Supplementary Materials of PortraitDirector: A Hierarchical Disentanglement Framework for Controllable and Real-time Facial Reenactment

Chaonan Ji Jinwei Qi Sheng Xu Peng Zhang Bang Zhang
Tongyi Lab

1. Train Details

Our training procedure consists of two stages. In the first stage, we focus exclusively on the face reenactment task. While adopting the core architecture of Wan-Animate [1], our key modification is the introduction of a PoseAdapter to independently inject the pose latent. This PoseAdapter shares the same architecture as the FaceAdapter. Furthermore, we omit the body control module. Consequently, the input to our Diffusion Transformer (DiT) consists solely of the noised ground-truth latent, the temporal latent, and a conditioning mask. Following Wan-Animate, both the FaceAdapter and PoseAdapter are inserted every 5 DiT blocks. The MotionEncoder adopts the same architecture as the LIA encoder [12], while the PoseEncoder is composed of a 4-layer MLP.

Upon completing the first stage, we proceed to the facial decomposition task. During this stage, we freeze the parameters of the pre-trained DiT backbone and the PoseEncoder. The MotionEncoders for individual facial components are initialized using the weights of the pre-trained MotionEncoder. Specifically, the MotionEncoders for the mouth and eyes are jointly optimized, while the MotionEncoder for the emotion remains frozen. We optimize the model for 20 k iterations on the entire dataset.

2. Test Details

To quantitatively evaluate *per-component control*, we measure the discrepancy in facial blendshape coefficients extracted by Mediapipe [4]. Specifically, Mediapipe provides a 52-dimensional expression coefficients, where distinct indices correspond to different facial regions as per its definition:

- **Mouth Control:** Indices [26–48]
- **Eye Control:** Indices [8–21]

Our metric is the L_2 distance between the generated and ground-truth coefficients, calculated exclusively over the subset of indices corresponding to the component being driven. For example, when only the mouth is driven, we compute the distance solely on indices [26–48].

Stage	**Time (ms)**
Data Preprocessing	10
Denoising Loop	327
VAE Decoder	62
Total	**399**

Table 1. Inference time breakdown of our method. The reported time is measured for a single inference pass on a chunk of 8 frames, benchmarked on a single NVIDIA RTX 5090 GPU.

Method	Pose ↑	Mouth ↑	Eye ↑	Emotion ↑
EDTalk	3.52	2.80	–	3.90
PDFGC	2.14	2.53	3.68	3.76
AniPortrait	3.86	–	–	–
Emoji	3.91	–	–	–
Ours	**4.65**	**4.23**	**4.10**	**4.59**

Table 2. User study results on perceived quality. We report the Mean Opinion Score on a 5-point scale (5=best). **Bold** indicates the best-performing method in each category.

3. Real-time Streaming Generation

During the training phase, the model takes video clips of 29 frames as input. All frames are first compressed into a sequence of 8 temporal latents using Wan-VAE [8]. During inference, inspired by TalkingMachines [3], we partition this sequence of latents into chunks to implement a hybrid attention mechanism. Within each chunk, bidirectional attention is employed, allowing latents to access the full local context. Between chunks, however, a causal attention mechanism is enforced to maintain the temporal autoregressive property. To enhance inference efficiency, we adopt a sparse attention pattern. Concretely, for a given latent, its receptive field is constrained to only three sources: 1) other latents within the same chunk, 2) all latents from the immediately preceding chunk, and 3) the reference image latent.

For real-time deployment, our model processes a live camera stream at approximately 20 FPS on a single

Method	PDFGC [9]	EDTalk [7]	AniPortrait [13]	Emoji [5]	Hunyuan Portrait [14]	Fantasy Portrait [11]	XPortrait2 [15]	Ours (vanilla)	Ours
Resolution	224	256	512	512	512	512	512	512	512
FPS	15	35	0.67	0.37	0.47	0.12	0.47	1.8	19

Table 3. Comparison of different methods in terms of resolution and inference speed (FPS). All results are tested on a single A100 GPU (80 G). "Ours (vanilla)" denotes the version of our model without the proposed acceleration strategies.

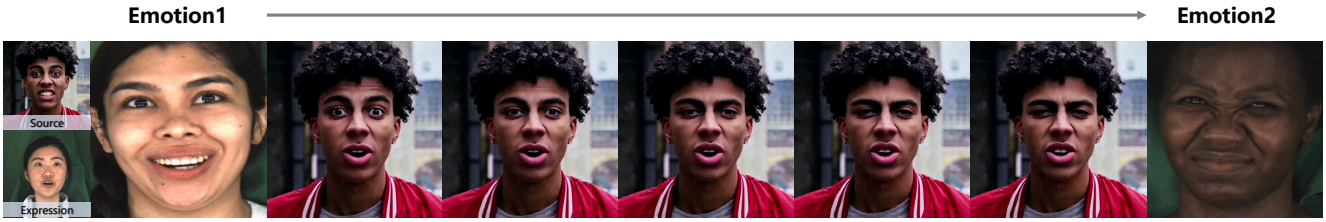

Figure 1. Emotion expression interpolation.

NVIDIA RTX 5090 GPU. The per-module latency for our real-time streaming inference pipeline is detailed in Tab.1. Our model processes inputs in chunks of 8 frames (2 latents) at a time. The 4-step denoising process within the network takes 327 ms, while the VAE decoder requires 62 ms. To accelerate the data preprocessing, we employ MediaPipe [4] to detect and crop the face from the live camera feed. This step yields both the facial bounding box of input image and its corresponding head pose. Crucially, MediaPipe is executed on the CPU to minimize GPU utilization, thereby freeing up resources for our main network. The entire preprocessing pipeline is highly efficient, accounting for only approximately 10 ms. This results in an average processing time of about 50ms per frame, achieving an average output frame rate of approximately 20 FPS. This results in a "glass-to-glass" latency of around 800 ms for interactive tasks like expression reenactment, which includes 400 ms for frame acquisition and 400 ms for model inference. This performance enables effective interactive use cases.

Tab.3 compares the inference efficiency of our method against competing approaches. To ensure a fair comparison, all methods were benchmarked on a single A100 GPU. Enabled by our novel acceleration strategies, our method significantly outperforms its counterparts in terms of efficiency at the same input resolution, thereby achieving real-time performance.

4. User Study

To further evaluate our model's facial component disentanglement capabilities from a human perception standpoint, we conducted a user study focusing on human likeness. For this study, we randomly selected 20 reference images and 20 driving videos, each with a unique identity, from the

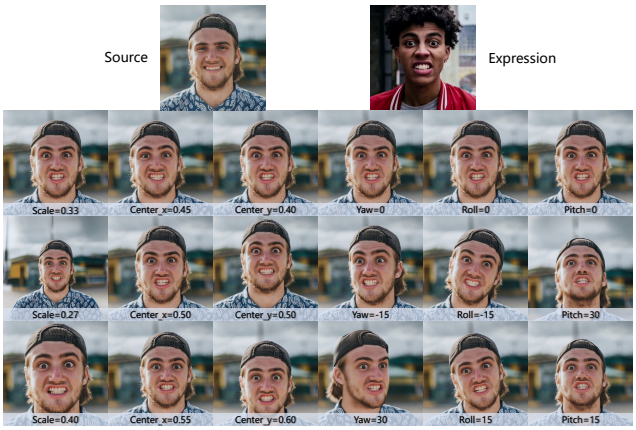

Figure 2. Head Pose Control.

MEAD [10] dataset. We then generated videos where a single facial component from the reference image (e.g., the mouth) was driven by the corresponding component in the driving video, while all other facial parts remained static.

We recruited 20 participants (19 provided valid responses) and asked them to rate the generated results on a 5-point Likert scale (where 1 indicates the worst and 5 the best). The evaluation consisted of two criteria: (1) Component Similarity: how closely the motion of the controlled component in the generated video matches that of the driving video, and (2) Generation Quality: the overall visual quality and realism of the generated video. The mean opinion scores are reported in Tab.2. As the results show, our method achieves higher scores across all controlled components, indicating that it provides more accurate and higher-quality component control in the perception of human observers.

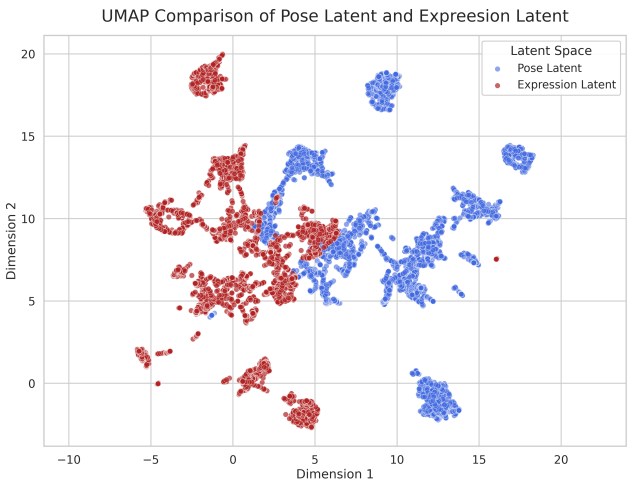

Figure 3. UMAP Visualization of Latent Feature Distributions.

5. Experiments

5.1. Pose and Expression Latent Distribution

To analyze the distributional properties of the latent features, we utilize the pre-trained PoseEncoder and MotionEncoder to extract pose and facial motion latents from 6400 images randomly sampled from the VFHQ dataset. We then project these latents onto a 2D plane using UMAP [6] for visualization. As illustrated in Fig.3, a significant distributional discrepancy is observed between the pose and motion latent spaces. This gap poses a considerable challenge for direct feature fusion, thereby corroborating the rationale behind our design of a dedicated injection pathway for pose information.

5.2. Emotion Interpolation

Our model enables fine-grained emotion expression manipulation by linearly interpolating within the emotion expression latent space. This is formulated as:

$$W = \alpha W_1 + (1 - \alpha) W_2 \tag{1}$$

where W_1 and W_2 represent emotion expression latents from two distinct reference clips, and $\alpha \in [0, 1]$ controls the interpolation. Fig.1 showcases this capability, demonstrating a seamless transition from Emotion1 to Emotion2.

5.3. Head Pose Control

Our explicit decoupling of head pose parameters enables precise, manual control over head orientation and scale. Fig.2 illustrates this unique advantage: we can freely adjust the head's rotation, position and scale while the model faithfully preserves the facial expression. This demonstrates a true disentanglement of pose from expression.

	tLPIPS↓	Temporal Flickering↑	FPS↑	Latency(s)↓
ours(chunk=1)	0.069	86.57%	24	0.36
ours(chunk=2)*	0.022	94.14%	20	0.80
ours(chunk=4)	0.018	95.46%	14	2.30
ours(non-distilled)	0.013	96.74%	1.8	16.11
Xportrait2	0.015	96.06%	0.47	61.70
EDTalk(256x256)	0.022	94.10%	35	0.06

Table 4. Speed–stability comparison (* default). Temporal Flickering is from VBench [2].

5.4. Face Reenactment Results

Fig.4 demonstrates the results of cross-identity face reenactment. Our method achieves performance comparable to existing approaches in terms of identity consistency and expression fidelity. Notably, it surpasses these methods in providing superior control over head pose and scale.

5.5. Fine-Grained Component Control

Fig.5 demonstrates our model's ability to independently control individual facial components. We showcase precise manipulation of the eyes, mouth, head pose, and expression. Notably, this control is maintained even under extreme head poses (up to 90 degrees) and for reference images with diverse visual styles, highlighting the model's robustness and generalization capabilities.

5.6. Speed-Stability Trade-off

We report a speed–stability table (Tab. 4) under matched settings by varying the chunk size to quantify the efficiency–stability trade-off. The results show that our non-distilled model achieves comparable efficiency and stability to prior methods, and under the same real-time budget, our streaming variant attains the best temporal stability.

References

[1] Gang Cheng, Xin Gao, Li Hu, Siqi Hu, Mingyang Huang, Chaonan Ji, Ju Li, Dechao Meng, Jinwei Qi, Penchong Qiao, Zhen Shen, Yafei Song, Ke Sun, Linrui Tian, Feng Wang, Guangyuan Wang, Qi Wang, Zhongjian Wang, Jiayu Xiao, Sheng Xu, Bang Zhang, Peng Zhang, Xindi Zhang, Zhe Zhang, Jingren Zhou, and Lian Zhuo. Wan-animate: Unified character animation and replacement with holistic replication. *CoRR*, abs/2509.14055, 2025. 1

[2] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 3

[3] Chetwin Low and Weimin Wang. Talkingmachines: Real-

Figure 4. Face reenactment results.

time audio-driven facetime-style video via autoregressive diffusion models. *CoRR*, abs/2506.03099, 2025. 1

[4] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019. 1, 2

[5] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, and Qifeng Chen. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers, SA 2024, Tokyo, Japan, December 3-6, 2024*, pages 110:1–110:12. ACM, 2024. 2

[6] Leland McInnes, John Healy, and James Melville. Umap:

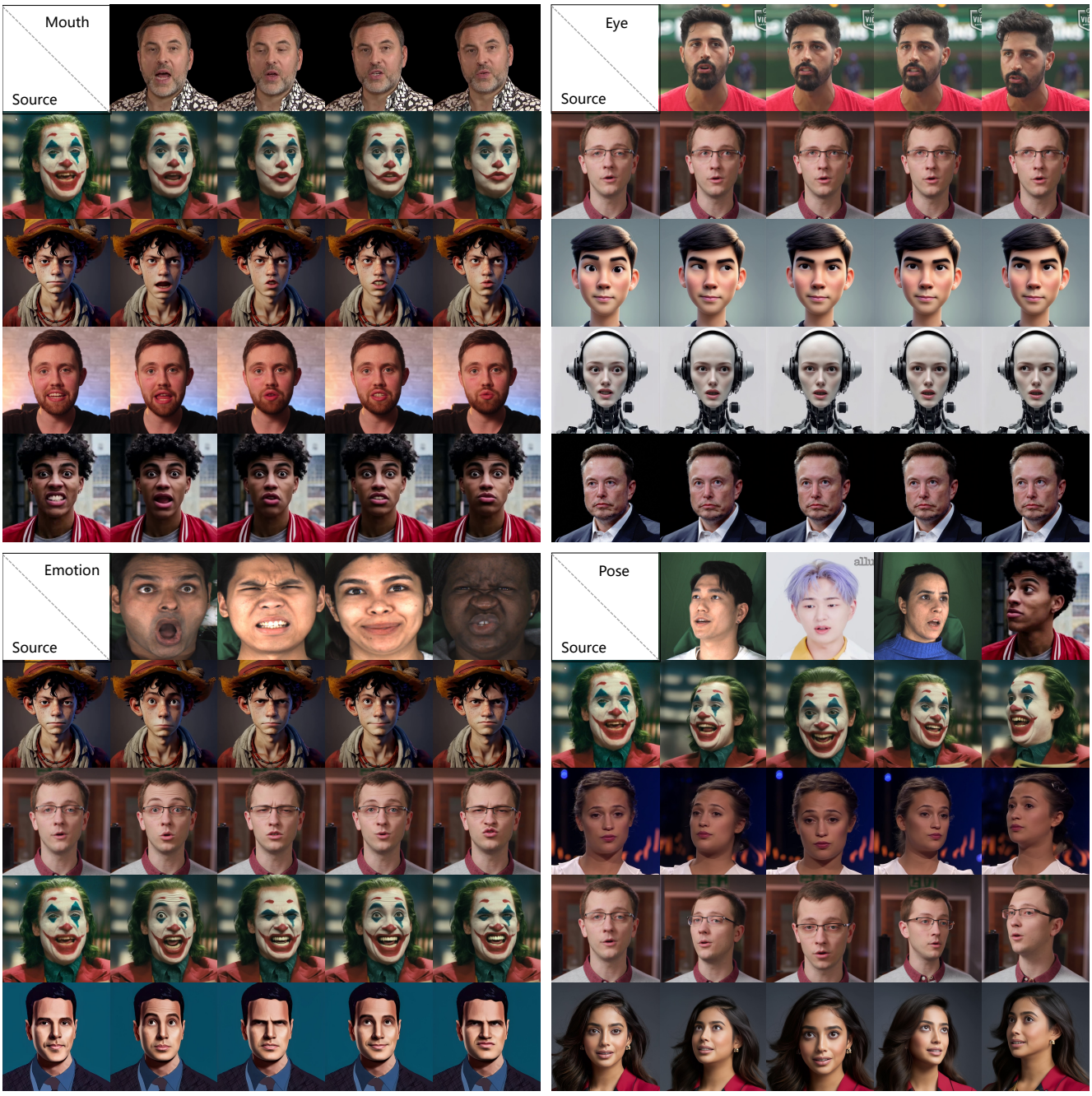

Figure 5. Fine-Grained facial component control.

Uniform manifold approximation and projection for dimension reduction, 2020. 3

[7] Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, pages 398–416. Springer, 2024. 2

[8] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yi-

tong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*, abs/2503.20314, 2025. 1

[9] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 17979–17989. IEEE, 2023. 2

[10] 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 *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*, pages 700–717. Springer, 2020. 2

[11] Qiang Wang, Mengchao Wang, Fan Jiang, Yaqi Fan, Yonggang Qi, and Mu Xu. Fantasyportrait: Enhancing multi-character portrait animation with expression-augmented diffusion transformers. *CoRR*, abs/2507.12956, 2025. 2

[12] Yaohui Wang, Di Yang, François Brémond, and Antitza Dantcheva. LIA: latent image animator. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(12):10829–10844, 2024. 1

[13] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *CoRR*, abs/2403.17694, 2024. 2

[14] Zunnan Xu, Zhentao Yu, Zixiang Zhou, Jun Zhou, Xiaoyu Jin, Fa-Ting Hong, Xiaozhong Ji, Junwei Zhu, Chengfei Cai, Shiyu Tang, Qin Lin, Xiu Li, and Qinglin Lu. Hunyuanportrait: Implicit condition control for enhanced portrait animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 15909–15919. Computer Vision Foundation / IEEE, 2025. 2

[15] Xiaochen Zhao, Hongyi Xu, Guoxian Song, You Xie, Chenxu Zhang, Xiu Li, Linjie Luo, Jinli Suo, and Yebin Liu. X-nemo: Expressive neural motion reenactment via disentangled latent attention. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 2