FixTalk: Taming Identity Leakage for High-Quality Talking Head Generation in Extreme Cases

Supplementary Material

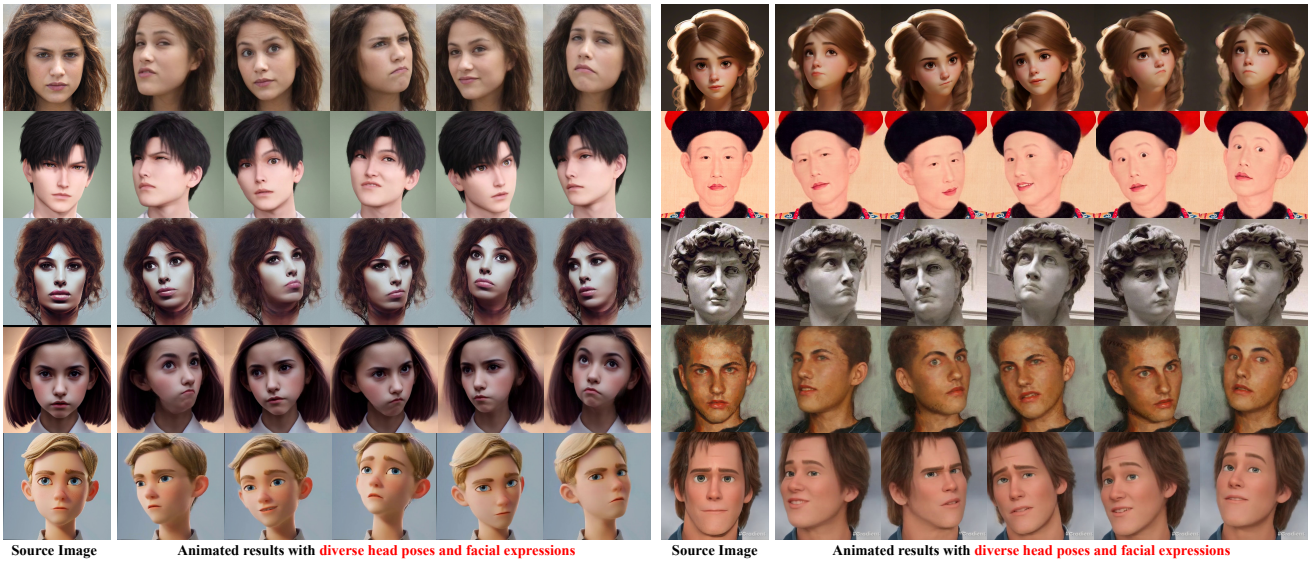

Figure 1. More animated results by FixTalk.

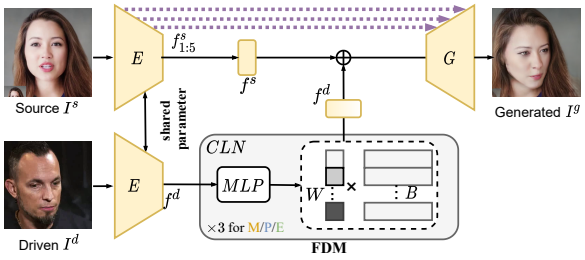

Figure 2. The detailed pipeline of Baseline, especially FDM.

In the main paper, we propose a novel framework, namely FixTalk, to enhance talking face generation in extreme cases by addressing the limitation of identity leakage and rendering artifacts. In this appendix, we provide more information about 1) Network Architecture and Training Details. 2) More Results. 3) More discussion. In addition, we strongly recommend watching the supplementary video.

1. Network Architecture and Training Details

1.1. Network Architecture

In Sec. 3, we briefly introduce the pipeline of the baseline. Here, we provide a more comprehensive explanation of the implementation details, especially for FDM. As shown in Fig. 2, the entire framework is based on an encoder-decoder structure. The Baseline simplifies the transformation $\mathcal{F}^{s\to d}$ from the source image I^s to the driven image I^d by introducing an abstract reference image I^r as transformation $\mathcal{F}^{s\to d} = \mathcal{F}^{s\to r} + \mathcal{F}^{r\to d}$ (r is omitted in the following statements). Therefore, Baseline employs a shared-parameter encoder E to extract multi-scale features $\{f^s_{1:5}, f^s\} = E(I^s)$ and $\{f^d_{1:5}, f^d\} = E(I^d)$ from both the source and driven images. To effectively disentangle different motion components from identity, Baseline defines three orthogonal motion spaces by three Componentaware Latent Navigation (CLN) (only one is depicted in Fig. 2), which together form the Face Decoupling Module (FDM). Specifically, each space employs a set of learnable orthogonal bases stored in bank $B = \{b_1, ..., b_n\}$, each representing one specific facial movement, ensuring that the motion spaces remain disentangled. Then each CLN utilizes a lightweight MLP to predict the weights $W = \{w_1, ..., w_n\}$ for each base, and further obtain the driving feature $f^d = \sum_{i=1}^{n} w_i b_i$, which is then added to f^s and fed into the generator G to produce the final result I^g. To preserve the identity information, G incorporates the multi-scale features $f^s_{1:5}$ of the source image, excluding the reference feature.

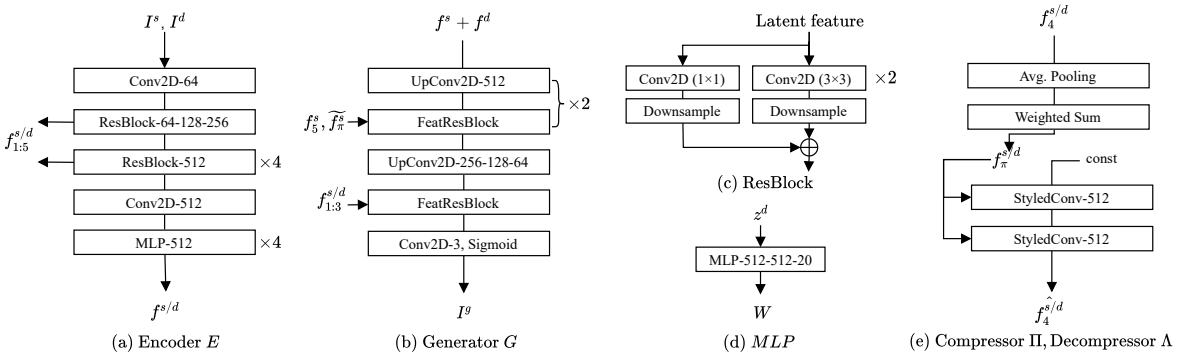

Figure 3. The detailed architectures of components in our model.

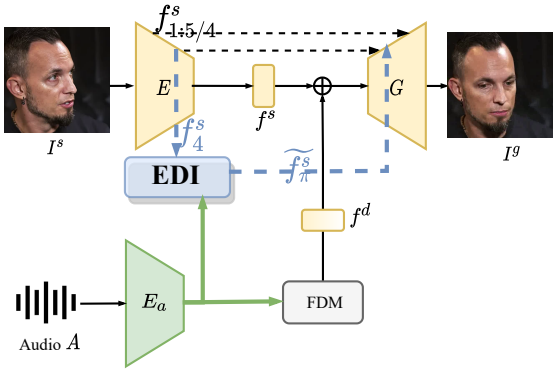

Figure 4. The detailed pipeline of audio-driven setting in FixTalk.

$$I^g = G(f^s + FDM(f^d), f^s_{1:5}) \quad (1)$$

We adopt the same Encoder E, Generator G as EDTalk [29], whose detailed structures are presented in Fig. 3. Please refer to their original paper for more textual descriptions. Besides, we have also provided a detailed description of the composition and workflow of EMI in the main paper, as well as the structure of the driven-identity memory M_{d} and the motion-source memory $M_{\text{m-s}}$ within EDI.

Here, we further detail the architecture of the Compressor Π and the Decompressor Λ in Fig. 3. Specifically, Compressor Π is composed of an average pooling layer and a weighted sum operation to effectively condense the input features. The Decompressor Λ is implemented using a StyleGAN2 [14] generator, where the compressed features f^s_π and f^d_π are injected into the layers via the style modulation operation, enabling efficient reconstruction and generation. This detailed design ensures robust feature processing and effective animation generation within our framework.

Our method not only supports video-driven talking head generation but also extends to audio-driven talking head generation, which is another significant contribution. As shown in Fig. 4, this is achieved by replacing the encoder E in the driven video branch with an audio encoder E_a. This audio-driven network takes audio feature sequences $A = a_{1:T}$ as input. These sequences are processed through a series of convolutional layers to generate audio features $f^a_{1:N}$, enabling the model to generate talking heads synchronized with the input audio. This design ensures flexibility in handling both video and audio inputs for talking head generation.

1.2. Training Details

Our method is trained on two datasets: VFHQ [40] and MEAD [34].

VFHQ. The VFHQ dataset is a comprehensive collection of over 16,000 high-quality video clips, meticulously curated from a diverse range of interview scenarios. Each clip in VFHQ is recorded in high resolution, ensuring precise details in facial features, skin texture, and movement dynamics. The clips feature subjects from diverse demographic backgrounds, including variations in age, gender, and ethnicity, ensuring that models trained on VFHQ are not biased toward specific groups.

MEAD. MEAD is a multimodal emotional audio-visual dataset consisting of 60 speakers recorded in a controlled laboratory environment. Of these, data from 43 speakers is publicly accessible. Each speaker delivers 30 sentences, encompassing expressions of eight distinct emotions: neutral, happy, sad, angry, surprise, fear, disgust, and contempt. These emotions are further captured at three varying intensity levels: low, medium, and high, resulting in a rich and diverse dataset tailored for emotion recognition and generation tasks.

The encoder E, Enhanced Motion Indicator (EMI), and generator G are pre-trained following a setting similar to

Figure 5. Additional video-driven talking head qualitative results, which are supplement to the main paper.

LIA [38]. This initialization ensures a robust latent representation, facilitating downstream tasks. After pre-training, the focus shifts to addressing rendering artifacts by training the Enhanced Detail Indicator (EDI). This step is crucial for enhancing output quality and achieving more natural and artifact-free renderings. All loss function weights are uniformly set to 1. The training employs a batch size of 16. A fixed learning rate of $2e-4$ is used. The training is conducted on 4 NVIDIA A100 GPUs, each equipped with 80GB of memory. Once both EMI and EDI modules are trained, all parameters are frozen to ensure consistency and prevent further updates. Subsequently, we follow the EDTalk framework to achieve decoupled facial control, enabling precise manipulation of facial attributes independently. This approach provides fine-grained control over expressions and motions, making it suitable for applications requiring high fidelity and flexibility in facial animation.

1.3. Evaluation Metrics

To comprehensively evaluate our method, we employ a variety of metrics that assess different aspects of generated talking head videos, including visual quality, audio-visual synchronization, identity preservation, rendering artifacts, and emotional expressiveness. Below, we provide detailed explanations of each metric:

Visual Quality Metrics Visual quality is a crucial aspect of talking head generation, reflecting how realistic and perceptually coherent the generated frames are. We use the following commonly adopted metrics:

- Peak Signal-to-Noise Ratio (PSNR) [11]: Measures the similarity between the generated and ground-truth images by evaluating pixel-wise differences. A higher PSNR indicates better reconstruction quality.
- Structural Similarity Index Measure (SSIM) [39]: Assesses the perceptual similarity between two images by considering luminance, contrast, and structural information. A higher SSIM value suggests better visual quality.
- Fréchet Inception Distance (FID) [24]: Quantifies the perceptual quality of generated images by comparing the distribution of deep feature representations extracted from the Inception network. Lower FID scores indicate higher realism.

Audio-Visual Synchronization Metrics Evaluating lip-sync accuracy is essential for assessing the temporal alignment between audio and facial movements. We employ:

- M/F-LMD (Landmark Deviation Metric) [1]: Measures the deviation of predicted facial landmarks from the ground truth in terms of mouth (M-LMD) and face (F-LMD). Lower values indicate better synchronization.
- SyncNet Confidence Score [3]: A widely used metric for audio-visual synchronization, computed based on the confidence of a pre-trained SyncNet model in determining whether the audio and video frames are temporally aligned. Higher confidence scores indicate better synchronization.

Identity Leakage Metric One of the main challenges in talking head generation is identity preservation, ensuring that the generated face retains the identity of the source image rather than being influenced by the driving image.

- Cosine Similarity (CSIM) for Identity Preservation: Measures the cosine similarity between the feature embeddings of the generated face and the source face, extracted using a face recognition model (e.g., ArcFace).

Figure 6. Additional audio-driven talking head qualitative results, which are supplement to the main paper.

Higher CSIM values indicate stronger identity preservation, meaning less identity leakage.

Rendering Artifacts Metrics Rendering artifacts, such as blurring, distortions, and unnatural textures, degrade the perceptual quality of generated images. To evaluate rendering quality, we use:

- Naturalness Image Quality Evaluator (NIQE) [19]: A no-reference image quality assessment metric that evaluates the naturalness of an image by analyzing statistical deviations from real images. Lower NIQE scores indicate fewer artifacts and higher realism.
- Cumulative Probability Blur Detection (CPBD): Measures the sharpness of an image based on edge contrast. Higher CPBD values indicate sharper images with fewer blur artifacts.

Emotional Expressiveness Metric Expressing emotions accurately is crucial for generating lifelike and engaging talking heads. We evaluate emotional expression fidelity using:

- Emotion Accuracy: Assesses how well the generated facial expressions match the target emotional category. This is computed by passing the generated frames through a

Method/Metric	PSNR↑	SSIM↑	LPIPS↓	$\mathcal{L}_1$ ↓	AKD↓	AED↓
PIRenderer [23]	22.13	0.72	0.22	0.053	2.24	0.032
Face-V2V [37]	23.29	0.74	0.17	0.037	1.83	0.025
LIA [38]	24.75	0.77	0.16	0.036	1.88	0.019
DaGAN [10]	23.21	0.74	0.16	0.041	1.93	0.023
MCNET [9]	21.74	0.69	0.26	0.057	2.05	0.037
StyleHEAT [43]	22.15	0.65	0.25	0.075	2.95	0.045
DPE [21]	26.16	0.83	0.18	0.059	2.05	0.031
EDTalk [29]	<u>26.50</u>	<u>0.85</u>	<u>0.13</u>	<u>0.031</u>	<u>1.74</u>	<u>0.017</u>
FixTalk	**27.16**	**0.86**	**0.11**	**0.024**	**1.62**	**0.009**

Table 1. The quantitative results compared with SOTA face reenactment methods on HDTF dataset.

pre-trained emotion recognition model and comparing the predicted emotion labels with the intended emotions.

2. More Experimental Results

2.1. More Video-Driven Comparison with SOTAs

More quantitative results. In addition to the quantitative comparison results with SOTA methods presented in the main paper, we provide a more comprehensive evaluation here, including additional methods such as PIRenderer [23], Face-V2V [37], LIA [38], DaGAN [10], MCNET [9], StyleHEAT [43], DPE [21], and more evaluation metrics, including LPIPS, L1, AKD, and AED. From the results in Tab. 1, it is evident that FixTalk outperforms all other methods across all metrics, demonstrating its robustness and effectiveness. Notably, FixTalk significantly surpasses EDTalk, highlighting the impact of the proposed EMI and EDI. These components effectively address the limitations of EDTalk, such as identity leakage and rendering artifacts, thereby validating the efficacy and superiority of our approach.

More qualitative results. To provide a more intuitive comparison of the differences between methods, we present visualized results in Fig. 5. When handling driven images with exaggerated expressions and extreme poses, most methods exhibit noticeable artifacts. In contrast, FixTalk not only generates normal results free from identity leakage and rendering artifacts but also achieves decoupled control of mouth poses and emotional expressions. This capability aligns well with the requirements of various application scenarios, showcasing the versatility and robustness of our approach.

Comparison to LivePortrait. We observe that among GAN-based models, LivePortrait achieves higher rendering quality than other methods, primarily due to its use of a larger and higher-quality dataset. Specifically, LivePortrait is trained on 69M video frames, including 200 hours of high-quality videos collected in-house, which is ten times more than the approximately 5M frames in our publicly available dataset. This significantly larger dataset undoubtedly contributes to the stronger performance of LivePortrait. However, despite this data advantage, our method achieves comparable image quality, particularly in teeth and background details, and outperforms LivePortrait in temporal consistency, as demonstrated in demo video (4:18–4:34) and Tab. 1 of main paper. For a fair comparison, we retrain LivePortrait on our dataset and re-evaluate the results, as shown in Fig. 7, where FixTalk demonstrates significant improvements over LivePortrait, particularly in reducing rendering artifacts, such as background inconsistencies and teeth details. To provide a more intuitive comparison, we compute quantitative metrics for rendering quality. The results show: LivePortrait: CPBD = 0.532, NIQE = 14.545; Our method: CPBD = 0.682, NIQE = 18.386. Both metrics indicate that under the same conditions, our method outperforms LivePortrait, further validating the effectiveness of our approach.

Figure 7. Comparison to retrained LivePortrait [8]. To ensure a fair comparison, we retrained LivePortrait using the same publicly available dataset we used.

2.2. More Audio-Driven Comparison with SOTAs

More quantitative results. Apart from the quantitative comparison with SOTA methods as detailed in the main paper, we present additional quantitative comparisons with more audio-driven talking head generation methods, including MakeItTalk [49], Audio2Head [35], PC-AVS [48], AVCT [36], MuseTalk [45], V-Epress [31], EAMM [12], StyleTalk [16], EmoGen [7], SAAS [27], DreamTalk [17]. The comparison results outlined in Tab. 2 demonstrate that our method outperforms state-of-the-art approaches across

Metric/Method	PSNR↑	SSIM↑	M-LMD ↓	F-LMD↓	FID↓	$\text{Sync}_{\text{conf}}$ ↑
MakeItTalk [49]	19.442	0.614	2.541	2.309	37.917	5.176
Wav2Lip [22]	19.875	0.633	1.438	2.138	44.510	**8.774**
Audio2Head [35]	18.764	0.586	2.053	2.293	27.236	6.494
PC-AVS [48]	16.120	0.458	2.649	4.350	38.679	7.337
AVCT [36]	17.848	0.556	2.870	3.160	37.248	4.895
IP-LAP [47]	19.832	0.627	2.140	2.116	46.502	4.156
TalkLip [33]	19.492	0.623	1.951	2.204	41.066	5.724
SadTalker [44]	19.042	0.606	2.038	2.335	39.308	7.065
AniTalker [15]	19.714	0.614	1.903	2.277	28.826	6.638
Hallo [42]	19.061	0.598	1.874	2.294	46.295	6.993
EchoMimic [2]	18.884	0.600	1.793	2.110	42.174	5.190
EAMM [12]	18.867	0.610	2.543	2.413	31.268	1.762
StyleTalk [16]	21.601	0.714	1.800	1.422	24.774	3.553
PD-FGC [32]	21.520	0.686	1.571	1.318	30.240	6.239
EMMN [26]	17.120	0.540	2.525	2.814	28.640	5.489
EAT [6]	20.007	0.652	1.750	1.668	21.465	7.984
FlowVQTalker [28]	21.572	0.709	1.551	1.304	23.453	6.901
EDTalk [29]	21.628	0.722	1.537	1.290	17.698	8.115
FixTalk	**22.382**	**0.743**	**1.314**	**1.215**	**15.851**	8.009

Table 2. Audio-driven quantitative comparisons on MEAD.

various metrics.

More qualitative results. Due to the space limitations in the main paper, we only present the qualitative results of FixTalk compared to several state-of-the-art (SOTA) methods. Here, we provide additional qualitative results, as shown in Fig. 6. Our method consistently outperforms SOTA methods in terms of mouth synchronization, consistency under extreme poses, and expressiveness of facial expressions. These observations align with the quantitative results presented in Tab. 2, further validating the superior performance of FixTalk.

2.3. User Study

To evaluate the performance of different methods from a human perspective, we randomly sample 10 videos generated by each method and conduct a **user study**. Specifically, we invite 20 participants and ask them to rate each video on a scale of 1-5 based on their motion consistency, the identity preservation, and the image quality. Since many participants do not have a background in computer vision, we provide detailed explanations for each criterion to assist their judgment.

- Motion Consistency: Participants assess how well the generated video aligns with the driven video in terms of motion accuracy.
- Identity Preservation: Participants evaluate the similarity between the generated face and the source image to assess how well the unique characteristics of the individual are preserved.
- Image Quality: Participants rate the clarity of the generated video and assess whether it contains fewer artifacts.

Method/Metric	PSNR↑	SSIM↑	LPIPS↓	$\mathcal{L}_1$ ↓	AKD↓	AED↓
Baseline	26.50	0.85	0.13	0.031	1.74	0.017
w/o EMI	26.68	0.86	0.13	0.029	1.69	0.013
w/o EDI	26.92	0.85	0.11	0.026	1.66	0.014
FixTalk	**27.16**	**0.86**	**0.11**	**0.024**	**1.62**	**0.009**

Table 3. The quantitative results of ablation study on HDTF dataset.

Metric/Method	Baseline	w EDI	w EMI	FixTalk
CSIM ↑	0.594	0.596	0.609	**0.613**
NIQE↓	42.41	19.26	36.64	**13.44**
CPBD↑	0.221	0.273	0.234	**0.282**

Table 4. More ablation studies on fixing identity leakage and rendering artifacts.

Metric/Method	LIA	LIA+EMI	LIA+EDI	LIA+EMI+EDI
CSIM ↑	0.522	0.543	0.525	**0.548**
NIQE ↓	46.76	38.72	25.14	**21.59**

Table 5. More ablation studies on LIA.

Metric/Method	Echomimic [2]	EchomimicV2 [18]	Hallo [42]	Hallo3 [4]	Ours
Computational cost (MB) ↓	7.32	11.60	9.64	37.71	**3.65**
Inference time (Frame/S) ↑	0.057	0.281	0.463	0.075	**27.601**
Dataset Size (hour)	∼540h	∼160h	∼164h	∼134h	**∼55h**
Public data	✗	✗	✗	✗	✓

Table 6. Computational cost, inference time, training data for each method.

The results are summarized in the main paper, where our method outperforms others in all aspects.

2.4. Ablation Study

In the main paper, we present qualitative ablation study results. Additionally, in Tab. 3, we further provide quantitative ablation study results, clearly illustrating the specific contributions of EMI and EDI to each aspect of the model's performance.

In particular, to emphasize that our method effectively fixes the issues of identity leakage and rendering artifacts in current GAN-based models for high-quality rendering, we provide quantitative results for both aspects in Tab. 4 and Tab. 5. Compared to the baseline, FixTalk shows significant improvements in both identity leakage and rendering artifacts. The improvement in identity leakage primarily stems from our proposed EMI module, which leads to a noticeable increase in CSIM. The reduction in rendering artifacts is mainly due to EDI, which results in significant improvements in both NIQE and CPBD.

2.5. Comparison on three key aspects

As mentioned in the main paper Sec. 1, to facilitate the application in various areas, modern talking head generation methods are expected to achieve three critical goals:

(1) *System Efficiency*: Given the potential real-time applications, fast inference and minimal computational overhead are necessary for broader adoption [8]. **(2)** *Decoupled Control*: Effective talking head generation involves finely controlling several key components, such as the mouth, head pose, and emotional expressions. To achieve realistic and expressive facial animations, it is crucial to disentangle and control these aspects independently, allowing for granular adjustments to each component without influencing the others [32]. **(3)** *High-Quality Rendering*: Producing high-quality video outputs is a primary goal for talking head generation, especially when creating extreme poses and expressions [41]. To comprehensively validate the superiority of our method, we conduct a detailed comparison across the three aspects mentioned above.

System Efficiency. Real-time generation is a critical component of talking head generation, as it enables the creation of dynamic, responsive avatars capable of interacting with users or real-time inputs. Here, we conduct an analysis of the model parameters size and the number of frames generated per second. To ensure a fair comparison, we run all models on an NVIDIA A100 GPU with 80GB memory and record the results in Tab. 6. As observed, diffusion-based models [2, 4, 18, 42] typically consume a significant amount of GPU memory. In contrast, our method requires only 3.65GB of memory, making it highly efficient and capable of running on devices with as little as 6GB of GPU memory. Additionally, our method achieves a much faster inference speed compared to other approaches, generating 27.6 frames per second. Given that standard talking head videos typically run at 25 frames per second, our method fully meets the requirements for real-time generation. Meanwhile, it also demonstrates that our proposed EDI and EMI modules are lightweight, as they do not affect the efficiency of the base model while effectively addressing its inherent identity leakage and rendering artifact issues.

Decoupled Control Our method is built upon Baseline [29], which is currently the state-of-the-art model for disentangled control. Since our EMI and EDI modules do not disrupt its disentangled semantic space, FixTalk retains the ability to perform disentangled control over talking heads. As shown in Fig. 5, Fig. 5 in the main paper, and the demo video (4:12–4:36), our method enables independent control over mouth shape, head pose, and facial expressions.

Rendering Quality The primary goal of this paper is to fix the common issues of identity leakage and rendering artifacts in GAN-based models when generating high-quality renderings. While our method is also GAN-based, we demonstrate that with the incorporation of our proposed EMI and EDI modules, these issues can be effectively mitigated. To highlight this, we compare our approach with the current state-of-the-art GAN-based models. Fig. 5 and Fig. 5 in the main paper qualitatively illustrate the improvements our method achieves in reducing identity leakage and rendering artifacts. For a more intuitive evaluation, we present quantitative results in Tab. 7, where our method consistently outperforms other GAN-based approaches across all metrics. We also provide the deviation values for each metric to ensure a more comprehensive evaluation. These results strongly validate the superiority and stability of our method in addressing identity leakage and rendering artifacts.

Metric	FOMM	LIA	DPE	EmoPort	EDTalk	Ours
CSIM ↑	0.503±0.11	0.522±0.09	0.567±0.09	0.493±0.19	0.594±0.08	**0.613±0.05**
NIQE ↓	34.85±15.24	46.76±19.97	42.96±21.48	29.88±10.36	42.41±18.71	**13.44±1.26**
CPBD ↑	0.176±0.13	0.161±0.07	0.183±0.07	0.178±0.05	0.221± 0.08	**0.282± 0.05**
Acc_{emo} ↑	38.72%	38.69%	35.99%	68.40%	68.85%	**70.93%**

Table 7. More quantitative metrics with deviation values.

2.6. More Animated Results by FixTalk

In Fig. 1, we showcase a diverse set of results generated by FixTalk, highlighting its robust performance across various input image types. These inputs include real human photographs, sculptures, AIGC-generated images, oil paintings, and more, demonstrating the model's versatility. The generated results also feature a wide range of extreme head poses and vivid facial expressions, showcasing FixTalk's ability to handle challenging scenarios with remarkable consistency and quality. This diversity of inputs and outputs underscores the model's capability to adapt to different visual domains while preserving realism and expressiveness in its generated talking heads.

2.7. More Exploration in GAN-based models

In Baseline, to generate results consistent with the identity of the source image and the motion of the driven image, the framework extracts f^s and $f^s_{1:5}$ (i.e., f^s_1, f^s_2, f^s_3, f^s_4, f^s_5) from the source image to represent identity information (e.g., appearance). Simultaneously, it extracts $\{f^d, f^d_{1:5}\}$ from the driven image, with motion features further refined through Conditional Layer Normalization (CLN).

To identify which variable contributes to identity leakage, we systematically replace features extracted from the source image with corresponding features from the driven image. In the main paper, we already presented results for replacing (1) $f^s_{1:5}$ with $f^d_{1:5}$, (2) f^s_4 with f^d_4, (3) $f^s_{1:5/4}$ with $f^d_{1:5/4}$. Here, we further demonstrate the results of replacing (4) f^d with f^s, as shown in Fig. 8. Combining these additional results with the analysis in the main text, we conclude that identity leakage primarily originates from the fourth layer's features, highlighting its critical role in encoding identity information.

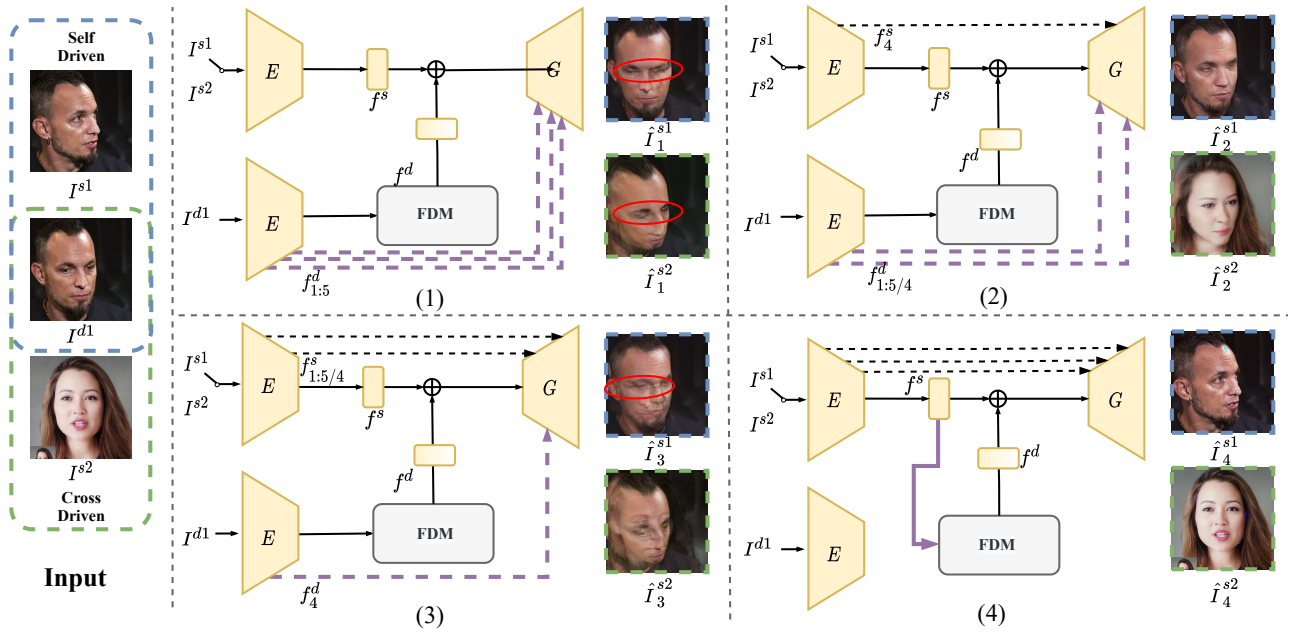

Figure 8. More Exploration in Baseline.

3. Discussion

3.1. Generalizability

To address the issues present in GAN-based models for high-quality rendering, we propose a set of exploratory experiments in Sec. 3 of the main paper. Based on the findings from these experiments, we introduce two modules, EMI and EDI, in Sec. 4 to specifically tackle identity leakage and rendering artifacts, respectively. Considering that EDTalk [29] is the current state-of-the-art among GAN-based methods, having already resolved various issues except for these two, we select it as the representative GAN-based model in our study. This choice allows us to better highlight the existing challenges in GAN-based talking head generation.

However, it is important to emphasize that the exploratory experiments and research methodologies proposed in Sections 3 and 4 are not specifically designed for EDTalk—they are also applicable to other GAN-based methods. To validate this, we use LIA as the backbone in Section 5.4, and the results demonstrate that our proposed method effectively addresses identity leakage and rendering artifacts in LIA as well. Next, we conduct a detailed analysis from the following three perspectives:

(1) **Identifying the Problem.** Our primary motivation stems from the investigative experiments conducted in Section 3, where we identified two key issues in Baseline: (a) Identity Leakage. (b) Self-driven results outperform cross-driven results. This leads us to pose the question: Can we mitigate identity leakage for high-quality talking head generation in extreme poses and expressions? To verify the generalizability of our exploratory experiments and research question, we further test recent state-of-the-art GAN-based models, including FOMM [25], LIA [38], DaGAN [10], DPE [21], MCNET [9], and EmoPor [5]. The results, shown in Fig. 5, confirm that these methods exhibit the same two findings (a) and (b). Furthermore, their respective papers also indicate that performance in the self-driven setting is consistently better than in the cross-driven setting. Therefore, our findings are not limited to Baseline but are also applicable to other GAN-based models.

(2) **Investigating the Problem.** Building on the findings in (1), we conduct systematic experiments on Baseline to identify the specific variables responsible for carrying identity information. Since most GAN-based models follow an encoder-decoder structure and involve extracting motion features from the driven video, our systematic experiments on Baseline can be extended to other methods to locate the variables contributing to identity leakage in each approach.

(3) **Solving the Problem.** Once the responsible variables are identified, we introduce EMI and EDI modules. As demonstrated in Figure 4 and Section 4, these modules are plug-and-play: EMI is inserted into the encoder, while EDI serves as a bridge between the encoder and decoder. As discussed in (2), most GAN-based models contain both an encoder and a decoder, meaning that EMI and EDI can be seamlessly integrated into other existing methods, such as LIA [38], DPE [21], and An-

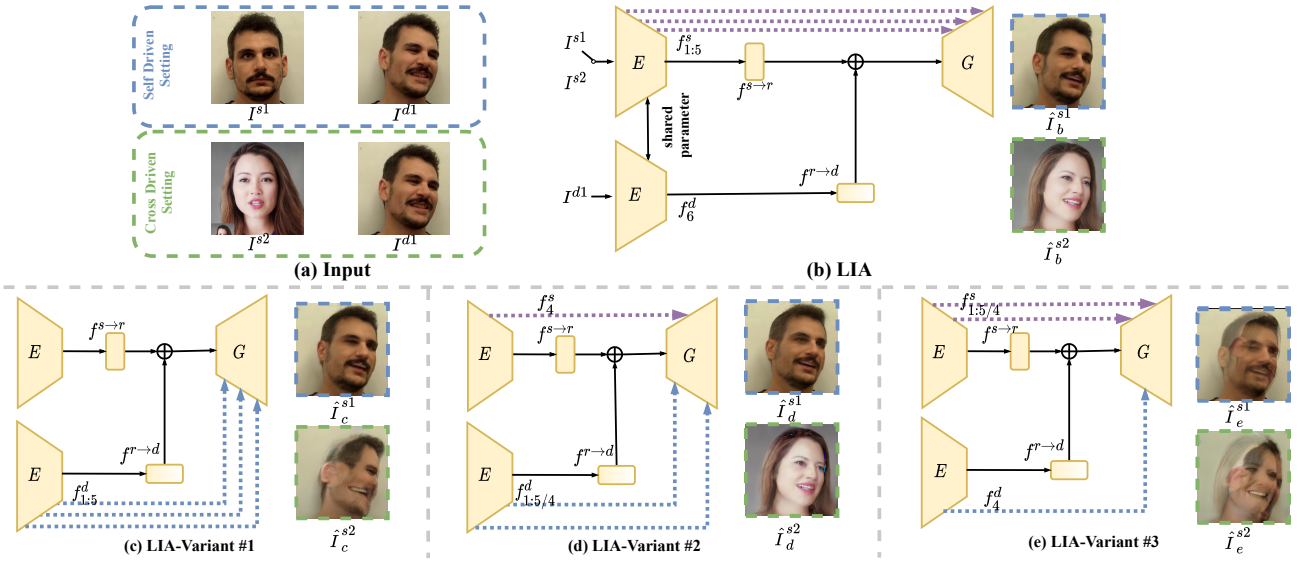

Figure 9. Exploration in other GAN-based model-LIA [38].

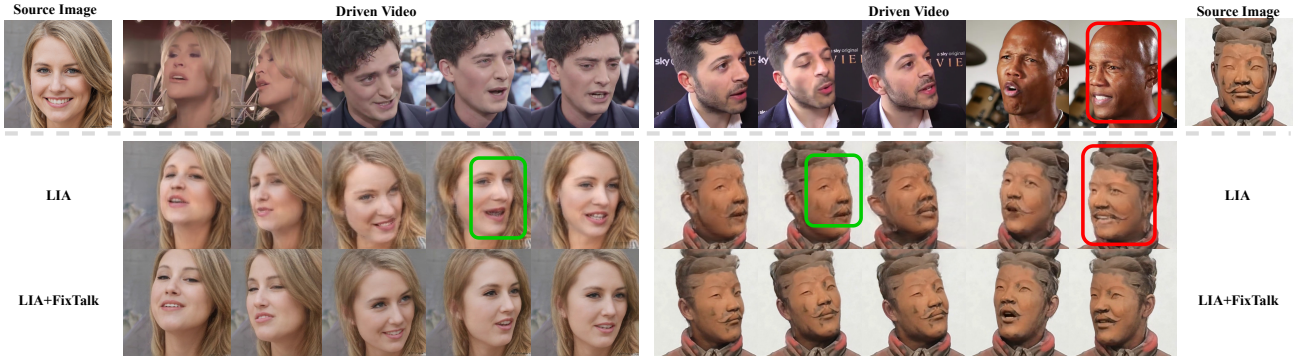

Figure 10. comparison results between original LIA and LIA + FixTalk.

iTalker [15].

Through the above three aspects of analysis, we theoretically demonstrate that our method can be generalized to other GAN-based models in terms of motivation, exploratory experiments, and research methodology. To further validate this, we conduct the same experiments on LIA [38]. As shown in Fig. 9, the findings from our exploratory experiments on LIA are consistent with the conclusions drawn in Section 3 of the main text.

Following this, we integrate our EMI and EDI modules into LIA, as illustrated in Fig. 10. The final experimental results confirm that our method effectively mitigates identity leakage and rendering artifacts in the original model (LIA). This further demonstrates that our approach is widely applicable to most GAN-based models.

3.2. Quality Improvement

Key contribution In our results, including figures and tables in both the main paper and supplementary materials, as well as the demo video, we showcase FixTalk's performance in both video-driven and audio-driven settings. Additionally, we demonstrate FixTalk's capability in face disentanglement control. However, it is important to emphasize that the **primary focus** of this paper is to address the prevalent issues of Identity Leakage (IL) and Rendering Artifacts (RA) in GAN-based models. Therefore, our method is primarily designed to improve these two aspects: (1) Identity Preservation: Our results show that the generated identity remains consistent with the source image rather than being influenced by the driven image, leading to higher CSIM scores. (2) Reduction of Artifacts: In extreme cases, such as side views or exaggerated facial expressions, our method produces fewer artifacts, maintains high image

quality, and avoids blurring, as reflected in higher CPBD and lower NIQE scores. Improving audio-to-lip synchronization is not the main focus of this work. Our goal is simply to ensure performance on par with existing methods. Qualitatively, our approach achieves satisfactory lip-sync results, and quantitatively, it matches or even surpasses current state-of-the-art (SOTA) methods. Thus, FixTalk not only ensures lip-sync quality but also effectively addresses identity leakage and rendering artifacts, validating our core contribution.

Comparison to diffision based models Inspired by the success of diffusion models in video generation, extensive efforts have been made to apply diffusion models to the talking head domain, achieving impressive results. Diffusion-based models [13, 30, 46] have demonstrated superior image quality and audio-lip synchronization compared to GAN-based models. However, these models heavily rely on large-scale, high-quality training data. As shown in Tab. 6, the current SOTA diffusion-based talking head models are trained on datasets that far exceed ours in size, with most of their data being privately collected high-quality datasets, whereas we use publicly available datasets. Additionally, their computational requirements are significantly higher than ours. Therefore, while a direct comparison with these models may not be entirely fair, our approach still achieves comparable results and even surpasses them in some aspects, demonstrating the effectiveness of our method.

3.3. Backbone Selection

GANs can be categorized into 2D GANs and 3D GANs. In this work, FixTalk adopts a 2D GAN as its backbone. Compared to 2D GANs, 3D GANs (e.g., StyleSDF [20]) introduce an additional dimension to model depth information, enabling multi-view synthesis. However, we chose not to use 3D GANs for the following reasons: (1) Our work primarily addresses identity leakage and rendering artifacts, where depth information provides limited benefits. (2) Real-time generation is crucial for talking head generation, but 3D GANs introduce additional parameters, requiring more computational resources and making real-time inference challenging. (3) Audio-driven animation prioritizes lip sync, which has minimal dependency on depth information. (4) Multi-view modeling in 3D GANs requires multi-view training data, which is currently scarce. (5) 3D GANs require additional loss functions to learn depth information, increasing training complexity. Considering these factors, we opted for a 2D GAN instead of a 3D GAN in our approach.

3.4. Limitation

While our current work, FixTalk, has made significant strides, it still has certain limitations that present opportunities for future improvement. (1) Lack of Decoupling for Eye Movement: Although our approach has successfully decoupled mouth movements, poses, and expressions, the eyes remain coupled with other facial features. Decoupling eye movements is a critical aspect, as they play a crucial role in conveying emotions and natural behavior, especially in expressive talking head generation. (2) Absence of Natural Emotion Prediction from Audio: Our method currently does not predict natural emotions, including subtle micro-expressions, directly from audio input. Incorporating emotional cues from audio could significantly enhance the realism and expressiveness of generated talking heads, but this remains an unexplored challenge. These limitations highlight the potential areas for further research and refinement in enhancing the realism and versatility of FixTalk.

3.5. Ethical Considerations

Our proposed talking head generation model demonstrates significant potential for creating realistic and expressive talking heads for a variety of applications, such as virtual assistants, language learning tools, and entertainment. This technology offers transformative benefits by enabling more accessible, engaging, and automated communication workflows. However, like any generative technology, talking head generation also raises important ethical considerations, particularly around potential misuse. Malicious actors could exploit this technology to produce fake or deceptive content, such as manipulated videos for propaganda, false endorsements, or defamatory material. Such misuse could erode trust in digital communications, amplify the spread of misinformation, and have damaging social, political, and economic consequences. Despite ongoing advancements in deepfake detection and content authentication technologies, challenges remain in identifying high-quality and contextually convincing synthetic videos. This makes it increasingly important to address the ethical implications of talking head generation proactively. To mitigate these risks, we are pleased to share our talking face generation results, which can help advance detection algorithms to effectively address increasingly sophisticated scenarios. Additionally, to prevent misuse of the generated videos, we will incorporate watermarks during the generation process.

References

[1] Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7832–7841, 2019. 3

[2] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024. 6, 7

[3] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017. 3

[4] Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with diffusion transformer networks. *arXiv preprint arXiv:2412.00733*, 2024. 6, 7

[5] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024. 8

[6] Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22634–22645, 2023. 6

[7] Sahil Goyal, Sarthak Bhagat, Shagun Uppal, Hitkul Jangra, Yi Yu, Yifang Yin, and Rajiv Ratn Shah. Emotionally enhanced talking face generation. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice*, pages 81–90, 2023. 5

[8] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024. 5, 7

[9] Fa-Ting Hong and Dan Xu. Implicit identity representation conditioned memory compensation network for talking head video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23062–23072, 2023. 5, 8

[10] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022. 5, 8

[11] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 3

[12] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, and Xun Cao. Eamm: One-shot emotional talking face via audio-based emotion-aware motion model. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 5, 6

[13] Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. *arXiv preprint arXiv:2411.16331*, 2024. 10

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[15] Tao Liu, Feilong Chen, Shuai Fan, Chenpeng Du, Qi Chen, Xie Chen, and Kai Yu. Anitalker: Animate vivid and diverse talking faces through identity-decoupled facial motion encoding. *arXiv preprint arXiv:2405.03121*, 2024. 6, 9

[16] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. *arXiv preprint arXiv:2301.01081*, 2023. 5, 6

[17] 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. 5

[18] Rang Meng, Xingyu Zhang, Yuming Li, and Chenguang Ma. Echomimicv2: Towards striking, simplified, and semi-body human animation. *arXiv preprint arXiv:2411.10061*, 2024. 6, 7

[19] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 4

[20] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13503–13513, 2022. 10

[21] Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, and Dong-ming Yan. Dpe: Disentanglement of pose and expression for general video portrait editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 427–436, 2023. 5, 8

[22] 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 *Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020. 6

[23] Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13759–13768, 2021. 5

[24] Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, 2020. Version 0.3.0. 3

[25] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019. 8

[26] Shuai Tan, Bin Ji, and Ye Pan. Emmn: Emotional motion memory network for audio-driven emotional talking face generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22146–22156, 2023. 6

[27] Shuai Tan, Bin Ji, Yu Ding, and Ye Pan. Say anything with any style. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5088–5096, 2024. 5

[28] Shuai Tan, Bin Ji, and Ye Pan. Flowvqtalker: High-quality emotional talking face generation through normalizing flow and quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26317–26327, 2024. 6

[29] Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pages 398–416. Springer, 2025. 2, 5, 6, 7, 8

[30] Linrui Tian, Siqi Hu, Qi Wang, Bang Zhang, and Liefeng Bo. Emo2: End-effector guided audio-driven avatar video generation. *arXiv preprint arXiv:2501.10687*, 2025. 10

[31] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024. 5

[32] Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, and Baoyuan Wang. Progressive disentangled representation learning for fine-grained controllable talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17979–17989, 2023. 6, 7

[33] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662, 2023. 6

[34] 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

[35] S Wang, L Li, Y Ding, C Fan, and X Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. In *International Joint Conference on Artificial Intelligence*. IJCAI, 2021. 5, 6

[36] Suzhen Wang, Lincheng Li, Yu Ding, and Xin Yu. One-shot talking face generation from single-speaker audio-visual correlation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2531–2539, 2022. 5, 6

[37] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10039–10049, 2021. 5

[38] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2021. 3, 5, 8, 9

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 3

[40] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022. 2

[41] You Xie, Hongyi Xu, Guoxian Song, Chao Wang, Yichun Shi, and Linjie Luo. X-portrait: Expressive portrait animation with hierarchical motion attention. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 7

[42] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Luc Van Gool, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024. 6, 7

[43] Fei Yin, Yong Zhang, Xiaodong Cun, Mingdeng Cao, Yanbo Fan, Xuan Wang, Qingyan Bai, Baoyuan Wu, Jue Wang, and Yujiu Yang. Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan. In *European conference on computer vision*, pages 85–101. Springer, 2022. 5

[44] 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*, pages 8652–8661, 2023. 6

[45] Yue Zhang, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. Musetalk: Real-time high quality lip synchorization with latent space inpainting. *arxiv*, 2024. 5

[46] Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. Memo: Memory-guided diffusion for expressive talking video generation. *arXiv preprint arXiv:2412.04448*, 2024. 10

[47] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023. 6

[48] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021. 5, 6

[49] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020. 5, 6