ID-Consistent, Precise Expression Generation with Blendshape-Guided Diffusion (Supplementary Material)

A. Implementation Details

Our method builds on Arc2Face [7], which uses a fine-tuned UNet and ID encoder derived from `stable-diffusion-v1-5`. For our **Expression Adapter**, we employ a two-layer MLP, as well as separate identical key/value matrices into the UNet's cross-attention layers. Training is performed with AdamW [4] using a learning rate of 1e-4, a batch size of 8 per GPU, across 8 NVIDIA A100 GPUs for 300K iterations. For the **Reference Adapter**, we use a copy of Arc2Face's UNet as the reference network and augment the original UNet with LoRA matrices of rank 128. The LoRA weights are trained using cross-paired frames from video datasets, as described in the main paper. We optimize them with a learning rate of 1e-5, using the same hardware configuration and a batch size of 8, for 15K iterations. For inference, we adopt DPM-Solver [5, 6] with 25 denoising steps and a classifier-free guidance scale of 3. For the LoRA weights in the **Reference Adapter**, we found a scale factor of 0.8 to provide a good balance between visual consistency with the input image and alignment with the target expression. Finally, for the baselines in our comparisons, we replace any additional text conditioning (if used by the method) with the prompt "photo of a person" to ensure a fair comparison that emphasizes identity and expression control.

B. Additional Results

Below, we provide additional visual comparisons in Fig. 1, along with further generations from our **Expression Adapter** in Figs. 2 to 4.

Figure 1. Visual comparison between our method and competing expression-conditioned models. **Top:** ID-driven results compared to [3, 8]. **Bottom:** Reference-driven generation compared to [1, 2, 9]. For the latter setting, our method is conditioned on both identity features and the reference image via the Reference Adapter.

References

[1] Bita Azari and Angelica Lim. Emostyle: One-shot facial expression editing using continuous emotion parameters. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 1

[2] Yue Han, Junwei Zhu, Keke He, Xu Chen, Yanhao Ge, Wei Li, Xiangtai Li, Jiangning Zhang, Chengjie Wang, and Yong Liu. Face adapter for pre-trained diffusion models with fine-grained id and attribute control. *arXiv preprint arXiv:2405.12970*, 2024. 1

[3] Chao Liang, Fan Ma, Linchao Zhu, Yingying Deng, and Yi Yang. Caphuman: Capture your moments in parallel universes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[4] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[5] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. 1

[6] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095*, 2022. 1

[7] Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 1

[8] Tuomas Varanka, Huai-Qian Khor, Yante Li, Mengting Wei, Hanwei Kung, Nicu Sebe, and Guoying Zhao. Towards localized fine-grained control for facial expression generation. *arXiv preprint arXiv:2407.20175*, 2024. 1

[9] Mengting Wei, Tuomas Varanka, Xingxun Jiang, Huai-Qian Khor, and Guoying Zhao. Magicface: High-fidelity facial expression editing with action-unit control. *arXiv preprint arXiv:2501.02260*, 2025. 1

Figure 2. Samples generated by our model, conditioned on the input identity and the corresponding target expression.

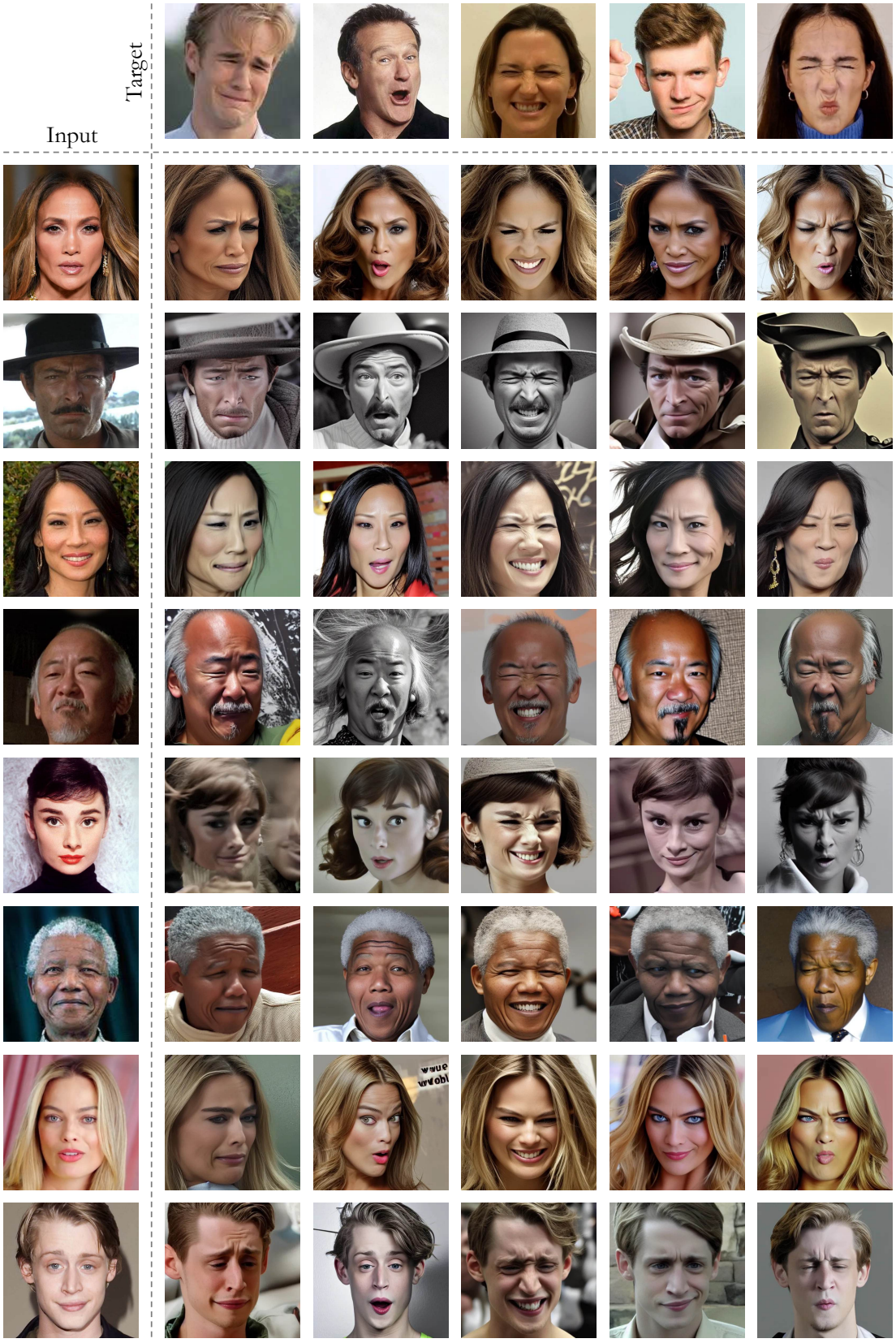

Figure 3. Samples generated by our model, conditioned on the input identity and the corresponding target expression (cont.).

Figure 4. Samples generated by our model, conditioned on the input identity and the corresponding target expression (cont.).