

– *Supplementary Material* –

Audio-visual Controlled Video Diffusion with Masked Selective State Spaces Modelling for Natural Talking Head Generation

Fa-Ting Hong^{1,2} Zunnan Xu^{2,3} Zixiang Zhou² Jun Zhou²
Xiu Li³ Qin Lin² Qinglin Lu² Dan Xu^{1,✉}

¹HKUST ²Tencent ³Tsinghua University



A. Experiment Detail

A.1. Implementation.

During the training process, we resize all images and videos to 640×640 . To optimize the framework, we use the AdamW optimizer with a learning rate of 1×10^{-5} . The Identity encoder [4] and VAE [7] are kept fixed, with their weights initialized from Stable Video Diffusion [1]. During training, we randomly select the gate states in the parallel-control mamba layer and manually set them during inference to enable flexible control.

A.2. Metrics

In this work, we evaluate our method and compare it with other approaches using comprehensive quantitative metrics. Our evaluation framework consists of three main categories: (1) audio-visual synchronization, (2) visual quality assessment, and (3) facial motion accuracy. Additionally, we assess temporal smoothness and identity preservation through specialized measures.

Audio-Visual Synchronization. We employ **Sync-C** (synchronization confidence) and **Sync-D** (synchronization distance) metrics from wav2lip[10] using a pretrained Sync-Net [2]. Sync-C measures the confidence level of lip-audio alignment through classifier outputs, where higher values indicate better synchronization. Sync-D calculates the L2 distance between audio and visual features, with lower values representing superior alignment.

Visual Quality Assessment. We utilize four complementary metrics:

- **PSNR:** Peak Signal-to-Noise Ratio quantifies pixel-level fidelity through a logarithmic decibel scale, where higher values reflect better reconstruction accuracy.
- **SSIM** [14]: Structural Similarity Index measures structural information preservation between generated and reference frames, ranging from 0 to 1, with higher values

indicating better quality.

- **LPIPS** [17]: Learned Perceptual Image Patch Similarity evaluates perceptual differences using VGG [12] features:

$$\text{LPIPS}(I_{gt}^t, I_{gen}^t) = \sum_l w_l \|\mathbf{F}_l(I_{gt}^t) - \mathbf{F}_l(I_{gen}^t)\|_2 \quad (1)$$

- **Fréchet Inception Distance (FID)** [5]: Measures feature distribution similarity between generated and real images using Inception-v3 features, with lower scores indicating better perceptual quality.
- **Fréchet Video Distance (FVD)** [13]: Assesses temporal coherence through pretrained network features:

$$\text{FVD} = \|\mu_{gen} - \mu_{gt}\|^2 + \text{Tr}(\Sigma_{gen} + \Sigma_{gt} - 2(\Sigma_{gen}\Sigma_{gt})^{1/2}) \quad (2)$$

Facial Motion Accuracy. For expression and pose evaluation:

- **Landmark Mean Distance (LMD):** Computes the average L2 distance between facial landmarks [8] of generated and reference frames, with lower values indicating better geometric accuracy.
- **Pose Distance:** Measures head pose discrepancies using EMOCA [3]-derived parameters through the mean L1 distance between generated and driving frames.
- **Expression Similarity:** Calculates the cosine similarity of expression parameters from EMOCA [3], with higher values indicating better emotional consistency.

Identity Similarity. We employ ArcFace [4] scores to measure identity similarity between generated frames and reference images through deep face recognition features, where higher scores indicate better identity preservation.

Temporal Smoothness. We evaluate motion temporal smoothness by computing the optical flow consistency using VBench metrics [6], where lower variance in motion vectors indicates smoother transitions.

A.3. Mask Design

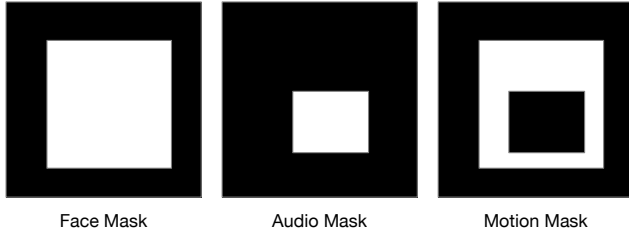


Figure 1. The type of masks we used in our framework.

In our framework, we utilize masks to indicate the control regions of each signal. Three types of masks are used in our framework. As illustrated in Figure 1, the face mask is used to indicate the rough position of the face in the source image. During training, we use RetinaFace [11] to calculate the bounding box for all frames in the ground truth segments and obtain the smallest enclosing rectangle of these bounding boxes. We then draw the face mask based on that rectangle to indicate the facial location in the desired video. Similarly, the audio mask is obtained by detecting the mouth bounding boxes, and the motion mask is generated by using the face mask to minimize the audio mask. During the inference stage, we detect the bounding box of the source image and apply the appropriate extension.

B. Visualization

B.1. Face reenactment

Figure 2 demonstrates that our approach achieves enhanced precision in replicating portrait motions that align closely with the driving video’s dynamics. For self-reenactment, the results generated by our framework better preserve intricate facial behaviors, particularly in eye movement patterns, ocular orientation, and lip articulation accuracy.

As illustrated in the third, fifth, and sixth rows of Figure 2, our method can achieve tracking of the overall rotation of the head, which cannot be achieved by previous top-performing warping-based methods, such as LivePortrait.

While previous diffusion-based methods demonstrate notable advantages in output fidelity, their reliance on facial keypoint tracking introduces limitations. As shown in the third and fifth rows of Figure 2, discrepancies in facial geometry between source and target identities, combined with the inherent limitations of keypoint representations in capturing detailed facial expressions, make previous state-of-the-art methods (e.g., AniPortrait [15], Follow Your Emoji [9]) less effective than our method in reconstructing facial contours, gaze direction, and lip synchronization accuracy. These keypoint-dependent methodologies remain susceptible to interference from driving video subjects’ facial geometries, resulting in incomplete motion-

identity separation. These methods face challenges in identity preservation due to changes in facial geometry resulting from misalignment of key points. Our framework overcomes these limitations through a parallel-control mamba layer (PCM), with an improved separation of facial identity characteristics from motion parameters, as evidenced in Figure 2. This enhanced decoupling enables superior identity retention while capturing nuanced facial dynamics. Although X-Portrait [16] utilizes a non-explicit keypoint control method, it does not completely decouple motion and appearance information. This limitation results in noticeable flaws in the generated results, particularly evident in the fourth and fifth lines of Figure 2.

Moreover, frameworks built upon Stable Diffusion’s image generation architecture typically under-perform our method in temporal coherence metrics. By integrating the stable video diffusion model with our framework, we achieve significant improvements in three critical aspects: identity consistency preservation, visual quality optimization, and micro-expression reproduction. This collectively produces more natural-looking and temporally stable animations. We provide video demos in the supplementary materials. In these video demos, we compare our method with other methods, and our method obviously achieves better results. Additionally, we found that when some of the reference images provide more details, the results can be even more realistic.

B.2. Audio Driven Talker Head Generation

We present a comprehensive comparison with all baseline methods in Figure 3. As shown in the figure, our method is able to produce accurate lip motion while containing fewer artifacts. Notably, our method generates natural head poses and expressions similar to the ground truth (please refer to the video demo in *Supplementary Material*), whereas other methods mainly manipulate the mouth shape and leave other regions static. These results confirm that our mamba design effectively aggregates audio signals with facial tokens to produce natural expressions and accurate lip synchronization, as we use the face mask as an audio mask to incorporate nearly all facial tokens in an audio-driven manner.

B.3. Audio-visual Joint Driven

We also present additional demonstrations in Figure 4, which displays the results produced by our method under audio-visual joint control. Our approach effectively maintains lip synchronization with the audio while accurately reflecting the expressions of the Motion Driving sources. We highly recommend watching the video demonstrations. Additional results can be found in the *supplementary materials*, where video demonstrations are also available.

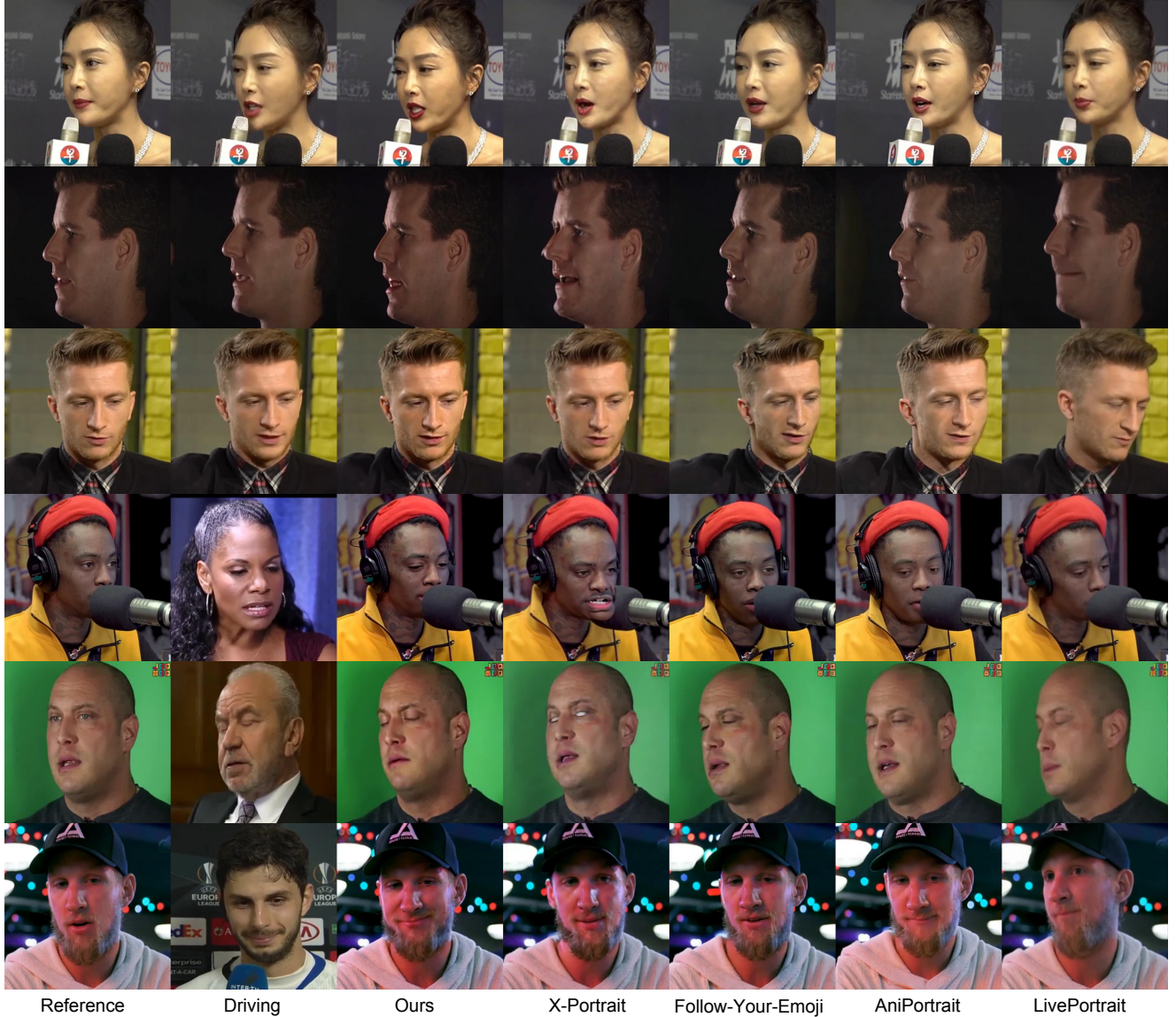


Figure 2. The results generated by our method under facial motion control.

B.4. User study.

To validate our method’s perceptual benefits, we conducted a user study (as shown in Table. 1) comparing six methods on lip-sync, naturalness, and quality. Participants selected the best video per criterion. As shown in the table below, our method was consistently preferred by a wide margin, demonstrating clear superiority.

C. Ethics Considerations and AI Responsibility

This study aims to develop artificial intelligence-driven virtual avatars with enhanced visual emotional expression ca-

pabilities, utilizing audio or visual inputs, for applications in positive and constructive domains. The technology is designed specifically for ethical purposes, focusing on applications that are beneficial to society, and is not intended for generating deceptive or harmful media content.

However, as with all generative approaches in this field, there remains a theoretical concern about potential misuse for identity replication or malicious purposes. The research team strongly condemns any attempts to use the technology for creating fraudulent, harmful, or misleading representations of real individuals. Rigorous technical evaluations of the current system indicate that the generated outputs exhibit clear artificial features, and quantitative comparisons

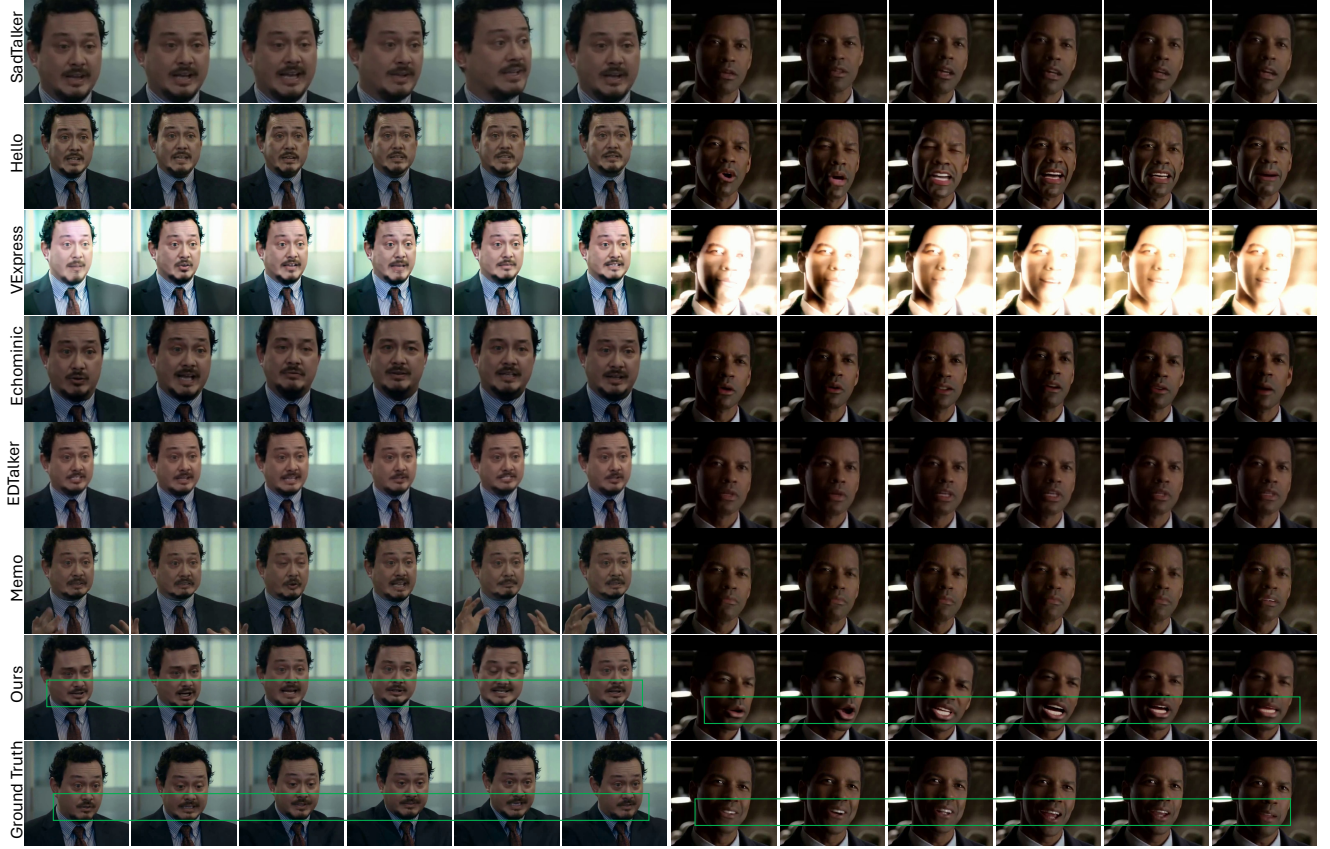


Figure 3. The results generated by our method under audio control.

Criterion	Ditto-Talking (%)	Echomimic (%)	Hallo (%)	Memo (%)	SadTalker (%)	VExpress (%)	Ours (%)
Lip-sync	0.00	5.89	12.74	14.71	1.96	3.92	60.78
Naturalness	1.11	6.67	11.11	12.22	0.00	0.00	68.89
Quality	2.78	5.56	11.11	15.74	0.00	0.00	64.81

Table 1. The user studies.

with genuine human recordings show measurable discrepancies, ensuring that the results remain distinguishable from authentic human expressions.

References

- [1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1
- [2] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *ACCV Workshops*, 2017. 1
- [3] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *CVPR*, 2022. 1
- [4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 1
- [5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017. 1
- [6] Ziqi Huang, Yanan 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 *CVPR*, 2024. 1
- [7] D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 1
- [8] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuoling Chang, Ming Guang Yong, Juhyun Lee, et al. Medi-



Figure 4. The results generated by our method under audio-visual joint control.

- apipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 1
- [9] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. *arXiv preprint arXiv:2406.01900*, 2024. 2
- [10] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM MM*, 2020. 1
- [11] Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *ASYU*, 2020. 2
- [12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1
- [13] 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. 1
- [14] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 1
- [15] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024. 2
- [16] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *SIGGRAPH*, 2024. 2
- [17] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman,

and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. [1](#)