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

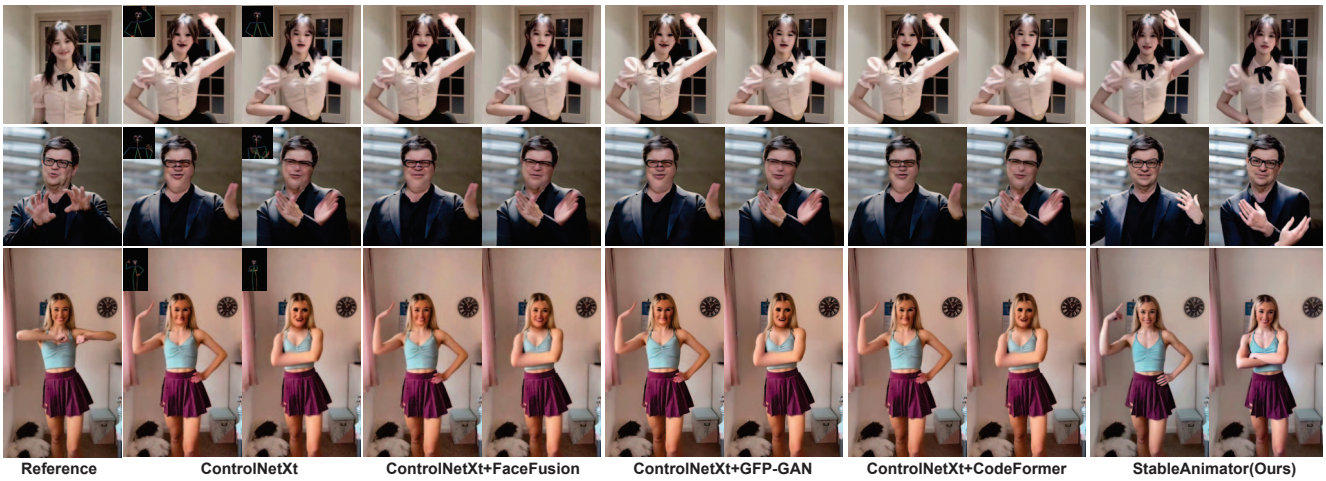

Figure 1. Pose-driven Human image animations generated by StableAnimator, showing its power to synthesize high-fidelity and ID-preserving videos. FaceFusion [15] is a face-swapping tool. GFP-GAN [51] and CodeFormer [69] are face restoration models.

Abstract

Current diffusion models for human image animation struggle to ensure identity (ID) consistency. This paper presents StableAnimator, the first end-to-end ID-preserving video diffusion framework, which synthesizes high-quality videos without any post-processing, conditioned on a reference image and a sequence of poses. Building upon a video diffusion model, StableAnimator contains carefully designed modules for both training and inference striving for identity consistency. In particular, StableAnimator begins by computing image and face embeddings with off-the-shelf extractors, respectively and face embeddings are further refined by interacting with image embeddings using a global content-aware Face Encoder. Then, StableAnimator introduces a novel distribution-aware ID Adapter that prevents interference caused by temporal layers while preserving ID via alignment. During inference, we propose a novel Hamilton-Jacobi-Bellman (HJB) equation-based optimization to further enhance the face quality. We demonstrate that solving the HJB equation can be integrated into the diffusion denoising process, and the resulting solution constrains the denoising path and thus benefits ID preservation. Experiments on multiple benchmarks show the effectiveness of StableAnimator both qualitatively and quantitatively.

*Corresponding authors.

1. Introduction

Diffusion models [8, 16–18, 32, 38, 42, 43, 47, 54, 55, 58–60] have achieved remarkable success in image/video generation, significantly inspiring researches in image animation [22, 35, 49, 61, 68, 70]. In particular, human image animation explores generative models [22, 35, 39, 40, 49, 52, 61, 68, 70] to animate a reference image conditioned on a sequence of poses through synthesizing controllable human animation videos, offering diverse applications in entertainment content creation and virtual reality experiences. However, when dealing with pose sequences that exhibit significant motion variation, current approaches suffer from significant distortions and inconsistencies, particularly in facial regions destroying identity information.

To address this issue, there are a number of approaches exploring identity (ID) preservation [14, 23, 48, 65] for image generation, yet limited effort has been made for videos. While one could add temporal modeling layers to image

diffusion models, doing so would inevitably affect the original spatial priors. As Image-domain ID-preserving methods rely on stable spatial priors, the interference caused by temporal layers leads to unsatisfactory results. Thus, for image animation, how to preserve identity information while ensuring video fidelity is extremely challenging. Furthermore, recent animation models [35, 68] rely on FaceFusion [15] for post-processing, which also degrades the quality of animated videos, particularly for facial areas.

In light of this, we propose StableAnimator, consisting of dedicated modules for both the training and inference to maintain ID consistency for high-quality human image animation. StableAnimator first uses off-the-shelf extractors [7, 37] to obtain face and image embeddings for the reference image, respectively. Face embeddings are further refined by a global content-aware Face Encoder to enable interaction with the reference, enhancing face embeddings' perception of the reference's overall layout, such as backgrounds. The refined face embeddings are fed to a video diffusion model with a novel distribution-aware ID Adapter that ensures video fidelity while preserving ID clues. In particular, diffusion latents perform separate cross-attention with refined face and image embeddings respectively, with their means and variances computed. We then use respective means and variances to conduct the alignment between the resulting outputs. This alignment effectively mitigates interference from the temporal layers by progressively bringing two distributions closer at each step, ensuring ID consistency without compromising video fidelity.

During inference, to further enhance face quality and reduce reliance on post-processing tools, StableAnimator solves the Hamilton-Jacobi-Bellman (HJB) equation [2, 36] for face optimization. We find that solving the HJB equation corresponds with the core principles of diffusion denoising. Therefore, we incorporate the HJB equation into the inference process, which allows a controllable variable to guide and constrain the direction of the denoising process. In particular, the solution of HJB is used to update the latents for each denoising step, constraining the denoising path and directing the model toward optimal ID consistency. Since this procedure always adapts to the current distribution of denoised latents, the simultaneous denoising and face optimization effectively eliminates detail distortions. Thus, it can replace the previous over-reliance on third-party post-processing tools, such as face-swapping tools.

As shown in Fig. 1, while the latest animation model ControlNeXt [35] suffers from dramatic face and body distortion even with face swapping/restoration tools, StableAnimator can accurately animate the reference based on given poses while preserving ID consistency. Experiments on TikTok dataset [25] also show that StableAnimator outperforms ControlNeXt by 47.1% in CSIM [12] while achieving the best result (140.62) in FVD. CSIM is the face similarity between the animated frame and the reference.

In conclusion, our contributions are as follows: (1) We propose a global content-aware Face Encoder and a novel distribution-aware ID Adapter to enable the video diffusion model to incorporate face embeddings without compromising video fidelity. (2) We propose a novel HJB equation-based face optimization method that further enhances face quality while conducting conventional denoising. It is only active in the inference without training any diffusion components. To our knowledge, we are the first to explore video diffusion for end-to-end ID-preserving human image animation. (3) Experimental results on benchmark datasets show the superiority of our model over the SOTA.

2. Related Work

Diffusion for Video Generation. Renowned for the capacity in diversity and high-fidelity, diffusion models [8, 16–18, 32, 33, 38, 42, 43, 47, 57] have demonstrated significant success in the video generation. Compared with image generation, video generation requires additional temporal smoothness and temporal consistency. Current video generation models [4, 13, 26, 41, 45, 46, 50, 53, 56] tend to add temporal layers to pre-trained image generation diffusion models for joint spatio-temporal modeling. Some researchers replace the diffusion U-Net with the transformer [1, 19, 31, 34, 44, 62, 66] for facilitating generative performance. Inspired by previous image animation models [35, 68], we utilize Stable Video Diffusion (SVD [3]) as the backbone.

ID Consistency Image Generation. Studies have explored ID preservation in the image domain. LoRA [21] applies a few additional weights for customized dataset training, but it requires individual training for each character, restricting its flexibility. IP-Adapter-FaceID [65] attempts to directly separate the cross-attention layers for text features and face features, which potentially introduces the misalignment among features. PhotoMaker [30], FaceStudio [63], and InstantID [48] present hybrid ID preservation mechanisms for refining face embeddings. ConsistentID [23] designs a facial prompt generator for capturing facial details. PuLID [14] introduces contrastive alignment loss and accurate ID loss, ensuring ID fidelity. However, these models cannot be directly integrated into video diffusion models, as the temporal layers may alter the spatial distribution, resulting in domain mismatching with diffusion latents. This conflict between video fidelity and ID consistency ultimately degrades the quality of animations. By contrast, our StableAnimator can integrate ID information into video diffusion models via a distribution-aware ID Adapter, effectively resolving the above conflict.

Pose-guided Human Image Animation. Human image animation aims to transfer motion from a given pose se-

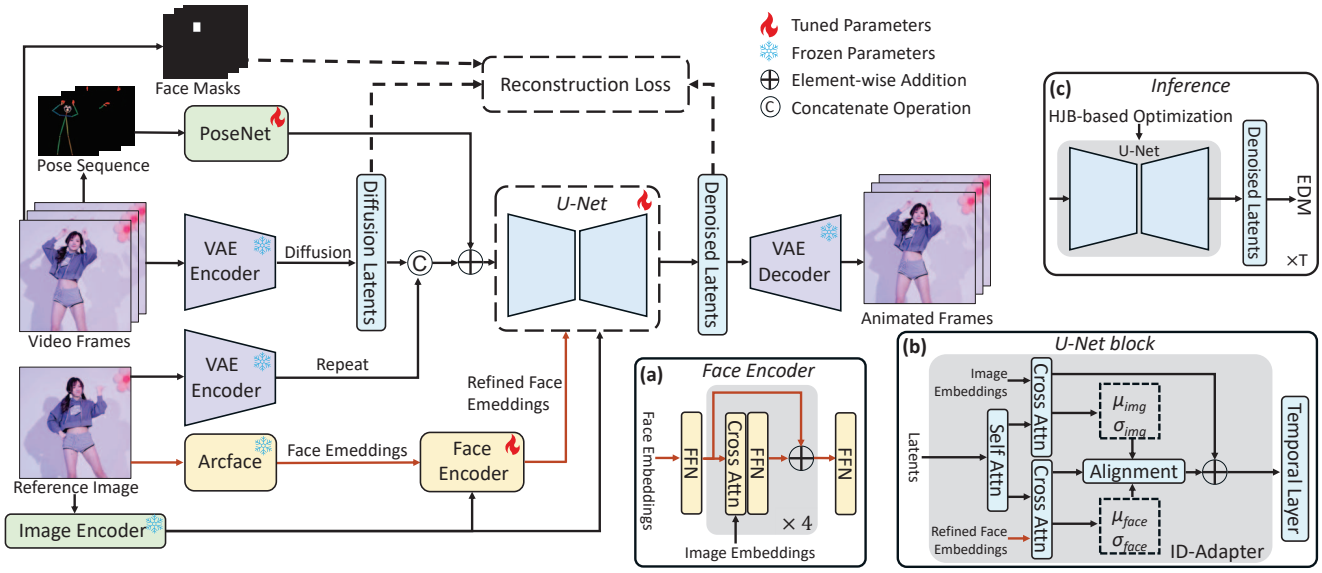

Figure 2. Architecture of StableAnimator. (a) and (b) refer to the structure of the Face Encoder and each block in the U-Net. Embeddings from the Image Encoder and Face Encoder are injected to each block of U-Net. Given the reference, we extract the image embeddings and face embeddings utilizing Image Encoder and Arcface. The face embeddings are fed into the FaceEncoder to enhance ID. Then, image embeddings and refined face embeddings are injected into the U-Net through the ID Adapter to ensure ID consistency.

quence to a reference human image. Early works [24, 39, 40] basically apply GANs [11] to animate the reference. However, animations of GAN-based models always encounter various artifacts. Recently, some studies have applied diffusion models to this field. Disco [49] is the first to use the diffusion model for image animation. MagicAnimate [61] and AnimateAnyone [22] both design their reference nets and pose nets to model poses and appearances independently. Champ [70] introduces 3D signal SMPL to enhance controllable capability. Unianimate [52] introduces Mamba [6] to the diffusion model for efficiency. MimicMotion [68] proposes the regional loss to reduce distortion. ControlNeXt [35] designs a convolution-based pose net to replace the heavy ControlNet [67]. However, previous animation models suffer from face distortion. MimicMotion [68] and ControlNeXt [35] utilize the third-party face-swapping tool FaceFusion [15] as post-processing to address this issue, yet this approach can degrade overall video quality. In this paper, our StableAnimator performs end-to-end human image animation that maintains ID consistency without relying on any post-processing tools.

3. Method

Illustrated in Fig. 2, StableAnimator is based on the commonly used SVD [3] following previous works [35, 68]. A reference image is processed through the diffusion model via three pathways: (1) Transformed into a latent code by a frozen VAE Encoder [28]. The latent code is duplicated to match video frames, then concatenated with main latents. (2) Encoded by the CLIP Image Encoder [37] to obtain image embeddings, which are fed to each cross-attention block of a denoising U-Net and our Face Encoder, respectively, to modulate the synthesized appearance. (3) Input to Arcface [7] to gain face embeddings, which are subsequently refined for further alignment via our Face Encoder. Refined face embeddings are then fed to the denoising U-Net. More details are described in Sec. 3.1. A PoseNet with a similar architecture as AnimateAnyone [22] extracts the features of the pose sequence, which are then added to the noisy latents.

We replace the original input video frames with random noise during inference, while the other inputs stay the same. We propose a novel HJB-equation-based face optimization to enhance ID consistency and eliminate reliance on third-party post-processing tools. It integrates the solution process of the HJB equation into the denoising, allowing optimal gradient direction toward high ID consistency as detailed in Sec. 3.2.

3.1. ID-preserving During Training

Global Content-aware Face Encoder. Our goal is to animate the reference image under the guidance of the pose sequence while preserving the ID of the reference image. Directly feeding face embeddings into the U-Net can enrich the diffusion model with face-related information, but lacks awareness of the global context (layout and background) in the reference image before being injected into the U-Net. As a result, ID-irrelevant elements in the reference image bring noise to face modeling, degrading the overall quality of animations. To address this, we propose a Global Content-Aware Face Encoder, in which the face embeddings go through multiple cross-attention blocks to interact with the reference image embeddings as shown in Fig. 2.

Distribution-aware ID Adapter. The outputs of the Face Encoder are then fed to our ID Adapter for further alignment to avoid the distortion of spatial features occurring when directly incorporating image-domain ID-preserving methods [14, 23, 39, 48] into video diffusion model. Feature distortion describes the misalignment between face embeddings and spatial diffusion latents, caused by distribution shifts when temporal layers are added at each denoising step. Image-domain ID-preserving methods rely heavily on a stable spatial distribution of diffusion latents, but temporal layers often alter this distribution, leading to instability in ID preservation. Such distortion causes a conflict between maintaining high video fidelity and preserving ID consistency. Thus, animated videos often suffer from noticeable blurring effects and can even lose background details. The Distribution-aware ID Adapter modifies each spatial layer of the U-Net, as shown in Fig. 2 (b). Before each temporal modeling, our ID Adapter aligns refined face embeddings with diffusion latents based on their feature distributions, effectively avoiding feature distortion.

Concretely, following the standard operation of spatial layers in the diffusion model, we first apply spatial self-attention on latents $\boldsymbol{z}_i$. The latents of the U-Net perform cross-attention with image embeddings $\boldsymbol{emb}_{img}$ and refined face embeddings $\boldsymbol{emb}_{face}$, respectively:

$$\begin{aligned} \boldsymbol{z}_i &= \texttt{SAttn}(\boldsymbol{z}_i), \\ \boldsymbol{z}_i^{img} &= \texttt{CAttn}(\boldsymbol{z}_i, \boldsymbol{emb}_{img}), \\ \boldsymbol{z}_i^{face} &= \texttt{CAttn}(\boldsymbol{z}_i, \boldsymbol{emb}_{face}), \end{aligned} \tag{1}$$

where $\texttt{SAttn}(\cdot)$ and $\texttt{CAttn}(\cdot)$ refer to self-attention and cross-attention operations. To align $\boldsymbol{z}_i^{img}$ and $\boldsymbol{z}_i^{face}$, we enforce $\frac{\boldsymbol{z}_i^{img}-\boldsymbol{\mu}_{img}}{\boldsymbol{\sigma}_{img}} = \frac{\boldsymbol{z}_i^{face}-\boldsymbol{\mu}_{face}}{\boldsymbol{\sigma}_{face}}$, where $\boldsymbol{\mu}_{img/face}$ and $\boldsymbol{\sigma}_{img/face}$ refer to the mean and standard deviation of $\boldsymbol{z}_i^{img/face}$, respectively. If the equation above holds, the feature distributions on both sides are basically in the same domain. Thus, the aligned $\boldsymbol{z}_i^{face}$ is element-wise added to $\boldsymbol{z}_i^{img}$ for maintaining ID consistency:

$$\begin{aligned} \bar{\boldsymbol{z}}_i^{face} &= \frac{\boldsymbol{z}_i^{face} - \boldsymbol{\mu}_{face}}{\boldsymbol{\sigma}_{face}} \times \boldsymbol{\sigma}_{img} + \boldsymbol{\mu}_{img}, \\ \bar{\boldsymbol{z}}_i &= \bar{\boldsymbol{z}}_i^{face} + \boldsymbol{z}_i^{img}. \end{aligned} \tag{2}$$

The outputs of our ID Adapter $\bar{\boldsymbol{z}}_i$ are further fed to temporal layers for temporal modeling. When spatial distribution is altered by temporal layers, the aligned $\bar{\boldsymbol{z}}_i^{face}$ remains in the same domain as $\boldsymbol{z}_i^{img}$, enabling the original $\boldsymbol{z}_i^{face}$ to reduce reliance on the unstable spatial distribution. Thus, subsequent temporal modeling does not impede the injection of ID information into the U-Net.

3.2. ID-preserving During Inference

To improve ID consistency, the latest animation works [35, 68] use a third-party face-swapping tool FaceFusion [15] for post-processing faces. However, animations suffer from overall quality degradation due to excessive reliance on post-processing tools. The reason is that post-processing tools can disrupt the original pixel distribution, as faces generated by third-party tools are clearly not aligned with the domain of original animations. To address this issue, inspired by the HJB equation [2, 5, 36], we propose the HJB Equation-based Face Optimization. The HJB equation guides optimal variable selection at each moment in a dynamic system to maximize the cumulative reward. In our setting, this reward refers to ID consistency, which we aim to enhance by integrating the HJB equation with the diffusion denoising process. The variable refers to the predicted sample by the diffusion model at each denoising iteration. We first introduce the process of our face optimization and then demonstrate its rationale.

Algorithm 1 Face Optimization ($\sigma(t) = t$ and $s(t) = 1$)

Input: $\mathtt{D}_\theta(\boldsymbol{x};\boldsymbol{\sigma}), t_{i\in\{0,\dots,N\}}, \boldsymbol{\gamma}_{i\in\{0,\dots,N-1\}}, \boldsymbol{y}$
Sample $\boldsymbol{x}_0 \sim \mathcal{N}(0, t_0^2\boldsymbol{I})$ ▷ $\mathtt{D}_\theta(\boldsymbol{x};\boldsymbol{\sigma})$ is a diffusion model
For $i \in \{0,\dots,N-1\}$ **do** ▷ $t_{i\in\{0,\dots,N\}}$ are timesteps
  $\boldsymbol{\gamma}_i = 0$ ▷ $\boldsymbol{\gamma}_{i\in\{0,\dots,N-1\}}$ are pre-defined factors.
  if $t_i \in [\boldsymbol{S}_{t_{\text{min}}}, \boldsymbol{S}_{t_{\text{max}}}]$: ▷ $\boldsymbol{y}$ is the reference image.
    $\boldsymbol{\gamma}_i = \min\left(\frac{\boldsymbol{S}_{\text{churn}}}{N}, \sqrt{2}-1\right)$
  Sample $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{S}_{\text{noise}}^2\boldsymbol{I})$
  $\hat{t}_i = t_i + \boldsymbol{\gamma}_i t_i$
  $\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \sqrt{\hat{t}_i^2 - t_i^2}\boldsymbol{\epsilon}_i$
  $\boldsymbol{x}_{\text{pred}} = \mathtt{D}_\theta(\hat{\boldsymbol{x}}_i; \hat{t}_i)$
  $\boldsymbol{x}_{\text{op}} = \boldsymbol{x}_{\text{pred}}.\texttt{clone}().\texttt{detach}()$ ▷ Starting optimization
  $\boldsymbol{op} = \texttt{Adam}([\boldsymbol{x}_{\text{op}}], \boldsymbol{\eta})$ ▷ Adam optimizer
  $\boldsymbol{x}_{\text{op}}$.requires_grad = True ▷ $\boldsymbol{x}_{\text{op}}$ is a HJB variable
  For $k \in \{1,2,\dots,10\}$ **do** ▷ k is the optimization step
    $\boldsymbol{f}_{\text{pred}} = \texttt{Decoder}(\boldsymbol{x}_{\text{op}})$ ▷ Decoder is a VAE decoder
    $\boldsymbol{loss} = (1 - \texttt{Cos}(\texttt{Arc}(\boldsymbol{f}_{\text{pred}}), \texttt{Arc}(\boldsymbol{y}))).\text{abs}().\text{mean}()$
    $\boldsymbol{op}$.zero_grad()
    $\boldsymbol{loss}$.backward(retain_graph=True)
    $\boldsymbol{op}$.step()
  $\boldsymbol{x}_{\text{pred}} = \boldsymbol{x}_{\text{op}}$ ▷ End of Optimization
  $\boldsymbol{d}_i = (\hat{\boldsymbol{x}}_i - \boldsymbol{x}_{\text{pred}})/\hat{t}_i$
  $\boldsymbol{x}_{i+1} = \hat{\boldsymbol{x}}_i + (t_{i+1} - \hat{t}_i)\boldsymbol{d}_i$
  if $t_{i+1} \neq 0$:
    $\boldsymbol{d}'_i = (\boldsymbol{x}_{i+1} - \mathtt{D}_\theta(\boldsymbol{x}_{i+1}; t_{i+1}))/t_{i+1}$
    $\boldsymbol{x}_{i+1} = \hat{\boldsymbol{x}}_i + (t_{i+1} - \hat{t}_i)\left(\frac{1}{2}\boldsymbol{d}_i + \frac{1}{2}\boldsymbol{d}'_i\right)$
return $\boldsymbol{x}_N$

In particular, we optimize the predicted sample $\boldsymbol{x}_{\text{pred}}$ by minimizing the face similarity distance between $\boldsymbol{x}_{\text{pred}}$ and the reference before employing denoising (EDM [27]) at each step. The details are in the Algorithm 1, following the structure of the Algorithm 2 in the EDM paper [27]. $\boldsymbol{S}_{\text{noise}}$,$\boldsymbol{S}_{\text{churn}}$, $\boldsymbol{S}_{t_{\text{min}}}$, and $\boldsymbol{S}_{t_{\text{max}}}$ are the pre-defined values of EDM. $\texttt{Arc}(\cdot)$ and $\boldsymbol{\eta}$ are Arcface [7] and a learning rate. We employ our optimization to refine the prediction of the diffusion regarding the face similarity with the reference.

The optimized $\boldsymbol{x}_{\text{pred}}$ can steer the denoising process forward in a way that maximizes ID consistency. As our optimization relies on the current distribution of denoised latents from diffusion, this parallel operation of denoising and optimizing ID consistency effectively reduces detail distortions, enhancing face quality.

Furthermore, we prove that the solving process of the HJB equation [2, 5, 36] can be integrated with the diffusion denoising process, as demonstrated below. The basic HJB Equation can be described as:

$$\frac{\partial \mathtt{V}(\boldsymbol{x},t)}{\partial t} + \mathtt{max}_{\boldsymbol{c}}[\mathtt{f}(\boldsymbol{x},\boldsymbol{c}) + \frac{\partial \mathtt{V}(\boldsymbol{x},t)}{\partial \boldsymbol{x}} \cdot \mathtt{g}(\boldsymbol{x},\boldsymbol{c})] = 0, \tag{3}$$

where $\mathtt{V}(\boldsymbol{x},t)$ refers to the value function, representing the minimum cost from state $\boldsymbol{x}$ at time t. $\mathtt{f}(\boldsymbol{x},\boldsymbol{c})$ is the immediate cost under the condition $\boldsymbol{c}$ in state $\boldsymbol{x}$. $\mathtt{g}(\cdot)$ depicts the system dynamics. In our settings, the condition $\boldsymbol{c}$ indicates the face-aware variable. Following the previous work [5], the solving process is formulated as:

$$\mathtt{min}_{\boldsymbol{c}_t} \int_0^1 \frac{1}{2} \left\|\boldsymbol{c}_t\right\|_2^2 dt + \frac{\boldsymbol{r}}{2} \left\|\boldsymbol{X}_1 - \boldsymbol{x}_1\right\|_2^2, \boldsymbol{X}_1 \sim \boldsymbol{p}_{data}, \tag{4}$$

s.t. $d\boldsymbol{X}_t = \boldsymbol{c}_t dt$ and $\boldsymbol{X}_0 = \boldsymbol{x}_0$ (Gaussian noise). $\boldsymbol{r}$ is the terminal cost coefficient. In our work, we normalize denoising timesteps t' (from $\boldsymbol{T}$ to 0) to $[0,1]$ and set $t = 1 - t'$. $\boldsymbol{T}$ is the maximum denoising timestep. $\boldsymbol{X}_t$ and $\boldsymbol{x}_t$ refer to the groundtruth sample and the predicted sample by the model. Thus, $\boldsymbol{x}_{\text{pred}}$ in Algorithm 1 is equivalent to $\boldsymbol{x}_1$. Following the Pontryagin Maximum Principle [29], we can construct the Hamiltonian equation:

$$\mathtt{H}(t, \boldsymbol{X}, \boldsymbol{c}_t, \boldsymbol{\gamma}) = -\frac{1}{2} \left\|\boldsymbol{c}_t\right\|_2^2 + \boldsymbol{\gamma} \boldsymbol{c}_t, \tag{5}$$

where $\boldsymbol{\gamma}$ refers to a coefficient. To minimize Eq. 5, we set $\frac{\partial \mathtt{H}}{\partial \boldsymbol{c}_t} = 0$. The optimal Hamiltonian is described as:

$$\mathtt{H}^* = \mathtt{H}(t, \boldsymbol{X}, \boldsymbol{c}_t^*, \boldsymbol{\gamma}) = \frac{1}{2}\boldsymbol{\gamma}^2, \text{where } \boldsymbol{c}_t^* = \boldsymbol{\gamma}. \tag{6}$$

Then we solve the Hamiltonian equation of motion:

$$\begin{aligned} \frac{d\boldsymbol{X}_t}{dt} &= \frac{\partial \mathtt{H}^*}{\partial \boldsymbol{\gamma}} = \boldsymbol{\gamma}, \\ \frac{d\boldsymbol{\gamma}}{dt} &= \frac{\partial \mathtt{H}^*}{\partial \boldsymbol{X}} = 0. \end{aligned} \tag{7}$$

At the final step $t = 1$, from Eq. 4 and Eq. 5, we can obtain $\boldsymbol{\gamma}_1 = -\boldsymbol{r} \cdot (\boldsymbol{X}_1 - \boldsymbol{x}_1)$. From Eq. 7, we can see that $\boldsymbol{\gamma}$ is a variable independent of t, thereby obtaining $\boldsymbol{\gamma} = \boldsymbol{\gamma}_1 = -\boldsymbol{r} \cdot (\boldsymbol{X}_1 - \boldsymbol{x}_1)$. We can also get $\boldsymbol{X}_t = \boldsymbol{X}_0 + \boldsymbol{\gamma} t \rightarrow \boldsymbol{X}_1 = \boldsymbol{X}_0 + \boldsymbol{\gamma}$ and $\boldsymbol{X}_0 = \boldsymbol{X}_t - \boldsymbol{\gamma} t$. We then obtain $\boldsymbol{c}_t^*$:

$$\begin{aligned} & \boldsymbol{X}_1 = \boldsymbol{X}_0 + \boldsymbol{\gamma} = \boldsymbol{X}_t - \boldsymbol{\gamma} t + \boldsymbol{\gamma} \\ \rightarrow \quad & \boldsymbol{\gamma} = -\boldsymbol{r} \cdot (\boldsymbol{X}_1 - \boldsymbol{x}_1) = -\boldsymbol{r} \cdot (\boldsymbol{X}_t - \boldsymbol{\gamma} t + \boldsymbol{\gamma} - \boldsymbol{x}_1), \\ \rightarrow \quad & \boldsymbol{c}_t^* = \boldsymbol{\gamma} = \frac{\boldsymbol{r}(\boldsymbol{x}_1 - \boldsymbol{X}_t)}{1 + \boldsymbol{r}(1-t)}. \end{aligned} \tag{8}$$

When $\boldsymbol{r} \rightarrow \infty$, following Eq. 4 ($d\boldsymbol{X}_t = \boldsymbol{c}_t dt$) and certainty equivalence [5, 10] (the stochastic case), we have

$$d\boldsymbol{X}_t = \frac{\boldsymbol{x}_1 - \boldsymbol{X}_t}{1-t} dt + d\boldsymbol{w}_t, \tag{9}$$

where $\boldsymbol{w}_t$ is Brownian motion [5]. According to EDM [27] in SVD [3], where $\boldsymbol{X}_{t'} = \boldsymbol{X}_{data} + t'\boldsymbol{\varepsilon}$ and $\boldsymbol{X}_{data} \sim \boldsymbol{p}_{data}$, the current state $\boldsymbol{X}_{t'}$ is converted to $\boldsymbol{X}_t = \boldsymbol{X}_1 + (1-t)\boldsymbol{\varepsilon}$ in our settings. We use the following Tweedie's formula [9]

$$\mathtt{E}[\boldsymbol{\theta}|\boldsymbol{x}] = \boldsymbol{x} + \boldsymbol{\sigma}^2 \cdot \nabla \log \mathtt{p}(\boldsymbol{x}), \tag{10}$$

where $\boldsymbol{x}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ and $\mathtt{p}(\cdot)$ is the marginal density of $\boldsymbol{x}$, to reform $\boldsymbol{X}_1$:

$$\boldsymbol{X}_1 = \mathtt{E}[\boldsymbol{X}_1|\boldsymbol{X}_t] = \boldsymbol{X}_t + (1-t)^2 \nabla \log \mathtt{p}(\boldsymbol{X}_t). \tag{11}$$

$\boldsymbol{x}_1$ aims to approximate $\boldsymbol{X}_1$. Thus, we substitute Eq. 11 in Eq. 9 for obtaining the ultimate formula:

$$\begin{aligned} d\boldsymbol{X}_t &= \frac{\boldsymbol{X}_t + (1-t)^2 \nabla \log \mathtt{p}(\boldsymbol{X}_t) - \boldsymbol{X}_t}{1-t} dt + d\boldsymbol{w}_t \\ &= (1-t) \cdot \nabla \log \mathtt{p}(\boldsymbol{X}_t) dt + d\boldsymbol{w}_t. \end{aligned} \tag{12}$$

It is evident that Eq. 12 and SDE formulation [43] are structurally the same, thus we can seamlessly incorporate the solution process of the HJB equation into the diffusion denoising for ID preservation.

3.3. Training

As illustrated in Fig. 2, we use the reconstruction loss to train our model, with trainable components including a U-Net, a FaceEncoder, and a PoseNet. We introduce face masks $\boldsymbol{M}$, extracted by ArcFace [7] from the input video frames to enhance the modeling of face regions:

$$\mathcal{L} = \mathbb{E}_{\varepsilon}(\|(\boldsymbol{z}_{gt} - \boldsymbol{z}_{\varepsilon}) \odot (1 + \boldsymbol{M})\|^2), \tag{13}$$

where $\boldsymbol{z}_{gt}$ and $\boldsymbol{z}_{\varepsilon}$ refer to diffusion latents and denoised latents in Fig. 2, respectively.

4. Experiments

4.1. Implementation Details

Since previous works do not open-source their training datasets, we collect 3K videos (60-90 seconds long) from the internet to train our model. We utilize DWPose [64] and Arcface [7] to extract skeleton poses and face embeddings/masks. Following previous works [22, 49, 52, 61, 70], we evaluate our model on TikTok dataset [25]. We conduct additional experiments on 100 unseen videos, referred to the Unseen100 dataset, selected from the internet to assess the robustness of our model. Following recent animation models [35, 68], the U-Net utilizes pre-trained weights of SVD [3], while the PoseNet and Face Encoder are trained from scratch. Our ID-Adapter uses pre-trained weights of spatial cross-attention blocks in SVD. Our model is trained for 20 epochs on 4 NVIDIA A100 80G GPUs, with a batch size of 1 per GPU. The learning rate is set to $1e$-5.

Table 1. Quantitative comparisons on TikTok dataset and Unseen100. Mem refers to GPU memory when manipulating 16 frames (576×1024). In the table elements a / b, a, and b refer to the result on the TikTok dataset and Unseen100, respectively. We reference competitors' results on the TikTok dataset from their papers, with $-$ indicating missing reports.

Model	L1 (E-4)↓	PSNR [20]↑	PSNR* [49]↑	SSIM↑	LPIPS↓	CSIM [12]↑	FVD↓	Mem↓
MRAA [40]	3.21 / 3.62	- / 26.62	18.14 / 17.28	0.672 / 0.692	0.296 / 0.313	0.248 / 0.221	284.82 / 540.35	**5.4G**
DisCo [49]	3.78 / 3.74	29.03 / 25.23	16.55 / 15.21	0.668 / 0.702	0.292 / 0.302	0.315 / 0.267	292.80 / 544.64	18.7G
MagicAnimate [61]	3.13 / 3.23	29.16 / 27.03	- / 17.11	0.714 / 0.746	0.239 / 0.264	0.462 / 0.338	179.07 / 398.94	20.84G
AnimateAnyone [22]	- / 3.15	29.56 / 27.14	- / 17.14	0.718 / 0.759	0.285 / 0.251	0.457 / 0.316	171.90 / 383.45	11.18G
Champ [70]	2.94 / 3.02	29.91 / 27.78	- / 17.35	0.802 / 0.772	0.234 / 0.234	0.350 / 0.304	160.82 / 373.77	13.2G
Unianimate [52]	**2.66** / 2.82	30.77 / 27.46	20.58 / 18.64	**0.811** / 0.778	**0.231** / 0.253	0.479 / 0.347	148.06 / 394.32	6.11G
MimicMotion [68]	5.85 / 3.55	- / 22.94	14.44 / 13.97	0.601 / 0.733	0.414 / 0.370	0.262 / 0.242	326.57 / 604.13	8.6G
ControlNeXt [35]	6.20 / 2.90	- / 25.28	13.83 / 14.84	0.615 / 0.743	0.416 / 0.262	0.360 / 0.264	326.57 / 389.45	12.23G
StableAnimator	2.87 / **2.71**	**30.81 / 28.85**	**20.66 / 18.85**	0.801 / **0.784**	0.232 / **0.223**	**0.831 / 0.805**	**140.62 / 349.94**	12.50G

4.2. Comparison with State-of-the-Art Methods

Quantitative results. We compare with recent human image animation models, including GAN-based models (MRAA [40]) and diffusion-based models (DisCo [49], AnimateAnyone [22], MagicAnimate [61], Champ [70], Unianimate [52], MimicMotion [68], ControlNeXt [35]). Based on previous studies that assess quantitative results using the self-driven and reconstruction approach, we perform quantitative comparisons with the above competitors on the TikTok dataset [25] and Unseen100, comprising complex motion and appearance information. Notably, all competitors are trained on our dataset before evaluating on Unseen100 to ensure a fair comparison. The results are shown in Table 1. CSIM [12] evaluates the cosine similarity between the facial embeddings of two images. We observe that our StableAnimator surpasses all competitors regarding face quality (CSIM) and video fidelity (FVD) while maintaining relatively high single-frame quality. Specifically, StableAnimator outperforms the leading competitor, Unianimate, by 36.9% and 45.8% in CSIM across two datasets, without sacrificing video fidelity and single-frame quality.

Qualitative Results. The qualitative results are shown in Fig. 3. Notably, qualitative results in the paper are in the cross-ID setting [70]. Disco [49], MagicAnimate [61], AnimateAnyone [22], and Champ [70] exhibit face/body distortion and clothing changes, while Unianimate [52] accurately modifies the reference motion, and MimicMotion [68] and ControlNeXt [35] effectively preserve clothing details. However, all competitors struggle to maintain reference identities. In contrast, our StableAnimator accurately animates images based on the given pose sequences while preserving reference identities, highlighting the superiority of our model in identity retention and in generating precise, vivid animations.

4.3. Ablation Study

ID Consistency Preservation. We conduct an ablation study to demonstrate the contributions of core components in StableAnimator, as shown in Table 2 and Fig. 4. Notably, all quantitative ablation studies are on the Unseen100 dataset. We can see that removing the core components significantly degrades performance, particularly in face-related regions (CSIM), highlighting that our components significantly enhance both video fidelity and single-frame quality while preserving high ID consistency.

Table 2. Ablation study on core components. Face Masks and Alignment refer to face masks in the loss and distribution alignment of our ID Adapter.

Model	L1↓	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	FVD↓
w/o Face Masks	3.01E-4	24.10	0.665	0.281	0.639	382.25
w/o Face Encoder	3.08E-4	22.25	0.674	0.282	0.594	385.91
w/o Alignment	3.11E-4	23.45	0.713	0.276	0.716	412.52
w/o Optimization	2.86E-4	27.17	0.769	0.245	0.782	365.43
Ours	**2.71E-4**	**28.85**	**0.784**	**0.223**	**0.805**	**349.94**

We further conduct an ablation study regarding current face enhancement approaches, as shown in Table 3 and Fig. 5. We replace our components with the commonly used IP-Adapter and FaceFusion. By analyzing the results, we can gain the following observations: (1) IP-Adapter can improve the ID consistency, while the video fidelity and single-frame quality dramatically degrade. The plausible reason is that directly inserting the IP-Adapter hinders its ability to adapt to spatial representation distribution variations during temporal modeling, thereby deteriorating the capacity of the video diffusion model. (2) The third-party post-processing face-swapping tool FaceFusion refines the face quality but relatively degrades the video fidelity. The underlying reason is that the third-party post-processing operates in a different domain from the diffusion model, leading to a loss of semantic details and disrupting video fidelity. (3) StableAnimator can significantly refine the face quality while maintaining high video fidelity since our model remains in the same domain as the video diffusion model due to the distribution-aware end-to-end pipeline.

Feature Alignment. We conduct a comparison between our distribution alignment in the ID-Adapter and other types of feature injection, as shown in Table 4 and Fig. 5. Norm

Figure 3. Qualitative comparisons with state-of-the-art methods. More examples can be found in the supplementary material.

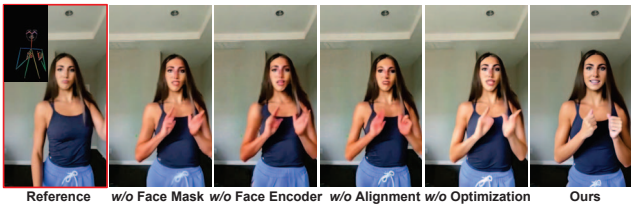

Figure 4. Ablations on core components of StableAnimator.

Table 3. Ablation study on face enhancement methods. *w/o* Face refers to the exclusion of any face-related strategies.

Model	L1↓	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	FVD↓
w/o Face	2.83E-4	26.75	0.741	0.264	0.324	371.38
IP-Adapter [65]	3.88E-4	18.86	0.672	0.287	0.511	484.77
FaceFusion [15]	3.31E-4	23.05	0.734	0.265	0.798	405.16
Ours	**2.71E-4**	**28.85**	**0.784**	**0.223**	**0.805**	**349.94**

refers to $\bar{z}_i^{face} = \frac{z_i^{face} - \mu_{face}}{\sigma_{face}}$. We can see that Addition and Norm fail to eliminate the interference of spatial feature distortion after temporal modeling, thereby achieving suboptimal results. By contrast, our alignment integrates the mean and standard deviation from both cross-attention features, significantly mitigating the impact of feature distortion.

Table 4. Ablation study on the distribution-based alignment. Addition and Norm refer to element-wise addition and normalization.

Model	L1↓	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	FVD↓
Addition	3.11E-4	23.45	0.713	0.276	0.716	412.52
Norm	2.73E-4	26.67	0.758	0.257	0.776	382.49
Ours	**2.71E-4**	**28.85**	**0.784**	**0.223**	**0.805**	**349.94**

Face Optimization. To validate the significance of our face optimization strategy, we conduct an ablation regarding different diffusion backbones. The results are in Table 5 and Fig. 6. MagicAnimate is based on SD [38]+AnimateDiff [13]. We have the following observations: (1) Common face enhancement strategies (IP-Adapter and FaceFusion) also degrade the video fidelity and single-frame quality of MagicAnimate, indicating that spatial feature distortion indeed occurs across different diffusion-based backbones. (2) Magic+Opt boosts overall performance, showing that our face optimization enhances the diffusion model even without any explicit introduction of face-related adapters. The results of Magic+IP+Opt indi-

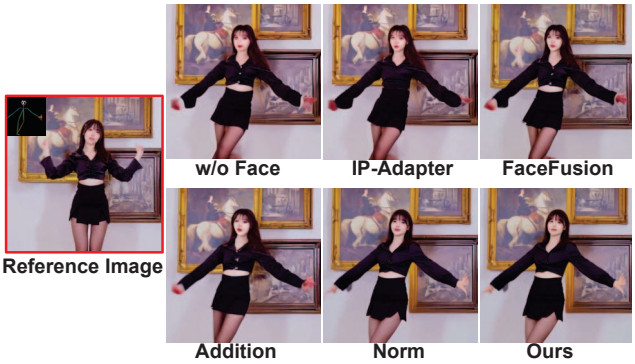

Figure 5. Ablation study on face enhancement strategies.

Table 5. Ablation study on the optimization. Magic, IP, ID, FE, and Opt refer to MagicAnimate, IP-Adapter, our ID Adapter, our Face Encoder, and our Optimization, respectively.

Model	L1↓	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	FVD↓
Magic+IP	3.85E-4	23.14	0.689	0.286	0.541	836.33
Magic+FaceFusion	3.31E-4	26.42	0.725	0.268	0.796	412.40
Magic+Opt	3.02E-4	27.56	0.762	0.258	0.480	381.61
Magic+IP+Opt	3.61E-4	26.12	0.714	0.279	0.624	754.34
Magic+FE+ID	2.85E-4	27.89	0.767	0.248	0.775	376.43
Magic+FE+ID+Opt	**2.69E-4**	**28.13**	**0.775**	**0.241**	**0.798**	**355.23**

cate that our optimization can mitigate the deterioration in fidelity due to the introduction of IP-Adapter while improving face quality to some extent. (3) The last two rows of Table 5 show that our face optimization can still work in the different diffusion-based backbone.

4.4. Applications and User Study

Long Animation. We conduct a qualitative comparison between our StableAnimator and current animation models in long animation generation. Inspired by MimicMotion [68], we follow the same pipeline for synthesizing long videos. The results are shown in Sec. D of the Supp. Each pose sequence contains over 300 frames with complex motion, while the references include intricate details of appearances and backgrounds. The results show that competitors suffer from blurry noises and face distortion. By contrast, our model can effectively handle long human image animation in high fidelity while preserving identities.

Multi-Person Animation. We experiment on multi-person animation, as shown in Sec. E of the Supp. We can see that our model is capable of animating multiple individuals.

User Study. We conduct a user study on 30 selected videos to evaluate the human preference between our StableAnimator and other competitors. The participants are basically university students and faculties. In each case, participants are first presented with the reference image and the pose sequence. Then we provide two videos (one is generated by our StableAnimator and the other is synthesized by a competitor) in random orders. Participants are then asked to answer the following questions: M-A/A-A/B-A/I-A: "Which one has better motion/appearance/background/ID alignment with the reference". Table 6 shows the superiority of our model regarding subjective evaluation.

Figure 6. Ablation study on different backbones.

Table 6. User preference of StableAnimator compared with other competitors. Higher indicates users prefer more to our model.

Model	M-A	A-A	B-A	I-A
DisCo [49]	95.6%	96.8%	94.2%	98.7%
MagicAnimate [61]	94.5%	94.8%	92.7%	97.4%
AnimateAnyone [22]	94.8%	93.1%	92.5%	98.3%
Champ [70]	95.0%	91.3%	91.7%	96.6%
Unianimate [52]	89.7%	88.4%	90.5%	95.8%
MimicMotion [68]	95.3%	95.5%	94.1%	97.6%
ControlNeXt [35]	93.6%	92.4%	90.3%	96.2%

5. Conclusion

In this paper, we proposed StableAnimator, a video diffusion model with dedicated modules for training and inference to generate ID-preserving human image animations. StableAnimator first used off-the-shelf models to gain image and face embeddings. To capture the global context of the reference, StableAnimator introduced a Face Encoder to refine face embeddings. StableAnimator further designed an ID-Adapter, which applied alignment to mitigate the interference from temporal modeling, enabling seamless face embedding integration without video fidelity loss. During inference, to further enhance face quality, StableAnimator incorporated the HJB equation alongside diffusion denoising for face optimization. Experimental results across various datasets demonstrated the superiority of our model in producing high-quality ID-preserving human animations.

Acknowledgement. This project was supported by NSFC under Grant No.62032006 and No. 62472098.

References

[1] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024. 2

[2] Martino Bardi, Italo Capuzzo Dolcetta, et al. *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Springer, 1997. 2, 4, 5

[3] 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. 2, 3, 5

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[5] Tianrong Chen, Jiatao Gu, Laurent Dinh, Evangelos A Theodorou, Joshua Susskind, and Shuangfei Zhai. Generative modeling with phase stochastic bridges. In *ICLR*, 2024. 4, 5

[6] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *ICML*, 2024. 3

[7] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2, 3, 4, 5

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 1, 2

[9] Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 2011. 5

[10] Wendell H Fleming and Raymond W Rishel. *Deterministic and stochastic optimal control*. Springer Science & Business Media, 2012. 5

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 3

[12] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 2, 6

[13] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 2, 7

[14] Zinan Guo, Yanze Wu, Zhuowei Chen, Lang Chen, and Qian He. Pulid: Pure and lightning id customization via contrastive alignment. In *NeurIPS*, 2024. 1, 2, 4

[15] Ruhs Henry. Facefusion. https://github.com/facefusion/facefusion, 2024. 1, 2, 3, 4, 7

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[18] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *JMLR*, 2022. 1, 2

[19] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 2

[20] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, 2010. 6

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021. 2

[22] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*, 2024. 1, 3, 5, 6, 8

[23] Jiehui Huang, Xiao Dong, Wenhui Song, Hanhui Li, Jun Zhou, Yuhao Cheng, Shutao Liao, Long Chen, Yiqiang Yan, Shengcai Liao, et al. Consistentid: Portrait generation with multimodal fine-grained identity preserving. *arXiv preprint arXiv:2404.16771*, 2024. 1, 2, 4

[24] Zhichao Huang, Xintong Han, Jia Xu, and Tong Zhang. Few-shot human motion transfer by personalized geometry and texture modeling. In *CVPR*, 2021. 3

[25] Yasamin Jafarian and Hyun Soo Park. Learning high fidelity depths of dressed humans by watching social media dance videos. In *CVPR*, 2021. 2, 5, 6

[26] Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *CVPR*, 2025. 2

[27] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 4, 5

[28] Diederik P Kingma. Auto-encoding variational bayes. In *ICLR*, 2014. 3

[29] Donald E Kirk. *Optimal control theory: an introduction*. Courier Corporation, 2004. 5

[30] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, 2024. 2

[31] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 2

[32] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2021. 1, 2

[33] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 2

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2

[35] Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. Controlnext: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024. 1, 2, 3, 4, 5, 6, 8

[36] Shige Peng. Stochastic hamilton–jacobi–bellman equations. *SIAM Journal on Control and Optimization*, 1992. 2, 4, 5

[37] 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 *ICML*, 2021. 2, 3

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 7

[39] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, 2019. 1, 3, 4

[40] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, 2021. 1, 3, 6

[41] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 1, 2

[43] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 1, 2, 5

[44] Rui Tian, Qi Dai, Jianmin Bao, Kai Qiu, Yifan Yang, Chong Luo, Zuxuan Wu, and Yu-Gang Jiang. Reducio! generating 1024×1024 video within 16 seconds using extremely compressed motion latents. *arXiv preprint arXiv:2411.13552*, 2024. 2

[45] Shuyuan Tu, Qi Dai, Zhi-Qi Cheng, Han Hu, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motioneditor: Editing video motion via content-aware diffusion. In *CVPR*, 2024. 2

[46] Shuyuan Tu, Qi Dai, Zihao Zhang, Sicheng Xie, Zhi-Qi Cheng, Chong Luo, Xintong Han, Zuxuan Wu, and Yu-Gang Jiang. Motionfollower: Editing video motion via lightweight score-guided diffusion. *arXiv preprint arXiv:2405.20325*, 2024. 2

[47] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 1, 2

[48] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 1, 2, 4

[49] 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. In *CVPR*, 2024. 1, 3, 5, 6, 8

[50] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024. 2

[51] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[52] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024. 1, 3, 5, 6, 8

[53] Yanhui Wang, Jianmin Bao, Wenming Weng, Ruoyu Feng, Dacheng Yin, Tao Yang, Jingxu Zhang, Qi Dai, Zhiyuan Zhao, Chunyu Wang, et al. Microcinema: A divide-and-conquer approach for text-to-video generation. In *CVPR*, 2024. 2

[54] Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. Art-v: Auto-regressive text-to-video generation with diffusion models. In *CVPRW*, 2024. 1

[55] Zejia Weng, Xitong Yang, Zhen Xing, Zuxuan Wu, and Yu-Gang Jiang. Genrec: Unifying video generation and recognition with diffusion models. *arXiv preprint arXiv:2408.15241*, 2024. 1

[56] 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 *CVPR*, 2023. 2

[57] Zhen Xing, Qi Dai, Zihao Zhang, Hui Zhang, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Vidiff: Translating videos via multi-modal instructions with diffusion models. *arXiv preprint arXiv:2311.18837*, 2023. 2

[58] Zhen Xing, Qi Dai, Han Hu, Zuxuan Wu, and Yu-Gang Jiang. Simda: Simple diffusion adapter for efficient video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7827–7839, 2024. 1

[59] Zhen Xing, Qi Dai, Zejia Weng, Zuxuan Wu, and Yu-Gang Jiang. Aid: Adapting image2video diffusion models for instruction-guided video prediction. *arXiv preprint arXiv:2406.06465*, 2024.

[60] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024. 1

[61] 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. In *CVPR*, 2024. 1, 3, 5, 6, 8

[62] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[63] Yuxuan Yan, Chi Zhang, Rui Wang, Yichao Zhou, Gege Zhang, Pei Cheng, Gang Yu, and Bin Fu. Facestudio: Put your face everywhere in seconds. *arXiv preprint arXiv:2312.02663*, 2023. 2

[64] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, 2023. 5

[65] 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. 1, 2, 7

[66] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, pages 10459–10469, 2023. 2

[67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3

[68] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 1, 2, 3, 4, 5, 6, 8

[69] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022. 1

[70] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *EECV*, 2024. 1, 3, 5, 6, 8