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

Supplementary Material

A. Related Work

A.1. Talking Face Generation

Talking face generation methods [6, 19, 28, 35] generate human videos with various expressions and poses conditioned on a given audio or motion, which can be categorized into two types: editing facial regions or generating dynamic head videos. Editing-based techniques, such as Wav2Lip [19] or VideoRetalking [6], face the problem of lip-gesture inconsistency. Usually, the gestures are fixed for different talking content. Generating dynamic head video requires methods to create head videos conditioned on given audio or motion, where head motions are accomplished by different manners such as motion flow [34], 3D landmarks [28, 35], self-supervised training [18], etc. Although producing high-quality and highly realistic facial videos, talking face generation is limited by its interest area and cannot achieve full-body human video generation.

A.2. Pose-guided Human Video Generation

To generate human video with body and hand, pose-guided methods are the most popular approaches. Early work focuses on the problem of motion transfer [1, 4, 15, 23–25, 29, 31, 36] methods. Balakrishnan et al. [1] separate a scene and transform each part to synthesize. FOMM [23] and MRAA [24] propose unsupervised body representation and warping to transfer, while LIA [29] employs latent space. TPS [36] introduces thin plate spline transformation into a motion transfer task, and UVA [25] presents a differential volumetric representation. Besides the coarse-grained motion transfer setting, researchers [9, 14, 21, 37] apply these kinds of methods into human video generation with face, body, and hand. However, constrained to the capability of these models, these methods potentially generate human videos with apparent artifacts.

With the progress of diffusion models [11, 22], some works introduce them into pose-guided human video generation. Follow-your-pose [17] introduces a two-stage training to get a pose-guided video diffusion model. DreamPose [13] proposes a vision and pose controlled diffusion on fashion dataset to animate a body image. DisCo [27] focuses on human dance generation, utilizes multiple ControlNet on pose and background, and introduces a pretraining strategy to improve generalizability. Nonetheless, these methods concentrate on coarse-grained body video generation, which is limited to the poor quality of face and hands. Besides, due to the randomness of the diffusion model, these methods struggle with temporal consistency. The proposed system, by proposing a simple yet efficient multi-frame inference strategy, could improve the temporal consistency of image-based diffusion models.

A.3. Video Diffusion Models

Due to the powerful capabilities of diffusion models, researchers in recent years have started to explore their potential in video generation, and much progress has been made in video generation [2, 5, 8, 12, 26] and video editing [3, 20, 32]. GEN-1 [7] extends the image diffusion model with a temporal module and utilizes depth to control the structure. Tune-A-Video [30] fine-tunes 3D U-Net in image diffusion model on a one-shot video to learn the motion and then edits the video content with text prompts. AnimateDiff [10] trains a temporal module with a fixed image diffusion model and can be applied to personalized weights. While VDMs possess strong video generation capabilities, the ability to control human motion and maintain appearance needs further improvement. In contrast, we tune the foundation diffusion model to learn the mapping from motion to a specific anchor appearance in a "binding" fashion following a pretrain-finetuning paradigm.

B. Video Results

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

Furthermore, we show the results of audio-driven digital avatar generation. We utilize TalkSHOW [33] to drive the 3D human mesh, and the examples are shown in the video as well as Fig. S1. By combining our method with existing audio-driven motion generation techniques, we create a system capable of automatically generating 2D avatar videos.

C. User Study

We conduct user studies under our experimental setting in the main text. The videos generated in Section 5.2 were

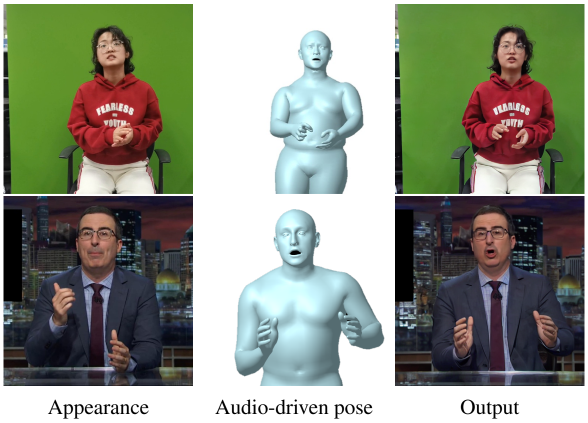


Figure S1. Examples of audio-driven results.

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

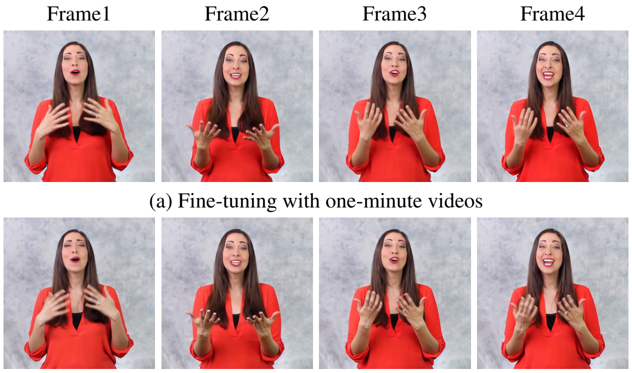
The statistics are listed in Table S1. As the results show, our method achieves the best scores of over four points in the user study. Pose2Img scored above three points in appearance preservation, and the results from Dreampose and DisCo are slightly inferior to Pose2Img. Our method has achieved a significant advantage in structure preservation compared to other methods, which is not apparent by LMD in Table 1 of the main text for inaccurately estimated landmarks.

Method	Appearance Preservation	Temporal Consistency	Structure Preservation	Overall Quality
Pose2Img [21]	3.36	2.38	1.59	2.01
TPS [36]	2.03	2.10	1.18	1.37
DreamPose [13]	2.78	1.94	2.27	1.94
DisCo [27]	2.29	2.32	1.45	1.72
Ours	**4.23**	**3.85**	**3.91**	**4.03**

Table S1. User study scores. The rating score is on a scale from one to five, where five is the highest score, and one is the lowest.

D. Ablation on Video Length

We conducted an analysis of the required duration for video data needed for fine-tuning. Compared to the one-minute videos used in the main text, we utilize five-minute videos in the fine-tuning stage. The quantitative results are demonstrated in Table. S2 and the qualitative results are displayed in Fig. S2. The numerical results demonstrate a slight improvement in all measurements. For instance, with LMD (hand), additional data allows the model to encompass a broader range of angles, enabling more accurate generation outcomes, and showing better results. In qualitative results, one minute of fine-tuning data already yields satisfactory outcomes.



(a) Fine-tuning with one-minute videos

(b) Fine-tuning with five-minute videos

Figure S2. Qualitative results with different video lengths for fine-tuning.

Fine-tuning with	FID↓	FVD↓	LMD (Face)↓	LMD (Body)↓	LMD (Hand)↓
One-minute videos	40.33	139.82	1.44	4.88	5.41
Five-minute videos	38.94	134.55	1.29	4.42	4.95

Table S2. Ablation analysis of video length used for fine-tuning.

E. Additional Time Cost.

Time cost of different methods is shown in Table S3. Our method is comparable to other diffusion methods. Most time is spent on diffusion. Off-the-shelf speed-up approaches such as LCM [16] could be further engaged.

In Algorithm 1, *count* is an array storing the computing counts for each frame, where only overlapped frames are calculated twice times. When the $ws = 16$ and $os = 4$ for 300 frames/10s to generate, the total computing times in *count* is 400, which means an additional one-third of the cost. The additional time cost brings consistency between batches to generate long duration. A comparison of time cost and ablation without batch-overlapped temporal denoising (TD) is listed in Table. S3.

Method	Pose2Img	TPS	DreamPose	DisCo	Ours	Ours w/o TD	300 frames w/o TD
Time cost	77s	42s	310s	154s	407s	301s	OOM
Method type	non-diffusion	non-diffusion	diffusion	diffusion	diffusion	diffusion	diffusion
Resolution	640px	384px	512px	256px	512px	512px	512px

Table S3. Comparison of time cost to generate 300 frames.

Method	FID	FVD↓	LMD (Face)↓
Ours	40.33	139.82	1.44
FE w 512px	39.73	140.18	1.35

Table S4. Numerical results of FE with 512px.

F. High-Resolution FE Model

We observed that our method sometimes generate inaccurate lip and expression movements, this could be dated back to the mouth's small size which constrains face enhancement's ability. We make a improvement that training a higher-pixel model from 256px to 512px. As in Fig. S3 and Table S4, the lip and expression movements becomes accurate and the quality of the face region is enhanced.

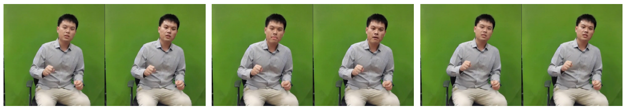


Figure S3. Improved FE. Left 256px, right 512px. Click the last image to play the embedded clips with Acrobat Reader.

G. Liminations

The generated video results are based on the input 3D mesh sequence. When the input mesh sequence is of large pose variations or even inaccurate, the visual quality of the results will be decreased. As shown in Fig. S4a, when the driven facial pose largely varies from the training data, the generated facial expressions are unsatisfactory. It could be improved with more fine-tuning data as shown in Section D. Furthermore, due to limitations in the precision of motion capture, especially regarding the accuracy of hand capture, some obtained motion meshes exhibit inaccuracies. As shown in Fig. S4b, when the input mesh is inaccurate, the generated frame is confusing. The inaccurate meshes are reflected in the results, leading to phenomena such as limb misalignment.

H. Border Impacts

Our method could be used maliciously to create DeepFake videos, which may bring adverse social impacts. We deeply understand the possible negative influence of human video generation technology, and will strictly prevent the spread of our method. Nevertheless, we believe that our approach can yield positive societal impacts, particularly in applications within fields such as education, entertainment, healthcare, and e-commerce.

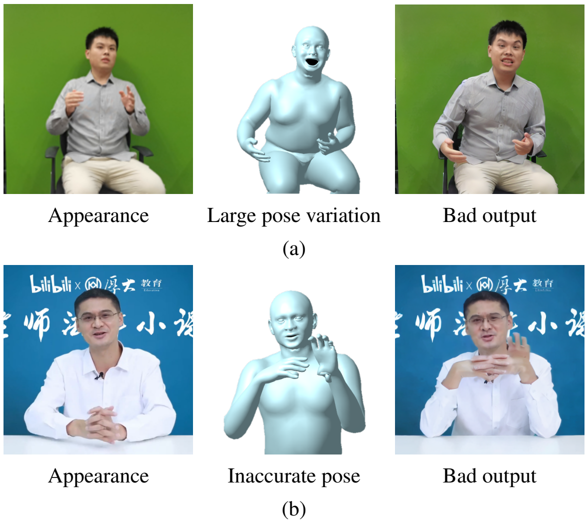


Figure S4. Limitations. When the driven mesh is of large variation or inaccurate, the generated frame will be unsatisfactory.

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. 1

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023. 1

[3] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. Stablevideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23040–23050, 2023. 1

[4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pages 5933–5942, 2019. 1

[5] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 1

[6] Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 1

[7] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models.

In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7346–7356, 2023. 1

[8] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22930–22941, 2023. 1

[9] 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. 1

[10] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 1

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 33:6840–6851, 2020. 1

[12] 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. 1

[13] 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. 1, 2

[14] Miao Liao, Sibo Zhang, Peng Wang, Hao Zhu, Xinxin Zuo, and Ruigang Yang. Speech2video synthesis with 3d skeleton regularization and expressive body poses. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2020. 1

[15] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5904–5913, 2019. 1

[16] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2

[17] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023. 1

[18] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong 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

[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] Chenyang QI, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15932–15942, 2023. 1

[21] 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, 2

[22] 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. 1

[23] 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

[24] 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

[25] Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Jian Ren, Hsin-Ying Lee, Menglei Chai, and Sergey Tulyakov. Unsupervised volumetric animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4658–4669, 2023. 1

[26] 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. 1

[27] 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. 1, 2

[28] 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

[29] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations (ICLR)*, 2022. 1

[30] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7623–7633, 2023. 1

[31] Guang Yang, Wu Liu, Xinchen 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

[32] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *ACM SIGGRAPH Asia Conference Proceedings*, 2023. 1

[33] 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

[34] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision (ECCV)*, pages 85–101. Springer, 2022. 1

[35] 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

[36] 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, 2

[37] 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