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.

1. Motivation and Contributions

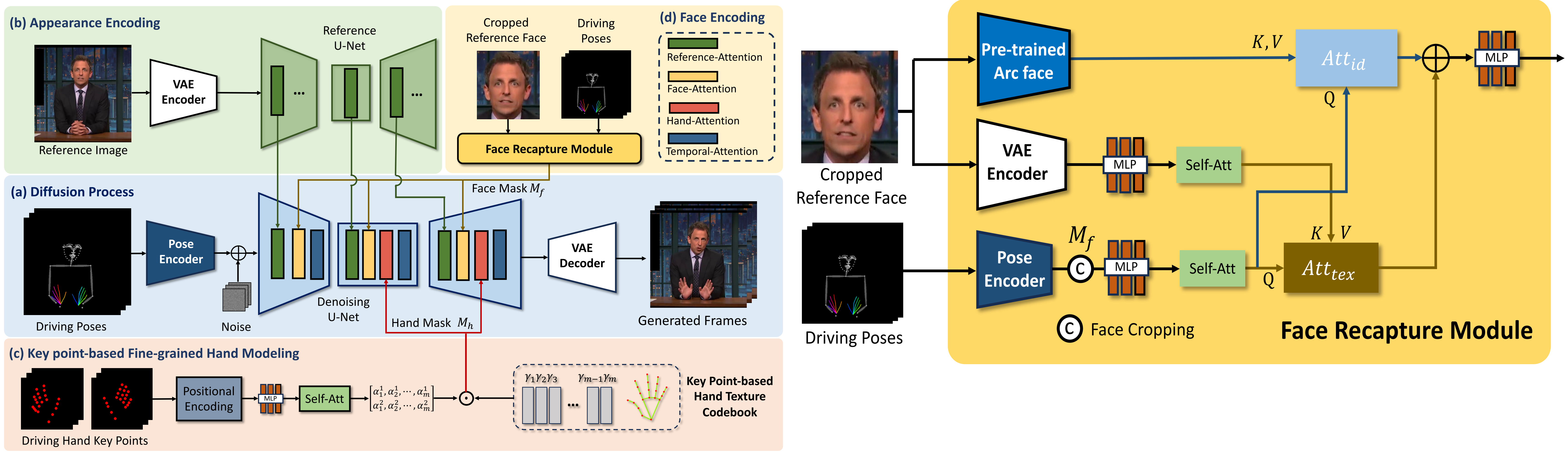
Challenges:

- ✓ **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.**
- ✓ **Facial identity preservation is another problem that has not been well investigated.**

Contributions:

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 effectively recover richer facial details and recapture the identity of the target subject.

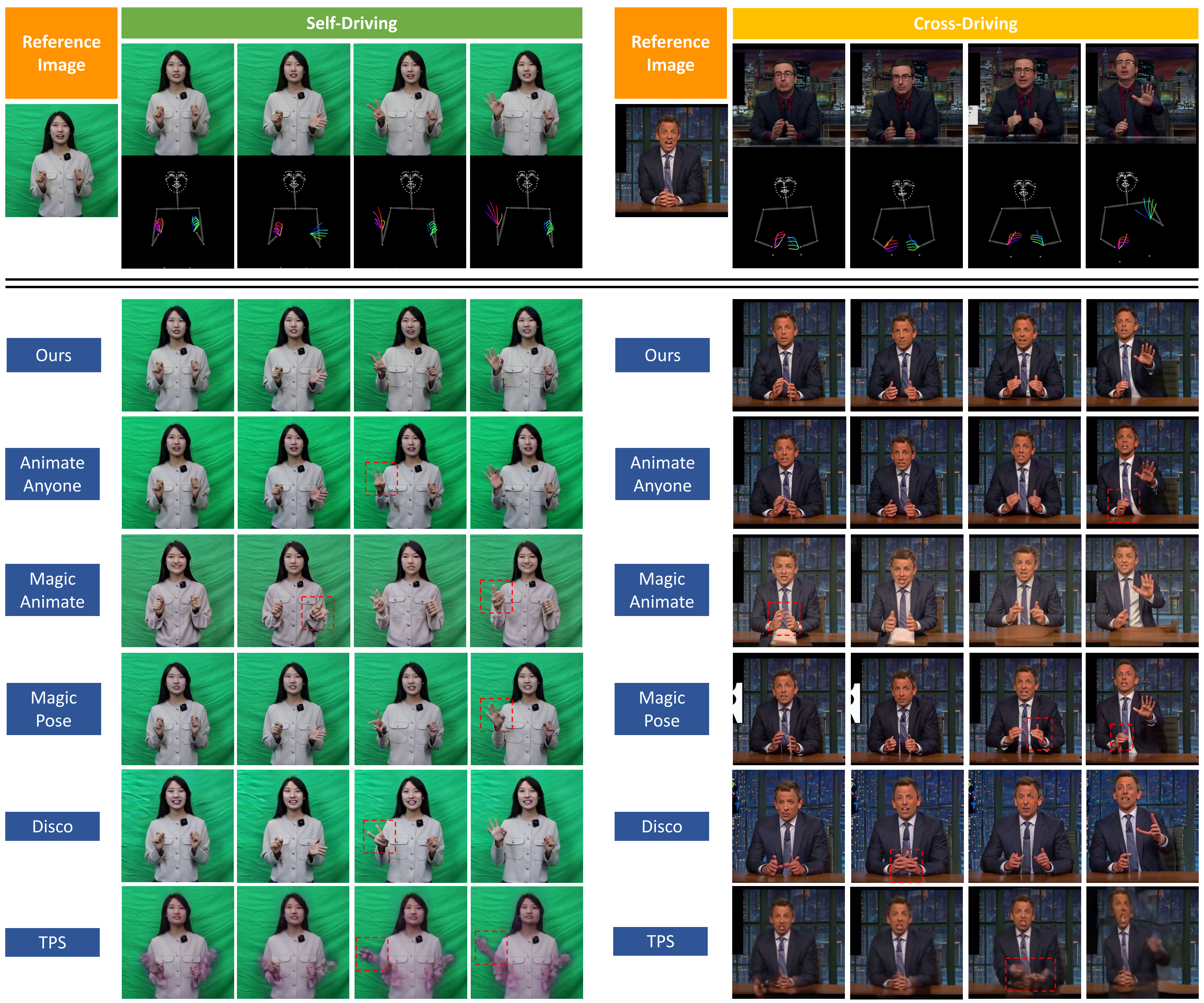
2. Framework



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.

3. Experiments

Qualitative results on the talkshow and collected dataset



Quantitative results on the talkshow and collected dataset

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**

Conclusion & Limitations

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. Quantitative and qualitative evaluation has indicated the superiority of our framework beyond the existing approaches.

Limitations： DWPose 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.