

StableAnimator++: Overcoming Pose Misalignment and Face Distortion for Human Image Animation

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, *Member, IEEE*, Chong Luo, *Senior Member, IEEE*, Zuxuan Wu, *Member, IEEE*, Yu-Gang Jiang, *Fellow, IEEE*

Abstract—Current diffusion models for human image animation often struggle to maintain identity (ID) consistency, especially when the reference image and driving video differ significantly in body size or position. We introduce StableAnimator++, the first ID-preserving video diffusion framework with learnable pose alignment, capable of generating high-quality videos conditioned on a reference image and a pose sequence without any post-processing. Building upon a video diffusion model, StableAnimator++ contains carefully designed modules for both training and inference, striving for identity consistency. In particular, StableAnimator++ first uses learnable layers to predict the similarity transformation matrices between the reference image and the driven poses via injecting guidance from Singular Value Decomposition (SVD). These matrices align the driven poses with the reference image, mitigating misalignment to a great extent. StableAnimator++ then computes image and face embeddings using off-the-shelf encoders, refining the face embeddings via a global content-aware Face Encoder. To further maintain ID, we introduce a distribution-aware ID Adapter that counteracts interference caused by temporal layers while preserving ID via distribution alignment. During the inference stage, we propose a novel Hamilton-Jacobi-Bellman (HJB) based face optimization integrated into the denoising process, guiding the diffusion trajectory for enhanced facial fidelity. Experiments on benchmarks show the effectiveness of StableAnimator++ both qualitatively and quantitatively. Project website: <https://francis-rings.github.io/StableAnimator++/>.

Index Terms—Video Diffusion Model, Video Generation, Human Image Animation

I. INTRODUCTION

HUMAN image animation [1], [2], [3], [4], [5], [6], [7], [8], [9], [10] aims to animate a reference image based on the motion pattern of a pose sequence, enabling diverse applications in entertainment and virtual reality. The phenomenal successes of diffusion models [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23] in video generation significantly inspire the advancement of human image animation. However, when dealing with pose sequences that exhibit significant motion variation, current approaches suffer from substantial distortions and inconsistencies, particularly in facial regions, destroying ID information. Misalignment in body size and position between the reference image and the driving video, which is common in real-world applications, further exacerbates this issue.

S. Tu, Z. Xing, Z. Wu, Y-G. Jiang are with School of Computer Science, Fudan University. Email: sytu23@m.fudan.edu.cn, {zxing20, zwxu}@fudan.edu.cn

Q. Dai, C. Luo are with Microsoft Research Asia. Emails: {qid, cluo}@microsoft.com

X. Han is with Tencent Inc. Emails: pathan@tencent.com

Z. Cheng is with University of Washington. Emails: zhiquics@uw.edu

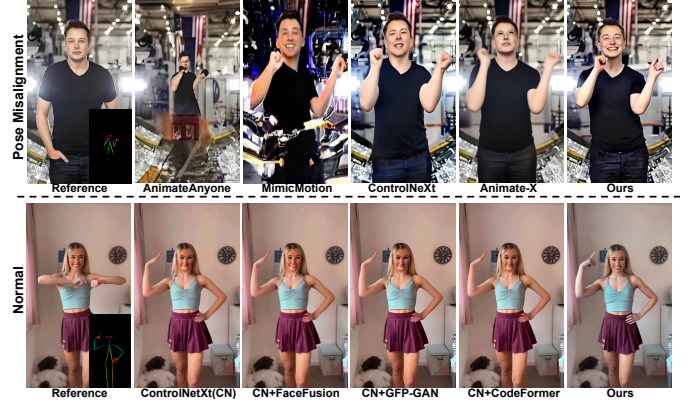


Fig. 1. Pose-driven Human image animations generated by our StableAnimator++ and compared methods, showing its power to synthesize ID-preserving videos even in scenarios with significant pose misalignment between the reference and driven poses. AnimateAnyone [3], MimicMotion [6], ControlNetXt [7], and Animate-X [8] are existing open-source animation models. FaceFusion [28] is a face-swapping tool. GFP-GAN [29] and CodeFormer [30] are face restoration models. Normal refers to the pose-aligned scenario.

To address this issue, there are numerous methods exploring identity (ID) preservation [24], [25], [26], [27] 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 disrupt the original spatial priors essential for identity preservation. Since image-based ID-preserving methods depend on these stable priors, introducing temporal layers often leads to poor results. This makes maintaining identity while ensuring video quality a major challenge for image animation. Furthermore, recent animation models [6], [7] rely on FaceFusion [28] for post-processing, which also degrades the quality of animated videos, particularly for facial areas.

Regarding the pose misalignment, previous methods [31], [5], [6], [7] utilize a pose alignment algorithm to align the driven pose with the reference image before animation, which roughly calculates the scaling factor and offset based on the relative size ratio between the reference image and the driven pose to scale and translate the driven pose. Champ [4] leverages the parametric shape alignment to align the 3D signal SMPL. However, in scenarios with significant discrepancies in body size and protagonist’s position, these approaches become highly inaccurate, negatively impacting the quality of the animated video. Furthermore, while Animate-X [8] claims to be insensitive to body size and protagonist’s position gaps between the reference image and driven poses, our experiments show that dramatic pose misalignment still significantly

degrades the quality of animations in such scenarios.

In light of this, we propose StableAnimator++, consisting of dedicated modules for both training and inference to maintain ID consistency for high-quality human image animation in various scenarios, including dramatic pose misalignment. StableAnimator++ first introduces learnable layers to predict similarity transformation matrices (rotation, scaling, and translation) between the reference image and driven poses, guided by Singular Value Decomposition (SVD). Since directly predicting aligned poses is challenging, SVD provides an intermediate transformation state to guide the learnable layers via cross-attention, significantly enhancing the model’s ability to capture the projection relationship between the reference image and driven poses. Trained layers offer greater robustness and accuracy in alignment in various scenarios than conventional methods. It uses the similarity transformation matrices to align driven poses with the reference image, reducing gaps in body size and protagonist position. Then, StableAnimator++ uses off-the-shelf extractors [32], [33] 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 distribution alignment between the resulting outputs. It 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 [34], [35] 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 Animate-X [8] suffers from dramatic body distortion, StableAnimator++ can effectively animate the reference image based on the pose sequence in the significant pose misalignment scenario. In the normal scenario, while ControlNeXt [7] exhibits severe facial and body distortions despite using face swapping or restoration tools, StableAnimator++ can accurately animate the reference based on given poses while preserving ID consistency.

In conclusion, our contributions are as follows: (1) We propose a novel learnable SVD-guided pose alignment model, which takes scaling, rotation, and translation into account, significantly reducing gaps from misalignment issues. To our knowledge, we are the first to explore learnable pose alignment for ID-preserving human image animation across various scenarios. (2) 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. (3) 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. (4) Experimental results on benchmark datasets show the superiority of our model over the SOTA.

A preliminary version of this paper appeared in [36]. The present paper includes a complete literature review on robust human image animation models, with a focus on handling pose misalignment commonly observed in real-world applications; an updated solution that utilizes learnable layers to predict similarity transformation matrices (rotation, scaling, and translation) between the reference image and driven poses, guided by Singular Value Decomposition.

II. RELATED WORK

Diffusion for Video Generation. Diffusion models have achieved remarkable success in video generation [11], [17], [37], [38], [12], [13], [15], [14], [18], driven by their superior diversity and high fidelity. Current video generation models [39], [40], [41], [42], [31], [43], [44], [36] capture spatio-temporal representations by adding temporal layers to pre-trained image generation models. Some works [45], [46], [47], [48], [49], [50], [51] replace the U-Net with transformers to scale up, showing a significant advancement in large video generation models. Following recent animation models [7], [6], we adopt Stable Video Diffusion [52] as the backbone.

Pose-guided Human Image Animation. Human image animation transfers motion from a given pose sequence to a reference image. Early works [53], [54], [55] primarily relied on GANs [56], but GAN-based models often suffer from flickering issues. Sparked by the diffusion models in video generation, recent animation models are basically based on diffusion models. Disco [1] is the first to try the diffusion model in human animation. MagicAnimate [2] and AnimateAnyone [3] both introduce transformer-based temporal attention modules for temporal smoothness. Champ [4] uses 3D signal SMPL to model motion patterns. Unianimate [5] inserts Mamba [57] into the diffusion U-Net for efficiency. MimicMotion [6] introduces the regional loss to address hand distortion. ControlNext [7] proposes a convolution-based PoseNet. Animate-X [8] aims to animate various character types. However, previous animation models suffer from face distortion. As they utilize the third-party face-swapping tool FaceFusion [28] as post-processing to address this issue, yet this approach can degrade overall video quality. This issue becomes more severe when there is a misalignment in body size and position between the reference image and the driven

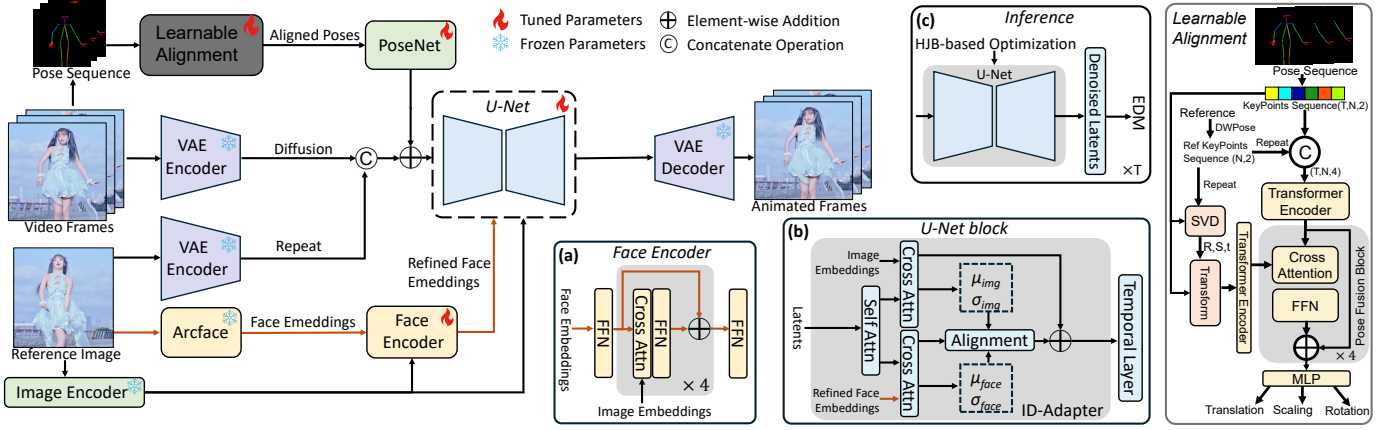


Fig. 2. Architecture of StableAnimator++. (a) and (b) refer to the structure of the Face Encoder and each block in the U-Net. We first apply our learnable alignment to the driving pose sequence and feed the aligned results into the PoseNet for motion modeling. Embeddings from the Image Encoder and Face Encoder are injected into 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.

pose. Our StableAnimator++ can still synthesize ID-preserving videos even when confronting dramatic pose misalignment scenarios without relying on any post-processing tools.

ID Consistency Image Generation. Recent studies have explored identity (ID) preservation in the image domain. LoRA [58] injects a few trainable parameters for personalized tuning but requires separate training for each identity, limiting scalability. IP-Adapter-FaceID [24] decouples cross-attention for text and facial features, potentially causing feature misalignment. PhotoMaker [59], FaceStudio [60], and InstantID [25] refine facial embeddings through hybrid mechanisms, while ConsistentID [26] leverages a facial prompt generator for detail preservation. PuLID [27] introduces contrastive and ID-specific losses to enhance identity fidelity. However, these approaches are not readily compatible with video diffusion models, where temporal layers may disrupt spatial consistency, leading to domain mismatch and degraded animation quality. In contrast, our StableAnimator++ integrates ID information into video diffusion models via a distribution-aware ID Adapter, effectively resolving the conflict between ID consistency and video fidelity.

III. METHOD

As shown in Fig. 2, inspired by previous works [6], [7], StableAnimator++ is based on the commonly used Stable Video Diffusion [52]. The driven pose sequence is aligned using our learnable alignment block and then processed by a PoseNet, as depicted in Sec. III-A. A PoseNet with a similar architecture to AnimateAnyone [3] encodes the aligned poses, which are then added to the noisy latents. A reference image is fed to the diffusion model through three pathways: (1) Converted into a latent code using a frozen VAE Encoder [61]. This latent code is then duplicated to align with the number of video frames and concatenated with the diffusion latents. (2) Encoded by the CLIP Image Encoder [33] to obtain image embeddings, which are then fed to each cross-attention block of a denoising U-Net, guiding the synthesized appearance. (3) Input to Arcface [32] 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. III-B.

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. III-C.

A. Learnable Alignment During Training

Previous pose alignment methods [31], [5], [6], [7] in animation basically calculate the scaling factor and offset [31] based on the relative size ratio between the reference image and the driven pose to adjust the driven skeleton keypoints. Champ [4] utilizes the parametric shape alignment to align the 3D signal SMPL. The above approaches are particularly inaccurate in cases of significant body size misalignment or positional discrepancies between the reference image and the driven video, thereby degrading the animation quality. While Animate-X [8] claims to be pose-agnostic and alignment-free, it still suffers from body distortions in cases of significant misalignment. To address this, we introduce a novel learnable alignment that uses learnable layers to predict accurate similarity transformation matrices (rotation, scaling, and translation) between the reference image and driven videos, guided by Singular Value Decomposition (SVD). Employing learnable layers to predict the aligned poses is relatively more effective and robust compared with conventional methods, as it is trained on diverse misalignment scenarios.

Fig. 2 illustrates the overall framework of our alignment block. Given a reference image $Ref \in \mathcal{R}^{3 \times H \times W}$ and a driven video $V \in \mathcal{R}^{T \times 3 \times H \times W}$, we leverage DWPose [62] to extract their pose keypoint sequences $P_r \in \mathcal{R}^{2 \times N}$ and $P_d \in \mathcal{R}^{2 \times N \times T}$, respectively. N refers to the keypoint number, with a default value of 18 in DWPose. T is the frame number of the driven video. We first repeat P_r to obtain P_r^* and concatenate them with P_d along the channel dimension, then

feed them into a Transformer Encoder $\text{Encoder}_m(\cdot)$ to model their motion patterns and relative positional relationships:

$$\mathbf{F}_m = \text{Encoder}_m(\text{Concat}(\text{Repeat}(\mathbf{P}_r), \mathbf{P}_d)), \quad (1)$$

where $\text{Concat}(\cdot)$ is the concatenation operation. Furthermore, inspired by ICP [63], we use SVD to obtain intermediate aligned poses, guiding the learnable layers to model the projection relationship between $\mathbf{P}_r$ and $\mathbf{P}_d$, as directly predicting the aligned keypoints is challenging for the learnable layers. ICP, designed for point clouds, iteratively optimizes transformation matrices without considering point correspondences (e.g., hand to hand), and its accuracy is unstable, making it impractical for animation. Thus, we use SVD only once in alignment as guidance for an intermediate state. In particular, we center $\mathbf{P}_r^*$ and $\mathbf{P}_d$ as follows:

$$\mathbf{X}_r = \mathbf{P}_r^* - \frac{1}{N} \sum_{i=1}^N \mathbf{P}_r^*[:, i, :], \mathbf{X}_d = \mathbf{P}_d - \frac{1}{N} \sum_{i=1}^N \mathbf{P}_d[:, i, :]. \quad (2)$$

Notably, we describe the case where only one person is present in the input image for readability. To determine the optimal rotation matrix $\mathbf{R}$, we first construct the covariance matrix $\mathbf{K}$, which captures the correlation between the two centered point sets:

$$\mathbf{K} = \mathbf{X}_d \mathbf{X}_r^T. \quad (3)$$

We then apply Singular Value Decomposition ($\text{SVD}(\cdot)$) to decouple $\mathbf{K}$ as follows:

$$\mathbf{U}, \mathbf{s}, \mathbf{V}^T = \text{SVD}(\mathbf{K}), \quad (4)$$

where orthogonal matrices $\mathbf{U}$ and $\mathbf{V}^T$ describe the principal axes of variation. We can obtain the rotation matrix $\mathbf{R}$:

$$\mathbf{R} = \mathbf{V} \mathbf{U}^T. \quad (5)$$

Furthermore, we use $\mathbf{R}$ to obtain the scale factor:

$$\mathbf{S} = \frac{\text{Trac}(\mathbf{R} \mathbf{K})}{\sum_{i=1}^N \mathbf{X}_d^{i^2}}, \quad (6)$$

where $\text{Trac}(\cdot)$ and $\sum_{i=1}^N \mathbf{X}_d^{i^2}$ refer to the trace operator and the dispersion of the body shape of the driven frame in space. The translation vector $\mathbf{t}$ describes the displacement between the centroids of the reference body and the driven frame body after rotation and scaling as:

$$\mathbf{t} = \frac{1}{N} \sum_{i=1}^N \mathbf{P}_r^*[:, i, :] - \mathbf{S} * (\mathbf{R} \frac{1}{N} \sum_{i=1}^N \mathbf{P}_d[:, i, :]). \quad (7)$$

We can further use the above $\mathbf{R}$, $\mathbf{S}$, and $\mathbf{t}$ to transform the initial driven keypoints sequence $\mathbf{P}_d$ into an intermediate state $\bar{\mathbf{P}}_d$ as follows:

$$\bar{\mathbf{P}}_d = \mathbf{S} * (\mathbf{R} \cdot \mathbf{P}_d) + \mathbf{t}. \quad (8)$$

We then input $\bar{\mathbf{P}}_d$ to another Transformer Encoder $\text{Encoder}_{svd}(\cdot)$ to extract motion-aware features, which are subsequently passed to a Pose Fusion Block $\text{PFusion}(\cdot)$ for guidance injection as follows:

$$\bar{\mathbf{F}}_m = \text{PFusion}(\text{Encoder}_{svd}(\bar{\mathbf{P}}_d), \mathbf{F}_m), \quad (9)$$

where $\text{PFusion}(\cdot)$ contains 4 modules, each comprising a cross-attention layer and an FFN. Although the SVD output from a single interaction may not be strictly accurate, injecting

this guidance into the main features via cross-attention still significantly enhances the model's ability to capture discrepancies in body size and position between the reference and driven poses, thereby facilitating their learning. We then use an MLP to predict the rotation/scaling/translation matrices ($\mathbf{R}'$, $\mathbf{S}'$, $\mathbf{t}'$) as follows:

$$\mathbf{R}', \mathbf{S}', \mathbf{t}' = \text{MLP}(\bar{\mathbf{F}}_m). \quad (10)$$

The above operation is set as $\mathbf{R}'_a, \mathbf{S}'_a, \mathbf{t}'_a = \text{Align}(\mathbf{P}_a, \mathbf{P}_b)$, where $\mathbf{P}_a$ is the keypoints to be aligned and $\mathbf{P}_b$ is the reference keypoints. The ultimate aligned driven poses $\mathbf{P}_d^{\text{align}}$ can be obtained as follows via applying $\text{Align}(\mathbf{P}_d, \mathbf{P}_r)$:

$$\mathbf{P}_d^{\text{align}} = \mathbf{S}'_d * (\mathbf{R}'_d \cdot \mathbf{P}_d) + \mathbf{t}'_d. \quad (11)$$

We train the alignment block from scratch at the image level for 50 epochs using 5K collected videos before training the entire StableAnimator++. With an average video length of 60 seconds and 30 FPS, the total number of training images exceeds 9 M. We first select two frames from a training video: one as the reference image and the other as the driven pose. For each driven pose, we modify it by applying random scaling, rotation, and translation matrices to simulate misalignment. We then feed the modified driven pose $\mathbf{P}_d$ and the reference image to our alignment block for predicting accurate transformation matrices ($\mathbf{R}'_d, \mathbf{S}'_d, \mathbf{t}'_d$). We calculate the average Euclidean distance $\text{Dis}_{\text{Euc}}(\cdot)$ between aligned poses and ground-truths $\mathbf{P}_d^{\text{gt}}$ as the loss function:

$$\mathcal{L}_{\text{align}} = \text{Avg}(\text{Dis}_{\text{Euc}}(\mathbf{P}_d^{\text{gt}}, \mathbf{S}'_d * (\mathbf{R}'_d \cdot \mathbf{P}_d) + \mathbf{t}'_d)). \quad (12)$$

B. ID-preserving During Training

Global Content-aware Face Encoder. To synthesize ID-preserving animations guided by a pose sequence, it's essential to retain both the facial details and the global context of the reference image. Although directly injecting face embeddings into the U-Net enhances facial fidelity, it fails to capture the global context (layout and background) in the reference image before being injected into the U-Net. Consequently, ID-irrelevant elements in the reference image bring noise to face modeling, impairing the overall animation quality. To overcome this, we introduce a Global Content-Aware Face Encoder, which refines face embeddings by allowing them to interact with the full reference image through a series of cross-attention blocks, enabling more context-aware modeling as shown in Fig. 2.

Distribution-aware ID Adapter. To mitigate the distortion of spatial features occurring when directly incorporating image-domain ID-preserving methods [53], [27], [26], [25] into the video diffusion model, the outputs of the Face Encoder are further fed to our ID Adapter. 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. This results in a conflict between preserving high video fidelity and maintaining identity integrity, often manifesting as facial

blurring or background degradation in the animations. As shown in Fig. 2 (b), our Distribution-aware ID Adapter is incorporated into each spatial layer of the U-Net. It performs distribution alignment between refined face embeddings and diffusion latents before each temporal modeling, effectively mitigating feature distortion.

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

$$\begin{aligned} z_i &= \text{SAttn}(z_i), \\ z_i^{img} &= \text{CAtn}(z_i, emb_{img}), \\ z_i^{face} &= \text{CAtn}(z_i, emb_{face}), \end{aligned} \quad (13)$$

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

$$\begin{aligned} \bar{z}_i^{face} &= \frac{z_i^{face} - \mu_{face}}{\sigma_{face}} \times \sigma_{img} + \mu_{img}, \\ \bar{z}_i &= \bar{z}_i^{face} + z_i^{img}. \end{aligned} \quad (14)$$

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

C. ID-preserving During Inference

To improve ID consistency, recent animation works [6], [7] use a third-party face-swapping tool FaceFusion [28], 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 not aligned with the domain of the original animations. To address this issue, inspired by the HJB equation [34], [35], [64], 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.

In particular, we optimize the predicted sample x_{pred} by minimizing the face similarity distance between x_{pred} and the reference before employing denoising (EDM [65]) at each step. The details are in the Algorithm 1, following the structure of the Algorithm 2 in the EDM paper [65]. S_{noise} , S_{churn} , $S_{t_{\min}}$

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

Input: $D_\theta(x; \sigma)$, $t_i \in \{0, \dots, N\}$, $\gamma_i \in \{0, \dots, N-1\}$, y
Sample $x_0 \sim \mathcal{N}(0, t_0^2 I)$ $\triangleright D_\theta(x; \sigma)$ is a diffusion model
For $i \in \{0, \dots, N-1\}$ **do** $\triangleright t_i \in \{0, \dots, N\}$ are timesteps
 $\gamma_i = 0$ $\triangleright \gamma_i \in \{0, \dots, N-1\}$ are pre-defined factors.
if $t_i \in [S_{t_{\min}}, S_{t_{\max}}]$: $\triangleright y$ is the reference image.
 $\gamma_i = \min\left(\frac{S_{\text{churn}}}{N}, \sqrt{2} - 1\right)$
Sample $\epsilon_i \sim \mathcal{N}(0, S_{\text{noise}}^2 I)$
 $\hat{t}_i = t_i + \gamma_i t_i$
 $\hat{x}_i = x_i + \sqrt{\hat{t}_i^2 - t_i^2} \epsilon_i$
 $x_{\text{pred}} = D_\theta(x_i; \hat{t}_i)$
 $x_{\text{op}} = x_{\text{pred}}.\text{clone}().\text{detach}()$ $\triangleright$ Starting optimization
 $op = \text{Adam}([x_{\text{op}}], \eta)$ $\triangleright$ Adam optimizer
 $x_{\text{op}}.\text{requires\_grad} = \text{True}$ $\triangleright x_{\text{op}}$ is a HJB variable
For $k \in \{1, 2, \dots, 10\}$ **do** $\triangleright k$ is the optimization step
 $f_{\text{pred}} = \text{Decoder}(x_{\text{op}})$ $\triangleright$ Decoder is a VAE decoder
 $loss = (1 - \text{Cos}(\text{Arc}(f_{\text{pred}}), \text{Arc}(y))).\text{abs}().\text{mean}()$
 $op.\text{zero\_grad}()$
 $loss.\text{backward}(\text{retain\_graph}=\text{True})$
 $op.\text{step}()$
 $x_{\text{pred}} = x_{\text{op}}$ $\triangleright$ End of Optimization
 $d_i = (\hat{x}_i - x_{\text{pred}})/\hat{t}_i$
 $x_{i+1} = \hat{x}_i + (t_{i+1} - \hat{t}_i)d_i$
if $t_{i+1} \neq 0$:
 $d'_i = (x_{i+1} - D_\theta(x_{i+1}; t_{i+1}))/t_{i+1}$
 $x_{i+1} = \hat{x}_i + (t_{i+1} - \hat{t}_i)(\frac{1}{2}d_i + \frac{1}{2}d'_i)$
return x_N

and $S_{t_{\max}}$ are the pre-defined values of EDM. $\text{Arc}(\cdot)$ and η are Arcface [32] and a learning rate. We employ our optimization to refine the prediction of the diffusion regarding the face similarity with the reference.

The optimized x_{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 optimization effectively reduces detail distortions, enhancing face quality.

Furthermore, we prove that the solving process of the HJB equation [34], [35], [64] can be integrated with the diffusion denoising process, as demonstrated below. The basic HJB Equation can be described as:

$$\frac{\partial V(x, t)}{\partial t} + \max_c [f(x, c) + \frac{\partial V(x, t)}{\partial x} \cdot g(x, c)] = 0, \quad (15)$$

where $V(x, t)$ refers to the value function, representing the minimum cost from state x at time t . $f(x, c)$ is the immediate cost under the condition c in state x . $g(\cdot)$ depicts the system dynamics. In our settings, the condition c indicates the face-aware variable. Following the previous work [64], the solving process is formulated as:

$$\min_{c_t} \int_0^1 \frac{1}{2} \|c_t\|_2^2 dt + \frac{r}{2} \|X_1 - x_1\|_2^2, X_1 \sim p_{data}, \quad (16)$$

s.t. $dX_t = c_t dt$ and $X_0 = x_0$ (Gaussian noise). r is the terminal cost coefficient. In our work, we normalize denoising timesteps t' (from T to 0) to $[0, 1]$ and set $t = 1 - t'$. T is the maximum denoising timestep. X_t and x_t refer to the groundtruth sample and the predicted sample by the model. Thus, x_{pred} in Algorithm 1 is equivalent to x_1 . Following

the Pontryagin Maximum Principle [66], we can construct the Hamiltonian equation:

$$\mathcal{H}(t, \mathbf{X}, \mathbf{c}_t, \gamma) = -\frac{1}{2} \|\mathbf{c}_t\|_2^2 + \gamma \mathbf{c}_t, \quad (17)$$

where γ refers to a coefficient. To minimize Eq. 17, we set $\frac{\partial \mathcal{H}}{\partial \mathbf{c}_t} = 0$. The optimal Hamiltonian is described as:

$$\mathcal{H}^* = \mathcal{H}(t, \mathbf{X}, \mathbf{c}_t^*, \gamma) = \frac{1}{2} \gamma^2, \text{ where } \mathbf{c}_t^* = \gamma. \quad (18)$$

Then we solve the Hamiltonian equation of motion:

$$\begin{aligned} \frac{d\mathbf{X}_t}{dt} &= \frac{\partial \mathcal{H}^*}{\partial \gamma} = \gamma, \\ \frac{d\gamma}{dt} &= \frac{\partial \mathcal{H}^*}{\partial \mathbf{X}} = 0. \end{aligned} \quad (19)$$

At the final step $t = 1$, from Eq. 16 and Eq. 17, we can obtain $\gamma_1 = -\mathbf{r} \cdot (\mathbf{X}_1 - \mathbf{x}_1)$. From Eq. 19, we can see that γ is a variable independent of t , thereby obtaining $\gamma = \gamma_1 = -\mathbf{r} \cdot (\mathbf{X}_1 - \mathbf{x}_1)$. We can also get $\mathbf{X}_t = \mathbf{X}_0 + \gamma t \rightarrow \mathbf{X}_1 = \mathbf{X}_0 + \gamma$ and $\mathbf{X}_0 = \mathbf{X}_t - \gamma t$. We then obtain $\mathbf{c}_t^*$:

$$\begin{aligned} \mathbf{X}_1 &= \mathbf{X}_0 + \gamma = \mathbf{X}_t - \gamma t + \gamma \\ \rightarrow \gamma &= -\mathbf{r} \cdot (\mathbf{X}_1 - \mathbf{x}_1) = -\mathbf{r} \cdot (\mathbf{X}_t - \gamma t + \gamma - \mathbf{x}_1), \\ \rightarrow \mathbf{c}_t^* &= \gamma = \frac{\mathbf{r}(\mathbf{x}_1 - \mathbf{X}_t)}{1 + \mathbf{r}(1 - t)}. \end{aligned} \quad (20)$$

When $\mathbf{r} \rightarrow \infty$, following Eq. 16 ($d\mathbf{X}_t = \mathbf{c}_t dt$) and certainty equivalence [67], [64] (the stochastic case), we have

$$d\mathbf{X}_t = \frac{\mathbf{x}_1 - \mathbf{X}_t}{1 - t} dt + d\mathbf{w}_t, \quad (21)$$

where $\mathbf{w}_t$ is Brownian motion [64]. According to EDM [65] in SVD [52], where $\mathbf{X}_{t'} = \mathbf{X}_{data} + t'\epsilon$ and $\mathbf{X}_{data} \sim \mathbf{p}_{data}$, the current state $\mathbf{X}_{t'}$ is converted to $\mathbf{X}_t = \mathbf{X}_1 + (1 - t)\epsilon$ in our settings. We use the following Tweedie's formula [68]

$$\mathbb{E}[\theta | \mathbf{x}] = \mathbf{x} + \sigma^2 \cdot \nabla \log p(\mathbf{x}), \quad (22)$$

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

$$\mathbf{X}_1 = \mathbb{E}[\mathbf{X}_1 | \mathbf{X}_t] = \mathbf{X}_t + (1 - t)^2 \nabla \log p(\mathbf{X}_t). \quad (23)$$

$\mathbf{x}_1$ aims to approximate $\mathbf{X}_1$. Thus, we substitute Eq. 23 in Eq. 21 for obtaining the ultimate formula:

$$\begin{aligned} d\mathbf{X}_t &= \frac{\mathbf{X}_t + (1 - t)^2 \nabla \log p(\mathbf{X}_t) - \mathbf{X}_t}{1 - t} dt + d\mathbf{w}_t \\ &= (1 - t) \cdot \nabla \log p(\mathbf{X}_t) dt + d\mathbf{w}_t. \end{aligned} \quad (24)$$

It is evident that Eq. 24 and SDE formulation [14] are structurally the same, thus we can seamlessly incorporate the solution process of the HJB equation into the diffusion denoising for ID preservation.

D. 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 $\mathbf{M}$, extracted by ArcFace [32] from the input video frames to enhance the modeling of face regions:

$$\mathcal{L} = \mathbb{E}_\epsilon (\|(\mathbf{z}_{gt} - \mathbf{z}_\epsilon) \odot (1 + \mathbf{M})\|^2), \quad (25)$$

where $\mathbf{z}_{gt}$ and $\mathbf{z}_\epsilon$ are diffusion latents and denoised latents.

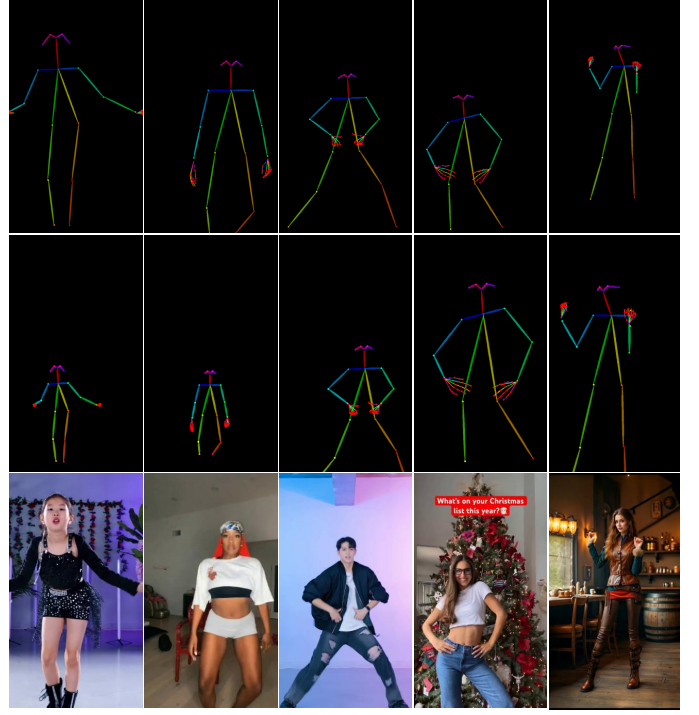


Fig. 3. Examples from MisAlign100. The first row, the second row, and the third row refer to the original driven poses, modified driven poses, and corresponding reference image, respectively.

IV. EXPERIMENTS

A. Implementation Details

As previous works do not open-source their training datasets, we collect 5K videos (60-90 seconds long) from the internet to train our model. We use DWPose [62] to extract skeleton poses. Following [1], [3], [2], [4], [5], [6], [7], [8], we evaluate our model on TikTok dataset [69]. We also select 100 unseen videos (the MisAlign100 dataset) from the internet, featuring scenarios with significant misalignment. Following recent works [6], [7], the U-Net uses pre-trained weights of Stable Video Diffusion [52], while the PoseNet, Face Encoder, and alignment block are trained from scratch. Regarding the Transformer Encoders ($\text{Encoder}_m(\cdot)$ and $\text{Encoder}_{svd}(\cdot)$) in our learnable alignment, they all share the same architecture, comprising two modules, each containing a self-attention block and an FFN. Notably, since $\text{Encoder}_m(\cdot)$ and $\text{Encoder}_{svd}(\cdot)$ both model keypoint sequences, where each token corresponds to a skeleton node, we apply position embeddings to the input sequences before passing them to the self-attention layers of the encoders. Our ID-Adapter uses pre-trained weights of spatial cross-attention blocks in Stable Video Diffusion. Our model is trained for 20 epochs on 8 NVIDIA A100 80G GPUs, with a batch size of 1 per GPU. The learning rate is set to $1e-5$. Our HJB-based face optimization is applied exclusively during the first 10 denoising steps at inference.

B. Data Collection

We collect our training videos from YouTube and TikTok. The raw videos are fed to the InsightFace [32] and Cotracker [72] models to filter out those with low facial



Fig. 4. Animation results generated by StableAnimator++. The images with red borders are the reference images. The presented pose skeletons are dramatically misaligned with the reference image in body size and position.

TABLE I
QUANTITATIVE COMPARISONS ON TIKTOK DATASET AND MISALIGN100.

Model	L1(E-4)↓	PSNR [70]↑	PSNR* [1]↑	SSIM↑	LPIPS↓	CSIM [71]↑	FVD↓	Mem↓
MRAA [54]	3.21/3.88	-18.12	18.14/9.81	0.672/0.285	0.296/0.637	0.248/0.163	284.82/1782.57	5.4G
DisCo [1]	3.78/3.84	29.03/18.58	16.55/9.84	0.668/0.293	0.292/0.634	0.315/0.202	292.80/1745.13	18.7G
MagicAnimate [2]	3.13/3.32	29.16/18.94	-10.06	0.714/0.315	0.239/0.623	0.462/0.268	179.07/1342.66	20.84G
AnimateAnyone [3]	-3.27	29.56/19.28	-10.16	0.718/0.324	0.285/0.619	0.457/0.261	171.90/1287.42	11.18G
Champ [4]	2.94/3.04	29.91/22.88	-12.17	0.802/0.389	0.234/0.522	0.350/0.307	160.82/1046.48	13.20G
Unianimate [5]	2.66/2.87	30.77/25.85	20.58/14.52	0.811/0.467	0.231/0.465	0.479/0.324	148.06/768.05	6.11G
MimicMotion [6]	5.85/3.80	-17.73	14.44/9.88	0.601/0.298	0.414/0.628	0.262/0.245	232.95/1652.78	8.60G
ControlNeXt [7]	6.20/2.92	-24.69	13.83/13.41	0.615/0.482	0.416/0.516	0.360/0.278	326.57/687.34	12.23G
Animate-X [8]	2.70/2.83	30.78/26.82	20.77/16.38	0.806/0.512	0.232/0.429	0.475/0.391	139.01/675.26	14.3G
StableAnimator++ (Ours)	2.90/ 2.74	30.81/30.17	20.79/18.22	0.816/0.709	0.230/0.375	0.831/0.802	122.47/384.27	11.40G

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 MisAlign100, respectively. We reference competitors' results on the TikTok dataset from their papers, with — indicating missing reports.

quality or significant camera motion (such as shot changes or background variance). We further apply DWPose to remove any videos where the skeletons lack more than 70% keypoints, thus obtaining our dataset.

Regarding the MisAlign100 dataset, we collect 100 unseen videos (10-20 seconds long) from the internet to construct the testing dataset MisAlign100. Some cases are shown in Fig. 3. The videos originate from various social media platforms, including YouTube, TikTok, and Bilibili, featuring individuals of diverse backgrounds and genders. They are captured in full-body, half-body, and close-up shots across a range of indoor and outdoor environments. In contrast to the existing open-source animation testing dataset (TikTok dataset), our MisAlign100 involves more complex motion patterns and intricate appearance information. Each driven pose is randomly rotated,

scaled, and translated to simulate the misalignment which is commonly encountered in real-world scenarios, making it more challenging to maintain ID consistency.

C. Animation Results

We demonstrate the animation results in Fig. 4. We can observe that our StableAnimator++ can perform a wide range of human image animation while simultaneously preserving the reference consistency, including the protagonist's appearance details and the background layouts. Each case involves a protagonist with complex appearance and intricate motion dynamics, while the reference image and driven video exhibit significant discrepancies in body shape and position. More cases are shown in the Sec.VIII of the Supp.



Fig. 5. Qualitative comparisons with state-of-the-art methods. The skeletons in the third and fourth rows are misaligned with the reference image in terms of body size or position. More examples can be found in the supplementary material.

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

Quantitative results. We compare with recent human image animation models, including GAN-based models (MRAA [54]) and diffusion models (AnimateAnyone [3], MagicAnimate [2], Champ [4], Unianimate [5], MimicMotion [6], ControlNeXt [7], Animate-X [8]), as shown in Table I. CSIM [71] evaluates the cosine similarity between the facial embeddings of two images. 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 [69] and MisAlign100. Notably, we randomly scale / translate / rotate the driven poses before evaluating on MisAlign100 to simulate misalignment. All competitors are trained on our dataset before evaluating on MisAlign100 to ensure a fair comparison. Since AnimateAnyone lacks a default alignment operation, we apply ControlNeXt’s alignment to it. We can see that our StableAnimator++ outperforms all competitors on MisAlign100 in both video fidelity and single-frame quality under significant misalignment scenarios while achieving relatively promising performance on the TikTok dataset. In particular, StableAnimator++ outperforms the leading competitor, Animate-X, by

TABLE II
ABLATION STUDY ON CORE COMPONENTS.

Model	L1↓	PSNR↑	SSIM↑	LPIPS↓	CSIM↑	FVD↓
w/o Pose Align (SA[36])	3.58E-4	18.51	0.298	0.630	0.448	1635.24
w/o Prediction	2.82E-4	26.84	0.542	0.424	0.726	552.13
w/o Face Masks	2.79E-4	27.11	0.653	0.386	0.694	458.91
w/o Face Encoder	2.82E-4	27.03	0.647	0.390	0.572	441.16
w/o Distribution Align	2.85E-4	25.98	0.496	0.435	0.707	587.36
w/o Optimization	2.78E-4	27.72	0.685	0.382	0.778	404.28
Ours	2.74E-4	30.17	0.709	0.375	0.802	384.27

w/o Prediction removes learnable layers in our pose alignment, directly applying SVD outputs to align poses. Face Masks and Distribution Align refer to face masks in the loss and distribution alignment of our ID Adapter. SA refers to StableAnimator [36].

35.6% and 41.1% in CSIM across two datasets, without sacrificing video fidelity and single-frame quality.

Qualitative Results. The qualitative results are shown in Fig. 5. All qualitative results in the paper are in the cross-ID setting [4]. MagicAnimate [2], AnimateAnyone [3], and Champ [4] exhibit face / body distortion and clothing changes, while Unianimate [5] and Animate-X [8] accurately modify the reference motion, and MimicMotion [6] and ControlNeXt [7] effectively preserve clothing details. However, all competitors still struggle with face/body distortion and

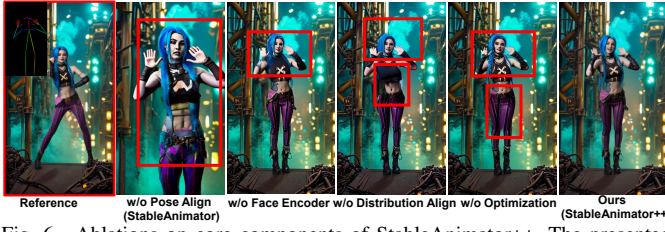


Fig. 6. Ablations on core components of StableAnimator++. The presented skeleton is misaligned with the reference image in body size and position.



Fig. 7. Ablations on the alignment. The poses in the last two columns are aligned by the respective methods.

TABLE III
ABLATION STUDY ON THE ALIGNMENT.

Model	L1↓	PSNR↑	SSIM↑	CSIM↑	Dis↓	FVD↓
w/o Pose Alignment (SA[36])	3.58E-4	18.51	0.298	0.448	0.597	1635.24
w/ ControlNeXt	2.88E-4	25.32	0.488	0.661	0.430	625.97
w/o Prediction	2.84E-4	26.94	0.510	0.702	0.345	571.56
Ours	2.74E-4	30.17	0.709	0.802	0.105	384.27

Dis is the average Euclidean distance between aligned poses and ground-truths.

blurry noises in both normal and pose-misaligned scenarios. In contrast, our StableAnimator++ accurately animates images based on the given pose sequences while preserving reference identities even in misalignment scenarios, showcasing the superiority of our model in identity retention and in generating precise, vivid animations.

E. Ablation Study

Pose Alignment. We conduct an ablation study to validate the contributions of core components in StableAnimator++, as shown in Fig. 6 and Table II. All quantitative ablation studies are conducted on the MisAlign100 dataset. The w/o Pose Align setting is equivalent to StableAnimator (SA) [36]. We can observe that removing core components dramatically deteriorates performance, particularly in face-related regions (CSIM), indicating that each core component can promote both single-frame quality and video fidelity while maintaining identity consistency even in misalignment scenarios.

We further compare our alignment with the current keypoint alignment approach [7], as shown in Fig. 7 and Table III. We replace our alignment with ControlNeXt’s alignment, which is also commonly used in current animation models [6], [5]. Fig. 8 ablates the effectiveness of the SVD-based transformation. By analyzing the results, we can gain the following observations: (1) StableAnimator [36] exhibits noticeable face/body distortions in scenarios with significant pose misalignment



Fig. 8. Ablations on the alignment. The poses in the last two rows are aligned by the respective methods. Only SVD removes learnable layers in our alignment, directly applying SVD outputs to align poses.

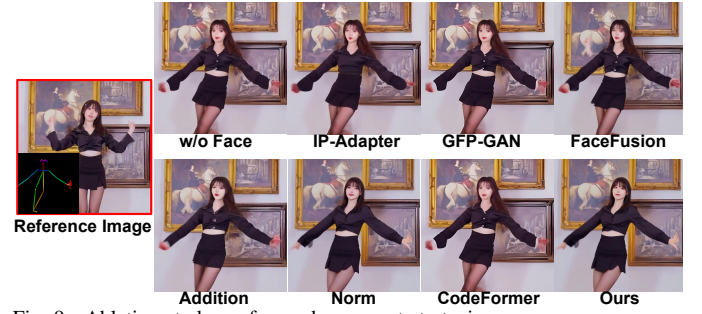


Fig. 9. Ablation study on face enhancement strategies.

between the reference and the driving video. (2) ControlNeXt’s alignment reduces body distortion but degrades video fidelity and reference consistency, as its aligned driven poses fail to match the reference image in body size and position, creating a conflict between appearance preservation and motion modeling. (3) Directly using SVD outputs for alignment enhances single-frame quality but compromises reference consistency. The plausible reason is that the transformation matrices of SVD are not particularly accurate, leading to a loss of semantic details. (4) StableAnimator++ can effectively preserve identity while achieving high video fidelity, as our pose alignment can dramatically reduce the gap between the reference and driven poses. More ablation studies are in Sec. IV of the Supp.

Face Enhancement Strategies. We conduct an ablation study regarding current face enhancement approaches, as shown in Table IV and Fig. 9. We replace our face-related components with the commonly used IP-Adapter and FaceFusion. **We temporarily apply our pose alignment to the MisAlign100 dataset to obtain aligned poses for a fair comparison in the following ablation studies.** 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

TABLE IV
ABLATION STUDY ON FACE ENHANCEMENT METHODS.

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 [24]	3.88E-4	18.86	0.672	0.287	0.511	484.77
FaceFusion [28]	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

w/o Face refers to the exclusion of any face-related strategies.

TABLE V
ABLATION STUDY ON THE DISTRIBUTION-BASED ALIGNMENT.

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

Addition and Norm refer to element-wise addition and normalization.

TABLE VI
ABLATION STUDY ON THE OPTIMIZATION.

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

Magic, IP, ID, FE, and Opt refer to MagicAnimate, IP-Adapter, our ID Adapter, our Face Encoder, and our Optimization, respectively.

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.

We further conduct a comparison between our StableAnimator++ and other facial restoration models (GFP-GAN [29] and CodeFormer [30]), as shown in Fig. 9. 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.

Feature Distortion. We conduct a comparison between our distribution alignment in the ID-Adapter and other types of feature injection, as shown in Table V and Fig. 9. Norm refers to $\tilde{z}_i^{face} = \frac{z_i^{face} - \mu_{face}}{\sigma_{face}}$. 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.

Face Optimization. To validate the significance of our face optimization strategy, we conduct an ablation regarding differ-

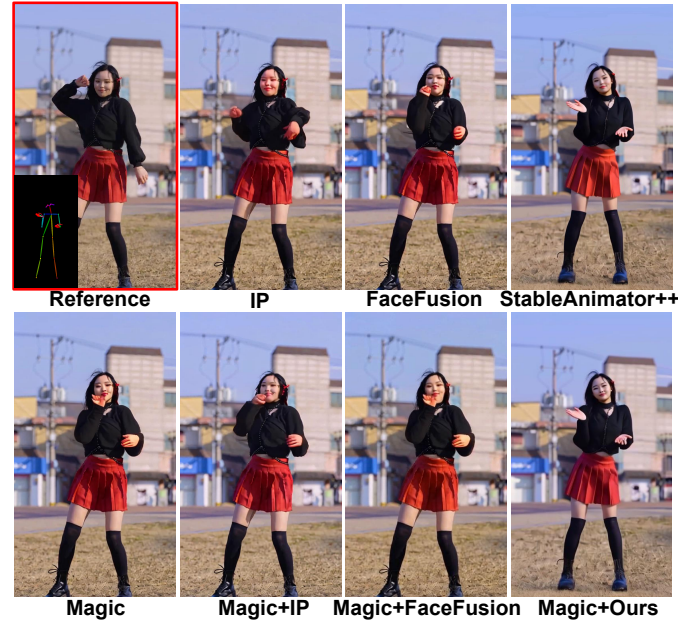


Fig. 10. Ablation study on different backbones.

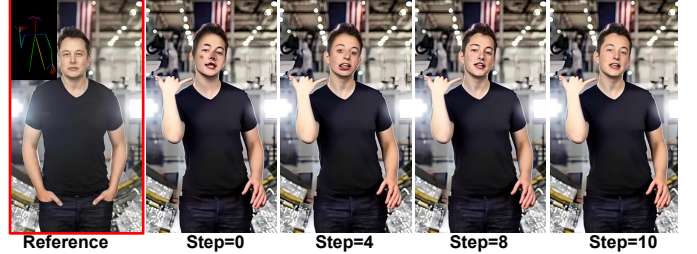


Fig. 11. Visual comparison of HJB-based face optimization at different denoising (optimization) steps.

ent diffusion backbones. The results are in Table VI and Fig. 10. MagicAnimate is based on SD [16]+AnimateDiff [40]. 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 indicate 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 VI show that our face optimization can still work in different diffusion-based backbones.

Fig. 11 shows a detailed visual comparison, where the step refers to the optimization step in HJB-based optimization. The facial quality progressively improves, which indicates the significance of our face optimization in terms of identity preservation. However, increasing the number of optimization steps introduces higher inference latency, and excessive steps tend to over-sharpen facial details. Thus, we empirically set the total number of steps to 10 as an optimal trade-off between quality and efficiency.

Speed. We compare our StableAnimator++ with current human image animation models in terms of inference latency and GPU memory consumption. Table VII describes the

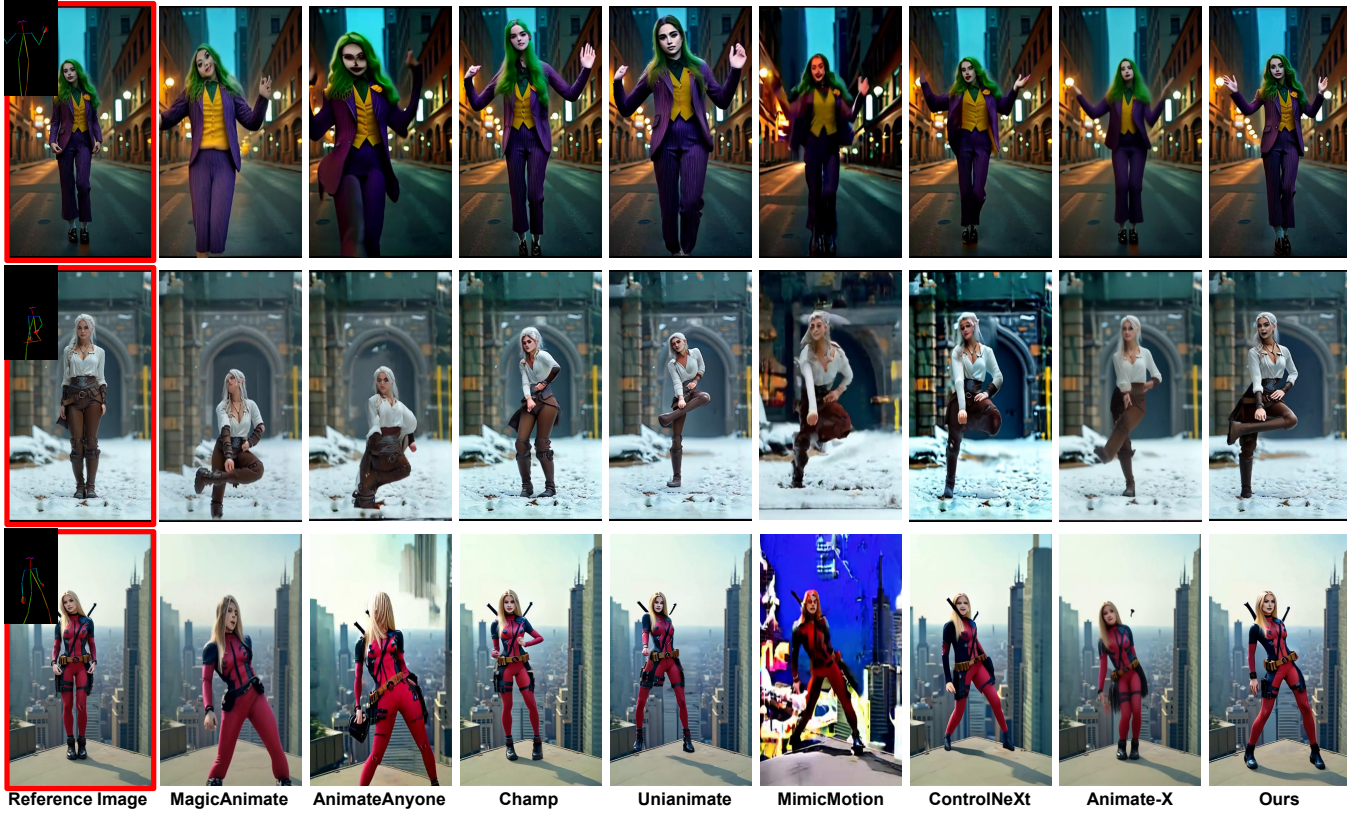


Fig. 12. Long animation results. The presented skeletons are misaligned with the reference image in body size and position.

TABLE VII
COMPARISON RESULTS ON INFERENCE LATENCY.

Model	PSNR $\uparrow$	FVD $\downarrow$	Mem $\downarrow$	Inference Latency $\downarrow$
MagicAnimate [2]	18.94	1342.66	20.84G	82s
AnimateAnyone [3]	19.28	1287.42	11.18G	75s
Champ [4]	22.88	1046.48	13.20G	145s
Unimate [5]	25.85	768.05	6.11G	86s
MimicMotion [6]	17.73	1652.78	8.60G	60s
ControlNeXt [7]	24.69	687.34	12.23G	139s
Animate-X [8]	26.82	575.26	14.30G	182s
Ours	30.17	384.27	11.4G	84s

TABLE VIII
COMPARISON RESULTS ON ANTHROPOMORPHIC CHARACTERS.

Model	L1 $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FVD $\downarrow$
Unimate	1.44E-4	10.05	0.325	0.617	1385.64
ControlNeXt	1.55E-4	9.84	0.296	0.620	1709.36
Animate-X	1.37E-4	10.45	0.368	0.592	1267.13
Ours	1.05E-4	14.13	0.488	0.425	830.10

comparison results. The inference latency and GPU memory consumption are measured when the model generates 16 frames at a resolution of 576 $\times$ 1024. We can observe that StableAnimator++ achieves better results at a faster speed with nearly the same GPU memory consumption as AnimateAnyone [3], demonstrating that our model is the best trade-off between efficiency and performance.

F. Application and User Study

Long Animation. We conduct qualitative comparisons between StableAnimator++ and current animation models in long animation generation, as shown in Fig. 12. Detailed comparisons are shown in Sec.VI of the Supp. Following



Fig. 13. (a), (b), and (c) refer to multiple-person animation results, general portrait generation results, and anthropomorphic character animation results, respectively. The image with the red border is the reference image.

MimicMotion [6], we follow the same pipeline to synthesize long animations. Each driven pose sequence consists of over 500 frames with complex motion, and the references show significant misalignment in terms of the protagonists' body sizes and positions relative to the driven poses. We can see that competitors encounter serious body distortion and blurry noise. By contrast, our model can effectively handle long animation in high fidelity while preserving identities even in scenarios involving dramatic misalignment.

TABLE IX

USER PREFERENCE OF ANIMATEMASTER COMPARED WITH OTHER COMPETITORS. HIGHER INDICATES USERS PREFER MORE TO OUR MODEL.

Model	M-A	A-A	B-A
MagicAnimate [2]	95.7%	98.5%	93.4%
AnimateAnyone [3]	94.8%	98.2%	92.3%
Champ [4]	92.3%	95.6%	91.8%
Unianimate [5]	91.2%	95.8%	90.6%
MimicMotion [6]	90.6%	96.9%	91.5%
ControlNeXt [7]	88.6%	93.1%	90.2%
Animate-X [8]	92.4%	92.2%	90.7%

Multi-Person Animation. We experiment with multiple-person animation, as shown in Fig. 13 (a). The results show that our model can animate multiple people.

General Text-to-Video Portrait Generation. To further validate the robustness of our core components, we integrate our face-related components (Face Encoder, ID-Adapter, and Face Optimization) into CogVideoX-I2V [73] to enable Text-to-Video generation, as shown in Fig. 13 (b), indicating that our core components effectively enable the base model to maintain identity consistency without compromising video fidelity.

Anthropomorphic Characters. We experiment with anthropomorphic characters, as shown in Table VIII and Fig. 13(c). As Animate-X does not release their A^2 Bench [8], we follow its method and use Kling AI to synthesize 100 anthropomorphic character videos for evaluation. We observe that ours outperforms current human image animation models.

User Study. We conducted a user study with 30 video-reference image pairs to evaluate human preferences between our model and competitors. The participants are roughly university students and faculty. In each case, participants are first shown a reference image and a pose sequence with significant misalignment. Then we present two videos (one is synthesized by StableAnimator++ and the other is generated by a competitor) in random order. Participants are asked to answer the questions: M-A/A-A/B-A: “Which one has better motion/appearance/background alignment with the reference”. Table IX shows the superiority of our model in subjective evaluation.

V. CONCLUSION

We propose StableAnimator++, a robust video diffusion model with dedicated training and inference modules for generating ID-preserving human animations, even under pose misalignment. It first uses SVD-guided learnable layers to predict transformation matrices that align driven poses, significantly reducing the body size and position gap with the reference. StableAnimator++ then 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. An ID-Adapter then performs distribution alignment to mitigate temporal interference, enabling seamless face embedding integration without degrading video fidelity. During inference, a hybrid of the HJB equation and diffusion denoising further enhances face quality. Experiments on multiple datasets demonstrate the model’s superiority in generating high-quality, ID-consistent animations, even in misalignment scenarios.

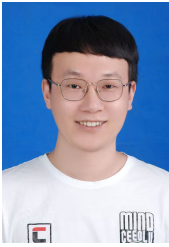
REFERENCES

- [1] T. Wang, L. Li, K. Lin, Y. Zhai, C.-C. Lin, Z. Yang, H. Zhang, Z. Liu, and L. Wang, “Disco: Disentangled control for realistic human dance generation,” in *CVPR*, 2024.
- [2] Z. Xu, J. Zhang, J. H. Liew, H. Yan, J.-W. Liu, C. Zhang, J. Feng, and M. Z. Shou, “Magicanimate: Temporally consistent human image animation using diffusion model,” in *CVPR*, 2024.
- [3] L. Hu, “Animate anyone: Consistent and controllable image-to-video synthesis for character animation,” in *CVPR*, 2024.
- [4] S. Zhu, J. L. Chen, Z. Dai, Y. Xu, X. Cao, Y. Yao, H. Zhu, and S. Zhu, “Champ: Controllable and consistent human image animation with 3d parametric guidance,” in *EECV*, 2024.
- [5] X. Wang, S. Zhang, C. Gao, J. Wang, X. Zhou, Y. Zhang, L. Yan, and N. Sang, “Unianimate: Taming unified video diffusion models for consistent human image animation,” *arXiv preprint arXiv:2406.01188*, 2024.
- [6] Y. Zhang, J. Gu, L.-W. Wang, H. Wang, J. Cheng, Y. Zhu, and F. Zou, “Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance,” *arXiv preprint arXiv:2406.19680*, 2024.
- [7] B. Peng, J. Wang, Y. Zhang, W. Li, M.-C. Yang, and J. Jia, “Controlnext: Powerful and efficient control for image and video generation,” *arXiv preprint arXiv:2408.06070*, 2024.
- [8] S. Tan, B. Gong, X. Wang, S. Zhang, D. Zheng, R. Zheng, K. Zheng, J. Chen, and M. Yang, “Animate-x: Universal character image animation with enhanced motion representation,” in *ICLR*, 2025.
- [9] S. Tu, T. Guan, and L. Kuang, “Multiple biological granularities network for person re-identification,” in *ICMR*, 2022.
- [10] S. Tu, Q. Dai, Z. Wu, Z.-Q. Cheng, H. Hu, and Y.-G. Jiang, “Implicit temporal modeling with learnable alignment for video recognition,” in *ICCV*, 2023.
- [11] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” in *NeurIPS*, 2021.
- [12] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” in *NeurIPS*, 2020.
- [13] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, “Cascaded diffusion models for high fidelity image generation,” *JMLR*, 2022.
- [14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” in *ICLR*, 2021.
- [15] J. Song, C. Meng, and S. Ermon, “Denoising diffusion implicit models,” in *ICLR*, 2021.
- [16] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *CVPR*, 2022.
- [17] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, “Sdedit: Guided image synthesis and editing with stochastic differential equations,” in *ICLR*, 2021.
- [18] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, “Plug-and-play diffusion features for text-driven image-to-image translation,” in *CVPR*, 2023.
- [19] Z. Weng, X. Yang, Z. Xing, Z. Wu, and Y.-G. Jiang, “Genrec: Unifying video generation and recognition with diffusion models,” *arXiv preprint arXiv:2408.15241*, 2024.
- [20] Z. Xing, Q. Feng, H. Chen, Q. Dai, H. Hu, H. Xu, Z. Wu, and Y.-G. Jiang, “A survey on video diffusion models,” *ACM Computing Surveys*, vol. 57, no. 2, pp. 1–42, 2024.
- [21] Z. Xing, Q. Dai, H. Hu, Z. Wu, and Y.-G. Jiang, “Simda: Simple diffusion adapter for efficient video generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7827–7839.
- [22] Z. Xing, Q. Dai, Z. Weng, Z. Wu, and Y.-G. Jiang, “Aid: Adapting image2video diffusion models for instruction-guided video prediction,” *arXiv preprint arXiv:2406.06465*, 2024.
- [23] Q. Li, Z. Xing, R. Wang, H. Zhang, Q. Dai, and Z. Wu, “Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance,” *arXiv preprint arXiv:2503.16421*, 2025.
- [24] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, “Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models,” *arXiv preprint arxiv:2308.06721*, 2023.
- [25] Q. Wang, X. Bai, H. Wang, Z. Qin, and A. Chen, “Instantid: Zero-shot identity-preserving generation in seconds,” *arXiv preprint arXiv:2401.07519*, 2024.

- [26] J. Huang, X. Dong, W. Song, H. Li, J. Zhou, Y. Cheng, S. Liao, L. Chen, Y. Yan, S. Liao *et al.*, “Consistentid: Portrait generation with multimodal fine-grained identity preserving,” *arXiv preprint arXiv:2404.16771*, 2024.
- [27] Z. Guo, Y. Wu, Z. Chen, L. Chen, and Q. He, “Pulid: Pure and lightning id customization via contrastive alignment,” in *NeurIPS*, 2024.
- [28] R. Henry, “Facefusion,” <https://github.com/facefusion/facefusion>, 2024.
- [29] X. Wang, Y. Li, H. Zhang, and Y. Shan, “Towards real-world blind face restoration with generative facial prior,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [30] S. Zhou, K. C. Chan, C. Li, and C. C. Loy, “Towards robust blind face restoration with codebook lookup transformer,” in *NeurIPS*, 2022.
- [31] S. Tu, Q. Dai, Z.-Q. Cheng, H. Hu, X. Han, Z. Wu, and Y.-G. Jiang, “Motioneditor: Editing video motion via content-aware diffusion,” in *CVPR*, 2024.
- [32] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *CVPR*, 2019.
- [33] A. Radford, J. W. Kim, C. Hallacy, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, “Learning transferable visual models from natural language supervision,” in *ICML*, 2021.
- [34] M. Bardi, I. C. Dolcetta *et al.*, *Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations*. Springer, 1997.
- [35] S. Peng, “Stochastic hamilton-jacobi-bellman equations,” *SIAM Journal on Control and Optimization*, 1992.
- [36] S. Tu, Z. Xing, X. Han, Z.-Q. Cheng, Q. Dai, C. Luo, and Z. Wu, “Stableanimator: High-quality identity-preserving human image animation,” *arXiv preprint arXiv:2411.17697*, 2024.
- [37] A. Q. Nichol and P. Dhariwal, “Improved denoising diffusion probabilistic models,” in *ICML*, 2021.
- [38] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, “Prompt-to-prompt image editing with cross attention control,” *arXiv preprint arXiv:2208.01626*, 2022.
- [39] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni *et al.*, “Make-a-video: Text-to-video generation without text-video data,” *arXiv preprint arXiv:2209.14792*, 2022.
- [40] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, “Animatediff: Animate your personalized text-to-image diffusion models without specific tuning,” in *ICLR*, 2024.
- [41] J. Z. Wu, Y. Ge, X. Wang, S. W. Lei, Y. Gu, Y. Shi, W. Hsu, Y. Shan, X. Qie, and M. Z. Shou, “Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation,” in *CVPR*, 2023.
- [42] W. Wang, J. Liu, Z. Lin, J. Yan, S. Chen, C. Low, T. Hoang, J. Wu, J. H. Liew, H. Yan *et al.*, “Magicvideo-v2: Multi-stage high-aesthetic video generation,” *arXiv preprint arXiv:2401.04468*, 2024.
- [43] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh, “Video generation models as world simulators,” 2024. [Online]. Available: <https://openai.com/research/video-generation-models-as-world-simulators>
- [44] S. Tu, Q. Dai, Z. Zhang, S. Xie, Z.-Q. Cheng, C. Luo, X. Han, Z. Wu, and Y.-G. Jiang, “Motionfollower: Editing video motion via lightweight score-guided diffusion,” *arXiv preprint arXiv:2405.20325*, 2024.
- [45] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *ICCV*, 2023.
- [46] W. Yan, Y. Zhang, P. Abbeel, and A. Srinivas, “Videogpt: Video generation using vq-vae and transformers,” *arXiv preprint arXiv:2104.10157*, 2021.
- [47] L. Yu, Y. Cheng, K. Sohn, J. Lezama, H. Zhang, H. Chang, A. G. Hauptmann, M.-H. Yang, Y. Hao, I. Essa *et al.*, “Magvit: Masked generative video transformer,” in *CVPR*, 2023, pp. 10 459–10 469.
- [48] X. Ma, Y. Wang, G. Jia, X. Chen, Z. Liu, Y.-F. Li, C. Chen, and Y. Qiao, “Latte: Latent diffusion transformer for video generation,” *arXiv preprint arXiv:2401.03048*, 2024.
- [49] F. Bao, C. Xiang, G. Yue, G. He, H. Zhu, K. Zheng, M. Zhao, S. Liu, Y. Wang, and J. Zhu, “Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models,” *arXiv preprint arXiv:2405.04233*, 2024.
- [50] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, “Cogvideo: Large-scale pretraining for text-to-video generation via transformers,” *arXiv preprint arXiv:2205.15868*, 2022.
- [51] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang *et al.*, “Hunyuanvideo: A systematic framework for large video generative models,” *arXiv preprint arXiv:2412.03603*, 2024.
- [52] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, “Stable video diffusion: Scaling latent video diffusion models to large datasets,” *arXiv preprint arXiv:2311.15127*, 2023.
- [53] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “First order motion model for image animation,” in *NeurIPS*, 2019.
- [54] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, “Motion representations for articulated animation,” in *CVPR*, 2021.
- [55] Z. Huang, X. Han, J. Xu, and T. Zhang, “Few-shot human motion transfer by personalized geometry and texture modeling,” in *CVPR*, 2021.
- [56] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” *Communications of the ACM*, 2020.
- [57] T. Dao and A. Gu, “Transformers are SSMS: Generalized models and efficient algorithms through structured state space duality,” in *ICML*, 2024.
- [58] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” in *ICLR*, 2021.
- [59] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan, “Photomaker: Customizing realistic human photos via stacked id embedding,” in *CVPR*, 2024.
- [60] Y. Yan, C. Zhang, R. Wang, Y. Zhou, G. Zhang, P. Cheng, G. Yu, and B. Fu, “Facestudio: Put your face everywhere in seconds,” *arXiv preprint arXiv:2312.02663*, 2023.
- [61] D. P. Kingma, “Auto-encoding variational bayes,” in *ICLR*, 2014.
- [62] Z. Yang, A. Zeng, C. Yuan, and Y. Li, “Effective whole-body pose estimation with two-stages distillation,” in *ICCV*, 2023.
- [63] P. J. Besl and N. D. McKay, “Method for registration of 3-d shapes,” in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.
- [64] T. Chen, J. Gu, L. Dinh, E. A. Theodorou, J. Susskind, and S. Zhai, “Generative modeling with phase stochastic bridges,” in *ICLR*, 2024.
- [65] T. Karras, M. Aittala, T. Aila, and S. Laine, “Elucidating the design space of diffusion-based generative models,” in *NeurIPS*, 2022.
- [66] D. E. Kirk, *Optimal control theory: an introduction*. Courier Corporation, 2004.
- [67] W. H. Fleming and R. W. Rishel, *Deterministic and stochastic optimal control*. Springer Science & Business Media, 2012, vol. 1.
- [68] B. Efron, “Tweedie’s formula and selection bias,” *Journal of the American Statistical Association*, 2011.
- [69] Y. Jafarian and H. S. Park, “Learning high fidelity depths of dressed humans by watching social media dance videos,” in *CVPR*, 2021.
- [70] A. Hore and D. Ziou, “Image quality metrics: Psnr vs. ssim,” in *2010 20th international conference on pattern recognition*, 2010.
- [71] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, and D. Zhang, “Liveportrait: Efficient portrait animation with stitching and retargeting control,” *arXiv preprint arXiv:2407.03168*, 2024.
- [72] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, “Cotracker: It is better to track together,” in *ECCV*, 2024.
- [73] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng *et al.*, “Cogvideox: Text-to-video diffusion models with an expert transformer,” *arXiv preprint arXiv:2408.06072*, 2024.



Shuyuan Tu is currently pursuing his Ph.D. degree in Computer Science in Fudan University. His research interests include conditioned video generation and video understanding. He has published several papers in top-tier conferences such as ICCV and CVPR, focusing on video generation and understanding, and is the main contributor and leader of the open-source model StableAnimator [36] (Github Star 1.3K+) for conditioned video generation.



Zhen Xing received his Ph.D. degree in Computer Science from Fudan University and has published over 20 papers in top AI conferences such as CVPR, ICCV, and ECCV. His research focuses on generative models and video understanding. He joined Alibaba Tongyi Lab in 2025 and is now a Research Scientist, focusing on video generation models.



Zhi-Qi Cheng was a project scientist at Carnegie Mellon University's Language Technologies Institute (School of Computer Science). He is currently an assistant professor at the University of Washington. He also serve as a part-time Research Scientist with the Meta AI AGI team. His research interests include Multimodal Generative AI, Embodied/Robotic Intelligence, and Intelligent Transportation.



Xintong Han is currently a Senior Researcher at Tencent Hunyuan. From 2019 to 2025, he served as a Senior Tech Lead Manager at Huya Inc. He received his Ph.D. degree in Electrical and Computer Engineering from the University of Maryland, College Park, MD, USA, under the supervision of Prof. Larry S. Davis. He also obtained his B.S. degree from Shanghai Jiao Tong University, Shanghai, China. His research interests include computer vision, deep learning, and multimodal understanding and generation.



Qi Dai (Member, IEEE) received the Ph.D. degree in computer science from Fudan University, China, in 2017. He is currently a Principal Researcher with Microsoft Research, Beijing. His research interests include image/video understanding, video generation, and multimedia.



Chong Luo (Senior Member, IEEE) received her B.Sc. degree from Fudan University in Shanghai, China, in 2000, her M.S. degree from the National University of Singapore in 2002 and her Ph.D. degree from Shanghai Jiao Tong University in 2012. She joined Microsoft Research Asia (MSRA) in 2003 and is now a Senior Principal Research Manager with the Visual Computing Group. Her current research interests include image/video understanding, video generation and editing, and intelligent multimedia systems. Her work has been recognized with the ICLR 2023 Outstanding Paper Award.



Zuxuan Wu (Member, IEEE) received the Ph.D. degree in computer science from the University of Maryland, College Park, MD, USA, with Prof. Larry S. Davis, in 2020. He is currently an Associate Professor with the Institute of Trustworthy Embodied AI, Fudan University, Shanghai, China. His research interests are in computer vision and deep learning. Dr. Wu's work has been recognized by the AI 2000 Most Influential Scholars Honorable Mention in 2021, the Microsoft Research Ph.D. Fellowship (ten people worldwide) in 2019, and the Snap Ph.D.

Fellowship (ten people worldwide) in 2017.



Yu-Gang Jiang (Fellow, IEEE) received the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2009. He worked as a Postdoctoral Research Scientist with Columbia University, New York, NY, USA, from 2009 to 2011. He is currently a Professor with the Institute of Trustworthy Embodied AI, Fudan University, Shanghai, China. His research lies in the areas of multimedia, compute vision, and robust and trustworthy AI. Dr. Jiang's work has led to many awards, including the inaugural ACM China Rising Star Award, the 2015

ACM SIGMM Rising Star Award, the Research Award for Excellent Young Scholars from NSF China, and the Chang Jiang Distinguished Professorship appointed by the Ministry of Education of China.