

Make-Your-Anchor: A Diffusion-based 2D Avatar Generation Framework

Ziyao Huang¹, Fan Tang¹, Yong Zhang², Xiaodong Cun², Juan Cao^{1*}, Jintao Li¹, Tong-Yee Lee³

¹ Institute of Computing Technology, Chinese Academy of Sciences ² Tencent AI Lab

³ National Cheng-Kung University

{huangziyao19f, tangfan}@ict.ac.cn {zhangyong201303, vinthony}@gmail.com

{caojuan, jtli}@ict.ac.cn tonylee@mail.ncku.edu.tw

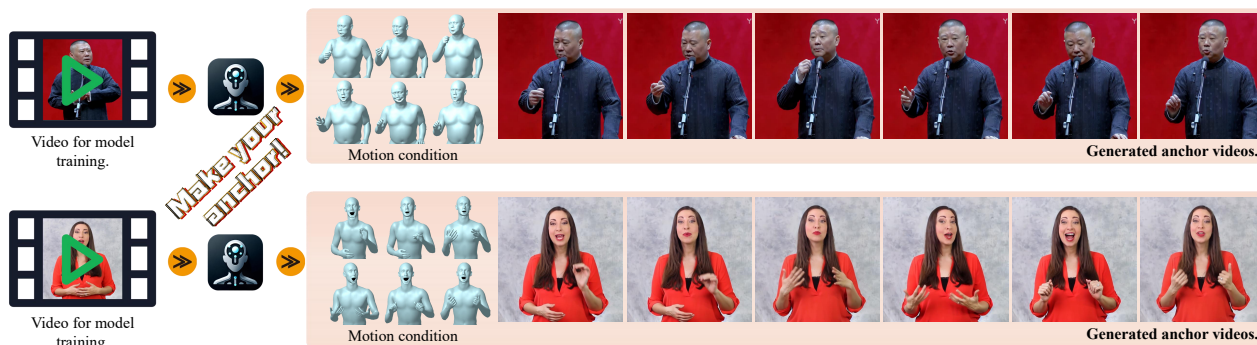


Figure 1. We propose Make-Your-Anchor, a diffusion-based 2D avatar generation framework, to map the animatable human 3D mesh sequence into realistic human video. By combing with motion capturing or audio-to-motion methods [20, 34], our framework could achieve anchor video auto-generation with temporal and lifelike outcomes.

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 binding movements with specific appearances. To produce arbitrary long temporal video, we extend the 2D U-Net in the frame-wise diffusion model to a 3D style without additional training 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 effectiveness and

superiority of the system in terms of visual quality, temporal coherence, and identity preservation, outperforming SOTA diffusion/non-diffusion methods.

1. Introduction

Automatically generating 2D human videos with vivid expressions, torsos, and gestures under various conditions such as audio, text, or motion has wide applications in e-commerce, online education/conferencing, VR, etc. While AI-powered digital anchors offer certain benefits, such as around-the-clock availability and the ability to generate new content quickly, they also raise significant quality concerns.

A common practice in popular 2D avatar systems is pronounced specifically as GAN-based talking face generation approaches [17, 19, 32, 36], where only the lip/facial region has been edited. This type of methods relies on the capability to modify faces based on GANs [11, 12, 29]. Despite the remarkable visual quality, these systems limit the anchor’s degrees of freedom. Beyond regional editing, researchers try to generate more natural digital avatars with the help of motion transfer [25, 26, 33, 38] and co-speech generation [20, 40]. By learning the mapping from

*Corresponding author: Juan Cao.

motion/speech to a person’s appearance, these methods could generate full-body talking videos while moving the torso and hands. However, the GAN-based solution limits the visual quality of the generated videos in case of fine details or fast movement.

Recently, diffusion models [7] have been proven to achieve promising generation quality on human image generation [23]. Compositing with pose or depth condition [16, 35], current text-to-image diffusion models, such as Stable Diffusion [23], can synthesize human body with a specific gesture, pose, and expression, and can keep identity by personalized fine-tuning methods [3, 9, 14, 24]. However, due to the randomness of the diffusion process, image diffusion models cannot be directly applied to generate temporal-consistent human videos.

Existing video diffusion models [8, 27], proven to generate temporal-consistent videos with simple motion and nature scenes. DreamPose [10] modifies Stable Diffusion to be controlled by input image and pose and animates a still human image conditioned on a given motion sequence. DisCo [31], which also adopts an image as input, applies pose and background ControlNet [35] to composite different motions and scenes with a pretraining strategy and achieves promising human dance generation results on generalization. However, these existing diffusion-based approaches still leave it challenging to synthesize facial details, exact gestures, and temporal frames, which is insufficient for making vivid digital anchors.

In this study, we propose “Make-Your-Anchor”, a diffusion-based 2D avatar generation system to create customized digital avatars. Despite the impressive visual quality of diffusion-based approaches, generating human images/videos has been challenging, in terms of identity preservation and motion consistency. To this end, we learn to link human appearance with parametric human representation. We replace the text clip inputs with image appearances, and a ControlNet-style modulation network is proposed to bind the human appearance to the pose control sequence frame-by-frame. After pertaining on the existing news anchor dataset [34], we finetune the diffusion model on a specific person with a one-minute video. To generate consistent temporal videos of any duration, we propose batch-overlapped temporal denoising by calculating the frame-wise latent noise in a cross-frame fashion during the inference phase. Furthermore, special attention was paid to the facial region by applying a novel inpainting-based enhancement operation.

With the frame-wise training and batch-wise inference, the proposed system can be performed within a practical time cost on a single 40G A100 GPU. Powered by existing technological solutions for generating corresponding 3D human mesh with lips, gestures, and body motions conditioned on music [39], audio [34] or motion capture [2], the

system could make AI-powered digital anchors with vivid expressions, torsos, and gestures. Compared with popular commercial software, which is accomplished by talking face generation, our system achieves natural human video generation with more flexibility. The main contributions are as follows:

- We propose “Make-Your-Anchor”, a 2D avatar customized system, to generate practical and applicable digital anchors with vivid lips, expressions, gestures, and body actions.
- We propose a frame-wise motion-to-appearance diffusion to bind movements into appearance with a two-stage training strategy, and batch-overlapped temporal denoising paradigm to generate temporal-consistent human video with long duration.
- Qualitative and quantitative experiments on ten anchors validate the effectiveness of the proposed system compared with SOTA GAN-based motion transfer and diffusion-based human video generation approaches.

2. System Overview

Setting. Despite the impressive progress in talking face [36] or fashion video generation [10], creating human video from a single or few images limits the ability to apply these methods in real-world applications. Talking head videos are restricted to face regions, while DreamPose could only output fashion videos with a limited movement range. In this study, we formulate 2D avatar generation as a learning paradigm from a video of one identity. The setting is similar to the traditional talking head methods or commercial services such as ZenVideo. However, these methods only learn from facial regions and use body motions of the input video as “templates” which are looped in the generated videos. 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. To this end, we tune the diffusion model to “bind” on the input video with appearance conditioned on motion conditions. To the best of our knowledge, the proposed system is the first to generate 2D human avatar videos that could be given full-body motions with high visual fidelity in terms of identity preservation and temporal consistency.

Input. Based on the above setting, our system requires an anchor-style source video \mathbb{S} of one person for training. The given video \mathbb{S} for training could contain lip, body, and gesture movements in anchor style, and the video length should be longer than one minute. Unlike DreamPose or ControlNet with OpenPose, we utilize human 3D mesh rendered from SMPL-X parameters [18] as motion conditions. The input pose condition’s smoothness and accuracy influence the output video’s temporal consistency, and the 3D human mesh has more structural information

about the motion to generate smooth video, especially in hand gestures. For inference, a sequence of pose $\mathbb{P} = \{p_1, p_2, \dots, p_n\}$ is given to guide the 2D avatar learned by diffusion models. Notably, our system can be effectively combined with existing human motion sequence generation methods such as audio-to-motion generation [34] or video-guided motion transfer [2]. In this case, the input could be a clip of audio or motion video reference. The final output is human motion video frames $\mathbb{X} = \{x_1, x_2, \dots, x_n\}$ with temporal appearance.

3. Methodology

Fig. 2 shows the network structure and the inference pipeline of the proposed system. In this section, we start by introducing the diffusion-based network architecture to accomplish image-level appearance and body structure control in Sec. 3.1. Then we move on to the design of batch-overlapped temporal denoising in Sec. 3.2, which is proposed to generate arbitrary long temporal consistent video without additional training efforts. Finally, we propose an enhancement-by-inpainting module to improve the visual quality of facial regions in Sec. 3.3.

3.1. Frame-wise Motion-to-Appearance Diffusing

Preliminaries. Given an initial noise $\epsilon \sim \mathcal{N}(0, 1)$ and a condition c , the training loss of vanilla frame-wise diffusion models can be written as:

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

where \hat{X}_θ is the diffusion model, s_i is the i_{th} frame in input source video \mathbb{S} , α_t , σ_t , w_t are hyperparameters, and t is the diffusion timestep. After training, the final output target frame $x_i = s_i^0$ is denoised from an initial noise s_i^T with T steps sampling. In this study, we follow latent diffusion [23] and use pretrained VAE encoder $\mathcal{E}(\cdot)$ to map the source frame s_i into latent code z_i and then conduct diffusing and denoising process in the latent space. Finally, a pretrained VAE Decoder $\mathcal{D}(\cdot)$ maps the latent noise into pixel space.

Structure-Guided Diffusion Model. We propose a structure-guided diffusion model (SGDM) to generate human videos under the control of 3D mesh conditions frame-by-frame, as shown in Fig. 3. Different from ControlNet [35], we embed the 3D mesh condition into the generation process to learn the corresponding mapping from pose \mathbb{P} to target video frames \mathbb{X} . An extra condition branch to the diffusion model’s U-Net is copied from U-Net’s encoder and a zero-convolution layer is added between the input condition image and encoder. The encoder’s layer-wise feature maps are added on the U-Net’s up layers:

$$F_{up} = F_{up} + W_c \times ControlN(p_i), \quad (2)$$

where F_{up} is the feature map of U-Net’s up layers, $ControlN(\cdot)$ is the additional module, p_i is the i_{th} 3D mesh in \mathbb{P} and W_c is a hyperparameter to weight condition.

However, adding a motion-control branch may still lead the diffusion model to lose control of the person’s identity. To solve this problem, we inject the appearance information of the same person by randomly choosing another source frame s_r from \mathbb{S} and then replace the text CLIP [22] with image CLIP feature c_{s_r} of s_r to the cross-attention module of the latent diffusion’s U-Net as well as the motion-control branch. Specifically, we utilize the CLIP feature map before the global pooling layer, which provides more fine-grained appearance information. After training, the generation of the human image can be abbreviated as:

$$z_i^0 = \hat{X}_\theta(ControlN(p_i), c_{s_r}, z_i^T), \quad (3)$$

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

Two-Stage Training Strategy. We adopt the pretrained text-to-image Stable Diffusion V1.5 [23] as a basic model and build the proposed SGDM. The training goal of our system is to bind the movement and appearance of the target identity by an “binding” style approach. Therefore, we propose a two-stage training strategy, pre-training on multiple identities to enhance the models’ ability for motion generation, and fine-tuning on a single identity to bind the movements to the appearance. A publicly available anchor video dataset [34] is used to train the new branch and tune the basic diffusion model using multiple identity data. Then, we fine-tune the whole network based on the given one-shot videos of an arbitrary identity.

To eliminate the interference of background on learning human information in the pre-training stage, we segment the foreground human body from the target images. During the fine-tuning stage on a single identity, we chose to fine-tune the model on the background and foreground together. This approach allows the model to quickly generalize to new individual identities while also remembering the background, thereby generating harmonious anchor videos.

3.2. Batch-overlapped Temporal Denoising

Adopting the two-stage train strategy to the proposed SGDM, our system could generate avatar videos frame-by-frame faithful to motion sequence. However, due to the inherent randomness in the diffusion model’s output, discontinuity still exists between frames. Previous studies [13, 37] have already demonstrated that extending self-attention in the image diffusion model to cross-frame attention enhances the consistency of generated videos. This study extends the 2D U-Net to a 3D U-Net to get a training-free video diffusion model \hat{X}_θ^V . Specifically, we propose an all-frame cross-attention module to replace the self-

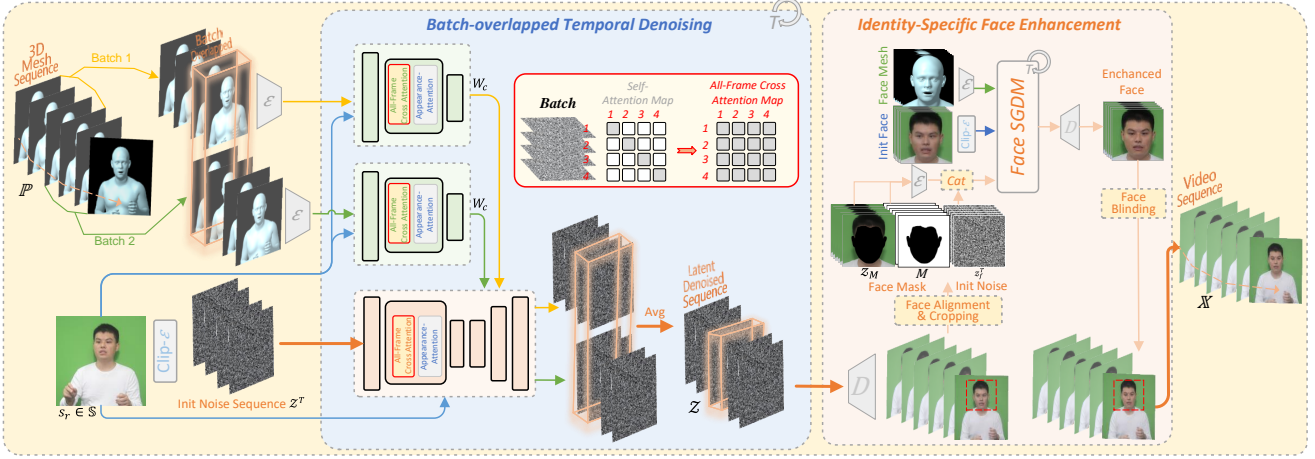


Figure 2. The inference pipeline of our system. An appearance condition and a 3D mesh sequence are inputted into the structure-guided diffusion, incorporating Batch-overlapped Temporal Denoising to accomplish video-level inference. Following the generation of arbitrary-length frame sequences, an inpainting-style module known as Identity-Specific Face Enhancement is utilized to enhance facial details.

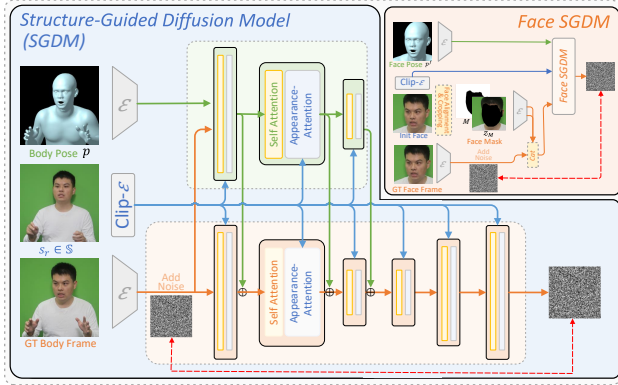


Figure 3. The network architecture of our proposed Structure-Guided Diffusion Model (SGDM) and Face SGDM. Our network achieves motion-to-appearance generation by embedding pose and appearance conditions into the pretrained diffusion model.

attention in a data batch \mathbb{B} , which can be written as:

$$Attention_{\mathbb{B}} := \text{Softmax}\left(\frac{Q_{\mathbb{B}}K_{\mathbb{B}}^T}{\sqrt{d}}\right) \cdot V_{\mathbb{B}}, \quad (5)$$

$$Q_{\mathbb{B}} = W_Q \cdot Z^t, \quad (6)$$

$$K_{\mathbb{B}} = W_K \cdot Z^t, \quad (7)$$

$$V_{\mathbb{B}} = W_V \cdot Z^t, \quad (8)$$

where $Z^t = \{z_1^t, z_2^t, \dots, z_N^t\}$ is the input latent frames on diffusion timestep t and W_Q, W_K, W_V are learnable weights for attention modules.

Furthermore, we propose an overlapped temporal denoising algorithm to generate an arbitrary length of anchor video as illustrated in Algorithm 1. Unlike existing methods that optimise the objective function between multiple batch

noise [30] or use window attention [21], we propose an effective and simple algorithm operated on multi-batch noise to achieve this goal. We partition a long motion sequence input into multiple overlapped windows. In each denoising step, the windows are fed one by one into our video diffusion model. Once all windows are handled, we normalize the overlapped noises between windows to ensure coherence, and then all video frames are denoised by the smoothing noises. After all denoising steps are executed, the final results present a long temporal video. We observe that this simple approach effectively extends the video diffusion model to generate videos of arbitrary length while keeping temporal consistency.

3.3. Identity-Specific Face Enhancement

Generating a satisfied face from a holistic human generation is challenging. From our perspective, this phenomenon arises due to the relatively diminutive size of the facial region within the entire image. With the training loss on holistic human image, the model encounters difficulty in fitting the small facial area. We believe that a feasible solution is adapting the model to the key information of the face. To address this issue, we opted for an inpainting-based approach with crop-and-blend operations to revising the facial region from the holistic generated body. In detail, we modify the proposed SGDM to Face SGDM, as shown in the right part of Fig. 3. The ready-to-revise face is cropped from the body generation, and the inpainting region is segmented via a facial mask M . We replace the input for SGDM’s U-Net with a concatenation of facial mask M , masked face latent z_M , and the original input latent z_f^T . The input appearance and motion condition are aligned with the corresponding reference and 3D mesh image. The

Algorithm 1: Overlapped Temporal Denoising

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

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

```
1  $count \leftarrow \{0, 0, \dots, 0\}$  ;
2  $Noise \leftarrow \{0, 0, \dots, 0\}$  ;
3 for  $i = 0, \dots, N/(ws - os)$  do
4    $wb \leftarrow i * (ws - os)$  ;
5    $we \leftarrow i * (ws - os) + ws$  ;
6    $Z_{wd}^t \leftarrow Z^t[wb : we]$  ;
7    $P_{wd} \leftarrow P[wb : we]$  ;
8    $N_{wd}^t \leftarrow \hat{X}_\theta^V(Z_{wd}^t, t, P_{wd})$  ;
9    $Noise[wb : we] \leftarrow Noise[wb : we] + N_{wd}^t$  ;
10   $count[wb : we] \leftarrow count[wb : we] + 1$  ;
11 end
12 # Normalize overlapped noises
13  $Noise \leftarrow Noise / count$  ;
14  $Z^{t-1} \leftarrow Scheduler(Noise, t, Z^t)$  ;
```

generation process can be abbreviated as:

$$z_f^0 = \hat{X}_\theta^f(\text{ControlN}(p^f), c_{s,r}^f, z_f^T, M, z_M), \quad (9)$$

where \hat{X}_θ^f is the face SGDM, z_f^0 is output, and z_f^T is input latent. The inpainting-style enhancement prevents the requirements of paired data and trains on only ground-truth.

4. Experiments

4.1. Experimental Settings

Dataset. For pretraining, we follow [4, 34] to utilize 27 hours of video with SMPL-X annotation on four identities. We apply robust video matting [15] on the pretraining videos to remove the background. For one-shot video fine-tuning, we collect a dataset with ten identities from diverse sources, each with a one to five-minute video. To validate the proposed method in various scenes, the ten identities are collected in three ways: four videos of the identities used for pertaining; three web videos of celebrities (Luo Xiang, Guo Degang, and one video from the YouTube channel ScienceOfPeople); three footages of invited individuals using a green screen. Videos are split into video clips, where each is 300 frames, 30 fps, and 10 seconds long, surpassing the processing capability of existing video diffusion models. We additionally collected some video clips for each identity to serve as their test motion samples.

<https://www.youtube.com/@ScienceOfPeople>

Comparison Methods. We compare our method with four state-of-the-art approaches: Pose2Img [20], TPS [38], DreamPose [10] and DisCo [31]. Pose2Img is part of the co-speech work SpeechDrivesTemplates [20], is a modified version of [1] and achieves person-specific human video generation conditioned on human landmarks, which is the most relevant GAN-based method to ours. TPS [38] is a generic motion-transfer method modeling motion via thin plate spline transformation. DreamPose [10] and DisCo [31] are diffusion-based human video generation systems. We train Pose2Img from scratch and fine-tune DreamPose and DisCo with our data while keeping the TPS’s generic weight.

Implementation Details. We manually crop the body region into 512×512 pixels, while face enhancement is 256×256. The pre-training stage with body generation lasted for 300K training steps with batch size 4, learning rate 1e-5, which took about seven days. Fine-tuning on one identity takes about one day, with 50K training steps for body generation and 80K steps for face enhancement. All experiments are operated on one Nvidia 40G A100 GPU. For inference, CFG [6] is set as 7.5. We found higher ControlNet condition scale W_c benefits preserving the structure of 3D mesh and hand gestures, and we set W_c as 2. SMPL-X annotations are produced by the tools from [34]. For face enhancement, FFHQ [11] preprocess method is utilized to align and crop the face images, and the inpainting mask is extracted by a facial segment model. For overlapped temporal denoising, the window size ws is set to 16, and the overlap size os is four. The number of denoising inference steps is set to 20.

Metrics. We use FID [5] to measure image quality, FVD [28] to measure temporal consistency. Landmark Mean Distances (LMD) are performed to calculate the preservation of body structure between the video frame and input 3D mesh. We compute the LMD on the face, body, and hands using OpenPose [2]. For the inaccuracy of landmarks prediction on hands, the outliers are ignored during computing LMD.

4.2. Main Results

We utilize the data mentioned above as our experiment dataset, where 10 identities from different sources are split into 9:1 as train and test data. Our experiment is set with 30 video clips as a test, as each identity takes 3 test data. Audio-driven results are further provided in supplementary materials, where we generate motions by TalkSHOW [34].

Quantitative Results. Quantitative results are presented in Table 1. Our approach achieves the best results on



Figure 4. 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.

FID and FVD, representing the image quality and video temporal consistency. Ours get the highest results on face and hand, while comparable to other methods. Pose2Img It is worth noting that hand landmarks are not accurate for complex gestures, which is not sufficient to reflect the performance. Besides, the landmarks used to compute LMD are also used for Pose2Img and DisCo inference, which causes their results to be somewhat biased. We conducted a user study in supplementary materials to provide complementary quantitative results.

Qualitative Results. The qualitative results of our experiment are shown in Fig. 4. In this figures, the blue boxes indicate distorted hand structures, and arrows label appearance artifacts. Our methods accomplish the best image quality and appearance preservation, and simultaneously faithfully reflect mouth movements and hand gestures. DisCo achieves high-quality generation through diffusion methods, yet it lacks focus on generating intricate details for hands and faces, resulting in noticeable imperfections

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
Ours	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

Table 1. Quantitative results of our method compared with SOTAs and ablation studies. Our method achieves better performance on image quality, temporal consistency, and structure preservation.

within our specific context. DreamPose focuses on generating videos within simple scenes, such as fashion-related contexts. In our scenario, it faces notable challenges in generating backgrounds and handling complex actions. TPS is a general warping-based method where, in unfamiliar domains, its generated poses are challenging to surpass beyond the given reference image’s pose. As for a person-specific method, Pose2Img achieves better results compared

to another comparison, while the artifacts are apparent in hands and lips because of the limitation of warping-based methods. Our approach learns the mapping from motion to appearance via a diffusion model, complemented by a temporal scheme, enabling the generation of sufficiently high-quality human videos.

More results. We further provide cross-person motion and full-body results. The cross-person motion-driven results are shown in Fig. 5. When the driven motion is in a similar style as the target person (a standing man to a standing man), the generated results are also in good visual quality. Full-body results are provided in Fig. 6. Although the full-body talking style differs from our training set, the results are still of good quality.



Figure 5. Cross-person motion results. For each image, left is pose, and right is output. Click the last images to play the embedded clips with Acrobat Reader.



Figure 6. Full-body results. Click the last images to play the embedded clips with Acrobat Reader.

4.3. Ablation Study

Validation of Batch-overlapped Temporal Denoising. To verify the effectiveness of Batch-overlapped Temporal Denoising (TD), we conduct analyses from multiple aspects. The numerical results of ablation are displayed in Table 1, where without TD, our method’s performance decreases significantly. We still get better results on temporal measures compared with DisCo because of the detailed pose condition and performs slightly worse compared to warping-based methods that leverage inter-frame advantages. The LMDs are not influenced much, as the generated structure is more influenced by SGDM.

Besides, we analyse the hyperparameters of TD, the window size wd and overlapped size os . The default setting is $ws = 16, os = 4$. we vary them as $ws = 8, os = 2$ and $ws = 32, os = 8$. The numerical results are displayed in Table 2, where temporal-related measures, FVD improve

with an increase in settings. We select $ws = 16, os = 4$ as the inference cost of the all-frame cross-attention module is increased as $O(ws^2)$.

Another analysis is the qualitative results, as shown in Fig. 7. We observed that without TD, the generated videos may exhibit temporal artifacts such as flickering and ghosting. This could be because the image model during training hasn’t accurately captured the distribution of human bodies within the dataset, leading to some unforeseen patterns. Introducing contextual information by Batch-Overlapped Temporal Denoising could better eliminate these unintended artifacts.

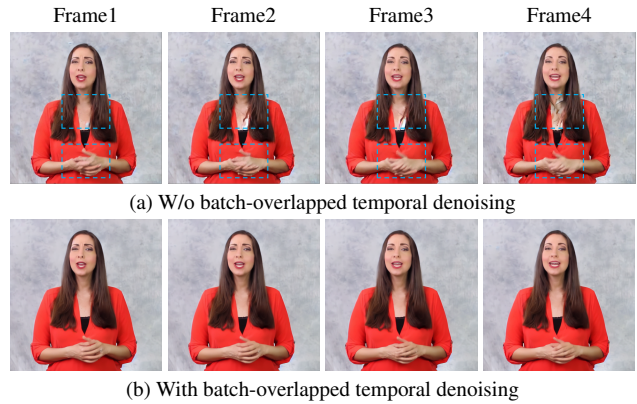


Figure 7. Ablation analysis of temporal denoising. The **first row** is the result without batch-overlapped temporal denoising, whereas the **second row** employed it. Without temporal denoising, the generated videos may exhibit discontinuities such as flickering and ghosting as the artifacts around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

Validation of Identity-Specific Face Enhancement. We take an analysis of the Identity-Specific Face Enhancement (FE) module. The quantitative results of the ablation study on this module are shown in Table 1. LMD (Face) is improved by our face enhancement results, which reveals the capability of FE to generate facial detail. Besides, FVD result is slightly lower down due to that we utilize FE without temporal denoising. The qualitative results of FE are shown in Fig. 8. We can distinctly observe that similar to other diffusion-based methods explicitly shown in Fig. 4, our approach sometimes struggles to directly generate high-quality facial features. FE can significantly improve facial details.

Validation of Two-Stage Training. Without two-stage training, there are two alternative approaches. Setting I: directly training our model with videos of the target subject; Setting II: training target with all other subject data. As shown in Fig. 9 (a) and Table 1, the result of Setting I lacks enough training data for pose-control ability. As shown in

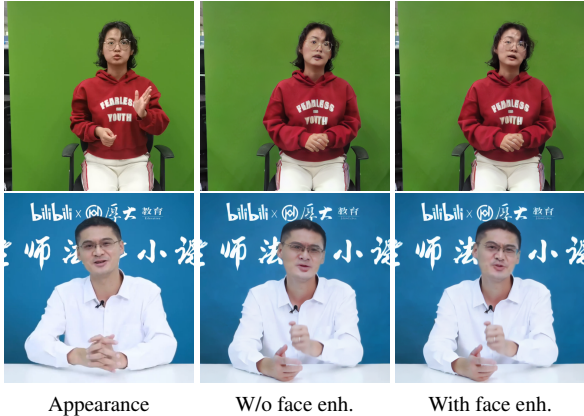


Figure 8. Ablations of identity-specific face enhancement. Directly training models to generate entire human body images often leads to the creation of facial features with poor realism. Our identity-specific face enhancement significantly improves the quality of generated content in the face region.

Method	FID↓	FVD↓	LMD (Face)↓	LMD (Body)↓	LMD (Hand)↓
$ws = 8, os = 2$	39.64	140.29	1.45	4.86	5.46
$ws = 16, os = 4$	40.33	139.82	1.44	4.88	5.41
$ws = 32, os = 8$	41.01	138.48	1.44	4.89	5.33

Table 2. Analysis of varying the hyperparameters of batch-overlapped temporal denoising, window size ws and overlapped size os . The default setting of our method is $ws = 16, os = 4$.

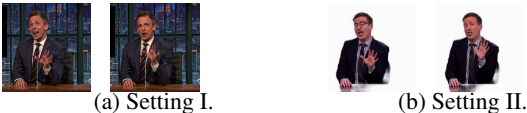


Figure 9. Validation of two-stage training. Click the last images to play the embedded clips with Acrobat Reader.

Fig. 9 (b) and Table 1, 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. With two-stage training, our model can control the pose and preserve the appearance of the target subject.

Validation of SMPL-X Parameters. Estimating SMPL inevitably comes with errors. We preprocess the extracted SMPL following TalkShow [34], optimizing SMPL through multiple estimations to reduce errors. We further explore imperfect SMPL by perturbing the SMPL parameters during training. Table 1 and Fig. 10 show that our method can still obey the input pose. The better statistics indicate that SMPL perturbing, serving as a kind of data augmentation, could further improve the performance.

5. Conclusion

In this paper, we propose “Make-Your-Anchor”, a diffusion-based 2D avatar generation framework to produce



Figure 10. Trained with SMPL perturbation. Click the last images to play the embedded clips with Acrobat Reader.

realistic and high-quality anchor-style human videos. In this framework, frame-wise motion-to-appearance diffusing is proposed to train a structure-guided diffusion model with a two-stage training strategy and accomplish binding specific appearance with movements via a binding style approach. To generate temporal consistent human video, we propose a training-free strategy to extend the image diffusion model into a video diffusion model with an all-frame cross-attention module and design a batch-overlapped temporal denoising algorithm to overcome the limitation on the generated video length. Identity-specific face enhancement is introduced from the observation that face details are difficult to reconstruct during holistic human generation. An inpainting-style enhancement module is applied to the human image to solve this problem. Through the amalgamation of our entire system approach, our framework successfully generates high-quality, structure-preserving, and temporal coherence anchor-style human videos, which may offer some reference value for the widely applicable technology of 2D digital avatars.

Limitation and further work. Despite the capability of our method to produce high-quality videos, if the human body orientation for inference significantly differs from what has been observed in the fine-tuning videos, there may be issues with preserving the appearance. This occurs due to the binding style training. From the other perspective, an increase in the quantity of pre-training and fine-tuning data may handle the generation of more complex movements and orientations. Another limitation is that our method does not model foregrounds with occlusions, potentially resulting in the ghosting of occluded elements. One possible solution could be segmenting occluded elements as a reference to enable the model to preserve the occlusions.

Acknowledgements

This work was partly supported by the National Natural Science Foundation of China under No. 62102162, Beijing Science and Technology Plan Project under No. Z231100005923033, and the National Science and Technology Council under No. 111-2221-E-006-112-MY3, Taiwan.

References

- [1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 8340–8348, 2018. 5
- [2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 7291–7299, 2017. 2, 3, 5
- [3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 2
- [4] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3497–3506, 2019. 5
- [5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems (NeurIPS)*, 30, 2017. 5
- [6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5
- [7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 33:6840–6851, 2020. 2
- [8] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2
- [9] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2
- [10] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion video synthesis with stable diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22680–22690, 2023. 2, 5
- [11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, Long Beach, CA, USA, 2019. IEEE. 1, 5
- [12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, Seattle, WA, USA, 2020. IEEE. 1
- [13] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint arXiv:2303.13439*, 2023. 3
- [14] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1941, 2023. 2
- [15] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 238–247, 2022. 5
- [16] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2
- [17] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xi-aodong Cun, Ying Shan, and Dong-ming Yan. Dpe: Disentanglement of pose and expression for general video portrait editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 427–436, 2023. 1
- [18] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10975–10985, 2019. 2
- [19] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia (ACM MM)*, pages 484–492, 2020. 1
- [20] Shenhan Qian, Zhi Tu, Yihao Zhi, Wen Liu, and Shenghua Gao. Speech drives templates: Co-speech gesture synthesis with learned templates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11077–11086, 2021. 1, 5
- [21] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023. 4
- [22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763. PMLR, 2021. 3
- [23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10684–10695, 2022. 2, 3

- [24] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023. [2](#)
- [25] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems (NeurIPS)*, 32, 2019. [1](#)
- [26] Aliaksandr Siarohin, Oliver J Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13653–13662, 2021. [1](#)
- [27] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. [2](#)
- [28] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. [5](#)
- [29] Cong Wang, Fan Tang, Yong Zhang, Tieru Wu, and Weiming Dong. Towards harmonized regional style transfer and manipulation for facial images. *Computational Visual Media (CVM)*, 9(2):351–366, 2023. [1](#)
- [30] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023. [4](#)
- [31] Tan Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for referring human dance generation in real world. *arXiv preprint arXiv:2307.00040*, 2023. [2](#), [5](#)
- [32] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10039–10049, 2021. [1](#)
- [33] Guang Yang, Wu Liu, Xinchun Liu, Xiaoyan Gu, Juan Cao, and Jintao Li. Delving into the frequency: Temporally consistent human motion transfer in the fourier space. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 1156–1166, 2022. [1](#)
- [34] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J Black. Generating holistic 3d human motion from speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–480, 2023. [1](#), [2](#), [3](#), [5](#), [8](#)
- [35] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. [2](#), [3](#)
- [36] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8652–8661, 2023. [1](#), [2](#)
- [37] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023. [3](#)
- [38] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3657–3666, 2022. [1](#), [5](#)
- [39] Qiu Zhou, Manyi Li, Qiong Zeng, Andreas Aristidou, Xiaojing Zhang, Lin Chen, and Changhe Tu. Let’s all dance: Enhancing amateur dance motions. *Computational Visual Media (CVM)*, 9(3):531–550, 2023. [2](#)
- [40] Yang Zhou, Jimei Yang, Dingzeyu Li, Jun Saito, Deepali Aneja, and Evangelos Kalogerakis. Audio-driven neural gesture reenactment with video motion graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3418–3428, 2022. [1](#)