ShowMaker: Creating High-Fidelity 2D Human Video via Fine-Grained Diffusion Modeling

Quanwei Yang [1†] **Jiazhi Guan** [2] **Kaisiyuan Wang** [3*] **Lingyun Yu** [1]
Wenqing Chu [3] **Hang Zhou** [3] **Zhiqiang Feng** [3] **Haocheng Feng** [3]
Errui Ding [3] **Jingdong Wang** [3] **Hongtao Xie** [1*]

[1] University of Science and Technology of China [2] Tsinghua University
[3] Department of Computer Vision Technology (VIS), Baidu Inc.

Abstract

Although significant progress has been made in human video generation, most previous studies focus on either human facial animation or full-body animation, which cannot be directly applied to produce realistic conversational human videos with frequent hand gestures and various facial movements simultaneously. To address these limitations, we propose a 2D human video generation framework, named *ShowMaker*, capable of generating high-fidelity half-body conversational videos based on 2D key points via fine-grained diffusion modeling. We leverage dual-stream diffusion models as the backbone of our framework and carefully design two novel components for crucial local regions (i.e., hands and face) that can be easily integrated into our backbone. Specifically, to handle the challenging hand generation caused by sparse motion guidance, we propose a novel Key Point-based Fine-grained Hand Modeling module by amplifying positional information from raw hand key points and constructing a corresponding key point-based codebook. Moreover, to restore richer facial details in generated results, we introduce a Face Recapture module, which extracts facial texture features and global identity features from the aligned human face and integrates them into the diffusion process for face enhancement. Extensive quantitative and qualitative experiments demonstrate the superior visual quality and temporal consistency of our method.[1]

1 Introduction

The recent advancement of generative models [5, 21, 11, 36, 33] has significantly propelled digital human technology, which is widely applied in business, education, and multimedia entertainment. Empowered by these generative models, numerous works [30, 40, 3, 52, 14, 32, 4, 24, 15, 48, 6, 16, 18, 34, 23, 35, 54, 49] have given their priority to 2D video synthesis to human generation. Though most of them focus on the head region, few studies [16, 18, 22, 34, 23, 35, 54, 37, 45, 25] attempt to generate full-body videos by animating a reference image with a sequence of driving motion signals via a warping paradigm, but their results fall short in terms of both generation quality and temporal consistency. However, the demand for creating high-fidelity 2D avatars under more challenging conversational scenarios (e.g., TV shows and stand-up comedy) is increasing rapidly, which cannot be fulfilled with such a fidelity state.

Recently, efforts have been made to investigate architectures built upon pre-trained diffusion models for controllable human body animation [1, 13, 47]. These methods adopt a dual-stream design to

† Work done during an internship at Baidu Inc.
* Correspondence to: htxie@ustc.edu.cn; wangkaisiyuan@baidu.com
[1] Project webpage: https://mumuwei.github.io/ShowMaker

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-1615

separately model the textural information from the reference image and the motion information from dynamic 2D skeletons [50] or 3D reconstructions [28, 55]. Effective interaction is also achieved between the two streams via widely used cross-attention [39]. Despite their captivating performance in full-body animation, we identify three major challenges that require rethinking. a) Creating a human-like conversational avatar requires a holistic synthesis of the human body. Though concurrent studies [17, 56] leverage 3D body reconstruction, which contains rich depth and shape information in face and hand regions, it suffers from occasional incorrect pose estimation for hands or temporal jittering problems. In addition, tedious pre-processing and time-consuming per-frame optimization [51] are required, which cannot be applied in a user-friendly system. We find using 2D representations more stable and efficient. b) It is quite intricate to generate human hands with sparse representations. Human hand regions occupy only a limited number of pixels in the original video frame. This easily leads to unstable hand synthesis with blurry texture or incorrect shape. Naive designs in [1, 13] cannot achieve detailed texture synthesis and delicate control in such local regions (e.g., faces and hands). c) Facial identity preservation is another problem that has not been well investigated. It becomes particularly challenging for these previous works [1, 13] due to the strong entanglement between identity information and 2D driving signals, which inevitably leads to identity degradation during cross-identity animation. The recent study [17] attempts to involve more controllable signals (i.e., 3D morphing parameters) for face enhancement, but the identity preservation performance is still not satisfying enough.

To tackle the above challenges, in this paper, we propose a holistic human video generation framework named *ShowMaker*, capable of generating an expressive conversational human video with fine-grained modeling conditioned solely on 2D key points. We adopt a similar dual-stream architecture as in [1, 13, 47, 56] as the backbone of our framework and introduce two new designs. We first introduce a novel *Key Point-based Fine-grained Hand Modeling* module to reliably recover hand regions in detail from sparse driving guidance. Our key insight is to leverage resolution-independent representations (i.e., the coordinates of the hand key points) to provide stable structure guidance. In terms of hand texture, we novelly design a key point-based hand texture codebook equipped with a set of learnable basis vectors, which has a one-to-one correspondence with the hand key point topology. Specifically, we predict a set of weights from the hand key points of each hand and produce key point-based hand textural compensation through the weighted combination, which can be directly injected into the diffusion process via cross-attention. Furthermore, to improve the facial region quality of the target subject, we propose a *Face Recapture* module to construct comprehensive face representations by performing structure encoding, texture encoding, and identity encoding simultaneously. This multi-level encoding strategy can significantly alleviate the identity leakage issue during cross-identity animation on 2D avatars, which is not available in other comparative methods. Extensive quantitative and qualitative experiments demonstrate that our proposed method can synthesize 2D human video with better visual quality and more accurate body movements, especially hand gestures.

The main contributions of this paper are summarized as follows: (1) We propose a novel holistic human video generation framework with fine-grained modeling, named *ShowMaker* for creating 2D human conversational videos conditioning on 2D key points. (2) We propose a *Key Points-based Fine-grained Hand Modeling* module, which achieves robust hand synthesis via a key point-based codebook. (3) We propose a *Face Recapture* module, which can effectively recover richer facial details and recapture the identity of the target subject.

2 Related Work

Talking Head Generation. The goal of talking head synthesis is to generate realistic face videos based on a driving video or audio. Depending on the driving signal, talking head synthesis can be divided into video-driven methods and audio-driven methods. Video-driven methods [43, 53, 3, 32, 4, 48, 42] aim to establish the mapping between the reference face and the driving head motion, such as head/mouth movements, expressions, etc. For instance, DaGAN [12] learns to face depth maps and 3D facial geometry in a self-supervised manner to estimate motion fields. While MetaPortrait [53] estimates the warping flow between the driving and reference face through predefined dense facial key points. Audio-driven methods[40, 52, 41, 7, 24, 15, 6, 46] focus on ensuring the generated mouth movements are synchronized with the given audio. For example, Wav2lip [30] edits the mouth area of the face video based on the input audio signal and employs a discriminator to ensure lip-sync. Similarly, StyleSync [6] generates lip-sync face videos by combining audio information with masked

facial spatial information in style space. Although these methods can produce high-quality face videos, they lack comprehensive pose movements and gestures, limiting their application scenarios.

Human Body Animation The human video synthesis task aims to generate a full-body human video with the reference person appearance and driving pose. According to the pipeline, these methods can be divided into two categories: implicit methods and explicit warp-based methods. For the first category, some early GAN-based algorithms [16, 18, 37] map or manipulate the reference appearance to the driving pose in the latent space, subsequently generating the target video. In recent years, significant progress has been made in human video generation by leveraging powerful diffusion models. Some approaches [44, 19, 47]integrate the reference appearance details with driving information through cross-attention mechanisms in the Denoising U-Net, resulting in enhanced generation quality. The warp-based methods [34, 23, 35, 54, 25, 45] warp the reference image or its features to the driving pose by various estimated flows, such as 2D optical flows, and 3D flow fields. In addition to the aforementioned methods, some efforts [29, 27] employ neural representations of 3D human mesh points in a canonical pose to model human appearance details, thereby achieving multi-view and temporal consistency in human videos. However, these methods primarily focus on appearance consistency and full-body pose accuracy, overlooking the importance of facial details and hand movements for video expressiveness and specific scenario suitability.

3 Method

The overview of our proposed framework ShowMaker is shown in Fig. 1, which can achieve high-quality video-driven conversational avatar synthesis. In the following, we first define our task in Sec. 3.1 and provide an overview of our framework pipeline in Sec. 3.2. Then detailed explanations of our novelly designed modules are demonstrated in Sec. 3.3 and Sec. 3.4. Finally, the training strategies are introduced in Sec. 3.5.

3.1 Task Formulation

For a conversational video V, we leverage the pre-trained DWPose [50] to extract human body key points of all frames, including face, body, and hands, and then we paint them into 2D heatmaps P as driving pose. Given a reference image $I_r \in \mathbb{R}^{H\times W\times 3}$ from a target person, and the driving poses $P_d = \{P_1, P_2, \ldots, P_F\} \in \mathbb{R}^{F\times H\times W\times 3}$, our task aims at generating target video sequence $\hat{V} = \{\hat{I}_1, \hat{I}_2, \ldots, \hat{I}_F\} \in \mathbb{R}^{F\times H\times W\times 3}$ with similar facial and body movements as in the driving poses P_d. The entire generative process can be formulated as $\hat{V} = \mathcal{G}\left(I_r, P_d\right)$, where $\mathcal{G}$ represents the generative model.

3.2 Pipeline Overview.

Fig. 1 provides an overview of our approach. We follow the recent works [1, 13, 47, 56] to adopt a dual-stream learning paradigm. The upstream branch (Green Area) processes the reference image, extracting the appearance information using a VAE encoder and Reference U-Net. In the downstream branch (Blue Area), the Pose Encoder takes the 2D pose heatmaps as input to extract human pose information, including face and hand. The subsequent Denoising U-Net effectively integrates the extracted appearance and driving information to predict noise intensity. Additionally, we introduce a novel Key Points-based Fine-grained Hand Modeling module to capture robust hand structure and texture features, and the comprehensive face representations are constructed by the Face Recapture module. The Denoising U-Net is derived from the pre-trained SD [33], replacing self-attention with reference-attention and CLIP [31] cross-attention with face attention, where hand attention integrates hand features from the Key Points-based Fine-grained Hand Modeling module, and face attention incorporates face features from the Face Recapture module. Additionally, temporal attention layers are introduced to promote the temporal consistency of the generated images.

Reference U-Net. Numerous studies have demonstrated that the U-Net network outperforms the CLIP image encoder in terms of appearance extraction. Firstly, the image resolution input to CLIP is limited to 224×224, which is insufficient for capturing detailed appearance information. Secondly, the features extracted by the CLIP image encoder are primarily oriented towards high-level semantic matching. We thus follow recent works [1, 13, 47, 56] to adopt a duplicate U-Net as our appearance extraction network named Reference U-Net, initializing its weights from the pre-trained SD.

https://doi.org/10.52202/079017-1615

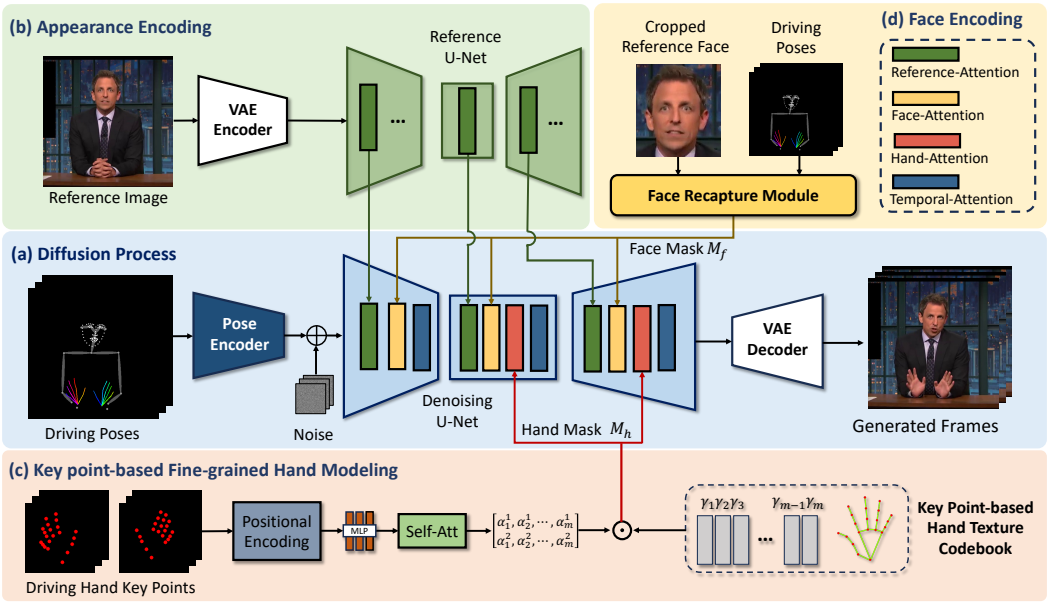

Figure 1: **Overview of our proposed framework ShowMaker.** Our framework adopts a dual-stream design including a Reference U-Net and a Denoising U-Net, where the former takes a reference image as input for appearance encoding and the latter takes noise latent and driving poses as input for diffusion processing. We further equip the backbone with a Key Point-based Fine-grained Hand Modeling module and a Face Recapture module for fine-grained avatar synthesis.

Specifically, as shown in Fig. 1, the reference image is first encoded into the latent space via the VAE Encoder and then passed to the Reference U-Net. The feature map $\mathbf{z}_r$ from each layer in Reference U-Net is utilized in the reference attention mechanism for appearance detail fusion. Initially, $\mathbf{z}_r$ is repeated F times along the temporal dimension to form $\mathbf{z}_r \in \mathbb{R}^{B\times F\times h\times w\times c}$ to match the size of the denoising feature map. Here, B indicates the batch size.

Pose Encoder. To reduce computational complexity, we employ a lightweight network as the pose encoder which consists of 8 convolutional layers initialized with Gaussian weights, with the final layer utilizing zero convolution. The pose information extracted by the pose encoder is then added to the noise latent and fed into the Denoising U-Net.

3.3 Key point-based Fine-grained Hand Modeling

Although the driving pose provides important guidance for hand synthesis, the hand regions only occupy a limited number of pixels. The convolution operation in the pose encoder and Denoising U-Net repeatedly downsamples the spatial size, which progressively weakens this guidance and results in unexpected structural and textural artifacts. Considering hand key points are clearly defined and cannot be constrained by the resolution of the hand region, we make our attempt to enhance this guidance by involving absolute coordinates of hand key points as additional inputs. Specifically, given the m key points of one hand $K = [k_1; k_2; \ldots; k_m]$, we first use Fourier positional encoding [26] to map their coordinates into a high-dimensional space, which can capture subtle differences in hand gestures. Formally, the frequency function $\mathcal{F}$ in positional encoding is defined as:

$$\mathcal{F}(k) = [(\sin\left(2^0\pi k\right), \cos\left(2^0\pi k\right), \cdots, \sin\left(2^{L-1}\pi k\right), \cos\left(2^{L-1}\pi k\right)), k]. \quad (1)$$

Here $k \in \mathbb{R}^2$ is the coordinate value of a single hand key point, which has been normalized to $[0, 1]$. Notably, we retain the original coordinate value in Eq. 1 and we set $L = 20$ by default. By using the function $\mathcal{F}$, the coordinates of hand key points $K_h \in \mathbb{R}^{B\times F\times 2\times 2m}$ can be mapped to high dimensional space $\mathbf{F}_h \in \mathbb{R}^{B\times F\times 2\times 2m*(2L+1)}$, which enables reliable enhancement on the guidance from the raw driving poses P_d. Here, the first 2 denotes two hands (both left and right hand).

Based on the enhanced hand structure guidance, we additionally design a key point-based hand texture codebook that achieves fine-grained hand texture synthesis. Particularly, as illustrated in Fig. 1 (c), the key point-based hand texture codebook consists of a set of learnable basis vectors $\mathbf{C}_{hand} = \{\gamma_i\}_{i=1}^{m}, \gamma_i \in \mathbb{R}^d$, which is built in the same topology as the m hand key points mentioned above. It is worth noting that though the left hand and right hand are symmetric in geometry, they share the same topology (i.e., the order of key points is the same). Moreover, these basis vectors are constrained to be orthogonal so that each basis vector represents a distinct hand texture pattern and only requires an emphasis on the texture modeling around its corresponding key point. Specifically, any two basis vectors γ_i, γ_j satisfy the following conditions:

$$< \gamma_i, \gamma_j >= \begin{cases} 0 & i \neq j, \\ 1 & i = j. \end{cases} \tag{2}$$

Subsequently, a self-attention layer followed by a linear projection layer is employed to predict the weights of hand texture patterns $\mathbf{A} \in \mathbb{R}^{B\times F\times 2\times m}$ from $\mathbf{F}_h$. By applying a weighted combination on all the basis vectors, we obtain the key point-based hand texture compensation defined as:

$$\mathbf{G}_h = \sum_{i=1}^{m} a_i \gamma_i. \tag{3}$$

Finally, the output feature $\mathbf{G}_h \in \mathbb{R}^{B\times F\times 2\times d}$ is sent to the hand attention layers in the Denoising U-Net for cross-attention calculation. Additionally, to explicitly inject hand motion information into the denoising latent feature maps, we apply a hand mask M_h to provide emphasis guidance on the hand regions, which can be expressed as:

$$\begin{aligned} latents &= \mathrm{Att}_{hand}(latents, \mathbf{G}_h, \mathbf{G}_h) * M_h + latents, \\ \mathrm{Att}_{hand}(latents, \mathbf{G}_h, \mathbf{G}_h) &= \mathrm{Softmax}\left(\frac{(\mathbf{W}_Q \cdot latents)(\mathbf{W}_K \cdot \mathbf{G}_h)^\top}{\sqrt{d}}\right) \cdot (\mathbf{W}_V \cdot \mathbf{G}_h). \end{aligned} \tag{4}$$

where $latents$ denotes the denoising latent feature maps, $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V$ represent learnable weights for the cross-attention modules.

3.4 Face Recapture Module

Similar to the challenge with hand generation, producing a satisfactory face in human video synthesis is difficult. Most approaches rely on post-processing strategies to address this issue, but these methods significantly increase training and inference costs. In this paper, we design a Face Recapture module to extract comprehensive face information and inject it into the face attention layer in the Denoising U-Net to improve the quality of the generated faces.

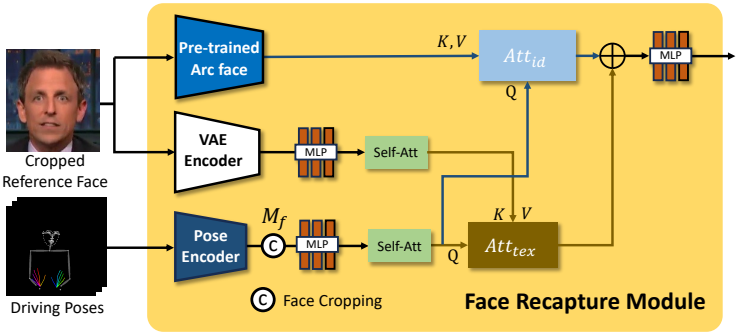

Figure 2: **Architecture of the proposed Face Recapture Module.**

Specifically, as shown in Fig. 2, we enhance the face area of the generated image from two aspects: texture details and global identity. First, we use a face detection model to crop and align the face from the reference image. Next, the pre-trained VAE Encoder and face recognition model ArcFace are leveraged to extract the facial texture feature $\mathbf{F}_{tex}$ and global identity feature $\mathbf{F}_{id}$, respectively. For the facial texture feature, we perform feature enhancement by combining a MLP and a self-attention

https://doi.org/10.52202/079017-1615

layer. We then repeat these features F times along the temporal dimension to match the size of driving poses and produce $\mathbf{F}_{id} \in \mathbb{R}^{B\times F\times 1\times d_{id}}$ and $\mathbf{F}_{tex} \in \mathbb{R}^{B\times F\times (hw)\times d_f}$.

In the meantime, we crop out the corresponding facial pose information $\mathbf{F}_d$ from the driving pose feature maps according to the face mask M_f. Similarly, a MLP and a self-attention are proposed for feature enhancement on $\mathbf{F}_d \in \mathbb{R}^{B\times F\times (hw)\times d_f}$. In order to equip $\mathbf{F}_d$ with textural and identity information, we adopt two separate cross-attention blocks to embed texture and identity information into the $\mathbf{F}_d$, respectively. Specifically, by taking $\mathbf{F}_d$ as query, $\mathbf{F}_{tex}$ as key and value, we define the texture fusing process as $\mathbf{G}_{tex} = \mathrm{Att}_{tex}(\mathbf{F}_d, \mathbf{F}_{tex}, \mathbf{F}_{tex})$. Similarly, by taking $\mathbf{F}_d$ as query, $\mathbf{F}_{id}$ as key and value, we can obtain the fused identity feature $\mathbf{G}_{id}$ through $\mathbf{G}_{id} = \mathrm{Att}_{id}(\mathbf{F}_d, \mathbf{F}_{id}, \mathbf{F}_{id})$. The comprehensive face information is obtained by

$$\mathbf{G}_f = \mathrm{MLP}(\mathbf{G}_{tex} + \mathbf{G}_{id}). \tag{5}$$

Finally, $\mathbf{G}_f \in \mathbb{R}^{B\times F\times (hw)\times d_f}$ is injected into the face attention layer in the Denoising U-Net. Similarly, the face mask is utilized to constrain the latent feature maps after face attention:

$$latents = \mathrm{Att}_{face}(latents, \mathbf{G}_f, \mathbf{G}_f) * M_f + latents. \tag{6}$$

3.5 Training strategies

During the diffusion process, the random noise is continuously added to the real image until it reaches a state of Gaussian noise. Conversely, the generation process within the diffusion model is the inverse of the diffusion process, which takes random Gaussian noise as input and generates real images through gradual denoising. In the training process, the diffusion model leverages the Denoising U-Net to predict the added noise at various time steps and the optimization objective can be defined as:

$$L = \mathbb{E}_{\mathbf{z}_t,\mathbf{c},\epsilon,t}\left(\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, \mathbf{c}, t\right)\|_2^2\right), \tag{7}$$

where $\mathbf{c}$ is the conditional features, ϵ_θ represents the Denoising U-Net, $\mathbf{z}_t$ is the denoising latent feature maps at timestep t.

In our work, we adopt a two-stage training strategy to separately perform appearance modeling and temporal consistency modeling. The first stage is image-level training. During this stage, we set $F = 1$, the training goal is to accurately map the appearance details from the reference image to the driving pose. The VAE encoder and CLIP Image encoder are fixed, and the rest of the network except for the temporal attention are all trainable. In terms of the second stage, the objective is to improve the temporal coherence of the generated frames. During this stage, we only perform training on the temporal attention layer with all other network weights fixed and the weights for the temporal attention layers in the Denoising U-Net are initialized by the pre-trained AnimateDiff [8]. The loss function can be reformulated as:

$$L = \mathbb{E}\left(\|\epsilon - \epsilon_\theta\left(\mathbf{z}_t, \mathbf{z}_r, \mathbf{G}_h, M_h, \mathbf{G}_f, M_f, t\right)\|_2^2\right), \tag{8}$$

where $\mathbf{z}_r$ represents the appearance feature maps extracted by Reference U-Net, $\mathbf{G}_h$ represents the hand movement feature maps from the Key Point-based Fine-grained Hand Modeling module, and $\mathbf{G}_f$ is the face feature maps obtained by the Face Recapture Module. Additionally, M_h and M_f denote the masks for the face and hand regions, respectively.

It is worth noting that in order to focus on the face and hand area generation, in the later phase of the first training stage, we adopt the hand mask and face mask to calculate the L_1 loss of the corresponding area as the final loss every 10 iterations.

4 Experiments

4.1 Experimental Settings

Datasets. In order to verify the effectiveness of the proposed method, we select the videos of two actors, Seth and Oliver, in the talkshow [51] dataset for training and testing. In addition, to enrich the diversity of characters and hand movements, we record videos of seven people in the indoor scene. The video length of each person is about 10 minutes and we divide the videos into multiple clips of

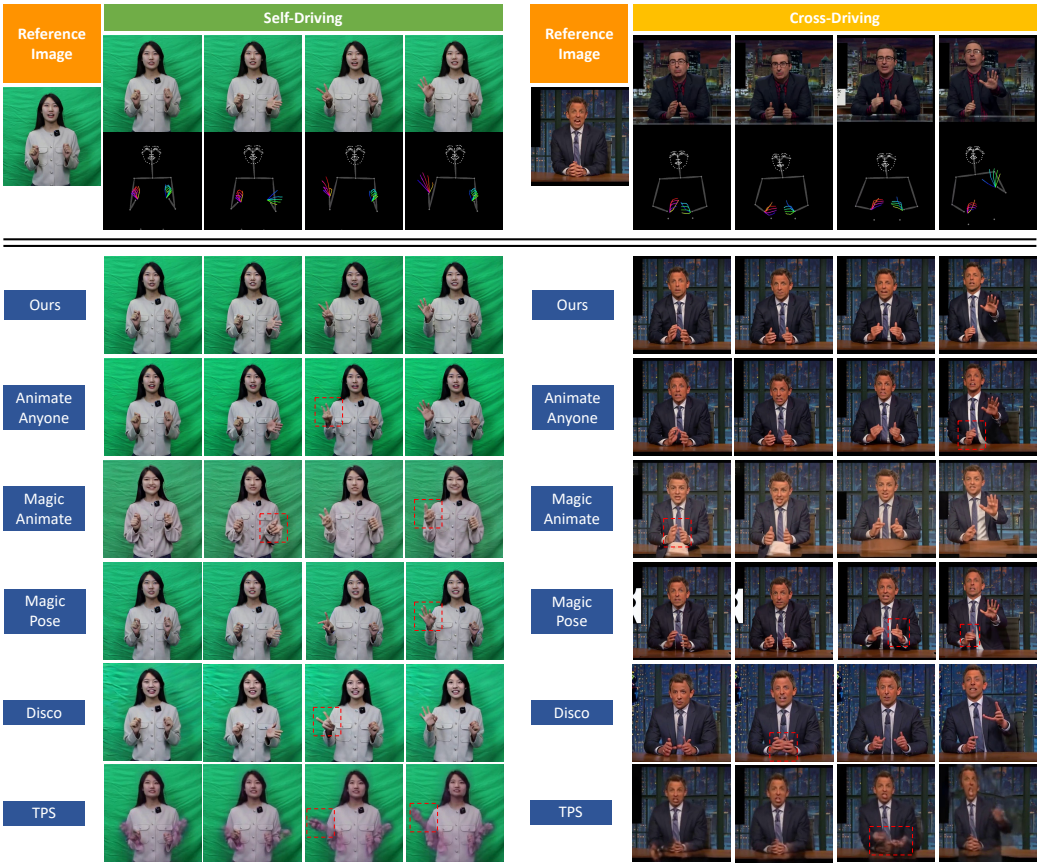

Figure 3: Qualitative results compared with other methods. Our approach achieves high-fidelity gesture details and satisfactory image quality in both self and cross-driving settings.

about 8 seconds for training convenience. In addition to simple rhythmic gestures, these recorded videos also include many complex gestures such as numbers. We randomly divide the training set and test set according to 9:1.

Implementation Details. For data processing, we crop out the body region with a resolution of 512×512 and utilize the pre-trained face detection model [2] to crop and align faces following FFHQ [20]. The resolution of the face image is 256×256. The hand and face masks are determined based on the largest circumscribed rectangle of the corresponding key points. All experiments were completed on 8 A800s, with a learning rate of 1e-5. For the first training stage, the batch size B is set to 24, and sequence length F is set to 1. The training step is 100k, which takes about six days. For the second stage, B and F are set to 1 and 24, respectively with a 30k training step, taking about one day. During inference, we adopt a CFG [10] of 7.5 and perform 30 denoising steps using the DDIM sampler. Additionally, to ensure temporal consistency between different clips of the same video, we use the same reference image, with an overlap of 4 frames between adjacent driving clips.

Comparison Methods. We select several state-of-the-art approaches, including TPS [54], Disco [44], AnimateAnyone [13], MagicAnimate [47], MagicPose [1], Make-Your-Anchor [17] as our comparison methods. TPS is a general animation model based on warping operation, which adopts GAN as the backbone, the remains are diffusion-based human video generation models. Among them, Make-Your-Anchor provides the specific person generation model. We finetune these models on the same training dataset with the official codes and pre-trained models.

Metrics. We comprehensively measure the quality of generated images from pixel space and feature space using SSIM, PSNR, and FID [9] and adopt FVD [38] to verify the temporal consistency of generated videos. In addition, the motion accuracies of body (L_{body}), face (L_{face}), and hand (L_{hand}) are measured by calculating the mean Euclidean distance between the key points of the generated images and real images.

https://doi.org/10.52202/079017-1615

Table 1: Quantitative results of our approach compared with SOTAs. Our method achieves the best performance on image quality, temporal consistency, and motion precision.

Method	SSIM ↑	PSNR ↑	FID ↓	FVD ↓	L_{body} ↓	L_{face} ↓	L_{hand} ↓
TPS	0.65	29.02	94.77	1120.37	5.99	1.26	17.99
Disco	0.69	29.13	80.76	540.76	5.85	1.52	4.33
AnimateAnyone	0.80	29.41	16.87	365.83	2.73	0.62	1.10
MagicAnimate	0.70	28.55	50.24	665.21	4.48	1.33	3.02
MagicPose	0.82	30.03	16.37	370.75	2.32	0.68	1.12
Ours	**0.85**	**32.23**	**15.43**	**197.43**	**2.27**	**0.19**	**0.77**
Make-Your-Anchor (Seth)	0.63	29.18	32.32	428.84	4.55	1.07	1.64
Ours (Seth)	**0.85**	**33.14**	**9.83**	**193.25**	**2.10**	**0.21**	**0.72**

4.2 Comparison with Other Methods

Quantitative Results. The quantitative results of our methods compared with SOTAs are shown in Tab. 1. Our method achieves the best results on SSIM, PSNR, and FID metrics, underscoring its clear advantages on image quality. In addition, our method obtains the best FVD indicating that the temporal consistency of generated videos is superior. Moreover, our approach outperforms others on the motion accuracy metric including (L_{body}), face (L_{face}), and hand (L_{hand}) which demonstrates the precision of generated poses and gestures. Benefiting from the proposed Key Points-based Fine-grained Hand Modeling module and Face Recapture module, our framework is able to generate accurate hand gestures and high-fidelity facial details. TPS is a general object animation approach and struggles with complex human poses and hand postures, resulting in the poorest overall generation quality. Disco, AnimateAnyone, MagicAnimate and MagicPose focus on generating coarse-grained human poses without fine-grained modeling of hands and faces, leading to lower generation quality, particularly in hand gesture accuracy. Note that Make-Your-Anchor [17] only provides a pre-trained model on Seth data and thus we also report the experimental results on the Seth for a fair comparison. It is observed that our method outperforms [17] on all metrics.

Qualitative Comparison. For qualitative comparison, we conduct two experimental settings, including self-identity and cross-identity driven. The self-identity driven comparison results are shown on the left of Fig. 3 and 4, where obvious artifacts are marked with red dotted boxes. Our method is able to generate accurate gestures and high-quality hand details while other approaches fail to convey complex gestures. In addition, face identity is well-preserved in our results, while other methods lead to unnatural face shapes, textures, and identity. Note that cross-identity driven is more challenging and our method still achieves high-fidelity gesture details and satisfactory image quality as indicated in the right of Fig. 3 and 4. The hand areas of other results are blurry and unrealistic. In summary, our method makes use of hand and face information through the well-designed Key Points-based Fine-grained Hand Modeling module and Face Recapture module, enhancing the video quality which could meet complicated requirements in conversational scenarios.

Human Evaluation. We conduct a user preference study to evaluate the performance of the proposed framework. There are 21 samples and 15 human voters in total. For each sample, we present six video results generated with ShowMaker and other SOTA methods to the human voter in a random order. The human voters are required to estimate the video results in three aspects: a) Motion Accuracy: Does the video accurately reproduce the motion in the driving video? b) Appearance Consistency: Does the subject in the video have a consistent appearance with the reference image? c) Temporal Consistency: How is the temporal coherence of this video? The rating score ranges from 1 to 5 and higher scores indicate better preference. As shown in Tab. 3, our method achieves the highest scores compared with its counterparts, which demonstrates that our method is preferred by a significant margin on motion accuracy and appearance consistency.

4.3 Ablation Study

To verify the contributions of different components, we train three variants by removing the Key Points-based Fine-grained Hand Modeling module (HM), the Face Recapture module (FR), and two-stage training (Stage 2), respectively.

Figure 4: Qualitative comparison between Make-Your-Anchor and our ShowMaker.

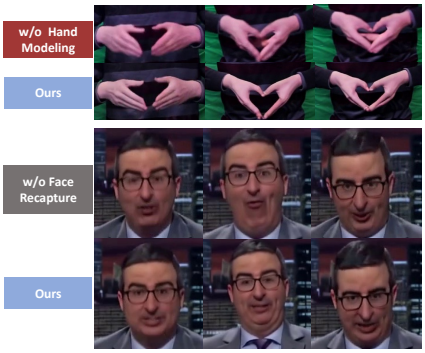

Figure 5: Qualitative ablation results when removing different components in our framework.

Table 2: Ablation analysis of HM, FR, and two-stage training.

Variations	FVD ↓	L_{face} ↓	L_{hand} ↓
w/o HM	212.25	-	1.12
w/o FR	208.28	0.23	-
w/o Stage 2	373.87	-	-
Ours	**197.43**	**0.19**	**0.77**

Table 3: User Study.

Methods	Ours	Magic Pose	Magic Animate	Animate Anyone	Make-Your-Anchor	Disco	TSP
Motion Accuracy	**4.53**	3.13	2.87	3.69	4.43	2.62	1.84
Appearance Consistency	**4.23**	3.47	2.00	3.62	4.18	2.16	1.91
Temporal Consistency	**4.33**	3.00	2.20	3.11	3.57	1.78	1.29

For quantitative results, we report the FVD and motion accuracy metrics in Tab. 2. It is observed that the HM and FR could enhance gesture accuracy and face quality by a significant margin. The two-stage training brings about notable improvements in FVD which indicates better temporal consistency. Additionally, we present qualitative comparisons to verify the effectiveness of the proposed HM and FR module. As shown in Fig. 5, the generated hands are more satisfactory, and the face shapes and textures are realistic and clear with the proposed modules.

5 Conclusion and Discussion

Conclusion. In this paper, we propose the framework ShowMaker, which achieves high-fidelity 2D human video synthesis with two novel designs to achieve fine-grained diffusion modeling. We first propose a Key Point-based Fine-grained Hand Modeling module for robust and fine-grained hand synthesis which takes advantage of 2D hand key points and a key point-based codebook. To further reconstruct the facial details and identity information, we introduce a Face Recapture module, which effectively equips the structure information with detailed textural information and global identity. Quantitative and qualitative evaluation has indicated the superiority of our framework beyond the existing approaches.

Limitations. Despite the success of our framework, we also recognize some limitations during the exploration. Firstly, our Key Point-based Fine-grained Hand Modeling module can robustly manage hand synthesis with occasional incorrect hand gestures. However, DWPose [50] suffers from performance degradation when handling videos with severe motion blur leading to considerable perturbation in the driving signal, which inevitably results in unexpected artifacts in our results. Secondly, our framework produces less satisfactory results when handling challenging cases, such as reflection on glasses. These will be part of our future work.

Broader Impact. Our method focuses on synthesizing realistic 2D avatars with rich facial expressions and complex hand gestures, which is intended for developing digital humans under more daily scenarios like TV shows. However, it may also be misused for some malicious purposes on social media, which leads to negative impacts on the whole society. Therefore, we will make our efforts to strictly oversee the dissemination of our models as well as the resulting content and also restrict access solely to research-oriented demands. We believe that the proper use of this technique will enhance positive societal development in both machine learning research and daily life.

https://doi.org/10.52202/079017-1615

ACKNOWLEDGMENTS

This work is supported by the National Nature Science Foundation of China (62121002, U23B2028, 62232006, 62032006, 62472395).

We acknowledge the support of the GPU cluster built by the MCC Lab of Information Science and Technology Institution, USTC. We also thank the USTC Supercomputing Center for providing computational resources for this project.

References

[1] Di Chang, Yichun Shi, Quankai Gao, Hongyi Xu, Jessica Fu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. In *ICML*, 2024.

[2] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *CVPR*, 2020.

[3] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *CVPR*, pages 5609–5619. IEEE, 2023.

[4] Yuan Gong, Yong Zhang, Xiaodong Cun, Fei Yin, Yanbo Fan, Xuan Wang, Baoyuan Wu, and Yujiu Yang. Toontalker: Cross-domain face reenactment. In *ICCV*, pages 7656–7666. IEEE, 2023.

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[6] Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, and Jingdong Wang. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *CVPR*, pages 1505–1515. IEEE, 2023.

[7] Jiazhi Guan, Zhiliang Xu, Hang Zhou, Kaisiyuan Wang, Shengyi He, Zhanwang Zhang, Borong Liang, Haocheng Feng, Errui Ding, Jingtuo Liu, et al. Resyncer: Rewiring style-based generator for unified audio-visually synced facial performer. In *European Conference on Computer Vision*, pages 348–367. Springer, 2025.

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

[9] 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. In *NIPS*, pages 6626–6637, 2017.

[10] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NIPS)*, 2020.

[12] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *CVPR*, pages 3387–3396. IEEE, 2022.

[13] Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *CVPR*. IEEE, 2024.

[14] Ricong Huang, Weizhi Zhong, and Guanbin Li. Audio-driven talking head generation with transformer and 3d morphable model. In *ACM Multimedia*, pages 7035–7039. ACM, 2022.

[15] Ricong Huang, Peiwen Lai, Yipeng Qin, and Guanbin Li. Parametric implicit face representation for audio-driven facial reenactment. In *CVPR*, pages 12759–12768. IEEE, 2023.

[16] Zhichao Huang, Xintong Han, Jia Xu, and Tong Zhang. Few-shot human motion transfer by personalized geometry and texture modeling. In *CVPR*, pages 2297–2306. Computer Vision Foundation / IEEE, 2021.

[17] Ziyao Huang, Fan Tang, Yong Zhang, Xiaodong Cun, Juan Cao, Jintao Li, and Tong-Yee Lee. Make-your-anchor: A diffusion-based 2d avatar generation framework. In *CVPR*, pages 6997–7006. IEEE, 2024.

[18] Subin Jeon, Seonghyeon Nam, Seoung Wug Oh, and Seon Joo Kim. Cross-identity motion transfer for arbitrary objects through pose-attentive video reassembling. In *ECCV (24)*, volume 12369 of *Lecture Notes in Computer Science*, pages 292–308. Springer, 2020.

[19] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *ICCV*, pages 22623–22633. IEEE, 2023.

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019.

[21] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[22] Fengyuan Liu, Lingyun Yu, Hongtao Xie, Chuanbin Liu, Zhiguo Ding, Quanwei Yang, and Yongdong Zhang. High fidelity face swapping via semantics disentanglement and structure enhancement. In *ACM Multimedia*, pages 6907–6917. ACM, 2023.

[23] Wen Liu, Zhixin Piao, Zhi Tu, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping GAN with attention: A unified framework for human image synthesis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(9):5114–5132, 2022.

[24] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. In *AAAI*, pages 1896–1904. AAAI Press, 2023.

[25] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. In *NeurIPS*, 2022.

[26] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV (1)*, volume 12346 of *Lecture Notes in Computer Science*, pages 405–421. Springer, 2020.

[27] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *CVPR*, pages 788–798. IEEE, 2024.

[28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[29] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, pages 14294–14303. IEEE, 2021.

[30] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM Multimedia*, pages 484–492. ACM, 2020.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

 https://doi.org/10.52202/079017-1615

[32] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H. Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *ICCV*, pages 13739–13748. IEEE, 2021.

[33] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685. IEEE, 2022.

[34] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *NeurIPS*, pages 7135–7145, 2019.

[35] Aliaksandr Siarohin, Oliver J. Woodford, Jian Ren, Menglei Chai, and Sergey Tulyakov. Motion representations for articulated animation. In *CVPR*, pages 13653–13662. Computer Vision Foundation / IEEE, 2021.

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

[37] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, pages 1526–1535. Computer Vision Foundation / IEEE Computer Society, 2018.

[38] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *CoRR*, abs/1812.01717, 2018.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[40] Jiayu Wang, Kang Zhao, Yifeng Ma, Shiwei Zhang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Facecomposer: A unified model for versatile facial content creation. In *NeurIPS*, 2023.

[41] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European Conference on Computer Vision*, pages 700–717. Springer, 2020.

[42] Kaisiyuan Wang, Changcheng Liang, Hang Zhou, Jiaxiang Tang, Qianyi Wu, Dongliang He, Zhibin Hong, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. Robust video portrait reenactment via personalized representation quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2564–2572, 2023.

[43] Kaisiyuan Wang, Hang Zhou, Qianyi Wu, Jiaxiang Tang, Zhiliang Xu, Borong Liang, Tianshu Hu, Errui Ding, Jingtuo Liu, Ziwei Liu, et al. Efficient video portrait reenactment via grid-based codebook. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–9, 2023.

[44] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *CVPR*. IEEE, 2024.

[45] Dongxu Wei, Xiaowei Xu, Haibin Shen, and Kejie Huang. C2F-FWN: coarse-to-fine flow warping network for spatial-temporal consistent motion transfer. In *AAAI*, pages 2852–2860. AAAI Press, 2021.

[46] Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, Chengjie Wang, Zhifeng Xie, and Yong Liu. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6609–6619, 2023.

[47] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *CVPR*. IEEE, 2024.

[48] Kewei Yang, Kang Chen, Daoliang Guo, Song-Hai Zhang, Yuan-Chen Guo, and Weidong Zhang. Face2face$^{\rho}$: Real-time high-resolution one-shot face reenactment. In *ECCV (13)*, volume 13673 of *Lecture Notes in Computer Science*, pages 55–71. Springer, 2022.

[49] Quanwei Yang, Xinchen Liu, Wu Liu, Hongtao Xie, Xiaoyan Gu, Lingyun Yu, and Yongdong Zhang. REMOT: A region-to-whole framework for realistic human motion transfer. In *ACM Multimedia*, pages 1128–1137. ACM, 2022.

[50] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV (Workshops)*, pages 4212–4222. IEEE, 2023.

[51] 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 *CVPR*, 2023.

[52] Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, and Baoyuan Wang. Talking head generation with probabilistic audio-to-visual diffusion priors. In *ICCV*, pages 7611–7621. IEEE, 2023.

[53] Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, HsiangTao Wu, Dong Chen, Qifeng Chen, Yong Wang, and Fang Wen. Metaportrait: Identity-preserving talking head generation with fast personalized adaptation. In *CVPR*, pages 22096–22105. IEEE, 2023.

[54] Jian Zhao and Hui Zhang. Thin-plate spline motion model for image animation. In *CVPR*, pages 3647–3656. IEEE, 2022.

[55] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and A benchmark. In *CVPR*, pages 20228–20237, 2022.

[56] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance, 2024.

 https://doi.org/10.52202/079017-1615

A Appendix

A.1 Pose Encoder Structure

Fig. 6 shows the detailed network structure of the pose encoder.

A.2 Dataset Details

Tab. 4 gives the number of training and testing clips for the indoor recording dataset and talkshow dataset. The length of each clip is 2-15 seconds.

When preparing the training data, we first produce the DWPose results from each frame and set the center of shoulders as the cropping center. Then we crop the original video frame using an adaptive cropping size, where the cropping size is designed as a fixed ratio (we set it to 2.65) of the shoulder width. This operation forces the human body to lie in a roughly consistent position. Finally, all cropped images are resized to 512×512.

In the inference stage, for the cross-identity driven setting, we further bridge the gap between body shapes by scaling the driving poses to match the reference pose. The scaling ratios of width and height are defined as W_r/W_d, and H_r/H_d, where W and H represent the shoulder width and the height of the human body, respectively. The shoulder width is measured as the Euclidean distance between the left and right shoulder key points, while the body height is determined by the Euclidean distance from the nose key point to the pelvis key point.

A.3 More Ablation Experiments

To further verify the effectiveness of the proposed Key point-based Fine-grained Hand Modeling (HM), we show more ablation experiments in Fig. 7, where **Vanilla** refers to removing the entire HM module from our proposed ShowMaker, **Hand image** refers to adopting the cropped hand image to extract texture features through the VAE encoder and then feeding it into the hand attention, and **w/o positional encoding** refers to not using positional encoding in the HM module. It can be seen that using the hand image as compensation information has no obvious effect on improving the hand texture and structure, As for HM, positional encoding can significantly improve the high-frequency details and structural accuracy of the generated hands.

A.4 Temporal consistency

Fig. 8 shows two generated video sequences, each sequence contains 3 adjacent video frames. It can be seen that the generated video frames at different times have consistent appearance and no temporal texture changes occur.

A.5 Challenging Examples

Fig. 9 shows two challenging sample generations. In the example on the left, the driving pose is extracted from a female video, while the reference image is a male, and the driving gestures and the reference gesture are very different. In the example on the right, the reference image is not a common frontal image instead of a side image. Our proposed ShowMaker can still generate target frames well in these two cases, which shows that our method has satisfied robustness.

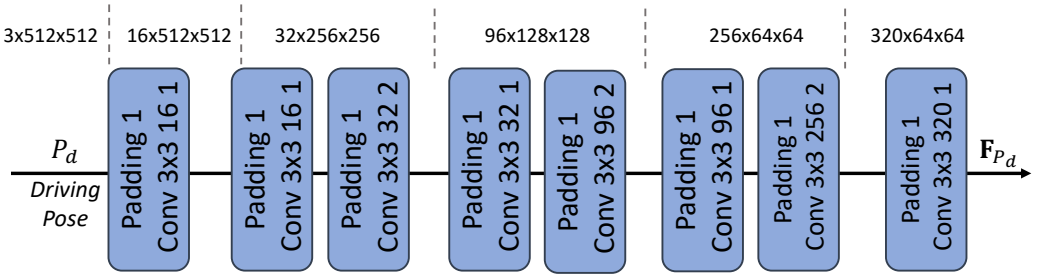

Figure 6: The pose encoder structure consists of 8 convolutional layers, where (Conv 3×3 16 1) represents the kernel size is $3 \times 3 \times 16$, and the stride is 1. Except for the last convolutional layer, each convolution is followed by GroupNormalization and the activation function silu.

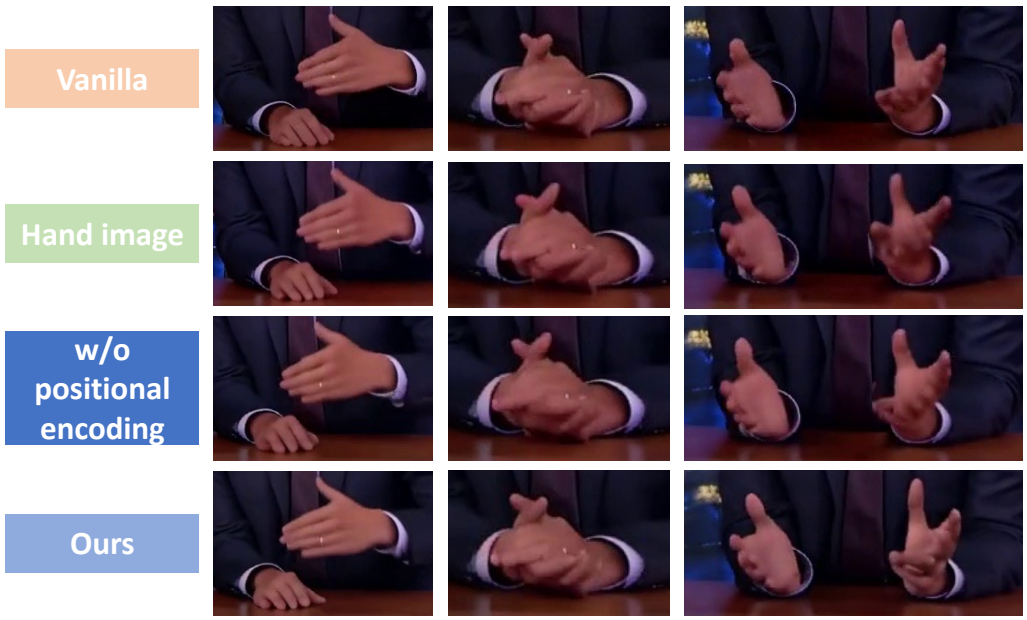

Figure 7: Ablation experiments of hand module.

Table 4: Dataset details. ID1-7 are datasets recorded indoors. Seth and Oliver belong to the talkshow dataset. All training and test sets do not overlap.

IDs	The number of clips in the training set	The number of clips in the test set
ID1	49	5
ID2	57	6
ID3	56	6
ID4	65	7
ID5	47	5
ID6	64	7
ID7	55	6
Seth	3306	100
Oliver	6746	200

https://doi.org/10.52202/079017-1615

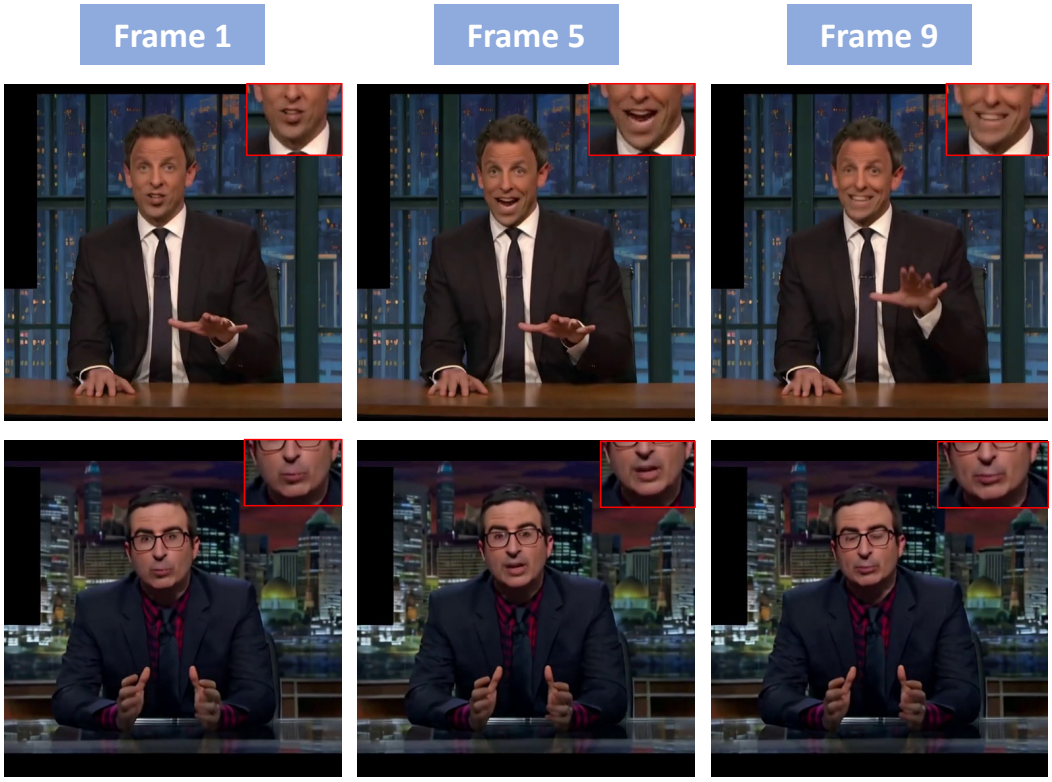

Figure 8: The generation results of adjacent frames.

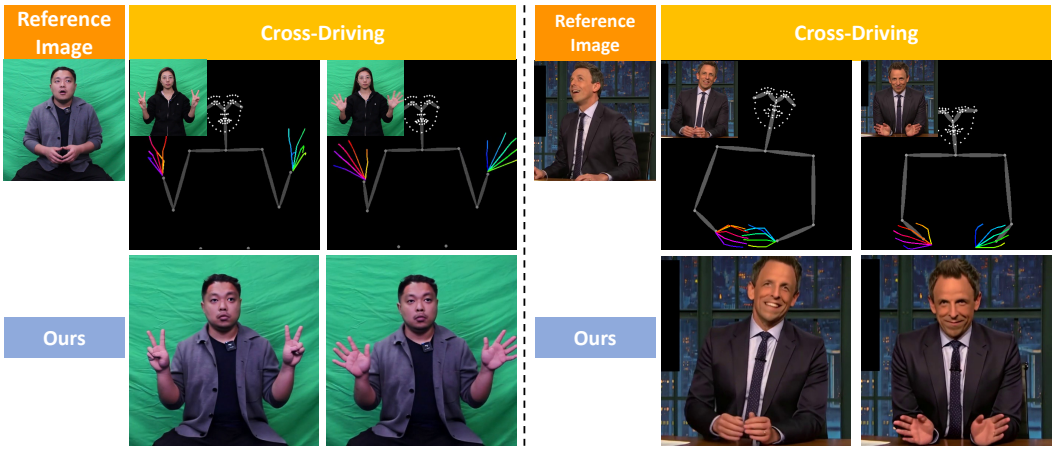

Figure 9: The generated results when the pose difference is obvious.

https://doi.org/10.52202/079017-1615

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claims in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

https://doi.org/10.52202/079017-1615

Justification: We discuss the limitations of the work in Section Conclusion and Limitation

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We fully disclose all the information needed to reproduce the main experimental results.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will provide the code in the camera-ready version.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

https://doi.org/10.52202/079017-1615

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We report the training and test details in the paper, including environment and dataset.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This paper does not include error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impacts in the paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We describe safeguards for responsible release of data or models by requiring that users adhere to usage guidelines.

    Guidelines:

https://doi.org/10.52202/079017-1615

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We explicitly mention and properly respect the license and terms of use in the paper and the creators or original owners of assets used in the paper are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

Justification: We report the full text of instructions.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: Our user study does not apply to it.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-1615

References for the Appendix