image animation models, as shown in Fig. 2, Fig. 3, and Fig. 4. Each video contains more than 300 frames, featuring complex appearances of the protagonists, complicated motion sequences, and intricate background information. The results highlight the superiority of our StableAnimator in generating long animations while competing methods experience dramatic distortion of human bodies and identities.

E. Multiple Person Animation

To demonstrate the robustness of our StableAnimator, we experiment on a particular video involving multiple protagonists, as shown in Fig. 5. We can see that our StableAnimator is also capable of handling multiple-person animations while preserving the original identity and achieving high video fidelity.

F. Optimization Details

We present a more detailed HJB Equation-based Face Optimization in Algorithm 1. Notably, the basic structure of

our algorithm closely resembles Algorithm 2 in the EDM paper. In the main paper, $\gamma_1 = -\boldsymbol{r} \cdot (\boldsymbol{X}_1 - \boldsymbol{x}_1)$ is derived from Eq.4 and Eq.5. In particular, this formula is obtained by calculating the transversality condition of Eq. 4 at the terminal time.

G. Additional Face Discussion

We further conduct a comparison between our StableAnimator and other facial restoration models (GFP-GAN and CodeFormer). The results are shown in Fig. 6. *w/o* Face refers to the baseline model of our StableAnimator without incorporating any face-related components. It is noticeable that our StableAnimator has the best identity-preserving capability compared with other competitors, demonstrating the superiority of our StableAnimator regarding identity consistency. By contrast, GFP-GAN and CodeFormer suffer from serious facial distortion and over-sharpening. The plausible reason is that *w/o* Face cannot synthesize the precise facial layout, which in turn undermines the effectiveness of subsequent facial restoration processes. This represents a fundamental limitation of post-processing-based face enhancement strategies.

H. Identity-Preserving Loss

In the image-domain identity-preserving methods, they often incorporate the ArcFace ID loss into the training process, which calculates the cosine similarity between the ArcFace face embeddings of the denoised result and the groundtruth. By contrast, during training, we introduce face masks extracted by Arcface to the conventional reconstruction MSE loss to improve modeling of face-related regions. The reason is that applying the ArcFace ID loss requires employing a VAE Decoder to convert the denoised latents into pixel level. The reason is that applying the ArcFace ID loss requires using a VAE Decoder to convert the denoised latents into the pixel level. Although the VAE Decoder is frozen during training, a gradient back propagation graph must be maintained within the VAE Decoder to allow gradients to flow back to the U-Net for weight updates. However, the VAE Decoder in SVD contains memory-intensive temporal layers, making this back propagation graph extremely resource-demanding. Since training the SVD U-Net already requires substantial computational resources, incorporating the ArcFace ID loss would result in an unaffordable computational cost and significantly slow down the training process. Therefore, we simply modify the reconstruction MSE loss by incorporating face masks to enable more explicit face modeling, making the training relatively lightweight.

I. Additional Comparison Results

Fig. 7 and Fig. 8 show additional comparison results. The provided pose sequences encompass complex motion information, and the initial poses of the reference images are two categories: one with the protagonist facing directly toward the camera, and another with the protagonist's profile turned toward the camera. We can observe that our StableAnimator can accurately modify the motion of the reference images and maintain the original identity, while other competitors encounter varying degrees of human body distortion and loss of facial details.

J. Animation Results

We show our animation results in Fig. 9. We can see that our StableAnimator can perform a wide range of human image animation while simultaneously preserving the protagonist's appearance, background, and identity. Fig. 10, Fig. 11, and Fig. 12 show additional animation results generated by our StableAnimator. Each cases contain complex protagonist's appearance and intricate motion information. For example, in the reference image in the fifth row of Fig. 10, the protagonist's closed eyes make it particularly challenging for the human animation model to preserve ID consistency. It is noticeable that our StableAnimator can accurately manipulate motion in the reference image while preserving high-quality identity consistency, even in specific cases involving significant motion variations, such as head shaking and body rotation. Even when the head of the protagonist is continuously shaking and the angle facing the camera is constantly changing during the animation process, StableAnimator can still maintain a high level of identity consistency in the animation results without sacrificing details of the protagonist and the background.

K. Additional Ablation Study

To validate the contribution of our proposed components, We conduct a more comprehensive qualitative ablation study on different diffusion backbones, as shown in Fig. 13. ControlNeXt and MagicAnimate are based on Stable Video Diffusion (SVD) and Stable Diffusion (SD), respectively. We can see that our proposed components can significantly facilitate the performance of different backbone-based models, particularly in the facial regions. Notably, our proposed HJB Equation-based Face Optimization can still enhance the overall quality of animations to some extents, even when the backbone models lack any face-related encoders or adapters. The plausible reason is that our proposed HJB Equation-based Face Optimization can update the diffusion latents based on the face embedding similarity at each denoising step, thereby progressively refining the overall quality of denoised results without introducing any explicit face-related components.

L. Limitation and Future Work

Fig. 14 shows one failure case of our StableAnimator. In the given reference image, the girl's hand covers most of her face. Our StableAnimator struggles to fill in the obscured face regions, thereby degrading the quality of the synthesized face. One potential solution is introducing an additional face-aware inpainting adapter to the diffusion backbone for refining the face quality of given reference images. This part is left as future work.

M. Ethical Concern

Our StableAnimator can animate the given reference image based on the given pose sequence, which can be implemented in various fields, including virtual reality and digital human creation. However, the potential misuse of this model, particularly for creating misleading content on social media platforms, is a concern. To mitigate this, it is essential to use sensitive content detection algorithms.

Figure 2. Long animation results (1/3). The images with red borders are the reference images.

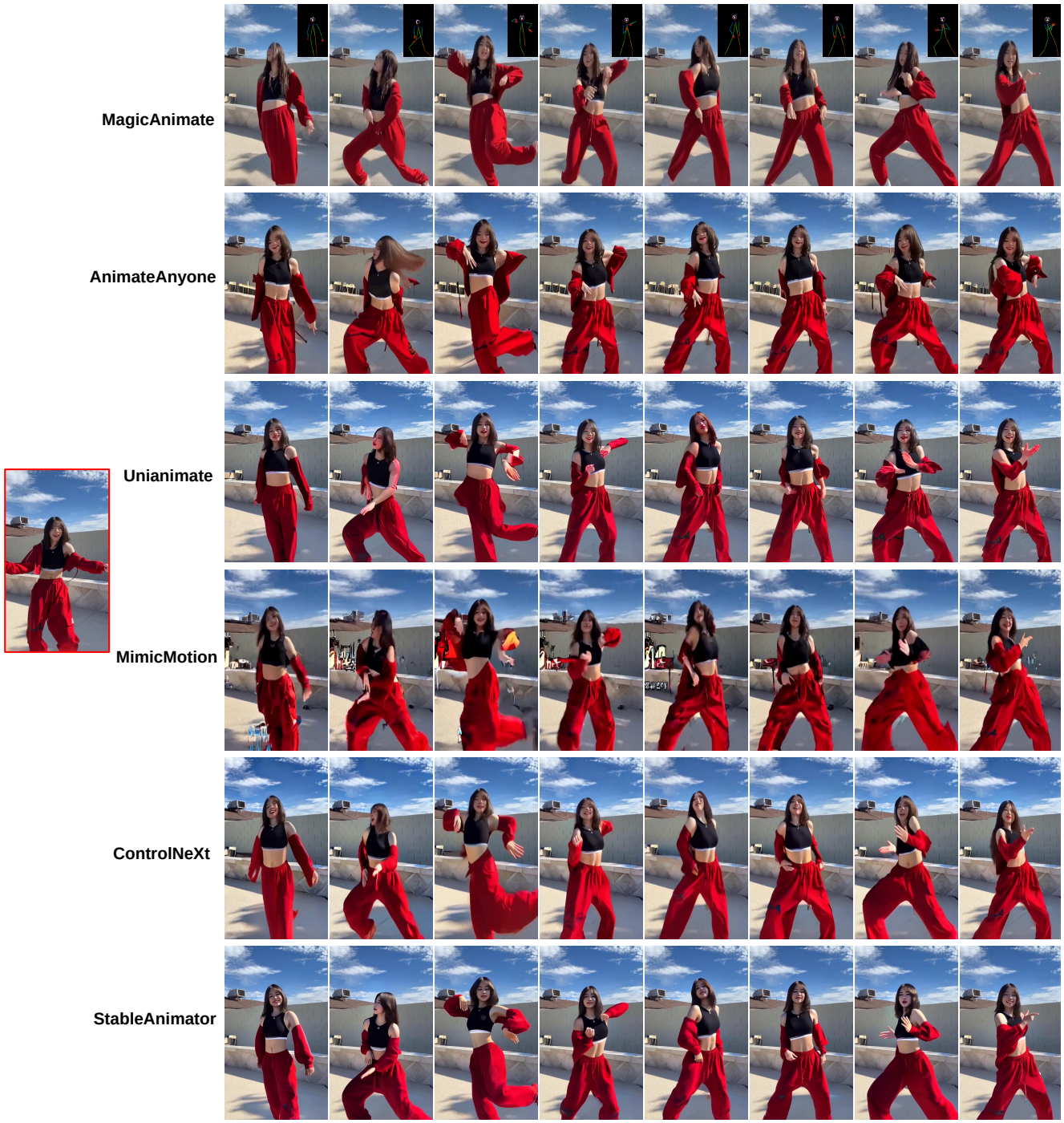

Figure 3. Long animation results (2/3). The images with red borders are the reference images.

Figure 4. Long animation results (3/3). The images with red borders are the reference images.

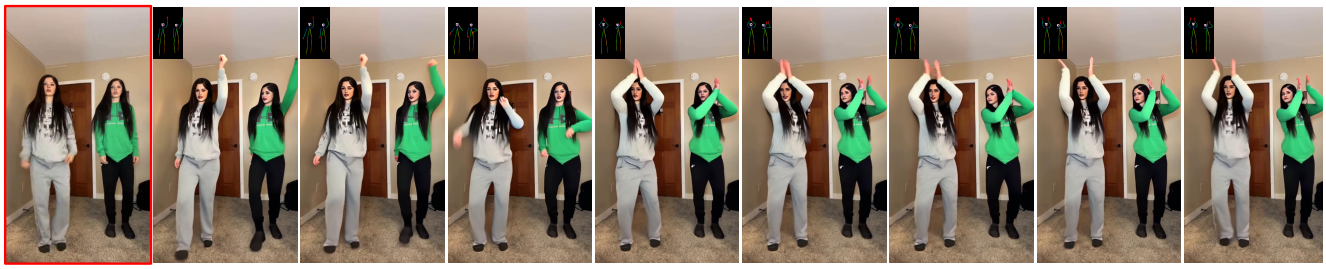

Figure 5. Multiple-person animation results.

Figure 6. Additional comparison results between our StableAnimator and current facial restoration models.

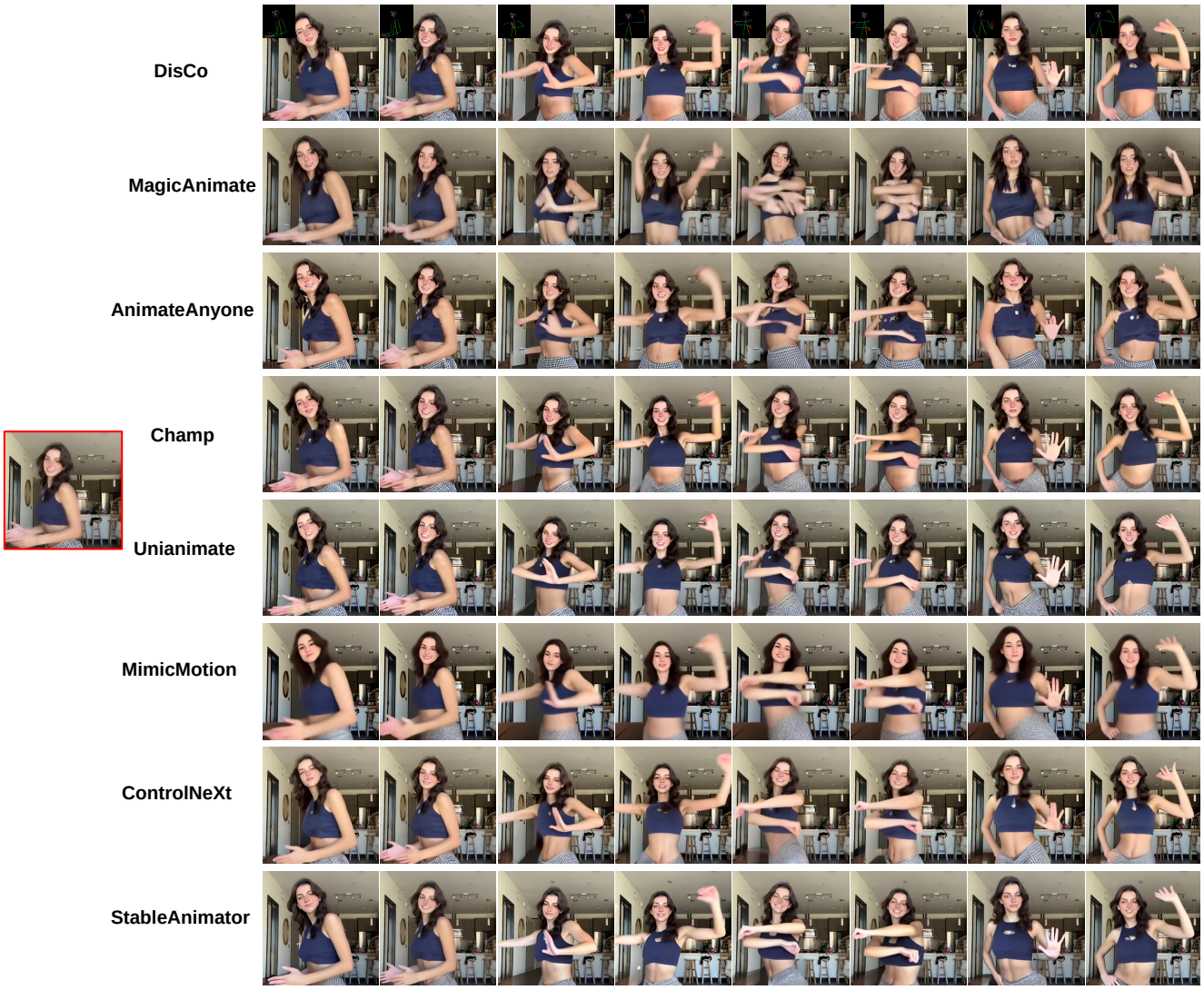

Figure 7. Additional comparison results (1/2), using the case presented in the paper of MagicAnimate. The images with red borders are the reference images.

Figure 8. Additional comparison results (2/2). The images with red borders are the reference images.

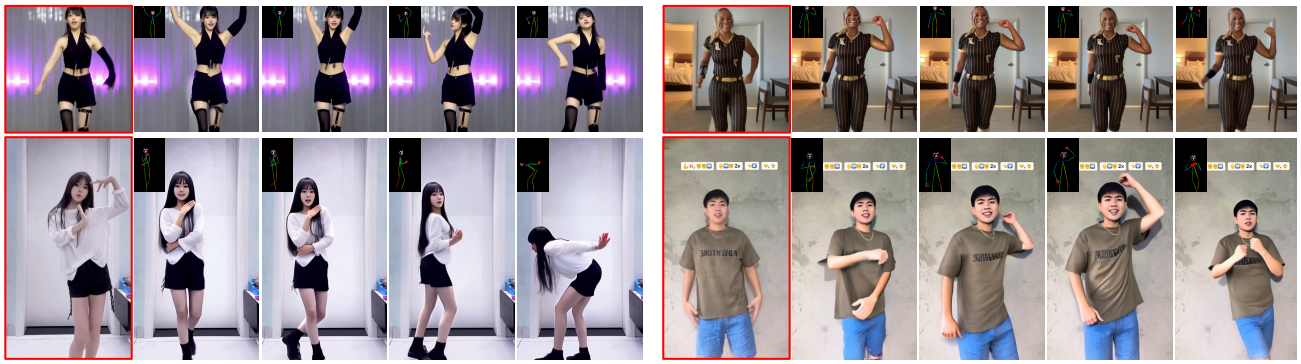

Figure 9. Animation results of our StableAnimator.

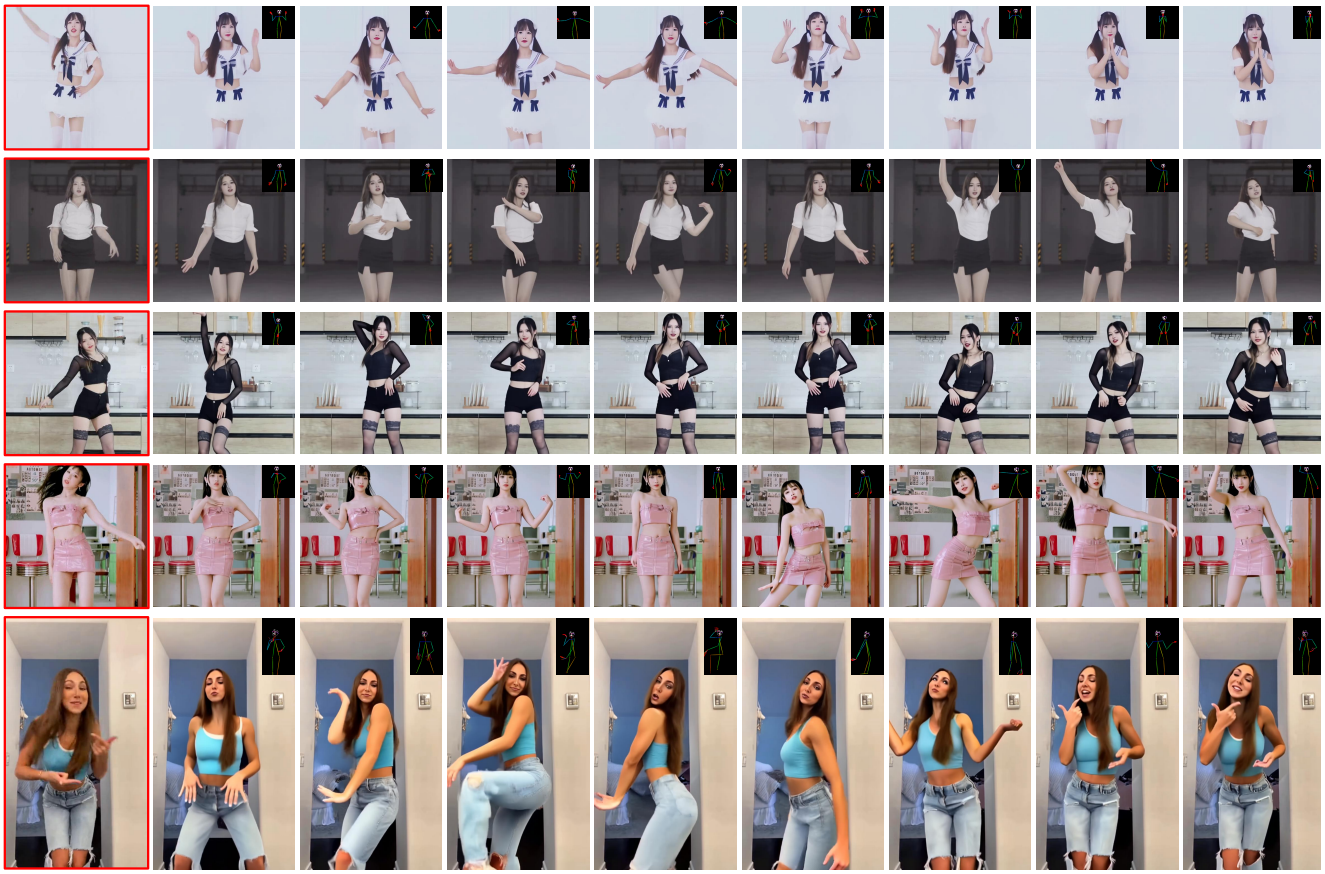

Figure 10. Additional animation results (1/3). The images with red borders are the reference images.

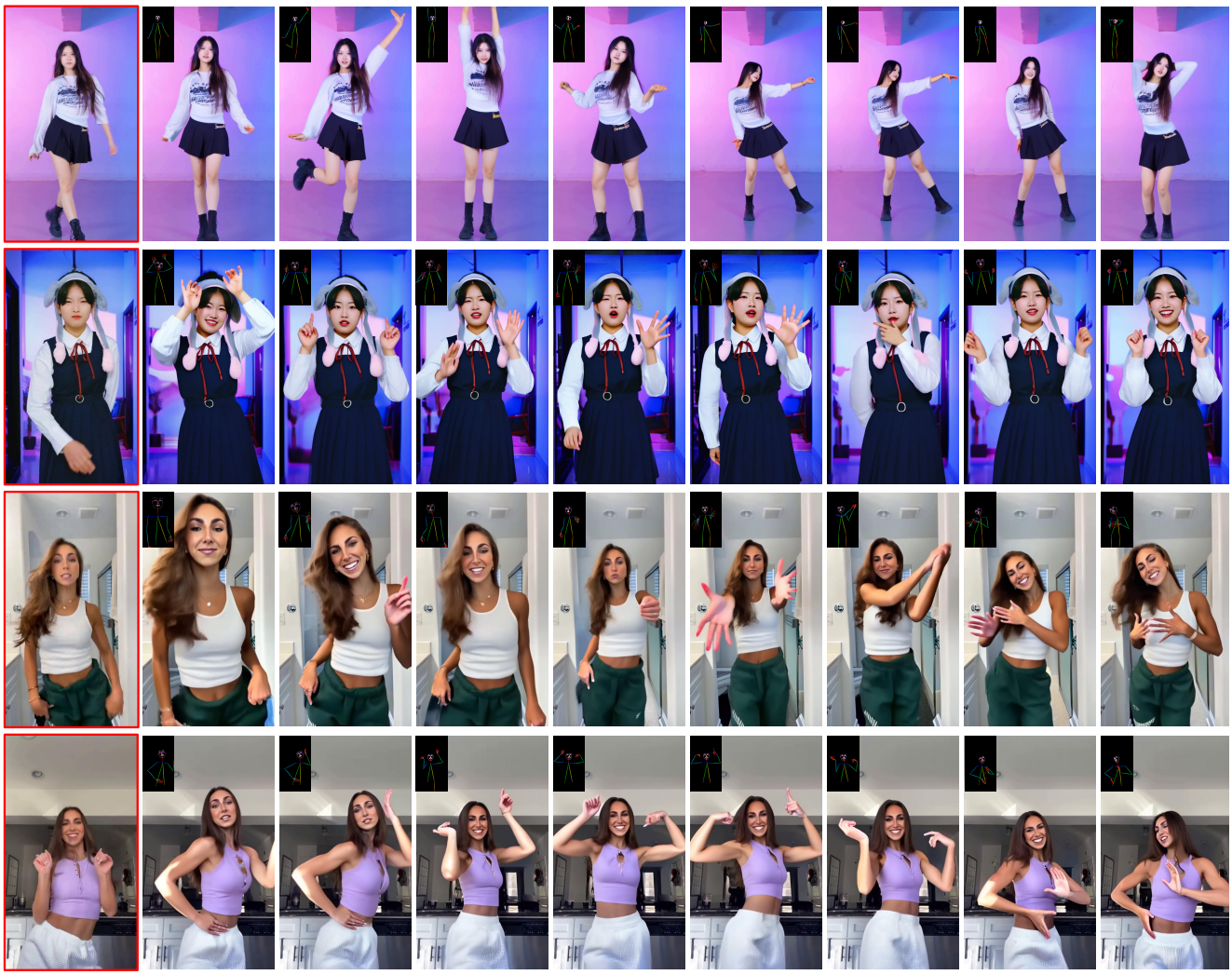

Figure 11. Additional animation results (2/3). The images with red borders are the reference images.

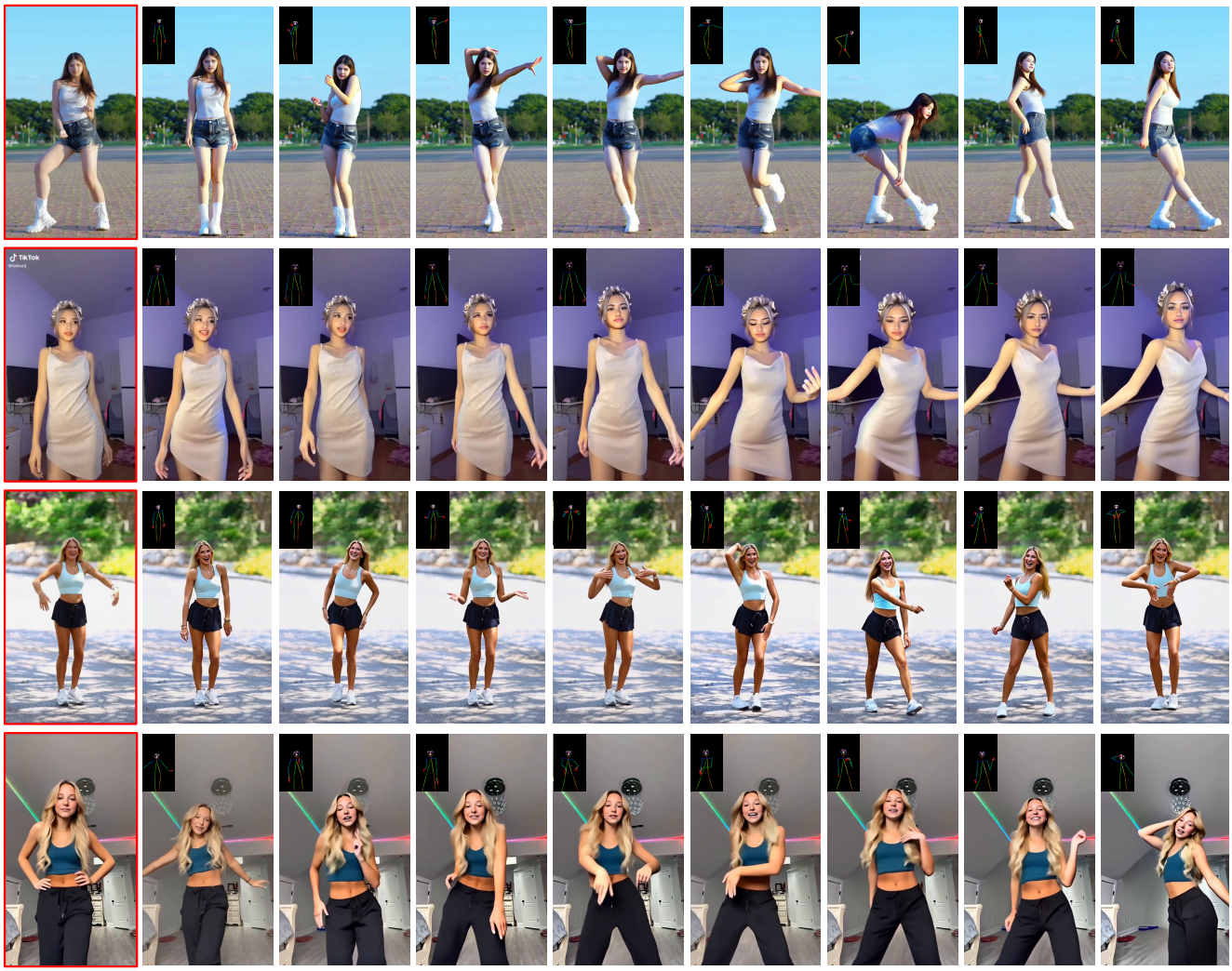

Figure 12. Additional animation results (3/3). The images with red borders are the reference images.

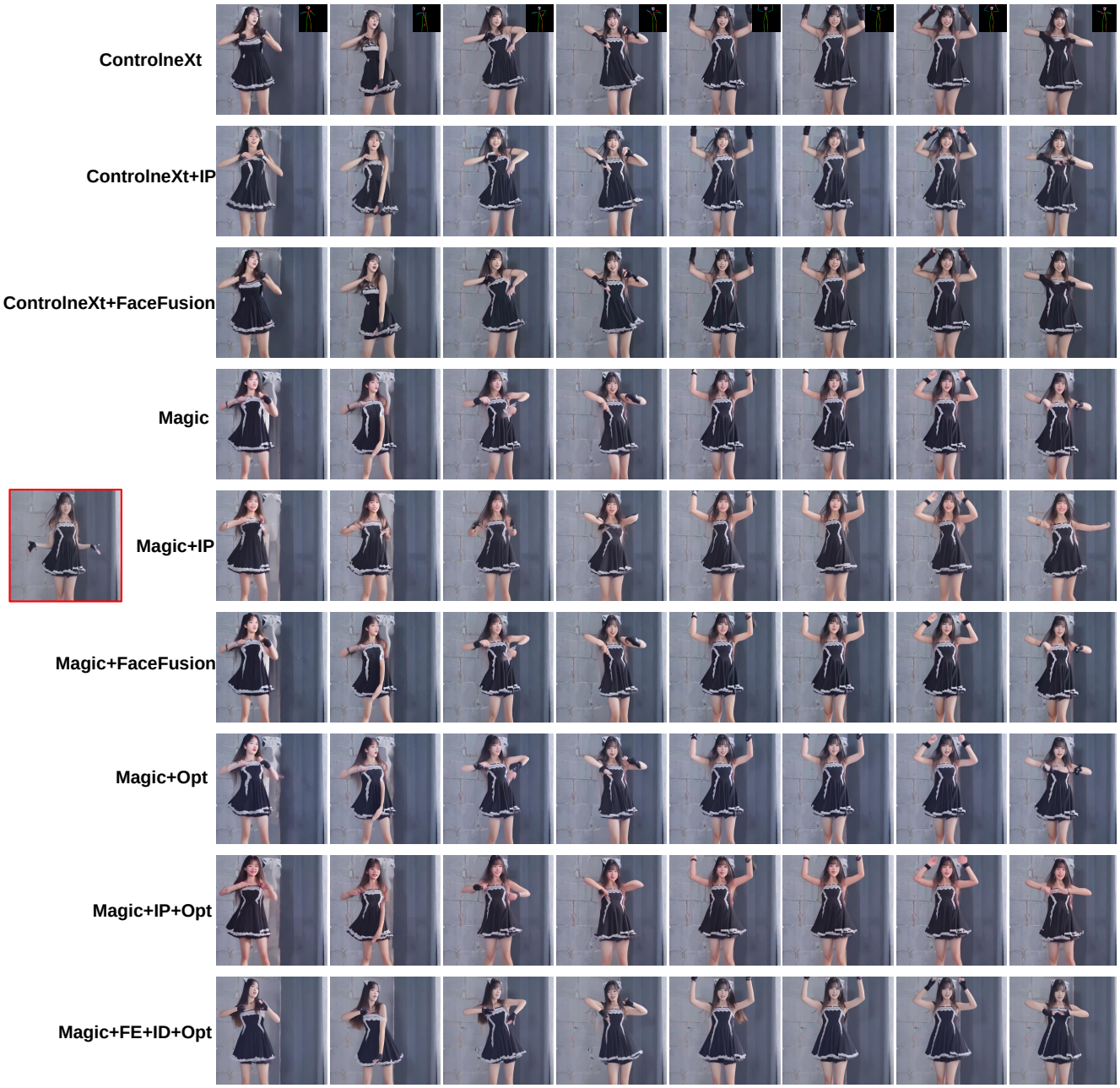

Figure 13. Additional ablation study results. IP, Magic, Opt, FE, and ID refer to IP-Adapter, MagicAnimate, our HJB Equation-based Face Optimization, our Global Content-Aware Face Encoder, and Distribution-Aware ID Adapter, respectively. The images with red borders are the reference images.

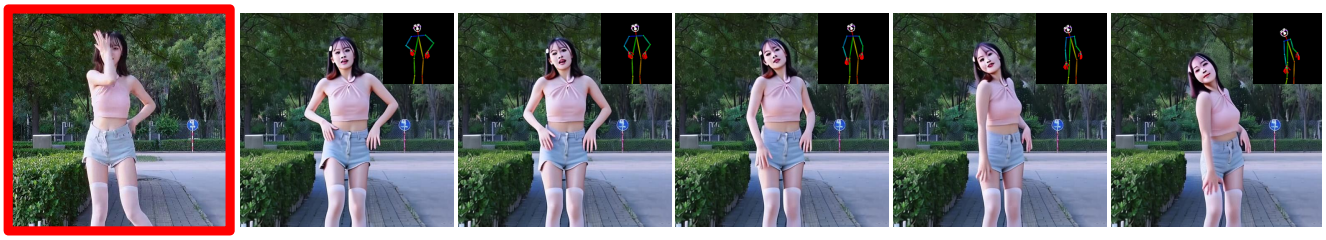

Figure 14. One failure case of our StableAnimator.