Make-Your-Anchor+: Temporal Consistent 2D Avatar Generation via Video Diffusion Prior

Ziyao Huang, Fan Tang, Juan Cao, Yong Zhang, Xiaodong Cun, *Member, IEEE*, Yihang Bo, Jintao Li, *Member, IEEE*, and Tong-Yee Lee, *Senior Member, IEEE*

***Abstract*—Despite the remarkable process of talking-head-based avatar-creating solutions, directly generating anchor-style videos with full-body motions remains challenging. In this study, we propose Make-Your-Anchor+, a novel system necessitating only a one-minute video clip of an individual for training, subsequently enabling the automatic generation of anchor-style videos with precise torso and hand movements. Specifically, we finetune a proposed structure-guided diffusion model on input video to render 3D mesh conditions into human appearances. We adopt a two-stage training strategy for the diffusion model, effectively mapping movements with specific appearances to create digital avatars for online streamers, live shopping hosts, and other applications. To produce arbitrary long temporal video, we extract human motion information from video diffusion prior by adapting the frame-wise diffusion model to pretrained video diffusion weights with lower cost, and a simple yet effective batch-overlapped temporal denoising module is proposed to bypass the constraints on video length during inference. Finally, a novel identity-specific face enhancement module is introduced to improve the visual quality of facial regions in the output videos. Comparative experiments demonstrate the system's effectiveness and superiority in visual quality, temporal coherence, and identity preservation, outperforming SOTA diffusion/non-diffusion methods.**

***Index Terms*—Human video generation, digital human synthesis.**

I. Introduction

AUTOMATICALLY generating 2D human videos with realistic expressions, body movements, and gestures under diverse conditions—such as audio, text, music, or motion—has broad applications in e-commerce, online education, virtual conferencing, and VR. While AI-powered digital anchors provide advantages such as continuous availability and rapid content generation, they also raise significant concerns regarding the quality and fidelity of the generated content.

A common practice in popular 2D avatar systems is GAN-based talking face generation [1], [2], [3], [4], which focuses exclusively on editing the lip or facial region. These methods rely on GANs [5], [6], [7] to manipulate facial appearances. Despite achieving remarkable visual quality, they restrict the anchor's degrees of freedom. To generate more natural and expressive digital avatars, researchers have explored motion transfer [8], [9], [10], [11] and co-speech-driven generation [12], [13]. By learning mappings from motion or speech to a person's appearance, these methods can produce full-body talking videos with torso and hand movements. However, GAN-based approaches often struggle to maintain high visual fidelity for fine details or rapid motion.

Recently, diffusion models [14] have demonstrated impressive generation quality for human images [15]. By conditioning on pose or depth [16], [17], text-to-image diffusion models such as Stable Diffusion [15] can synthesize human bodies with specific gestures, poses, and expressions, while identity can be preserved via personalized fine-tuning [18], [19], [20], [21]. However, the inherent stochasticity of the diffusion process prevents these models from generating temporally consistent human videos directly.

Existing video diffusion models [22], [23] have been proven to generate temporally consistent videos with simple motions and natural scenes. DreamPose [24] extends Stable Diffusion to be conditioned on an input image and pose, animating a still human image along a given motion sequence. DisCo [25] further incorporates pose and background ControlNets [16] to compose diverse motions and scenes via a pretraining strategy, enabling generalizable human dance generation. Despite these advances, diffusion-based approaches still struggle to synthesize fine facial details, accurate gestures, and temporally consistent frames—limiting their utility for realistic digital anchors. Recent methods [26], [27], [28], [29] introduce temporal modules to extend image diffusion to video diffusion, training all weights on video datasets. While effective, these approaches require carefully curated datasets and substantial computational resources. Moreover, most existing methods rely on a single input image as the appearance reference, which often results in identity distortion due to the inherently ill-posed nature of the problem [30], [31].

In this study, we propose "Make-Your-Anchor+", a diffusion-based 2D avatar generation framework for creating customized

Received 9 September 2024; revised 26 November 2025; accepted 8 January 2026. Date of publication 19 January 2026; date of current version 2 March 2026. This work was supported in part by the National Natural Science Foundation of China under Grant 62572458, in part by CCF-Tencent Rhino-Bird Open Research Fund, in part by Beijing Film Academy Research Base Project under Grant 7000452203/033, and in part by the National Science and Technology Council, Taiwan, under Grant 113-2221-E006-161-MY3. Recommended for acceptance by L. Liu. *(Corresponding author: Fan Tang.)*

Ziyao Huang, Fan Tang, Juan Cao, and Jintao Li are with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100045, China, and also with the University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: tfan.108@gmail.com).

Yong Zhang is with Tencent, Shenzhen 518054, China.

Xiaodong Cun is with Great Bay University, Guangdong 523000, China.

Yihang Bo is with Beijing Film Academy, Beijing 100088, China.

Tong-Yee Lee is with National Cheng Kung University, Tainan 701, Taiwan.

Digital Object Identifier 10.1109/TVCG.2026.3655478

digital avatars. While diffusion-based approaches achieve impressive visual quality, generating human images or videos remains challenging in terms of identity preservation [32] and motion consistency [26], [27], [29]. To address this, we learn to link human appearance with parametric human representations. Specifically, we replace text CLIP [33] inputs with image-based appearance features and introduce a ControlNet-style modulation network to map appearance to pose control sequences on a frame-by-frame basis. Our system is pretrained on an existing anchor dataset [34] and fine-tuned on a one-minute video of a target individual. To generate temporally consistent videos, we propose a novel adaptation of the video diffusion prior, integrating motion information from pretrained video diffusion models into a customized image diffusion model with minimal computational overhead. Building on this adapted video model, we introduce batch-overlapped temporal denoising, a training-free inference strategy that computes frame-wise latent noise in sliding windows, enabling long-duration anchor videos of arbitrary length. Finally, to improve facial fidelity, we apply a novel inpainting-based face enhancement operation that focuses on fine-grained facial details.

With frame-wise training and batch-wise inference, the proposed system can generate high-quality videos within a practical runtime on a single 40 GB A100 GPU. Leveraging existing methods to generate 3D human meshes with lips, gestures, and body motions conditioned on music [35], [36], [37], [38], audio [34], [39], text [40], or motion capture [41], our framework enables AI-powered digital anchors with vivid expressions, torso movements, and gestures. Compared to commercial face-generation software, our system produces natural full-body videos with greater flexibility.

The main contributions are summarized as follows:

- We introduce Make-Your-Anchor+, a 2D avatar customization system capable of generating digital anchors with expressive lips, facial expressions, gestures, and body motions.
- We propose a frame-wise motion-to-appearance diffusing strategy with a two-stage training paradigm, combined with batch-overlapped temporal denoising, to produce temporally consistent human videos of arbitrary length.
- We introduce temporal denoising via video diffusion prior, leveraging pretrained video priors to improve inter-frame motion modeling and achieve temporally coherent outputs.
- Extensive qualitative and quantitative experiments on ten anchors demonstrate the effectiveness of our approach compared with state-of-the-art GAN-based motion transfer and diffusion-based human video generation methods.

A preliminary version of this work was published at CVPR 2024 [42]. The journal article improves over the conference paper in multiple aspects. First, to produce temporally consistent human videos, we propose adapting the video diffusion prior strategy to effectively incorporate natural movement information from the video diffusion model into our proposed system. Second, we strengthen our evaluation by including two more recent human video generation methods, i.e., MagicAnimate [26], and Champ [29], which provide additional samples of the established type in the preliminary version.

II. Related Work

A. Talking Face Generation

Talking face generation methods [1], [2], [3], [43] generate human videos with various expressions and poses conditioned on a given audio or motion, which can be categorized into two types: editing facial regions or generating dynamic head videos. Editing-based techniques, such as Wav2Lip [1] or VideoRetalking [43], face the problem of lip-gesture inconsistency. Usually, the gestures are fixed for different types of talking content. Generating dynamic head videos requires methods that condition head videos on given audio or motion, where head motions are achieved in different ways, such as motion flow [44], 3D landmarks [2], [3], and self-supervised training [4], among others. Although producing high-quality and highly realistic facial videos, talking face generation is limited by its area of interest and cannot achieve full-body human video generation.

B. Pose-Guided Human Video Generation

Pose-guided methods are the most popular approaches for generating human videos with body and hand. Early work focuses on the problem of motion transfer [8], [9], [10], [11], [45], [46], [47], [48], [49] methods. Balakrishnan et al. [45] separate a scene and transform each part to synthesize. FOMM [10] and MRAA [9] propose unsupervised body representation and warping to transfer, while LIA [49] employs latent space. TPS [8] introduces thin plate spline transformation into a motion transfer task, and UVA [48] presents a differential volumetric representation. Besides the coarse-grained motion transfer setting, researchers [12], [13], [50], [51] apply these kinds of methods to human video generation with face, body, and hand. However, constrained by the capabilities of generative models, these methods may potentially generate human-like videos with apparent artifacts.

With the advancement of diffusion models [14], [15], some works have introduced them into pose-guided human video generation. Follow-your-pose [52] introduces a two-stage training to get a pose-guided video diffusion model. DreamPose [24] proposes a vision and pose-controlled diffusion on a fashion dataset to animate a body image. DisCo [25] focuses on human dance generation, utilizing multiple ControlNets for pose and background, and introduces a pretraining strategy to enhance generalizability. Besides, a temporal module is proposed to strengthen temporal consistency. MagicAnimate [26] and AnimateAnyone [27] leverage the feature maps from self-attention layers to capture appearance. To improve temporal consistency, these approaches extend image diffusion models into video diffusion models and train on privately collected datasets. Champ [29] further combined multiple pose conditions, including landmarks, depth, normal map, and DensePose [53]. The complementary structure guidance improves the capture of pose and shape variations. Nonetheless, these methods concentrate on coarse-grained body video generation, which is limited by the poor quality of the face and hands. MagicPose [54] utilizes human pose and expression retargeting as a plug-in for the human video diffusion model to improve body and face fidelity.

Besides, these approaches learn human motions by fine-tuning temporal layers, which increases the risk of losing motion fidelity if the dataset is incomplete. MagicAnimate performs joint image-video training with all weights updated simultaneously, while Animate Anyone and Champ adopt a two-stage training scheme, where only the temporal attention is fine-tuned in the second video stage. In contrast, our approach maintains the original motion capabilities by incorporating a video diffusion prior, even with fine-tuning on just a one-minute video.

C. Video Diffusion Models

Due to the powerful capabilities of diffusion models, researchers in recent years have started to explore their potential in video generation, and much progress has been made in video generation [22], [23], [55], [56], [57], [58], [59] and video editing [60], [61], [62]. GEN-1 [63] and Make-Your-Video [59] extend the image diffusion model with a temporal module and utilize depth to control the structure. Tune-A-Video [64] fine-tunes a 3D U-Net in an image diffusion model on a one-shot video to learn the motion and then edits the video content with text prompts. AnimateDiff [65] trains a temporal module with a fixed image diffusion model and can be applied to personalized weights. DynamiCrafter [58] animates open-domain images by leveraging a video diffusion prior. While VDMs possess strong video generation capabilities, the ability to control human motion and maintain appearance needs further improvement. In contrast, we tune the foundation diffusion model to learn mapping from motion to a specific anchor appearance in an "appearance mapping" fashion, following a pretraining-finetuning paradigm.

III. System Overview

A. Setting

Despite recent advances in talking-face [3] and fashion video generation [24], current approaches that rely on a single or a few reference images remain limited for practical human video generation. Talking-head methods operate only on the facial region, while models such as DreamPose generate fashion videos with a restricted range of movements. We instead formulate 2D avatar generation as a personalized learning problem built from a single-identity video. This setting resembles talking-head pipelines and commercial systems such as ZenVideo, but existing methods typically learn only facial appearance and reuse the input video's body motions as fixed "templates" often resulting in looped or repetitive motion patterns. The core of our system is to learn a personalized diffusion model that could generate a human video in the same scenario as the input video. We finetune a diffusion model to directly map to the input video under both appearance and motion conditions, enabling controllable, diverse body movements while maintaining consistency with the identity and scenario. To the best of our knowledge, this is the first system that enables 2D human avatar generation with controllable full-body motions while preserving high visual fidelity, identity consistency, and temporal coherence.

B. Input

Based on the above setting, our system requires an anchor-style source video $\mathbb{S}$ of a single person for training. The source video should include natural lip, body, and gesture movements, and typically needs to be longer than one minute. Unlike DreamPose or ControlNet variants conditioned on OpenPose, we employ human 3D mesh sequences rendered from SMPL-X parameters [66] as motion conditions. The smoothness and accuracy of the input pose conditions are crucial for achieving temporally consistent outputs. Leveraging 3D meshes provides richer structural information, enabling smoother motion synthesis, particularly in fine-grained gestures such as hand movements. During inference, a pose sequence $\mathbb{P} = p_1, p_2, \ldots, p_n$ is provided to guide the diffusion-based 2D avatar. Our system can also be seamlessly combined with existing human motion generation pipelines, such as audio-to-motion synthesis [34] or video-guided motion transfer [41]. In these settings, the input can be an audio clip or a motion reference video. The final output is a sequence of human video frames $\mathbb{X} = x_1, x_2, \ldots, x_n$ with temporally coherent appearance.

IV. Methodology

Fig. 1 presents the overall network structure and inference pipeline of our system. Section IV-A introduces the diffusion-based architecture that enables image-level appearance control and body structure conditioning. Section IV-B describes our two-stage training strategy. Section IV-C then details the temporal denoising module based on a video diffusion prior, which allows the generation of arbitrarily long and temporally consistent videos with reduced training cost. Finally, Section IV-D introduces an inpainting-based facial enhancement module that further improves visual fidelity in the facial regions.

A. Frame-Wise Motion-to-Appearance Diffusing

1) Preliminaries: Given an initial noise $\epsilon \sim \mathcal{N}(0, 1)$ and condition c, the training objective of a vanilla frame-wise diffusion model is defined as:

$$L_{\text{diff}} = \mathbb{E}_{s_i, c, \epsilon, t}\left[w_t ||\hat{X}_\theta(\alpha_t s_i + \sigma_t \epsilon, t, c) - \epsilon||_2^2\right], \quad (1)$$

where $\hat{X}_\theta$ denotes the diffusion model, s_i is the i-th frame of the source video $\mathbb{S}$, and α_t, σ_t, and w_t are predefined hyperparameters. After training, the target output frame $x_i = s_i^0$ is obtained by denoising an initial noise sample s_i^T through a T-step sampling process. Following latent diffusion [15], we employ a pretrained VAE encoder $\mathcal{E}(\cdot)$ to map each frame s_i into a latent code z_i, and perform diffusion and denoising directly in the latent space. A pretrained VAE decoder $\mathcal{D}(\cdot)$ subsequently transforms the denoised latent representation back into pixel space.

2) Structure-Guided Diffusion Model: We propose a structure-guided diffusion model (SGDM) to generate human videos under frame-wise 3D mesh control, as illustrated in Fig. 2. Unlike ControlNet [16], we embed the 3D mesh condition into the generation process to learn a mapping from pose $\mathbb{P}$ to target video frames $\mathbb{X}$. To accomplish this, we introduce an

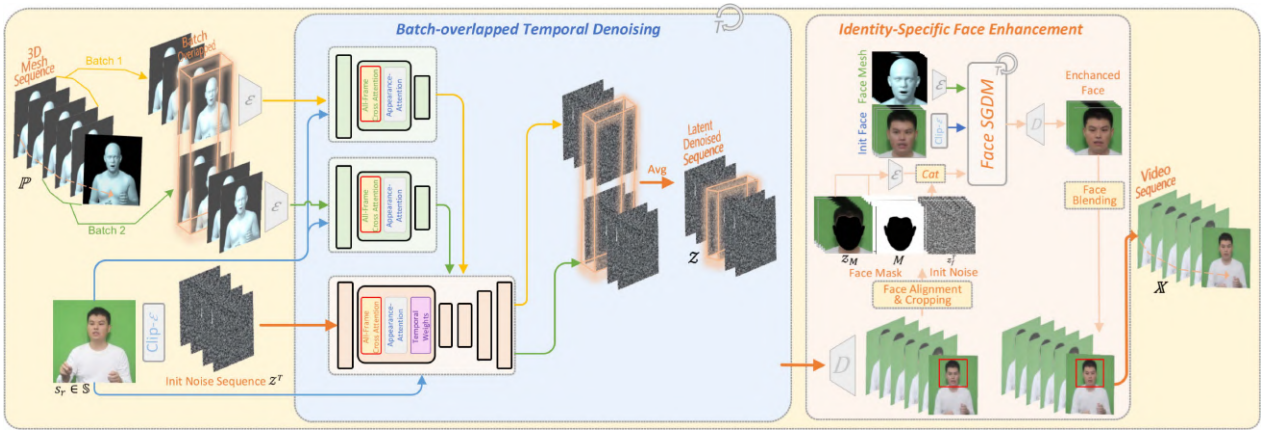


Fig. 1. The inference pipeline of our system. An appearance condition and a 3D mesh sequence are inputted into the structure-guided diffusion, incorporating Batch-overlapped Temporal Denoising to accomplish video-level inference. Following the generation of arbitrary-length frame sequences, an inpainting-style module known as Identity-Specific Face Enhancement is utilized to enhance facial details.

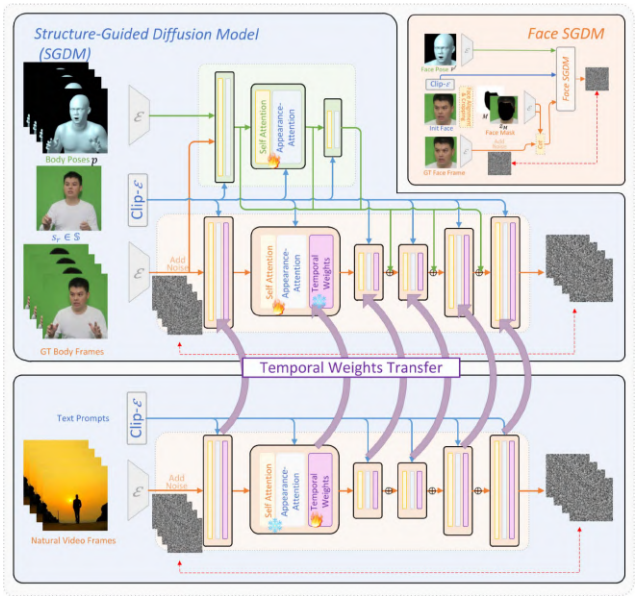


Fig. 2. The network architecture of our proposed Structure-Guided Diffusion Model (SGDM) and Face SGDM. Our network achieves motion-to-appearance generation by embedding pose and appearance conditions into the pretrained diffusion model. Temporal weights are transferred from the pretrained video diffusion model to generate consistent videos.

additional conditioning branch copied from the U-Net encoder and apply a zero-convolution layer to process the input mesh image. The layer-wise feature maps from this condition branch are fused into the U-Net decoder via additive modulation:

$$F'_{\text{up}} = F_{\text{up}} + W_c \times ControlN(p_i), \tag{2}$$

where F_{up} denotes the decoder feature map, $ControlN(\cdot)$ is the conditioning module, p_i is the i-th mesh in $\mathbb{P}$, and W_c is a weighting coefficient for the condition signal.

For the appearance input, we inject identity information by randomly selecting another source frame s_r from $\mathbb{S}$ and replacing the text-based CLIP [33] embeddings with an image CLIP feature c_{s_r} extracted from s_r. This appearance embedding is introduced into the cross-attention modules of both the latent diffusion U-Net and the motion-control branch. We specifically employ the CLIP feature map before the global pooling layer, as it retains fine-grained spatial details crucial for appearance preservation. After training, the generation of each human frame is given by

$$z_i^0 = \hat{X}_\theta(ControlN(p_i), cs_r, z_i^T), \tag{3}$$

$$x_i = \mathcal{D}(z_i^0). \tag{4}$$

B. Two-Stage Training Strategy

We adopt the pretrained text-to-image Stable Diffusion V1.5 [15] as our base model and build the proposed SGDM upon it. Our objective is to learn the movement and appearance of the target identity through an appearance-mapping paradigm. To this end, we introduce a two-stage training strategy: (1) pre-training on multiple identities to strengthen the model's motion generation capability, and (2) fine-tuning on a single identity to effectively bind the generated motion to the target appearance.

For the first stage, we use a publicly available anchor video dataset [34] to train the new control branch and adapt the base diffusion model across diverse identities. We segment human foregrounds in this stage to remove background interference and encourage the model to focus on human motion and structure. In the second stage, we fine-tune the entire network using the one-shot video of the target identity, jointly training on both foreground and background. This enables rapid adaptation to new identities while preserving the scene context, ultimately producing coherent anchor-style videos.

C. Temporal Denoising Via Video Diffusion Prior

1) Adaptation to Video Diffusion Prior: By adopting the two-stage training strategy, the proposed SGDM can generate anchor-style videos frame-by-frame conditioned on a given motion sequence. However, diffusion models inherently introduce stochasticity, leading to temporal discontinuities. While training-free extensions of image diffusion models [67],

[68] improve temporal consistency by adding cross-frame attention, models trained solely on image distributions still struggle to handle non-planar temporal variations such as clothing dynamics, hair motion, and large pose changes. As later discussed in Section V-C4, these approaches often produce unnatural cloth folds and temporal artifacts.

To address this limitation, we incorporate a video diffusion prior into our image diffusion model. We employ AnimateDiff [65], which extends the 2D U-Net into a 3D U-Net by inserting temporal attention layers $Attention_{\mathbb{T}}$ after each spatial attention block. These temporal layers are trained on large-scale video datasets [69] while keeping spatial weights frozen, allowing the model to encode temporal priors that match real video distributions.

Exploiting the motion priors from the temporal weights is critical to our framework. Unlike existing works [26], [27], [29], [65], our setting requires fine-tuning on single-identity videos. Directly updating the temporal layers in this stage degrades the learned temporal priors. Therefore, we design a training strategy that preserves these priors while adapting the spatial representation.

Specifically, we investigate three integration strategies: (1) directly inserting temporal weights, (2) fine-tuning temporal weights, and (3) fine-tuning spatial weights. Direct insertion improves consistency but introduces color shifts and sticking artifacts. Fine-tuning temporal weights causes flickering due to catastrophic forgetting of large-scale temporal knowledge. In contrast, tuning only the spatial weights maintains motion coherence while preserving appearance fidelity. Ablation results in Section V-C5 and supplementary videos verify these findings.

Building on this insight, we adopt a temporal-preserving adaptation strategy: we first fine-tune spatial weights on the multi-identity pre-training set to learn diverse motion patterns, and subsequently fine-tune on single-identity data to map motion dynamics to the target appearance.

2) Batch-Overlapped Temporal Denoising: We propose an overlapped temporal denoising algorithm to generate anchor videos of arbitrary length, as detailed in Algorithm 1. Unlike prior methods that optimize objectives across multiple noisy batches [70] or rely on windowed attention [71], our approach is simple yet effective, operating directly on multi-batch noise. Specifically, we partition a long motion sequence into overlapping windows. During each denoising step, these windows are processed sequentially by the video diffusion model. After all windows are handled, overlapping noise regions are normalized to ensure temporal coherence, and the entire video is denoised using the smoothed noise. Repeating this process across all steps produces a final long video with consistent temporal dynamics. We find that this simple algorithm effectively extends the video diffusion model to generate arbitrarily long sequences while maintaining temporal consistency.

Algorithm 1: Overlapped Temporal Denoising.

Input: Video Diffusion Model $\hat{X}_\theta^V$, Timestep t, Latent sequence $Z^t = \{z_1^t, z_2^t, ..., z_N^t\}$, Pose sequence $\mathbb{P} = \{p_1, p_2, ..., p_N\}$, Window size ws, Overlap size os, $Scheduler$.

Result: Denoised latent sequence $Z^{t-1} = \{z_1^{t-1}, z_2^{t-1}, ..., z_N^{t-1}\}$

$count \leftarrow \{0, 0, ..., 0\}$;
$Noise \leftarrow \{0, 0, ..., 0\}$;
for $i = 0, ..., N/(ws - os)$ **do**
  $wb \leftarrow i * (ws - os)$;
  $we \leftarrow i * (ws - os) + ws$;
  $Z_{wd}^t \leftarrow Z^t[wb : we]$;
  $P_{wd} \leftarrow P[wb : we]$;
  $N_{wd}^t \leftarrow \hat{X}_\theta^V(Z_{wd}^t, t, P_{wd})$;
  $Noise[wb : we] \leftarrow Noise[wb : we] + N_{win}^t$;
  $count[wb : we] \leftarrow count[wb : we] + 1$;
end
`# Normalize overlapped noises`
$Noise \leftarrow Noise/count$;
$Z^{t-1} \leftarrow Scheduler(Noise, t, Z^t)$;

D. Identity-Specific Face Enhancement

Generating a high-quality face within a holistic human frame remains challenging due to the relatively small proportion of the facial region. When trained with a global human-level loss, the model often fails to capture fine facial details. To address this, we guide the model to focus on key facial features using an inpainting-based approach.

As illustrated in the right part of Fig. 2, we extend the original SGDM into a dedicated Face SGDM. The process is as follows: the face region is first cropped from the fully generated body frame, and the inpainting area is segmented using a facial mask M. The input to the U-Net in SGDM is then replaced with a concatenation of the facial mask M, the masked face latent z_M, and the original input latent z_f^T. Appearance and motion conditions are aligned with their respective reference image and 3D mesh. The face generation is formulated as:

$$z_f^0 = \hat{X}_\theta^f(ControlN(p^f), {cs_r}^f, z_f^T, M, z_M), \tag{5}$$

where $\hat{X}_\theta^f$ denotes the Face SGDM, z_f^0 is the output latent, and z_f^T is the input latent. Algorithm 2 summarizes the procedure.

This inpainting-based refinement requires no paired training data and trains solely using the ground truth, enabling high-fidelity facial synthesis within full-body frames.

V. Experiments

A. Experimental Settings

1) Dataset: For pretraining, we follow [34], [50] and use 27 hours of video with SMPL-X annotations covering four identities. Robust video matting [72] is applied to remove backgrounds from the pretraining videos. For one-shot video fine-tuning, we collected a dataset of ten identities from diverse sources, each providing a one to five-minute video. To evaluate our method across various scenarios, the dataset includes: (1) four videos of identities used in pretraining, (2) three web videos of celebrities (Luo Xiang, Guo Degang, and a video from the YouTube channel ScienceOfPeople[1]), and (3) three green-screen recordings of

[1] https://www.youtube.com/@ScienceOf People

Algorithm 2: Identity-Specific Face Enhancement.

Input: Generated body frame X, facial mask M, original input latent z_f^T, appearance condition c_{s_r}, 3D mesh pose P

Output: Enhanced face latent z_f^0

```
# Crop face region from the generated body
```
$X_f = \text{CropFace}(X)$;
```
# Segment inpainting region
```
$R = X_f \odot M$;
```
# Encode masked face to latent
```
$z_M = \text{Encode}(R)$;
```
# Align face pose
```
$p^f = \text{CropFace}(\text{P})$
```
# Get face image embedding
```
$c^f_{s_r} = CLIP(X^f)$;
```
# Generate enhanced face latent
```
$z_f^0 = \hat{X}^f_\theta(ControlN(p^f), c^f_{s_r}, z_f^T, M, z_M)$;
```
# Generate enhanced face latent
```
$X_f^{\text{enh}} = \text{D}(z_f^0)$;
```
# Blend enhanced face back to body:
```
$X_{\text{final}} = M \odot X_f^{\text{enh}} + (1 - M) \odot X$;

invited individuals. All videos are split into clips of 300 frames (10 seconds at 30 FPS). For each identity, one-minute-long clips are used for training, and additional clips are reserved as test motion samples.

2) Comparison Methods: We compare our method with six state-of-the-art approaches: Pose2Img [12], TPS [8], DreamPose [24], DisCo [25], MagicAnimate [26], and Champ [29]. Pose2Img, derived from the co-speech work SpeechDrivesTemplates [12] and modified from [45], performs person-specific video generation conditioned on human landmarks and represents the most relevant GAN-based baseline. TPS [8] is a generic motion-transfer method based on thin plate spline transformations. DreamPose, DisCo, MagicAnimate, and Champ are diffusion-based human video generation systems. For fair comparison, we train Pose2Img from scratch, fine-tune DreamPose and DisCo on our dataset, and use the pretrained weights of TPS, MagicAnimate, and Champ without modification. Additionally, we evaluate MagicAnimate and Champ under fine-tuning on our dataset with their original training protocols: MagicAnimate fine-tunes all parameters, whereas Champ employs a two-stage process, first fine-tuning spatial weights with the motion module frozen, then fixing spatial weights and fine-tuning only the motion module.

3) Implementation Details: We crop the body region to 512×512 pixels and the face to 256×256 pixels for enhancement. A reference appearance image is randomly selected from another frame of the same video during training. The body generation pre-training lasts 300,000 steps with a video length of 4, batch size 1, and learning rate 1e-5, taking approximately seven days. The face enhancement model is trained for 50,000 steps and completes in under one day. Fine-tuning on a single identity requires around one day, with 50,000 steps for body generation and 80,000 steps for face enhancement. All experiments are conducted on a single NVIDIA A100 GPU (40GB). During inference, the classifier-free guidance scale (CFG) [73] is set to 3.5. We set the ControlNet condition scale W_c to 2, which improves preservation of the 3D mesh structure and hand gestures. SMPL-X annotations are generated using the tools from [34]. For face enhancement, FFHQ [5] preprocessing is applied for alignment and cropping, and facial segmentation is used to extract the inpainting mask. For overlapped temporal denoising, the window size ws is set to 16, with an overlap os of 4, and the number of denoising steps is 20. Using batch-overlapped temporal denoising, inference takes approximately ten minutes for a 300-frame video, compared to 30 minutes with frame-by-frame inference.

TABLE I
QUANTITATIVE RESULTS OF OUR METHOD COMPARED WITH SOTAS AND ABLATION STUDIES. OUR METHOD ACHIEVES BETTER PERFORMANCE ON IMAGE QUALITY, TEMPORAL CONSISTENCY, AND STRUCTURE PRESERVATION.

Method	FID↓	FVD↓	LMD (Face)↓	LMD (Body)↓	LMD (Hand)↓
Pose2Img	51.92	328.18	3.73	4.78	6.96
TPS	136.02	884.97	5.22	8.37	18.72
DreamPose	83.47	868.20	4.21	5.92	13.85
DisCo	60.95	390.77	4.15	5.11	11.21
MagicAnimate	97.97	310.95	3.43	**4.12**	11.39
w/ fine-tuned	464.95	1043.41	3.76	5.72	8.72
Champ	53.39	144.78	1.93	5.33	8.44
w/ stage1 fine-tuned	56.97	409.14	5.73	5.52	9.14
w/ stage2 fine-tuned	57.02	420.11	6.17	5.91	9.41
Ours	**35.91**	**53.38**	**1.45**	5.14	**5.12**
w/o Video Diffusion Prior	40.33	139.82	1.44	4.88	5.41
w/o TD	48.84	344.85	1.45	4.74	5.61
w/o FE	40.84	124.10	1.56	-	-
w/o Two-Stage Setting I	55.87	278.73	4.38	6.92	7.25
w/o Two-Stage Setting II	53.99	178.79	1.56	4.96	6.01
SMPL perturbation	39.56	136.21	1.40	4.33	5.25

4) Metrics: We evaluate image quality using FID [74] and temporal consistency using FVD [75]. To quantify body-structure preservation, we compute Landmark Mean Distances (LMD) between video frames and the input 3D mesh. LMD is calculated separately for the face, body, and hands using OpenPose [41]. To mitigate inaccuracies in hand landmark predictions, outliers are excluded from the computation.

B. Main Results

We utilize the data mentioned above as our experiment dataset. Three video clips are collected for each identity to be used as test data.

1) Quantitative Results: Quantitative results are summarized in Table I. Our method achieves the best performance on FID and FVD, indicating superior image quality and temporal consistency. For structure preservation, we obtain the highest scores on face and hands, while LMD(body) is comparable to other approaches. Although MagicAnimate attains the best LMD(body), its strict adherence to input conditions often leads to appearance distortions. Fine-tuning MagicAnimate and Champ on single-identity data according to their original protocols causes a substantial drop in FVD, as their training disrupts the learned temporal distribution, resulting in flickering or jitter in generated videos. In contrast, our fine-tuning strategy preserves the temporal distribution, producing more consistent and higher-quality outputs. We note that the landmarks used for LMD computation

are also employed for Pose2Img and DisCo inference, which may bias their results in structure evaluation.

2) Qualitative Results: Qualitative results are shown in Fig. 3, where blue boxes highlight distorted hand structures and arrows indicate appearance artifacts. Our method achieves the best image quality, appearance preservation, and facial detail, while accurately capturing mouth movements and hand gestures. CHAMP, using the same SMPL parameters as pose conditions, maintains structural consistency reasonably well, but issues such as finger fusion remain. MagicAnimate strictly follows input poses, achieving the best LMD(body) in Table I but failing to preserve identity, and its fine-tuning leads to significant mode collapse. Champ's two-stage, partial fine-tuning preserves appearance better, yet suffers from flickering as reflected in quantitative results. DisCo produces high-quality diffusion-based outputs but lacks fine detail in hands and faces, resulting in noticeable imperfections. DreamPose is limited to simple scenes, such as fashion videos, making it unsuitable for complex backgrounds or actions in our setting. TPS, a general warping-based method, struggles to generalize beyond reference poses in unfamiliar domains. Pose2Img, a person-specific warping-based approach, outperforms other baselines but still exhibits artifacts in hands and lips. In contrast, our approach learns a motion-to-appearance mapping via a diffusion model combined with a temporal scheme, enabling the generation of high-quality, temporally consistent human videos.

Cross-person Motion Results We further evaluate cross-person motion transfer, with results shown in Fig. 4. When the source motion closely matches the target person's pose (e.g., standing man to standing man), the generated videos maintain high visual quality and realistic motion.

Full-body Results Since most previous results focus on half-body videos, we further evaluate our method on the collected full-body videos, as shown in Fig. 5. Despite the full-body talking style differing from the training set, the generated results maintain high visual quality and coherent motion.

Audio-driven Results We demonstrate audio-driven digital avatar generation using TalkSHOW [34] to drive the 3D human mesh. Examples are provided in both the supplementary video and Fig. 6. By integrating our method with existing audio-driven motion generation techniques, we establish a system capable of automatically producing 2D avatar videos from audio input.

C. Ablation Study

1) Validation of Batch-Overlapped Temporal Denoising: To evaluate the effectiveness of Batch-overlapped Temporal Denoising (TD), we perform both quantitative and qualitative analyses. Ablation results are summarized in Table I. Removing TD leads to a significant drop in temporal performance. Although our method without TD still outperforms DisCo due to detailed pose conditioning, it performs slightly worse than warping-based methods that leverage inter-frame information. LMDs remain largely unaffected, as SGDM primarily governs structure generation.

We further study the TD hyperparameters: window size ws and overlap size os, with the default setting $ws = 16, os = 4$.

TABLE II
ANALYSIS OF VARYING THE HYPERPARAMETERS OF BATCH-OVERLAPPED TEMPORAL DENOISING, WINDOW SIZE ws AND OVERLAPPED SIZE os. THE DEFAULT SETTING OF OUR METHOD IS $ws = 16, os = 4$.

Method	FID↓	FVD↓	LMD (Face)↓	LMD (Body)↓	LMD (Hand)↓
$ws = 8, os = 2$	34.88	54.63	1.47	4.86	5.46
$ws = 16, os = 4$	35.91	53.38	1.45	5.14	5.12
$ws = 32, os = 8$	36.52	51.17	1.43	4.89	5.33

We also test $ws = 8, os = 2$ and $ws = 32, os = 8$. Results in Table II show that larger settings improve temporal metrics and FVD, but inference cost grows as $O(ws^2)$, leading us to select $ws = 16, os = 4$ for a balance between quality and efficiency.

Qualitative results are shown in Fig. 7. Without TD, generated videos exhibit temporal artifacts such as flickering and ghosting, likely because the image-level model fails to fully capture the human body distribution. Incorporating contextual information through Batch-overlapped TD effectively reduces these unintended artifacts, yielding smoother and more temporally consistent videos.

2) Validation of Identity-Specific Face Enhancement: We analyze the Identity-Specific Face Enhancement (FE) module. Quantitative ablation results are presented in Table I. LMD(Face) is improved by FE, demonstrating its effectiveness in generating fine facial details. FVD shows a slight decrease, as FE is applied without temporal denoising. Qualitative results are shown in Fig. 8. As observed in Fig. 3, our baseline method, like other diffusion-based approaches, sometimes struggles to generate high-quality facial features directly. The FE module significantly enhances facial detail, improving the visual fidelity of generated faces.

3) Validation of Two-Stage Training: We investigate alternatives to our two-stage training strategy. Setting I directly trains the model on videos of the target subject, while Setting II trains on data from all other subjects. As shown in Fig. 9(a) and Table I, Setting I suffers from insufficient data for robust pose control. Setting II, illustrated in Fig. 9(b), leads to appearance and identity entanglement with other subjects and is difficult to extend to new identities. In contrast, our two-stage training enables accurate pose control while preserving the target subject's appearance.

4) Comparison With Make-Your-Anchor: Compared to Make-Your-Anchor, Make-Your-Anchor+ significantly improves temporal consistency by incorporating a video diffusion prior. As shown in Table I, adding the prior substantially enhances FVD, indicating better video continuity. Qualitative results are presented in Fig. 10, with video comparisons included in the supplementary material. We observe that the video diffusion prior helps the model generate realistic clothing fold dynamics and reduces color flickering artifacts, demonstrating that Make-Your-Anchor+ produces videos with notably improved temporal consistency and visual fidelity.

5) Validation of Adaptation Strategies: We compare the strategies described in Section IV-C1 by analyzing temporal differences between generated frames. Fig. 10 visualizes the generated frames alongside their temporal differences. The first row shows ground-truth results, while frames with small and large

Fig. 3. Qualitative results compared with other methods. Our methods achieve accurate gestures and high-quality generation with facial details. More results are provided in supplementary materials.

Fig. 4. Cross-person motion results. For each pair, the left is the pose, and the right is the output.

Fig. 5. Full-body results with self-collected videos as training data.

Fig. 6. Examples of audio-driven results.

motions are displayed together for comparison. For large-motion frames, directly applying temporal weights produces minimal temporal differences relative to ground truth, indicating sticking artifacts, along with noticeable color aberrations. For small-motion frames, tuning temporal weights induces flickering, as highlighted by red boxes. Fixing the temporal weights while tuning spatial weights provides a favorable trade-off, maintaining image quality while preserving temporal consistency. Supplementary videos provide further visual comparisons.

6) Discussion of SMPL-X Parameters: Estimating SMPL inevitably introduces errors. We preprocess SMPL parameters following TalkSHOW [34], optimizing them via multiple estimations to reduce inaccuracies. To further study the effect of imperfect SMPL, we perturb SMPL parameters during training. Specifically, we add noise to the position coordinates of each frame, causing visible jitter in the rendered 3D human mesh. This serves as a form of spatial data augmentation, improving the model's generalization to SMPLX positional variations and enhancing overall accuracy. As shown in Table I and Fig. 11, our method still faithfully follows the input poses. Quantitative improvements indicate that SMPL perturbation as data augmentation further boosts performance.

7) Influences of Input Mesh Distributions: In Figs. 12 and 13, we present samples of fine-tuning poses and cross-person motion. Specifically, in Fig. 13, we demonstrate the effect of driving different identities with the same motion. The motion sequences of each individual are extracted using the preprocessing pipeline of SHOW [34]. During fine-tuning, human meshes of different individuals exhibit noticeable variations in shape,

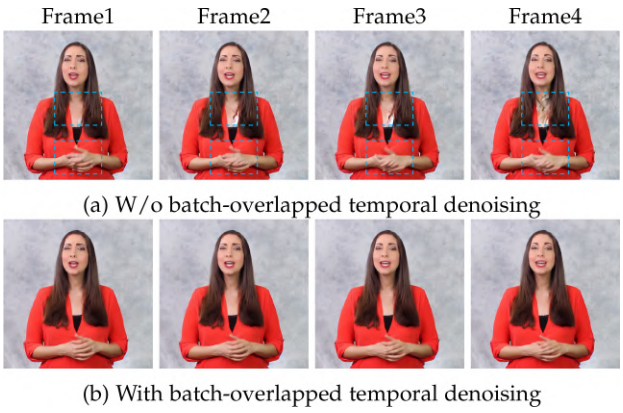


Fig. 7. Ablation analysis of temporal denoising. The **first row** shows the result without batch-overlapped temporal denoising, whereas the **second row** shows it with batch-overlapped temporal denoising. Without temporal denoising, the generated videos may exhibit discontinuities such as flickering and ghosting as the artifacts around the neck region on the first row.

Fig. 8. Ablations of identity-specific face enhancement. Directly training models to generate entire human body images often leads to the creation of facial features with poor realism. Our identity-specific face enhancement significantly improves the quality of generated content in the face region.

TABLE III
ABLATION ANALYSIS OF VIDEO LENGTH USED FOR FINE-TUNING

Fine-tuning with	FID↓	FVD↓	LMD (Face)↓	LMD (Body)↓	LMD (Hand)↓
One-minute videos	35.91	53.38	1.45	5.14	5.12
Five-minute videos	33.56	49.70	1.27	4.86	4.71

and each identity has distinct motion distributions [34]. At the inference stage, both shape and motion distribution differences need to be considered. As illustrated in Fig. 13, under similar motion distributions, cross-person motion transfer achieves promising results, and incorporating SMPL-X shape parameters further improves the fidelity of the generated body shape to the target identity.

8) Ablation on Video Length: We analyzed the required duration for video data needed for fine-tuning. Compared to the one-minute videos used in the main text, we utilize five-minute videos in the fine-tuning stage. The quantitative results are demonstrated in Table III and the qualitative results are displayed in Fig. 14. The numerical results demonstrate a slight improvement in all measurements. For instance, with LMD (hand), additional data allows the model to encompass a broader range of angles, enabling more accurate generation outcomes, and showing better results. In qualitative results, one minute of fine-tuning data already yields satisfactory outcomes.

9) Discussion of Time Cost: The time cost of different methods is shown in Table IV. Our process is comparable to other diffusion methods. Most time is spent on diffusion. In Algorithm 1, $count$ is an array storing the computing counts for each frame, where only overlapped frames are calculated twice. When $ws = 16$ and $os = 4$ for 300 frames of 10s each to generate, the total computing time in $count$ is 400, which means an additional one-third of the cost. The additional time cost ensures consistency between batches, allowing for a longer duration. A comparison of time cost and ablation without batch-overlapped temporal denoising (TD) is listed in Table IV.

10) High-Resolution FE Model: We observed that our method sometimes generates inaccurate lip and expression movements, which could be traced back to the mouth's small size, which constrains the face enhancement's ability. We improve by training a higher-pixel model from 256 px to 512 px. As in Fig. 15 and Table V, the lip and expression movements become accurate, and the quality of the face region is enhanced.

VI. VIDEO RESULTS

We show video results in the attached materials. In the videos, we compare with SOTA methods as well as present the results of ablation studies. Video results demonstrate the effectiveness of the proposed method. For example, we could observe slight flickering frames in the video results without overlapped batches. This is due to the absence of information transfer between different batches and the inherent randomness of the diffusion model, resulting in subtle variations in human structure across batches. The proposed overlapped-batch design in the denoising process allows for certain context exchanges between different batches, thereby reducing the occurrence of this phenomenon.

VII. USER STUDY

We conduct user studies within our experimental setting. The videos generated in Section V-B were collected, and we invited 30 participants to participate in this user study. For each participant, 15 video instances are randomly sampled from all results, and the corresponding reference appearance and input pose sequence are shown simultaneously. For each instance, we asked participants to rate four aspects: appearance preservation, temporal consistency, structure preservation, and overall quality. Appearance preservation measures the appearance between the reference image and the generated video. Structure preservation is asked to evaluate the structure similarity between the input pose and the output video, especially for the hand structure. The rating score of each question is on a scale from one to five, with five being the highest score and one being the lowest.

The statistics are listed in Table VI. As the results show, our method achieves scores of over four points in the user study, which is the best among all methods. Pose2Img scored above three points in appearance preservation, and the results from Dreampose and DisCo are slightly inferior to Pose2Img.

(a) Setting I. (b) Setting II.

Fig. 9. Validation of two-stage training. The result of Setting I lack enough training data for pose-control ability. The appearance and identity may mix with other subjects for Setting II. Meanwhile, it is hard to extend to a new subject for such a strategy.

Fig. 10. Temporal differences between different adaptation strategies for video diffusion prior and Make-Your-Anchor. The first two columns are frames with small motion and the fourth and fifth columns are with large motion. The third column is the temporal differences between the first two columns, similar to the last column. Red boxes indicate the flicking artifacts in the results of Make-Your-Anchor and tuning temporal weights with small motions. Blue boxes show the stitching artifacts caused by directly applying temporal weights with large motion. Additionally, directly applying temporal weights results in apparent color aberration artifacts compared to ground-truth. The video results can be found in the attached files.

TABLE IV
COMPARISON OF TIME COST TO GENERATE 300 FRAMES

Method	Pose2Img	TPS	DreamPose	DisCo	Ours	Ours w/o TD	300 frames w/o TD
Time cost	77s	42s	310s	154s	405s	303s	OOM
Method type	non-diffusion	non-diffusion	diffusion	diffusion	diffusion	diffusion	diffusion
Resolution	640px	384px	512px	256px	512px	512px	512px

Fig. 11. Trained with SMPL perturbation.

Fig. 12. Sample human meshes of different identities during training.

TABLE V
NUMERICAL RESULTS OF FE WITH 512 PX

Method	FID↓	FVD↓	LMD (Face)↓
Ours	35.91	53.38	1.45
FE w 512px	35.13	54.02	1.33

TABLE VI
USER STUDY SCORES. THE RATING SCORE IS ON A SCALE FROM ONE TO FIVE, WHERE FIVE IS THE HIGHEST SCORE, AND ONE IS THE LOWEST.

Method	Appearance Preservation	Temporal Consistency	Structure Preservation	Overall Quality
Pose2Img [12]	3.36	2.38	1.59	2.01
TPS [8]	2.03	2.10	1.18	1.37
DreamPose [24]	2.78	1.94	2.27	1.94
DisCo [25]	2.29	2.32	1.45	1.72
MagicAnimate [26]	2.52	2.75	2.07	2.43
Champ [29]	2.97	3.06	3.45	3.28
Ours	**4.25**	**4.02**	**3.93**	**4.10**

Our approach has achieved a significant advantage in structural preservation compared to others, which is not apparent in Table I of the main text due to inaccurately estimated landmarks.

VIII. CONCLUSION

In this paper, we present Make-Your-Anchor+, a diffusion-based framework for generating realistic, high-quality 2D

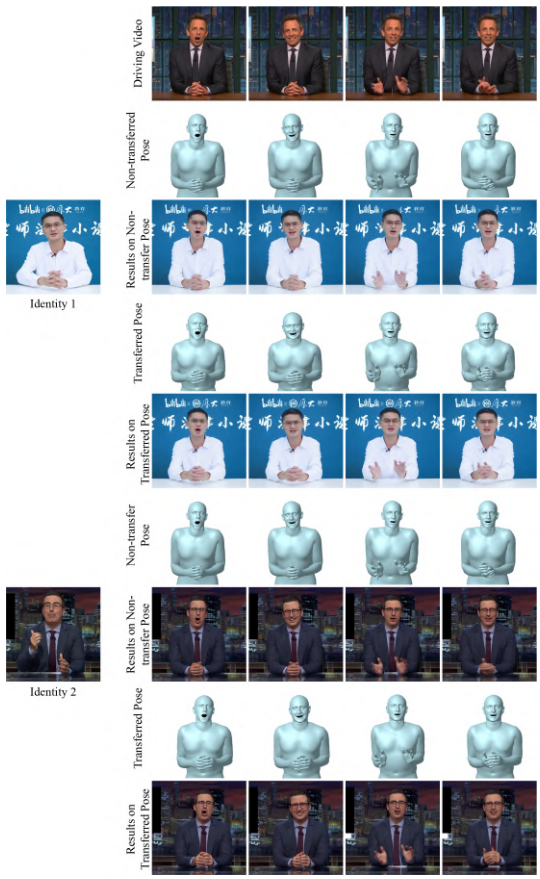


Fig. 13. Generated results with/without shape transfer under the same motion.

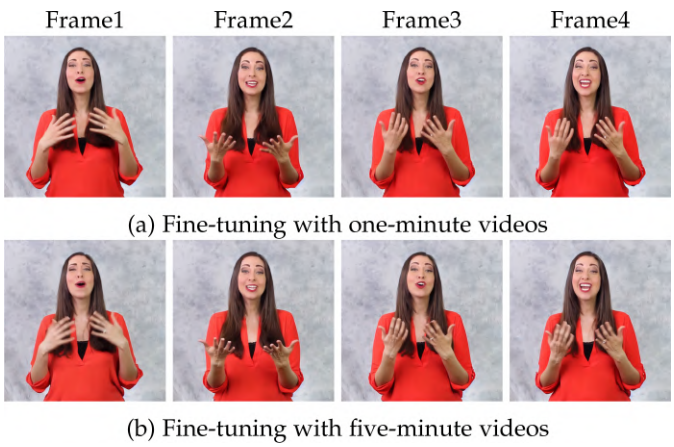


Fig. 14. Qualitative results with different video lengths for fine-tuning.

Fig. 15. Samples for the improved FE model with 512x512 pixels generate more accurate expressions. The Left is the FE model with 256x256 pixels, and the right is with 512x512 pixels.

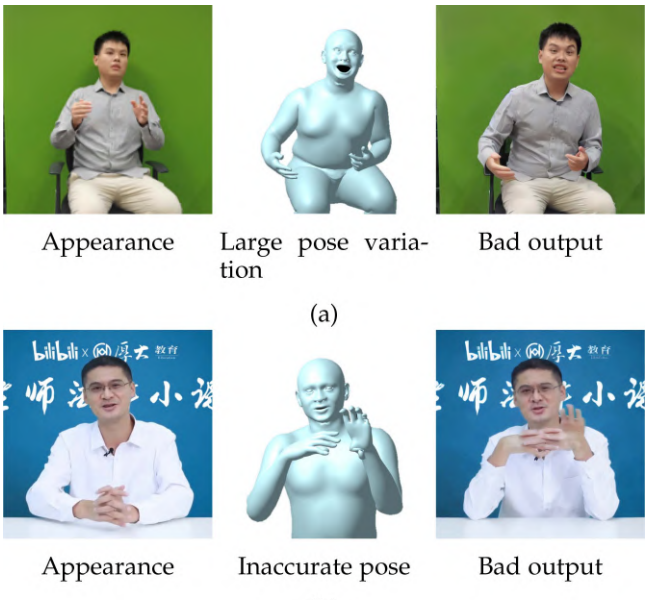


Fig. 16. Limitations. When the driven mesh is of large variation or inaccurate, the generated frame will be unsatisfactory.

anchor-style human videos. Our approach leverages frame-wise motion-to-appearance diffusion to train a structure-guided diffusion model via a two-stage training strategy, effectively mapping specific appearances to target motions. To ensure temporal consistency, we introduce an adaptation of the video diffusion prior and design a batch-overlapped temporal denoising algorithm, overcoming limitations on video length. We further propose an identity-specific face enhancement module to improve facial detail reconstruction, using an inpainting-style strategy to refine holistic human images. By integrating these components, our framework generates high-quality, structure-preserving, and temporally coherent anchor-style human videos, providing a valuable reference for the development of widely applicable 2D digital avatars.

Limitation and further work. Despite the capability of our method to produce high-quality videos, as shown in Fig. 16, if the human body orientation during inference differs significantly from that observed in the fine-tuning videos, there may be issues with preserving the appearance. This occurs due to the appearance-mapping style of training. From the other perspective, an increase in the quantity of pre-training and fine-tuning data may help generate more complex movements and orientations. Another limitation is that inaccurate input poses constraints on our generation results. Future work may address this issue by exploring more robust control strategies or adopting more accurate pose extraction methods. Finally, the current model requires one day for fine-tuning. Reducing the fine-tuning time remains an important direction for future work.

In future work, we also plan to integrate our framework into stronger video generation backbones, such as WAN [76] or next-generation models, further to enhance identity-specific video synthesis as more advanced training resources and pipelines become available.

References

[1] K. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar, "A lip sync expert is all you need for speech to lip generation in the wild," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 484–492.
[2] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 10039–10049.
[3] W. Zhang et al., "SadTalker: Learning realistic 3D motion coefficients for stylized audio-driven single image talking face animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8652–8661.
[4] Y. Pang et al., "DPE: Disentanglement of pose and expression for general video portrait editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 427–436.
[5] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 4401–4410.
[6] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2020, pp. 8110–8119.
[7] C. Wang, F. Tang, Y. Zhang, T. Wu, and W. Dong, "Towards harmonized regional style transfer and manipulation for facial images," *Comput. Vis. Media*, vol. 9, no. 2, pp. 351–366, 2023.
[8] J. Zhao and H. Zhang, "Thin-plate spline motion model for image animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3657–3666.
[9] A. Siarohin, O. J. Woodford, J. Ren, M. Chai, and S. Tulyakov, "Motion representations for articulated animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13653–13662.
[10] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, "First order motion model for image animation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 7137–7147.
[11] G. Yang, W. Liu, X. Liu, X. Gu, J. Cao, and J. Li, "Delving into the frequency: Temporally consistent human motion transfer in the fourier space," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 1156–1166.
[12] S. Qian, Z. Tu, Y. Zhi, W. Liu, and S. Gao, "Speech drives templates: Co-speech gesture synthesis with learned templates," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 11077–11086.
[13] Y. Zhou, J. Yang, D. Li, J. Saito, D. Aneja, and E. Kalogerakis, "Audio-driven neural gesture reenactment with video motion graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3418–3428.

[14] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10684–10695.

[16] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 3836–3847.

[17] C. Mou et al., "T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," in *Proc. AAAI Conf. Artif. Intell.*, 2024, pp. 4296–4304.

[18] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22500–22510.

[19] R. Gal et al., "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *Proc. 11th Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=NAQvF08TcyG

[20] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1931–1941.

[21] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," in *Proc. Int. Conf. Learn. Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[22] J. Ho et al., "Imagen Video: High definition video generation with diffusion models," 2022, *arXiv:2210.02303*.

[23] U. Singer et al., "Make-A-Video: Text-to-video generation without text-video data," in *Proc. 11th Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=nJfylDvgzlq

[24] J. Karras, A. Holynski, T.-C. Wang, and I. Kemelmacher-Shlizerman, "DreamPose: Fashion video synthesis with stable diffusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 22680–22690.

[25] T. Wang et al., "DisCo: Disentangled control for realistic human dance generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9326–9336.

[26] Z. Xu et al., "MagicAnimate: Temporally consistent human image animation using diffusion model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1481–1490.

[27] L. Hu, "Animate Anyone: Consistent and controllable image-to-video synthesis for character animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 8153–8163.

[28] H. Fang et al., "Dance your latents: Consistent dance generation through spatial-temporal subspace attention guided by motion flow," *IEEE Trans. Circuits Syst. Video Technol.*, 2025.

[29] S. Zhu et al., "Champ: Controllable and consistent human image animation with 3D parametric guidance," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 145–162.

[30] Z. Liu et al., "Fine-grained face swapping via regional gan inversion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 8578–8587.

[31] Z. Huang et al., "Identity-preserving face swapping via dual surrogate generative models," *ACM Trans. Graph.*, vol. 43, pp. 1–19, 2024.

[32] Z. Li, M. Cao, X. Wang, Z. Qi, M.-M. Cheng, and Y. Shan, "PhotoMaker: Customizing realistic human photos via stacked id embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 8640–8650.

[33] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[34] H. Yi et al., "Generating holistic 3D human motion from speech," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 469–480.

[35] Q. Zhou et al., "Let's all dance: Enhancing amateur dance motions," *Comput. Vis. Media*, vol. 9, no. 3, pp. 531–550, 2023.

[36] J. Chen et al., "A music-driven deep generative adversarial model for guzheng playing animation," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 2, pp. 1400–1414, Feb. 2023.

[37] A. Aristidou, A. Yiannakidis, K. Aberman, D. Cohen-Or, A. Shamir, and Y. Chrysanthou, "Rhythm is a dancer: Music-driven motion synthesis with global structure," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 8, pp. 3519–3534, Aug. 2023.

[38] Z. Yang et al., "Keyframe control of music-driven 3D dance generation," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 7, pp. 3474–3486, Jul. 2024.

[39] J. Li et al., "Audio2gestures: Generating diverse gestures from audio," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 8, pp. 4752–4766, Aug. 2024.

[40] X. Gao, Y. Yang, Z. Xie, S. Du, Z. Sun, and Y. Wu, "GUESS: Gradually enriching synthesis for text-driven human motion generation," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 12, pp. 7518–7530, Dec. 2024.

[41] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7291–7299.

[42] Z. Huang et al., "Make-your-anchor: A diffusion-based 2D avatar generation framework," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 6997–7006.

[43] K. Cheng et al., "VideoReTalking: Audio-based lip synchronization for talking head video editing in the wild," in *Proc. SIGGRAPH Asia Conf. Papers*, 2022, pp. 1–9.

[44] F. Yin et al., "StyleHEAT: One-shot high-resolution editable talking face generation via pre-trained StyleGAN," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 85–101.

[45] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. Guttag, "Synthesizing images of humans in unseen poses," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8340–8348.

[46] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody dance now," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5933–5942.

[47] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 5904–5913.

[48] A. Siarohin et al., "Unsupervised volumetric animation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 4658–4669.

[49] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, "Latent image animator: Learning to animate images via latent space navigation," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–18.

[50] S. Ginosar, A. Bar, G. Kohavi, C. Chan, A. Owens, and J. Malik, "Learning individual styles of conversational gesture," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3497–3506.

[51] M. Liao, S. Zhang, P. Wang, H. Zhu, X. Zuo, and R. Yang, "Speech2video synthesis with 3D skeleton regularization and expressive body poses," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 308–323.

[52] Y. Ma et al., "Follow your pose: Pose-guided text-to-video generation using pose-free videos," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 5, 2024, pp. 4117–4125.

[53] R. A. Güler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7297–7306.

[54] D. Chang et al., "MagicPose: Realistic human poses and facial expressions retargeting with identity-aware diffusion," in *Proc. 41st Int. Conf. Mach. Learn.*, 2024, pp. 6263–6285.

[55] A. Blattmann et al., "Align your latents: High-resolution video synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22563–22575.

[56] H. Chen et al., "VideoCrafter1: Open diffusion models for high-quality video generation," 2023.

[57] S. Ge et al., "Preserve your own correlation: A noise prior for video diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 22930–22941.

[58] J. Xing et al., "DynamiCrafter: Animating open-domain images with video diffusion priors," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2024, pp. 399–417.

[59] J. Xing et al., "Make-Your-Video: Customized video generation using textual and structural guidance," *IEEE Trans. Vis. Comput. Graph.*, vol. 31, no. 2, pp. 1526–1541, Feb. 2025.

[60] W. Chai, X. Guo, G. Wang, and Y. Lu, "StableVideo: Text-driven consistency-aware diffusion video editing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 23040–23050.

[61] C. QI et al., "FateZero: Fusing attentions for zero-shot text-based video editing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 15932–15942.

[62] S. Yang, Y. Zhou, Z. Liu, and C. C. Loy, "Rerender a video: Zero-shot text-guided video-to-video translation," in *Proc. ACM SIGGRAPH Asia Conf.*, 2023, pp. 1–11.

[63] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 7346–7356.

[64] J. Z. Wu et al., "Tune-A-Video: One-shot tuning of image diffusion models for text-to-video generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 7623–7633.

[65] Y. Guo et al., "AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning," in *Proc. 12th Int. Conf. Learn. Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=Fx2SbBgcte
[66] G. Pavlakos et al., "Expressive body capture: 3D hands, face, and body from a single image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10975–10985.
[67] L. Khachatryan et al., "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 15954–15964.
[68] Y. Zhang, Y. Wei, D. Jiang, X. Zhang, W. Zuo, and Q. Tian, "ControlVideo: Training-free controllable text-to-video generation," in *Proc. 12th Int. Conf. Learn. Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=5a79AqFr0c
[69] M. Bain, A. Nagrani, G. Varol, and A. Zisserman, "Frozen in time: A joint video and image encoder for end-to-end retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1728–1738.
[70] F.-Y. Wang, W. Chen, G. Song, H.-J. Ye, Y. Liu, and H. Li, "Gen-l-video: Multi-text to long video generation via temporal co-denoising," 2023, *arXiv:2305.18264*.
[71] H. Qiu et al., "FreeNoise: Tuning-free longer video diffusion via noise rescheduling," in *Proc. 12th Int. Conf. Learn. Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=ijoqFqSC7p
[72] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, "Robust high-resolution video matting with temporal guidance," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 238–247.
[73] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *Workshop Generative Models Downstream Appl.*, 2021. [Online]. Available: https://openreview.net/forum?id=qw8AKxfYbI
[74] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6629–6640.
[75] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," 2018, *arXiv:1812.01717*.
[76] T. Wan et al., "Wan: Open and advanced large-scale video generative models," 2025, *arXiv:2503.20314*.

Ziyao Huang received the BE degree from Beihang University, in 2017 and the MS degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2022. He is currently working toward the PhD degree with the Institute of Computing Technology, Chinese Academy of Sciences and University of Chinese Academy of Sciences, Beijing, China. His research interests include digital human and video generation.

Fan Tang received the BSc degree in computer science from North China Electric Power University, Beijing, China, in 2013, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2019. He is currently an associate professor with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include computer graphics, computer vision, and machine learning.

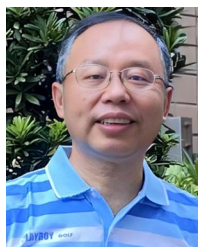
Juan Cao received the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2008. She is currently a professor with the Institute of Computing Technology, Chinese Academy of Sciences. She has more than 90 publications in international journals and conferences, including *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Image Processing*, KDD, and CVPR. Her research interests include multimedia content analysis and fake multimedia detection.

Yong Zhang received the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences, in 2018. From 2015 to 2017, he was a visiting scholar with Rensselaer Polytechnic Institute. He was with Tencent AI Lab. He is currently with Meituan. His research interests include computer vision and machine learning.

Xiaodong Cun (Member, IEEE) received the BSc degree in computer science from Xidian University and the MS and PhD degrees from the Department of Computer and Information Science, University of Macau, in 2018 and 2021, respectively. He was a senior researcher with Visual Computing Center, Tencent AI Lab, from 2021 to 2024. He is currently an assistant professor with Great Bay Universtiy. His research focuses on image/video generation, translation, and editing.

Yihang Bo received the doctor degree from Beijing Jiaotong University. She was a joint-training PhD with UC Irvine. She completed her postdoc research with the Institute of Automation, Chinese Academy of Sciences and Boston College. She is currently an associate professor with Fine Art Department, Beijing Film Academy. Her research interests include computer vision, computational film, and interactive art design.

Jintao Li (Member, IEEE) received the PhD degree from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 1989. He is currently a professor with the Institute of Computing Technology, Chinese Academy of Sciences. His research interests include multimedia analysis, virtual reality technology, and pervasive computing.



Tong-Yee Lee (Senior Member, IEEE) received the PhD degree in computer engineering from Washington State University, Pullman, in 1995. He is currently the chair professor with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan, where he leads Computer Graphics Group, Visual System Laboratory (http://graphics.csie.ncku.edu.tw). His research interests include computer graphics, nonphotorealistic rendering, medical visualization, virtual reality, and media resizing. He is a senior member of IEEE Computer Society and member of ACM.

He is also on the editorial boards of *IEEE Transactions on Visualization and Computer Graphics*.