

Supplementary Materials: Highly Dynamic and Realistic Portrait Image Animation with Diffusion Transformer Networks

Anonymous CVPR submission

Paper ID 11345

1. Network

1.1. Identity Reference Network Details

Maintaining subject consistency in diffusion transformer based video generation presents a significant challenge, particularly as video length increases. While integrating speech audio embeddings as conditional features aids in aligning facial movements with audio, prolonged generation often results in a degradation of facial identity fidelity.

To address this issue, we propose an identity reference network within the diffusion transformer framework designed to preserve facial identity coherence in realistic portrait animation. Figure 1 illustrates various strategies for identity preservation after 10 seconds of video generation:

(a) **No Identity Condition.** In the absence of identity conditions, the model struggles to maintain adequate portrait coherence after 10 seconds.

(b) **Face Attention.** Incorporating features from the face encoder InsightFace [3] into the cross-attention module effectively captures high-level features—such as the appearance of age indicated by wrinkles in the reference image. However, this method still results in noticeable alterations to the subject’s appearance.

(c) **Face Adaptive Norm.** Here, face embeddings obtained from InsightFace [3] are injected via an adaptive layer normalization technique. However, this approach also fails to preserve the subject’s identity by emphasizing overall visual context at the expense of specific portrait features, potentially leading to distortion.

(d) **Identity Reference Network.** Our proposed identity reference network comprises several transformer blocks, each containing adaptive layer normalization layers, a 3D full attention layer, and a feed-forward layer. We first employ a 3D VAE to encode the reference image, then input these latent features into the identity reference network to extract reference image features. These features are concatenated with the input features of the 3D full-attention layer in the denoising network, allowing the reference image features to be injected through the 3D full-attention module. Our identity reference network effectively encodes

reference images, preserving detailed identity and background features (e.g., the text “CPS”). However, it tends to introduce a smoothing effect that compromises finer details, such as wrinkles in the portrait.

(e) **Face Attention and Identity Reference Network.** Finally, we combine the identity reference network with the face encoder to incorporate higher-level semantic features. This integration enhances the portrait’s characteristic attributes while maintaining identity fidelity.

1.2. Training Details

The training process comprises two phases:

(1) **Identity Consistency Phase.** In this initial phase, we train the model to generate videos with consistent identity. The parameters of the 3D Variational Autoencoder (VAE) and face image encoder remain fixed, while the parameters of the 3D full attention blocks in both the reference and denoising networks, along with the face attention blocks in the denoising network, are updated during training. The model’s input includes a randomly sampled reference image from the training video, a textual prompt, and the face embedding. The textual prompt is generated using MiniCPM [14], which describes human appearance, actions, and detailed environmental background. The face embedding is extracted via InsightFace [3]. With these inputs, the model generates a video comprising 49 frames.

(2) **Audio-Driven Video Generation Phase.** In the second phase, we extend the training to include audio-driven video generation. We integrate audio attention modules into each transformer block of the denoising network, while fixing the parameters of other components and updating only those of the audio attention modules. Here, the model’s input consists of a reference image, an audio embedding, and a textual prompt, resulting in a sequence of 49 video frames driven by audio.

**Implementation Details.** We initialize the identity reference and denoising networks with weights derived from CogVideoX-5B-I2V [13]. During both training phases, we employ the v-prediction diffusion loss [6] for optimization. Each training phase comprises 20,000 steps, utilizing 64

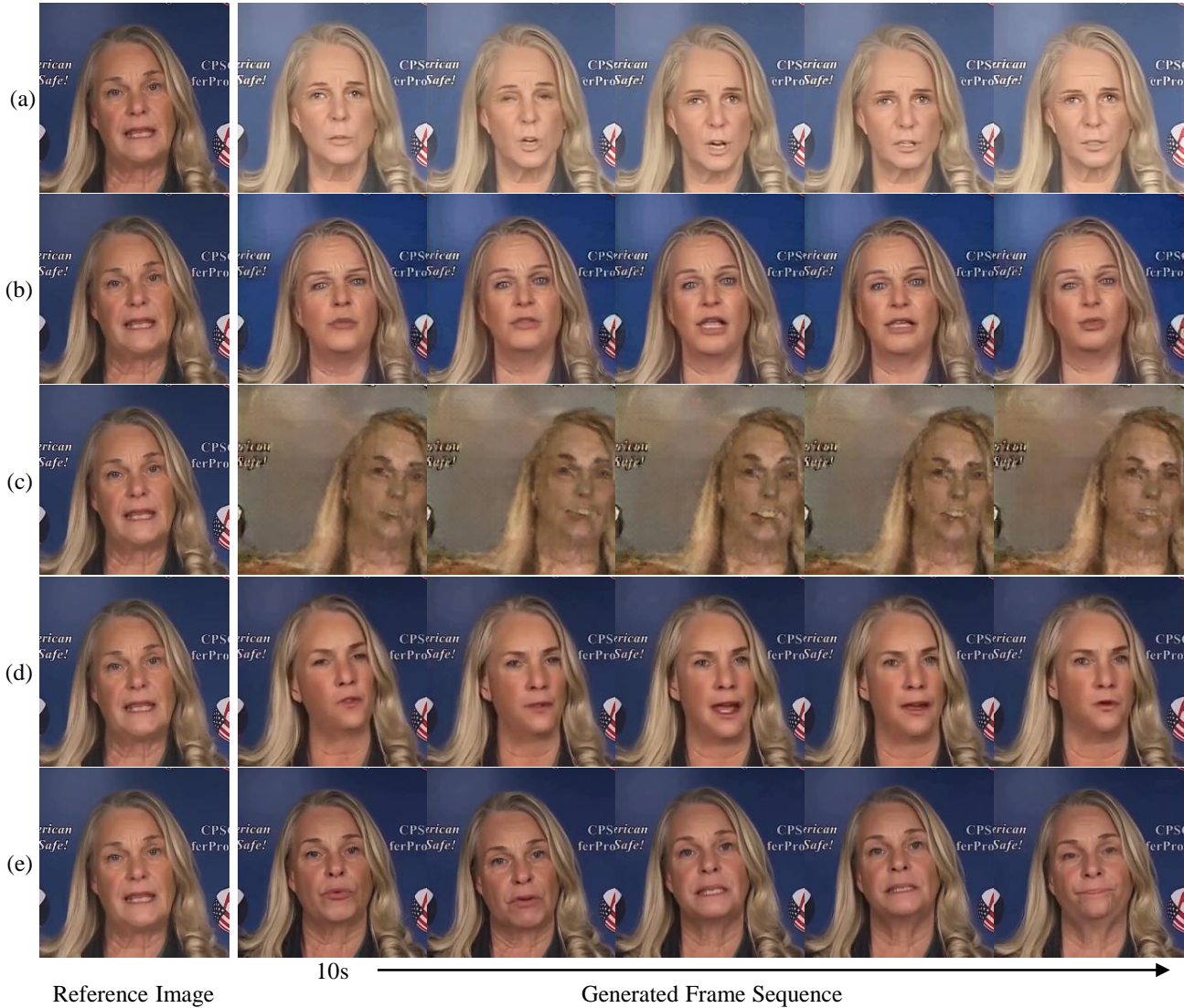


Figure 1. Qualitative study of different setting of the identity injection method. Selected frames begin at 10s. (a) **No identity condition**: No specific conditions are applied to control the subject’s appearance; (b) **Face attention**: Identity features from InsightFace [3] are processed through a cross-attention module; (c) **Face adaptive norm**: Identity features from InsightFace [3] are incorporated via adaptive layer normalization, applied through scaling and shifting.; (d) **Identity reference network**: Features are encoded using a reference network and integrated within the 3D full attention module.; (e) **Face attention and identity reference network**: Identity features are encoded separately via the InsightFace [3] and the reference network. These encoded features are then integrated within the 3D full attention module and processed through a cross-attention mechanism.

NVIDIA A100 GPUs. The batch size per GPU is set to 1, with a learning rate of  $1 \times 10^{-5}$ . The resolution of the training videos is 480 x 720 pixels. To enhance video generation variability, the reference image, guidance audio, and textual prompt are dropped with a probability of 0.05 during training.

## 2. Experiments

### 2.1. Experimental Setup Details

**Evaluation Metrics.** We employed a range of evaluation metrics for generated videos across benchmark

datasets, including HDTF and Celeb-V. These metrics comprise Fréchet Inception Distance (FID) [7], Fréchet Video Distance (FVD) [10], Synchronization-C (Sync-C) [1], Synchronization-D (Sync-D) [1], and E-FID [9]. FID and FVD quantify the similarity between generated images and real data, while Sync-C and Sync-D assess lip synchronization accuracy. E-FID evaluates image quality based on features extracted from the Inception network.

Besides, we introduced V-bench [4] metrics to enhance evaluation, focusing on dynamic degree and subject consistency. Dynamic degree is measured using RAFT [8] to quantify the extent of motion in generated videos, provid-

	Audio	Text	Image	Sync-C $\uparrow$	Sync-D $\downarrow$	Subject Dynamic $\uparrow$	Background Dynamic $\uparrow$	Subject FVD $\downarrow$	Background FVD $\downarrow$	Subject Consistency $\uparrow$
$\lambda_t \downarrow$	$\lambda_a = 3.5$	$\lambda_t = 1.0$	$\lambda_i = 1.0$	6.168	8.589	13.164	3.955 $\downarrow$	361.582	263.416	0.9813
Base	$\lambda_a = 3.5$	$\lambda_t = 3.5$	$\lambda_i = 1.0$	6.154	8.574	13.286	4.481	359.493	248.283	0.9810
$\lambda_t \uparrow$	$\lambda_a = 3.5$	$\lambda_t = 6.0$	$\lambda_i = 1.0$	6.044	8.861	13.616	4.659 $\uparrow$	342.894	235.307	0.9808
$\lambda_a \uparrow$	$\lambda_a = 6.0$	$\lambda_t = 3.5$	$\lambda_i = 1.0$	6.469 $\uparrow$	8.515	14.778	4.066	379.073	264.969	0.9809
$\lambda_i \uparrow$	$\lambda_a = 3.5$	$\lambda_t = 3.5$	$\lambda_i = 3.5$	6.023	8.654	12.599	4.219	367.225	265.414	0.9835 $\uparrow$

Table 1. Quantitative study of audio, text and image CFG scales on our proposed wild dataset.

ing a comprehensive assessment of temporal quality. Subject consistency is measured through DINO feature similarity, ensuring uniformity of a subject’s appearance across frames.

**Baseline Approaches.** We considered several representative audio-driven talking face generation methods for comparison, all of which have publicly available source code or implementations. These methods include SadTalker [15], DreamTalk [5], AniPortrait [11], and Hallo [2, 12]. The selected approaches encompass both GANs and diffusion models, as well as techniques utilizing intermediate facial representations alongside end-to-end frameworks. This diversity in methodologies allows for a comprehensive evaluation of the effectiveness of our proposed approach in comparison to existing solutions.

## 2.2. Ablation and Discussion

**CFG Scales for Diffusion Model.** Table 1 provides a quantitative analysis of video generations using various CFG scales for audio, text, and reference images. A comparison between the second and fourth rows demonstrates that increasing the audio CFG scale enhances the model’s ability to synchronize lip movements. The text CFG scale significantly influences the video’s dynamism, as indicated in the first three rows, where both the subject’s and the background’s dynamics increase with higher text CFG scales. Conversely, the reference image CFG scale primarily governs the subject’s appearance; higher values improve subject consistency, as illustrated by the second and fifth rows. Among the tested configurations, setting  $\lambda_a = 3.5$ ,  $\lambda_t = 3.5$ , and  $\lambda_i = 1.0$  yields a balanced performance. This interplay between visual fidelity and dynamics underscores the effectiveness of CFG configurations in generating realistic portrait animations.

## 2.3. Generation Controllability

**Textual Prompt for Subject Animation.** To evaluate whether textual conditional controllability is effectively preserved, we conducted a series of experiments comparing the performance of our method to that of the baseline model, CogVideoX [13], using same text prompts. As shown in Figure 2, the white number represents the BLIP score, which measures how well the generated videos align

with the textual prompts. A higher score indicates better alignment. The results shows that our model maintains its ability for textual control, achieving a BLIP score comparable to that of CogVideX, and effectively captures the interaction between different subjects as dictated by the textual prompts.

**Textual Prompt for Foreground and Background Animation.** We also explore model’s ability to follow the foreground and background textual prompt. As illustrated in Figure 3, our method animates the foreground and background subjects naturally, such as the ocean waves and flickering candlelight. The results demonstrates the model’s ability to control foreground, and background with the textual caption, which is maintained even after introducing the audio condition.

## 2.4. Limitations and Future Works.

Despite the advancements in portrait image animation techniques presented in this study, several limitations warrant acknowledgment. While the proposed methods improve identity preservation and lip synchronization, the model’s ability to realistically represent intricate facial expressions in dynamic environments still requires refinement, especially under varying illumination conditions. Future work will focus on enhancing the model’s robustness to diverse perspectives and interactions, incorporating more comprehensive datasets that include varied backgrounds and facial accessories. Furthermore, investigating the integration of real-time feedback mechanisms could significantly enhance the interactivity and realism of portrait animations, paving the way for broader applications in live media and augmented reality.

## 3. Safety Considerations.

The advancement of portrait image animation technologies, particularly those driven by audio inputs, presents several social risks, most notably concerning the ethical implications associated with the creation of highly realistic portraits that may be misused for deepfake purposes. To address these concerns, it is essential to develop comprehensive ethical guidelines and responsible use practices. Moreover, issues surrounding privacy and consent are prominent when utilizing individuals’ images and voices. It is imperative to establish transparent data usage policies, ensuring



Figure 2. Condition on Interactive Subjects. The white number represents the BLIP score, which measures the alignment between the generated video and the textual prompts. A higher value indicates a better alignment. Our method achieves alignment comparable to that of CogVidEX, maintaining the controllability of interactive subjects even after introducing the audio condition.

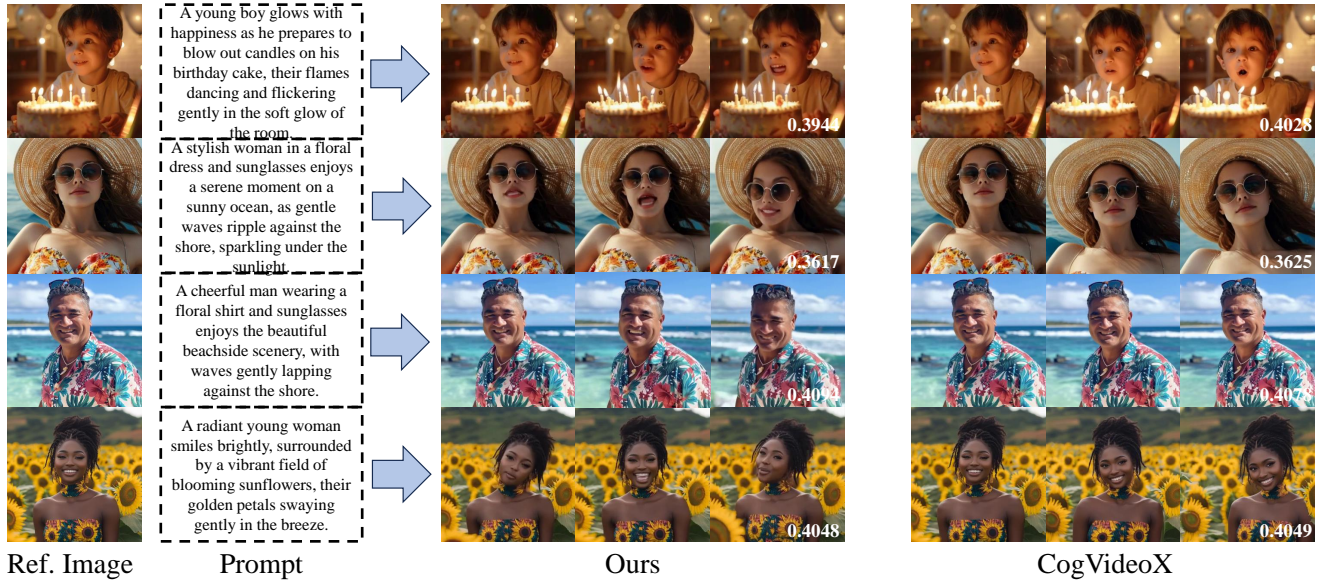


Figure 3. Textual Condition on Foreground and Background. The white number represents the BLIP score, which measures the alignment between the generated video and the textual prompts. A higher value indicates a better alignment. Our method achieves alignment comparable to that of CogVidEX, maintaining the controllability of foreground and background after incorporating the audio condition.

that individuals provide informed consent and that their privacy rights are fully protected. By acknowledging these risks and implementing appropriate mitigation strategies, this research aims to promote the responsible and ethical development of portrait image animation technology.

References

[1] J. S. Chung and A. Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. 2

[2] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024. 3

[3] DeepInsight. Insightface: An open-source 2d and 3d deep face analysis toolkit. <https://github.com/deepinsight/insightface>, 2024. 1, 2

[4] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In

*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[5] Yifeng Ma, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yingya Zhang, and Zhidong Deng. Dreamtalk: When expressive talking head generation meets diffusion probabilistic models. *arXiv preprint arXiv:2312.09767*, 2023. 3

[6] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 1

[7] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>, August 2020. Version 0.3.0. 2

[8] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 2

[9] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive-generating expressive portrait videos with audio2video diffusion model under weak conditions. *arXiv preprint arXiv:2402.17485*, 2024. 2

[10] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 2

[11] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 3

[12] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation, 2024. 3

[13] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 3

[14] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 1

[15] 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. 3