Supplementary Materials of Control-Talker

Anonymous Authors

A APPENDIX

In this supplement, we offer a detailed description of some implementation details of the proposed Control-Talker in Section A.1, and we also provide additional results in Section A.2.

A.1 Implementation Details

Model Architecture. The detailed architecture of the proposed Control-Talker is shown in Figure 1. And the configurations for training HF-ControlNet are presented in Table 1. As discussed in our main paper, we initialize the U-Net part of the diffusion model with DiffusionRig [1] model pretrained on the FFHQ dataset and we employ a structure akin to ControlNet [4] for the extraction of high-frequency texture features. Specifically, the HF-ControlNet initially employs the Input Hint Block shown in Figure 1 (c), to perform feature extraction on the concatenated input $C = \{I_{hf}, I_{pb}\}$, where I_{hf} represents the high-frequency feature map and I_{pb} denotes the combination of physical buffers. Subsequently, we employ a structure identical to that of the DiffusionRig Encoder for the ControlNet to ensure better feature integration. Finally, similar to the structure in ControlNet, we preserve the capabilities of the DiffusionRig model through the implementation of zero convolution.

Video Inference. During the video inference phase, we do not employ the Random Target Strategy which is used in the training process. Instead, we directly extracted the high-frequency feature map from the source image to serve as a supplement for textures in the video synthesis process. In addition, we perform inference with Denoising Diffusion Implicit Model-based (DDIM) [2] iterative denoising steps. The length of the denoising step T is set as 20.

Dataset Processing. We train our **HF-ControlNet** on the VFHQ dataset [3], which consists of 15,204 training videos and 50 test videos. This dataset is commonly used for video super-resolution tasks. In this paper, we obtain approximately 210k frames of dataset through the construction method of the multi-view dataset proposed in the main paper, including 10,000 identities. We present an example of the results obtained from processing the VFHQ video dataset, as shown in Figure 5, where a multi-view dataset comprising 20 frames is extracted from a video of approximately 200 frames. Specifically, the multi-view dataset includes aligned images, high frequency feature maps, masks, albedo, surface normals, and lambertian rendered images. HDTF dataset [5] comprises 396 videos with a total duration of approximately 16 hours. We randomly selected 300 of these to serve as the training dataset for our Audio2FLAME model.

Personal Dataset Collection. As discussed in our main paper, we identify two methods for collecting a personal dataset for a specific individual. These methods can be categorized into extracting images with different poses and expressions from a video, or collecting and downloading images from the Internet. We utilize the dataset depicted in Figure 2 to train a model on the identity of Barack Obama. It is observed that personalized images obtained from videos exhibit superior identity consistency, offering diverse expressions and poses within the same scene. Conversely, images collected from the Internet demonstrate greater variability, encompassing a broader range of lighting conditions and more extensive pose variations. Therefore, we can flexibly choose these two methods for data collection according to different application scenarios.

Table 1: Training and fine-tuning configurations of the proposed two-stage model Control-Talker.

	Stage 1	Stage 2
Image Size	256	
Optimizer	Adam	
Diffusion Steps	1000	
Channels	128	
Channels Multiple	1,1,2,2,4,4	
Attention Resolution	16	
High Frequency Threshold	10	
Batch Size	64	4
Iterations	50k	2k
Learning Rate	4×10^{-5}	4×10^{-6}
Dataset Size	210k frames	20 frames
Device	4 A40	1 A40

A.2 Additional Results

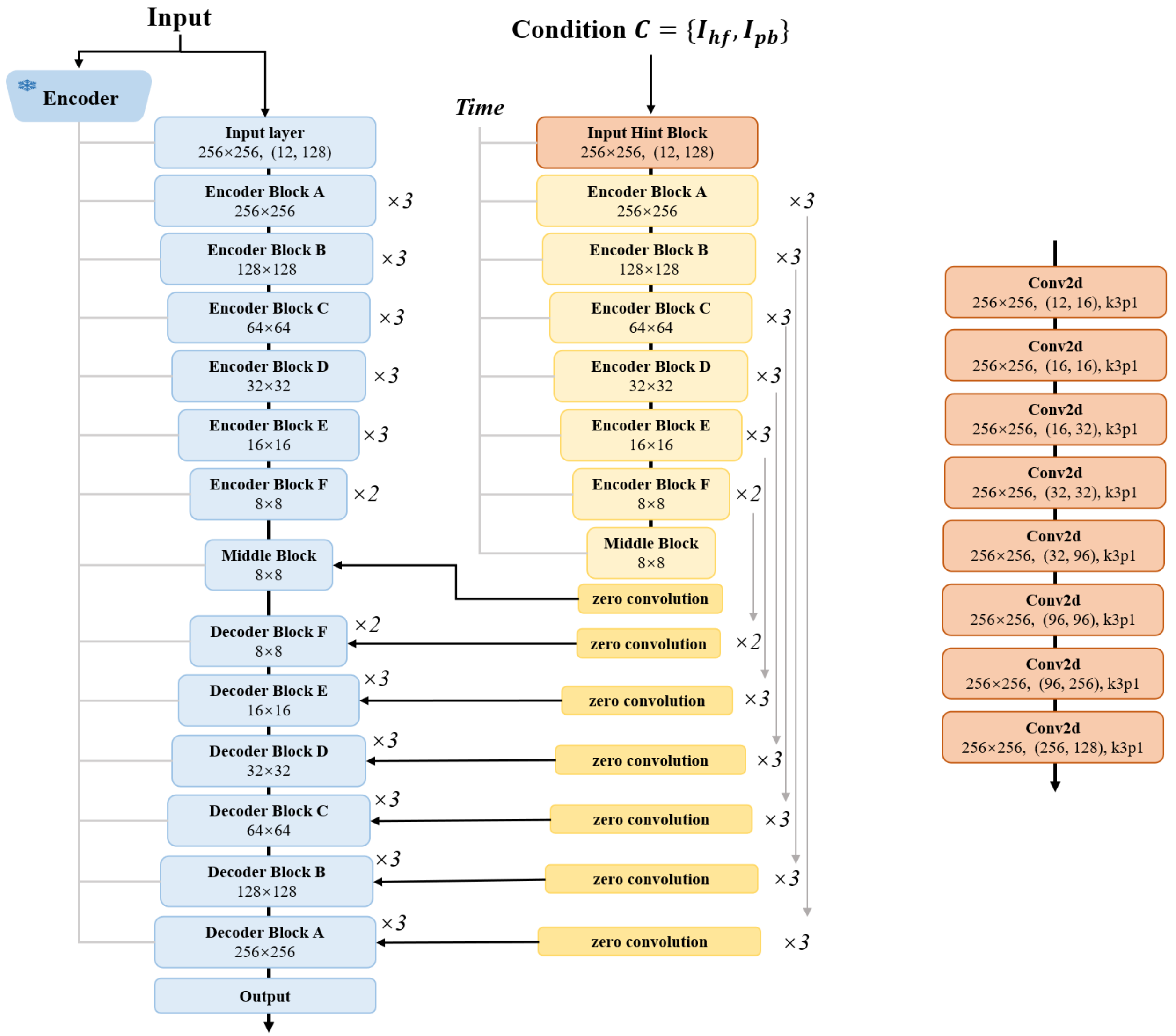
Supplementary Video. We provide a supplementary video that includes the following content:

- Results of the audio-driven talking head generation.
- Results of the video-driven face editing and self-reconstruction.

Please refer to the '.mp4' file in the supplementary materials.

Multi-Condition Face Editing. Our proposed model, Control-Talker, can achieve multi-condition facial editing by manipulating parameters such as expressions, poses, and lighting conditions within the parameter sets $\{\beta, \psi, \theta, \lambda\}$. As shown in Figure 3, we achieve the synthesis of talking head videos under varying lighting conditions by manipulating λ. Therefore, we can readily achieve control over the lighting in talking head videos to accommodate various application scenarios.

Furthermore, we attempt to use different driving videos to individually provide the parameters ψ, θ, and λ, for random combinations of expressions, poses, and lighting conditions. As illustrated in Figure 4, the videos in the first row provide the information for expressions, those in the second row for poses, and the images on the left side for lighting. The synthetic results successfully achieve a combination of multiple conditions while ensuring identity consistency and preserving the textural details of the source image.



(a) DiffusionRig (b) HF-ControlNet (c) Input Hint Block

Figure 1: Detailed architecture of the proposed Control-Talker. 'k3p1' denotes a convolutional kernel with a kernel size of 3 and padding of 1.

REFERENCES

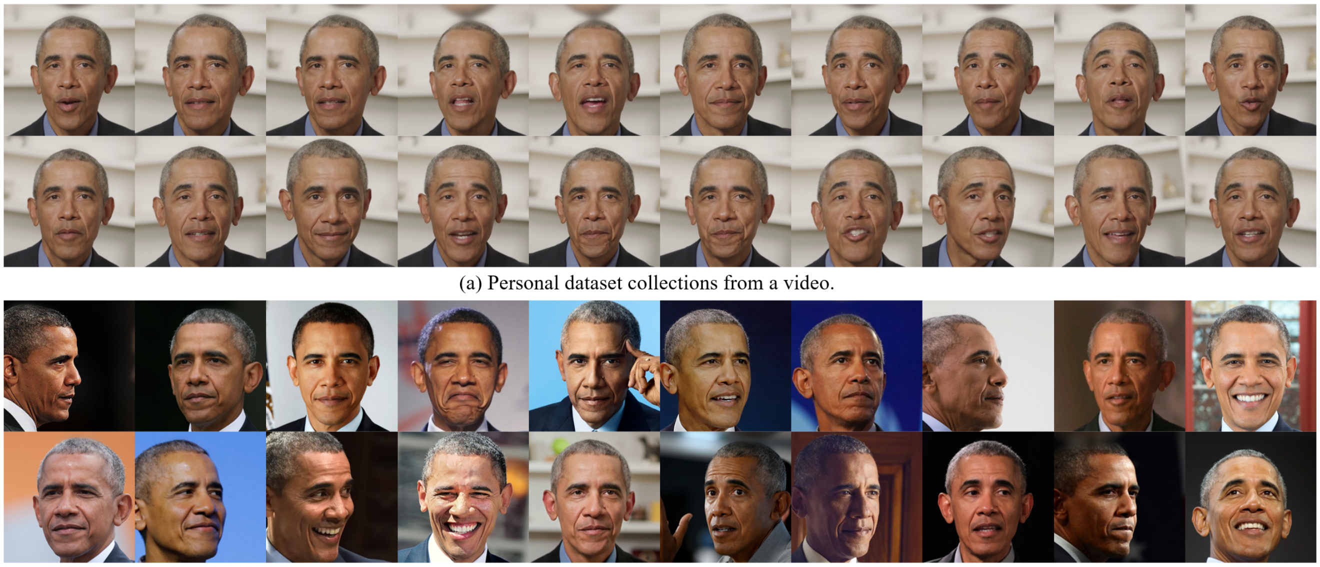
[1] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. 2023. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12736–12746.

[2] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[3] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. 2022. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 657–666.

[4] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[5] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3661–3670.



(a) Personal dataset collections from a video.

(b) Personal dataset collections from internet.

Figure 2: Examples of personal dataset collections.

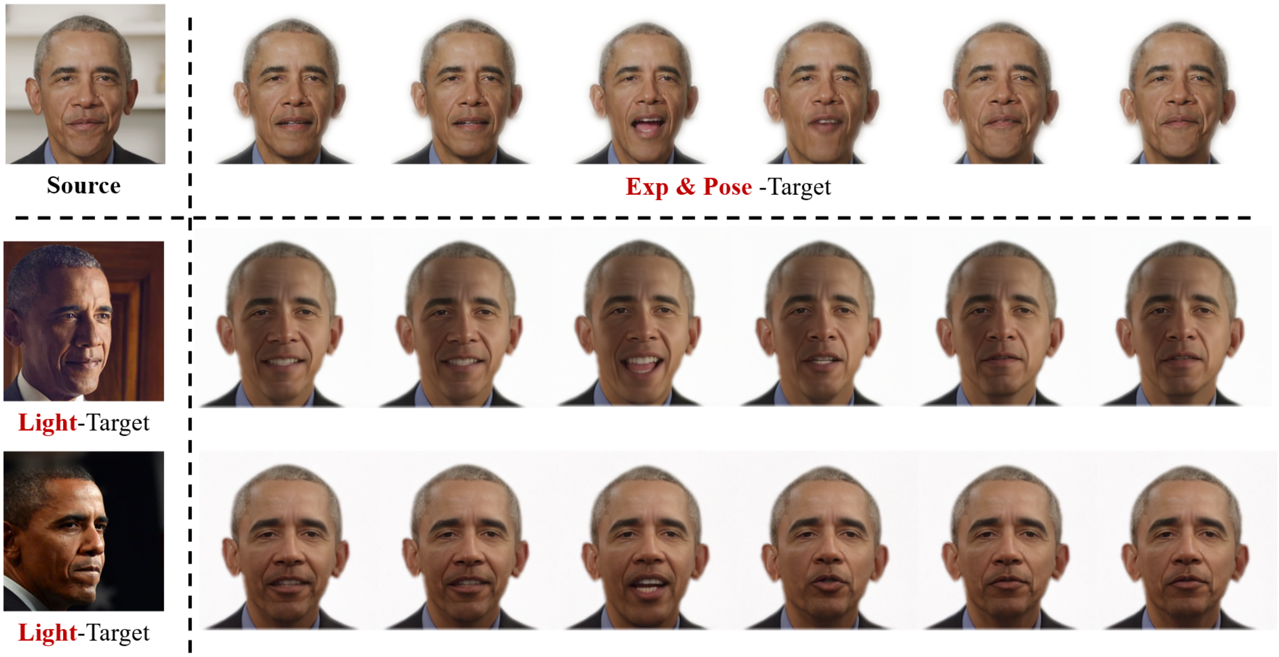


Figure 3: Examples of Mulit-Condition Control 1: Light + Exp & Pose.

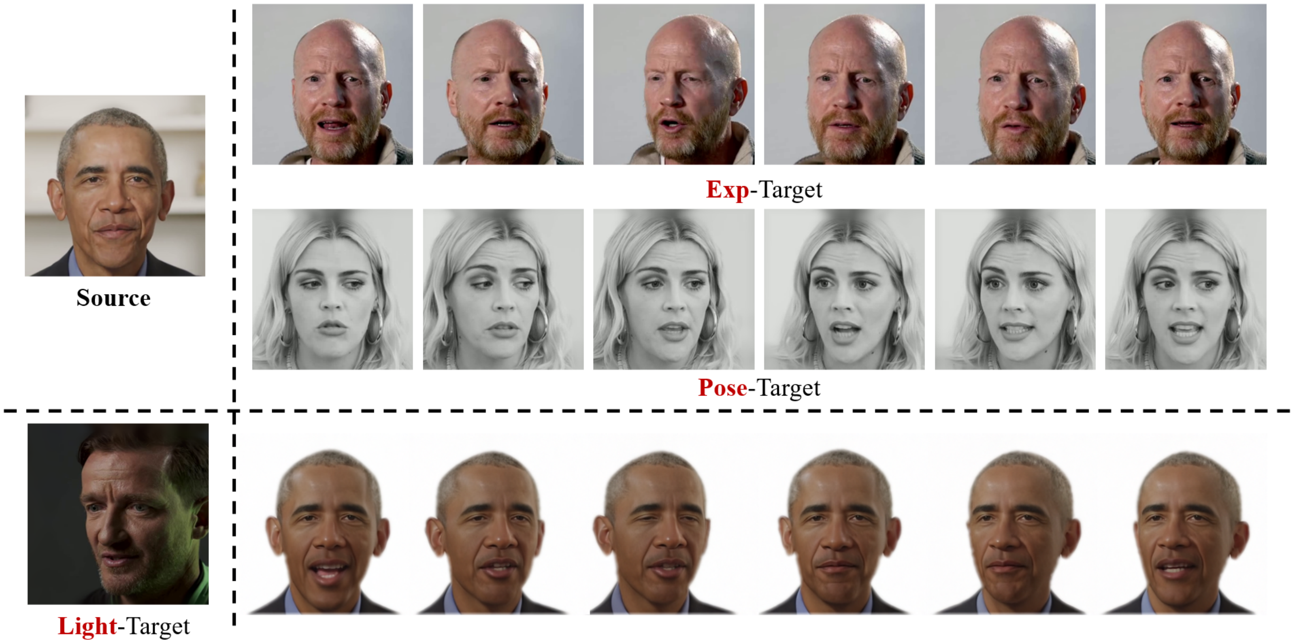
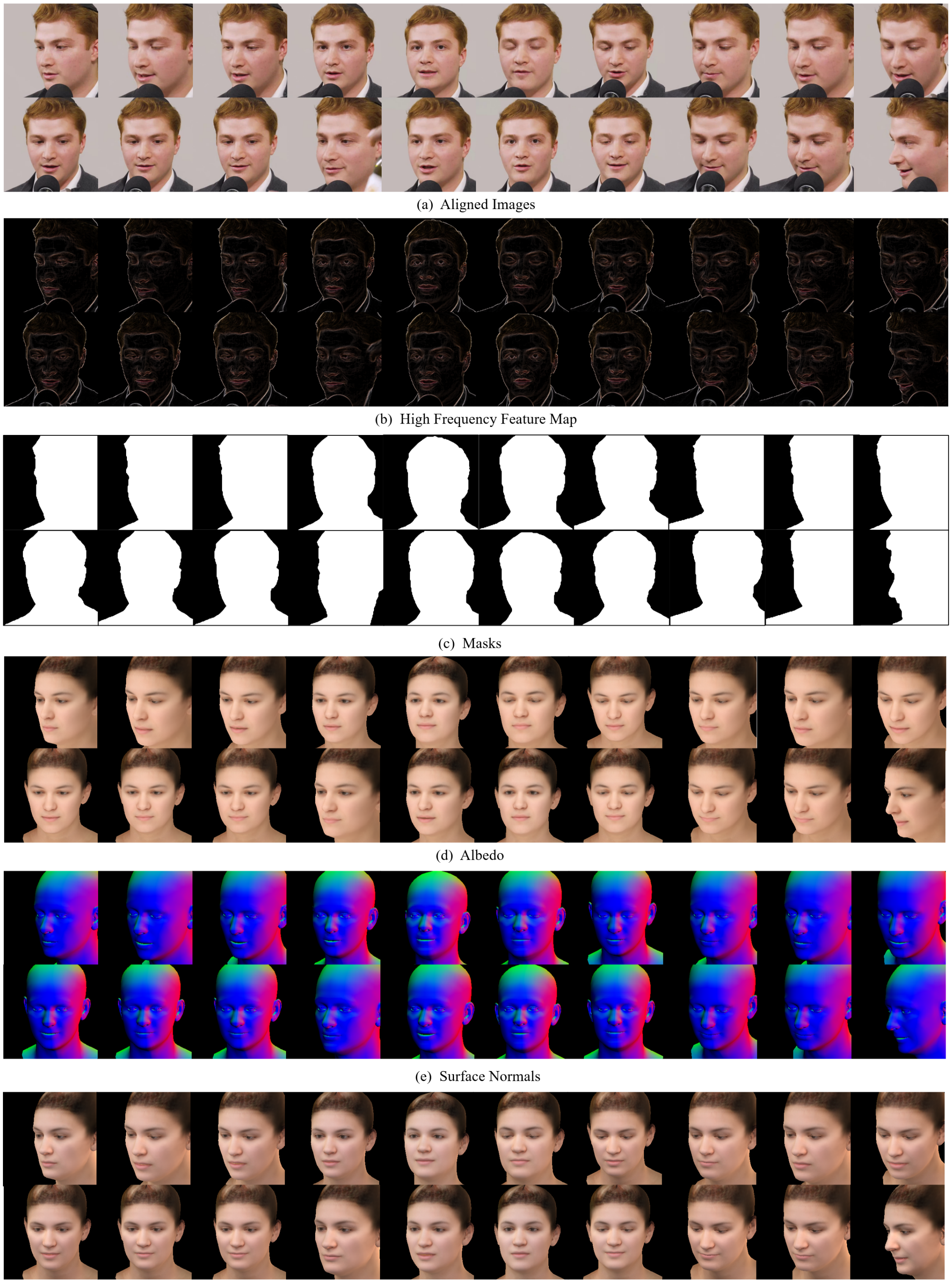


Figure 4: Examples of Mulit-Condition Control 2: Light + Exp + Pose.



(a) Aligned Images

(b) High Frequency Feature Map

(c) Masks

(d) Albedo

(e) Surface Normals

(f) Lambertian Render

Figure 5: Examples of multi-view dataset preprocessing results.