EmotiveTalk: Expressive Talking Head Generation through Audio Information Decoupling and Emotional Video Diffusion

Supplementary Material

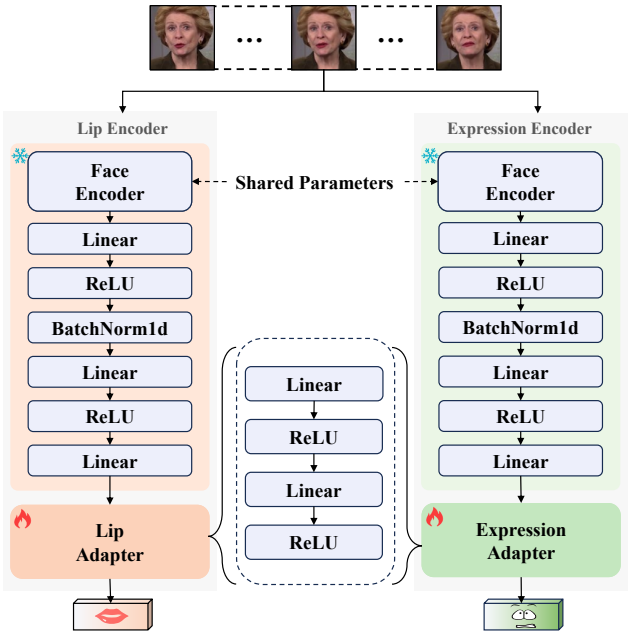

Figure 7. The structure of lip and expression encoder in EmotiveTalk.

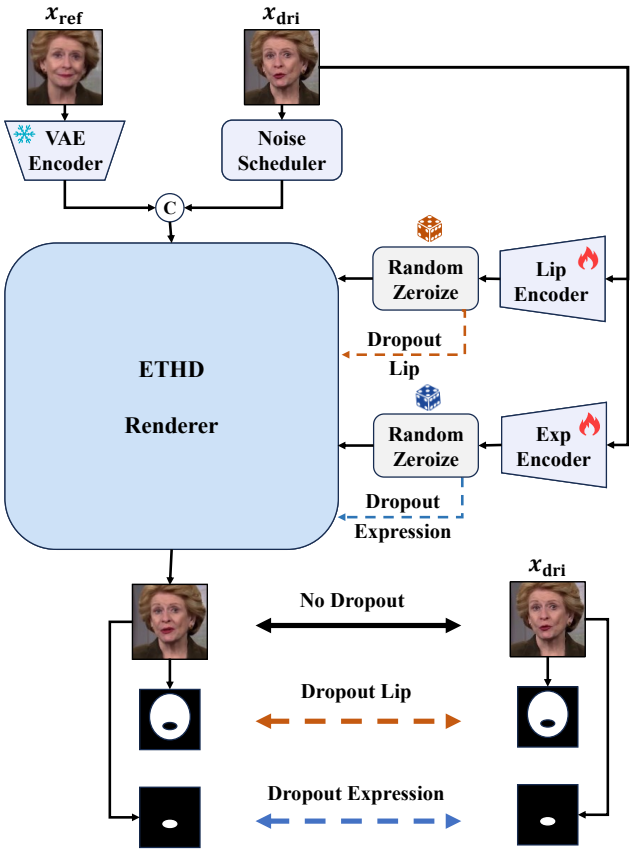

Figure 8. The self-driven dropout image reconstruction training process for lip and expression encoders pre-training.

6. Implementation Details of EmotiveTalk

6.1. Vision Encoders Pre-training

In talking head videos, the movements of different facial regions are often coupled, which increases the difficulty of independently controlling facial expressions and lip motions. Extracting decoupled representations related to expressions and lips is crucial for generating controllable expressions in talking head models. To obtain better performance and controllability in emotional talking head generation, we design lip and expression encoders to encode the lip motions and facial expressions independently to generate lip motion latent embedding $\vec{l}_v$ and expression latent embedding $\vec{e}_v$ from self-driven video.

Architecture of Lip and Expression Encoders. Our lip and expression encoders are based on PDFGC [41], which demonstrates remarkable facial motion decoupling results. We add trainable lip and expression adapters before the pre-trained PDFGC encoder to achieve better decoupling results based on the following tasks. Our network structure of lip and expression encoders are illustrated in Fig. 7.

Self-driven Dropout Image Reconstruction. We have pre-trained our lip and expression encoders on a self-driven task. We employ the ETHD backend described in Sec. 3.3 as the renderer to perform self-driven image reconstruction. The inputs consist of a reference image and another driving image from the same speaker. The driving image is processed by the lip and expression encoder to obtain latent embeddings of the lip and expression, which serve as conditional inputs to the renderer.

To more effectively combine the driving representations for controlling movements in different facial regions and to alleviate the coupling between lips and expressions, we adopt condition dropout training. As shown in Fig. 8, during training, we adopt a conditional dropout strategy with three configurations:

- Dropping the lip latent embedding $\vec{l}_v$;
- Dropping the expression latent embedding $\vec{e}_v$;
- Utilizing both latent embeddings simultaneously.

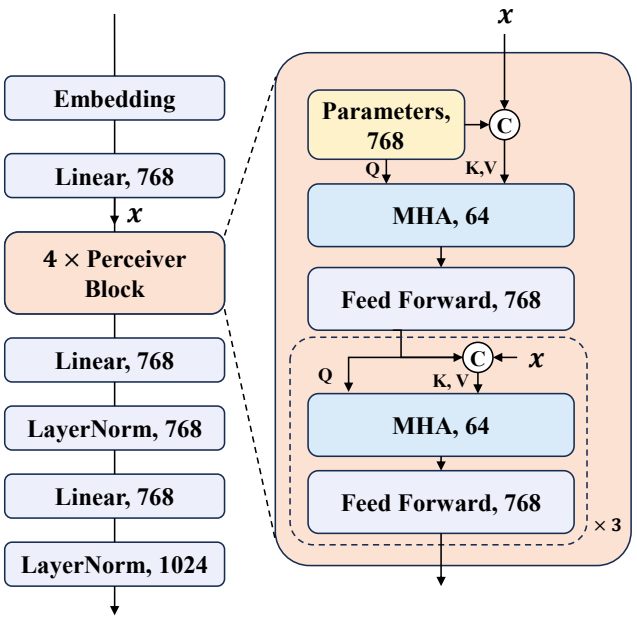

Figure 9. The structure of the audio-lip projector of V-AID in EmotiveTalk.

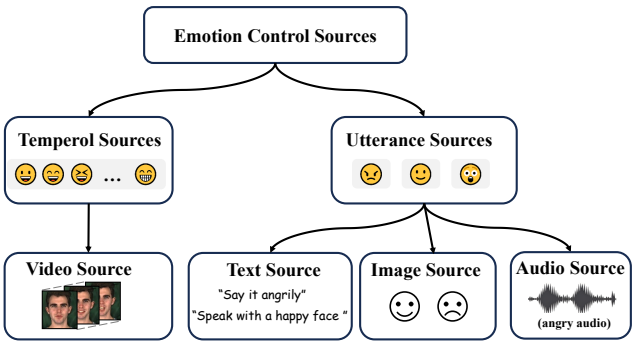

Figure 10. The categories of emotion control sources.

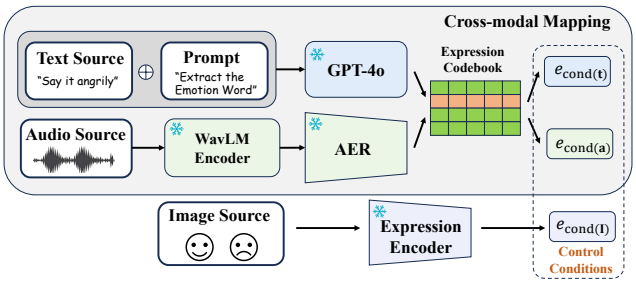

Figure 11. The designed mapping module of mapping different utterance sources to vision expression space to serve as emotion control conditions.

When dropping the expression latent embedding, we apply the facial mask described in Sec. 3.3 to both the ground-truth and the generated frame latents, ensuring that parameter updates focus exclusively on the reconstruction of facial expressions. Similarly, when dropping the lip latent embedding, we apply the lip mask from Sec. 3.3 to both ground-truth and the generated frame latents and focus only on lips reconstruction. When both latent embeddings are used simultaneously, we train the model to reconstruct the entire image.

6.2. Details of V-AID

Architecture of Audio-lip Projector. Our audio-lip projector demonstrated in Sec. 3.2 leverages a Perceiver Transformer [19] architecture, illustrated in Fig. 9. The input to the audio-lip projector consists of audio embedding window $\boldsymbol{A}_{\mathrm{w}} \in \mathbb{R}^{l \times w \times c}$ encoded by a pre-trained Wav2Vec [2] encoder, where l denotes the length of audio and w denotes the window size. The audio embedding is first passed through an embedding layer and a linear layer for feature projection. Subsequently, the processed embedding x is fed into the following four Perceiver Transformer blocks. In the first block, the query input y is a learnable vector, while the keys and values are derived by concatenating y with the original input x. After a matrix transformation, multi-head attention is applied along the window dimension to capture internal relationships within the speech feature window. The output of this process is further refined using a feed-forward module. As for the following blocks, the query input is the output of the preceding block, while the keys and values are constructed by concatenating the block's output with the original input x. This design iteratively enhances the feature representations by leveraging both the temporal dependencies and the shared information between the block outputs and the original input. Leveraging the vision-guided training strategy described in Sec. 3.2, the model effectively learns to map audio representation space to the corresponding lip motion representation space.

6.3. Details of MEC

To facilitate the use of multiple optional emotion-driven control sources for customizing speaking emotions, we propose a Multi-Source Emotion Control (MEC) pipeline tailored to different emotion control sources. This section provides a detailed elaboration of the concepts introduced in Sec. 3.5. Overall, the pipeline incorporates various types of emotion control inputs, as illustrated in Fig. 10, for temporal sources like driven video, we utilize the pre-trained expression encoder to obtain expression latent embedding from the emotion video source as the expression-driven representation to control the expression generation, as illustrated in Sec. 3.5. As for the utterance sources, we map the emotional information to the emotion condition $\boldsymbol{e}_{\mathrm{cond}}$, then we utilize the Di-CTE model in Sec. 3.2 to expanding the utterance condition to temporal driven condition. In this section, we provide a detailed explanation of how emotion condi-

tions are derived from various utterance control sources, as illustrated in Fig. 11.

Image Source. We utilize the pre-trained expression encoder to obtain expression latent embedding from the driven image as the emotion control condition, as illustrated in Sec. 3.5.

Text Source. We utilize GPT-4o to retrieve emotion keywords from specified emotion prompts. To facilitate this process, we define a set of eight commonly observed emotion keywords: happy, angry, sad, surprised, fear, disgusted, worried, and neutral. Based on these keywords, we construct an emotion-to-expression codebook, which stores latent representations of facial expressions corresponding to each emotion. These latent representations are extracted using the pre-trained expression encoder (Sec. 6.1) applied to facial images reflecting various emotional states. To enhance the extraction process, we prepend a retrieval prompt to the emotion prompt, which is "Extract the keyword representing emotion from the following sentence and return only the keyword. The keywords should be one of happy, angry, sad, surprised, fearful, disgusted, worried, or neutral."

Audio Source. To enable the retrieval of emotion-driven keywords from external emotional audio input, we have designed an audio-emotion retrieval module. This module employs a frozen audio encoder and a trainable speech emotion recognition module. In our implementation, we utilize WavLM as the audio encoder due to its superior performance in speech emotion recognition tasks. The architecture of our speech emotion recognition module leverages the audio branch of the state-of-the-art multimodal emotion recognition model. Similarly, we use an emotion-to-expression codebook to map the retrieved emotional keywords to latent embedding representations in the expression space. These representations serve as emotional control conditions for the subsequent Di-CTE module, enabling the generation of sequential emotional control conditions.

6.4. Training and Inference Details

This section provides a detailed explanation of the training and inference processes for EmotiveTalk, serving as a supplementary discussion to Sec. 3.4.

Training Details. EmotiveTalk is trained using eight NVIDIA A100 GPUs. The training of the V-AID module takes about 20 A100 GPU hours, while the backbone network requires around 200 A100 hours. Compared to the current mainstream talking head video generation models based on the Stable Diffusion [28] framework, our model's initialization weights are partially inherited from the pre-trained Stable Video Diffusion model [5]. We also utilize the pre-rained temporal VAE from Stable Video Diffusion [5]. Moreover, our model achieves a balance between performance and efficiency, supporting the training of long-duration, high-resolution videos. In our project, we train the audio-lip projector and Di-CTE module in V-AID in Sec. 3.2 with a learning rate of 1×10^{-4} and a batch size of 16 for each iteration by Adam optimizer and training length of each expression-related latent Di-CTE module is set to 220, and randomly cat the ground-truth expression latent for a ratio of 0.8. We train the ETHD backbone network under the following three configurations:

- 512-resolution: Training images and videos at a resolution of 512×512, with 120 frames per training iteration;
- 1024-resolution: Training images and videos at a resolution of 1024×1024, with 32 frames per training iteration.

The training length of 1024-resolution is much shorter than 512-resolution due to the training cost constraint of high resolution. The hyperparameters for all configurations are kept consistent. We use the Adam8bit optimizer with a learning rate of 1×10^{-5} and a batch size of 1 for each iteration. And the possibility of choosing ground-truth vision expression latents $\vec{e}_{\text{v}}$ when training is set to 0.6.

Inference Details. During inference, we first utilize the V-AID module to generate lip-related latent embeddings $\boldsymbol{l}_{\text{a}}$ and expression-related latent embeddings $\boldsymbol{e}_{\text{a}}$ based on the driving speech signal and emotional control source. For long-time expression-related latent embedding generation, we guide the inference process by appending the last 20 frames of the generated expression-related latent embeddings to the sampled noise as initialization.

Subsequently, the lip-related $\boldsymbol{l}_{\text{a}}$ and expression-driven latent embeddings $\boldsymbol{e}_{\text{dri}}$ obtained in the previous step are used as conditional inputs for the denoising process of the ETHD backbone network. $\boldsymbol{e}_{\text{dri}}$ is choose from $\boldsymbol{e}_{\text{a}}$ and $\boldsymbol{e}_{\text{v}}$ based on the category of the emotion source, $\boldsymbol{e}_{\text{v}}$ for temporal sources and $\boldsymbol{e}_{\text{a}}$ for utterance sources. We adopt the DDIM inference strategy with 25 denoising steps. The inference configurations for different resolutions are as follows:

- 512-resolution: Window size of 120 frames with an overlap of 24 frames;
- 1024-resolution: Window size of 32 frames with an overlap of 12 frames.

7. Detailed Evaluations of EmotiveTalk

This section serves as an extended discussion of Sec. 4 in the main paper, providing a more comprehensive and detailed analysis and comparison of EmotiveTalk.

7.1. Evaluation Details of the Main Paper

This section serves as a detailed explanation of the evaluation settings of our experiments in Sec. 4 of the main paper.

Evaluation Settings Details. We utiliz the publicly available HDTF [51] and MEAD [42] datasets for training and evaluation. The same data split is applied to both datasets: 90% of the data is allocated to the training set for model

training, while the remaining 10% is reserved as the evaluation set for evaluation. To ensure robust evaluation, we strictly maintain no overlap between the training set and the evaluation set, and the evaluation set data is entirely unseen during the training process. Specifically, for the MEAD dataset, due to limitations in training length, we filter the training set by excluding sequences shorter than 120 frames (equivalent to 4.8 seconds). For evaluation, MEAD sequences shorter than 120 frames are zero-padded to reach this duration of 120 frames. When generating videos from the processed test data, frames beyond the original sequence length are removed using the ffmpeg tool, ensuring the generated videos match the length of the original ground-truth video. We utilize these above-mentioned principles for all our evaluations in the main paper Sec. 3.4 and following experiments.

Comparison Settings Details. To ensure a fair comparison, we conduct an evaluation with audio of the same length when comparing our model with other state-of-the-art models in Tab. 1. On the HDTF dataset, we use audio clips of 5.76 seconds (approximately 144 frames in the ground-truth video) to drive the portrait. Since different models generate slightly varying numbers of frames for the same audio length, we standardize the evaluation by reporting the FVD metric for the first 128 frames (FVD_{128}), along with the average FID, and Sync-C Sync-D metrics across all generated frames. As for the E-FID metric, we follow the approach used in EMO [38], extracting 3D reconstructed expression coefficients from all frames of both the generated and ground truth videos. The E-FID is then computed as the FID between the expression coefficients of the generated and ground truth videos. And on the MEAD dataset, due to the shorter video length, we standardize the evaluation by generating 3.04 seconds (approximately 76 frames in the ground-truth video) and testing the FVD metic for the first 72 frames (FVD_{72}). The settings of other metrics are the same as those of the abovementioned HDTF testing. This methodology ensures the comprehensiveness and fairness of the evaluation, providing an objective comparison across all the models.

Users Study Details. As described in Sec. 4.5, our user study involved 26 participants, including 10 professional video evaluation engineers and 16 graduate students with experience in audio and video information processing. We select 10 video clips from the partitioned HDTF evaluation set and extract their audio to generate video clips using the models outlined in Sec. 4.5. For audio-video driven models that require video input, such as DreamTalk [23] and StyleTalk [22], the original videos are also provided as inputs to these models. In contrast, our model is operated solely in an audio-only driven manner without incorporating additional emotion control conditions, ensuring fairness in input information. We employed several subjective evaluation criteria and defined detailed quantitative metrics for each score level in the subjective standards. These requirements are thoroughly documented in an evaluation guide provided to the participants. Our designs ensured the validity and fairness of the user study.

7.2. Comparison with More Recent Methods

To further demonstrate the superiority of our method over the latest approaches, we include a comparison with other recent state-of-the-art methods. The baseline methods are: EchoMimic [6], AniPortrait [44], and EMOPortraits [10]. It is worth noting that the AniPortrait comparison includes both the audio-only driven and video-only driven modes. However, due to the audio-video driven model for EMO-Portraits is not open-sourced, we only compare the results from its video-driven mode. We conducted the experiment on the HDTF [51] test set, the testing procedure is strictly consistent with Sec. 4.2, and the testing metrics have been previously defined in Sec. 4.1. The results are shown in Tab. 5.

Methods	Driven	FID (↓)	Sync-C (↑)	Sync-D (↓)	E-FID (↓)
AniPortrait	A	17.71	3.75	10.63	1.21
EchoMimic	A	16.68	6.74	8.49	0.78
Ours	A	**16.64**	**8.24**	**7.09**	**0.54**
EMOPortraits	V	19.68	7.38	7.65	0.44
AniPortrait	V	17.34	6.82	7.96	0.40
Ours	A+V	**16.09**	**8.41**	**7.11**	**0.34**

Table 5. Comparison with recent methods on HDTF dataset. "A" denotes audio-only driven, "V" denotes video-only driven and "A+V" denotes audio-video driven. "↑" indicates better performance with higher values, while "↓" indicates better performance with lower values.

The results demonstrate that our approach outperforms the latest methods in audio-only driven tasks, particularly in terms of lip synchronization and expression similarity. Furthermore, for video-driven tasks where reference videos are available, our method, which can integrate audio information and video information, also surpasses the latest video-only driven approaches. These aforementioned results serve as a supplementary addition to Tab. 1 of the main paper, further substantiating the superiority of our method in generating expressive talking heads.

7.3. Qualitative Ablation Study on V-AID

Sec. 4.3 of the main text presents a quantitative ablation experiment on the proposed V-AID module in Tab. 2. As a supplement, we also conduct a qualitative ablation of the V-AID module to further test the effect of introducing the V-AID module. In our study, we employed 18 samples of paired reference images and audio clips to generate talking head videos using the following two methods:

Methods	Lip-Sync (↑)	Exp-Q (↑)	Preference (↑)
end-to-end (w/o V-AID)	3.98	3.59	0.21 (21%)
w/ V-AID	**4.28**	**4.19**	**0.79 (79%)**

Table 6. User study on the ablation of V-AID module. The study involved 20 participants to evaluate 18 paired samples.

Modules	V-AID	VAE (decoder)	Backbone (ETHD)
TFLOPs	**0.45**	180.51	3668.16

Table 7. Computation complexity of each module of EmotiveTalk.

- End-to-end (w/o V-AID): End-to-end driven mode without the V-AID module;
- w/ V-AID: Driven with the V-AID module.

Then, we conducted a user study of 20 participants. For each method, the participant is required to give a rating (from 1 to 5, 5 is the best) for each generated sample on two aspects: (1) the lip sync quality (Lip-Sync), (2) the quality of expressions (Exp-Q), and finally select a subjectively preferred sample from those generated using different methods with the same paired image-audio data. The results of the user study are shown in Tab. 6.

The results indicate that the samples generated with the V-AID module exhibited superior generation quality, including enhanced lip synchronization and improved expression quality, and are more preferred by the participants. These findings demonstrate the effectiveness of the V-AID module in the generation of expressive talking head videos. Representative ablation samples can be found in our page.

7.4. Complexity Analysis on the Overall Framework

To further investigate the impact of integrating the V-AID module on the overall system complexity, we conducted an additional complexity analysis of the entire system. We utilized our overall framework to generate a 120-frame, 512x512 resolution talking head video and analysis the computational complexity of each module during the generation process. The results are presented in Tab. 7.

Based on the results of Tab. 6 and Tab. 7, the proposed V-AID module occupies only a minimal proportion of the overall computational complexity. However, it markedly enhances the quality of talking head video generation, achieving superior lip synchronization and more expressive facial animation. This optimal balance between computational complexity and performance effectively validates the efficacy of our proposed V-AID module.

7.5. Ablation Studies on EDI

The design of the EDI module aims to achieve the automatic removal of the emotion information from the reference image while injecting target emotion information into the hidden states during both training and inference. This enables the transformation of facial expressions from the reference image to the target expression. To evaluate the effectiveness of the EDI module, we conduct the following ablation experiment. The experiments are performed on the test set of MEAD, as described in Sec. 4, due to the extensive diversity of speaker emotions in the MEAD dataset, with three different configurations:

- Without the EDI module (w/o EDI);
- Direct injection of target emotional information (DI);
- Using the EDI module for emotional information injection (w/ EDI).

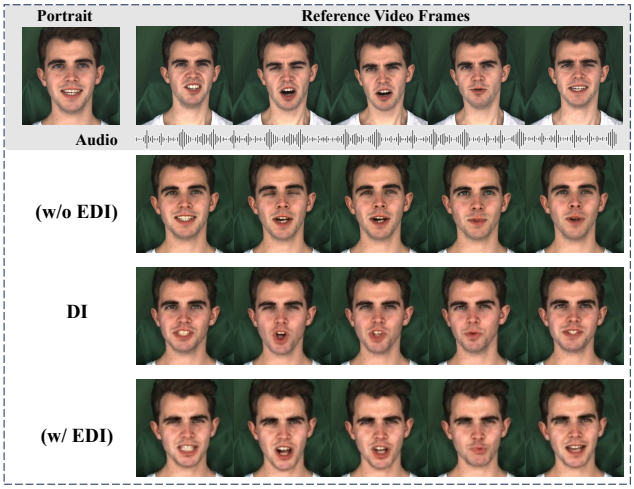

Figure 12. Ablation studies on EDI block.

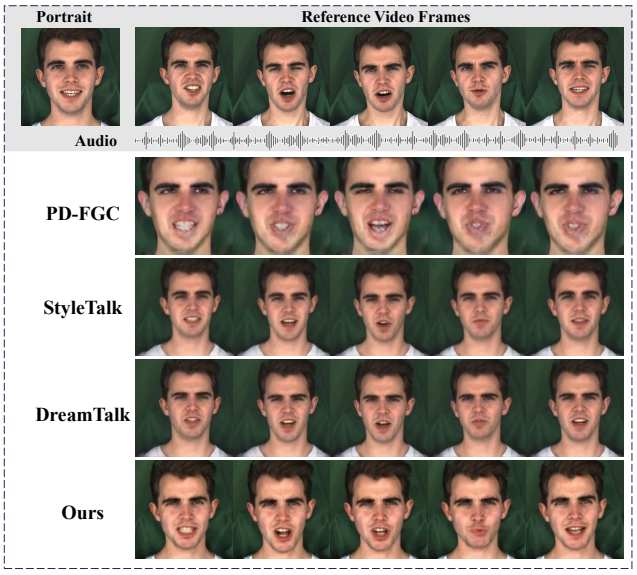

Figure 13. Comparison of emotion transferring with other methods.

Methods	Driven	HDTF / MEAD FID (↓)	FVD (↓)	Sync-C (↑)	Sync-D (↓)	E-FID (↓)
Ours (512)	A	16.64 / 53.21	140.96 / **207.67**	**8.24 / 6.82**	**7.09 / 7.43**	**0.54** / 0.57
Ours (1024)	A	**12.75 / 37.97**	**134.64** / 303.55	7.04 / 4.60	8.09 / 9.58	0.62 / **0.54**
Ours (512)	A+V	16.09 / 50.84	**120.70 / 153.71**	**8.41 / 6.79**	**7.11 / 7.58**	**0.34 / 0.40**
Ours (1024)	A+V	**12.68 / 38.33**	147.40 / 275.04	7.85 / 5.00	7.82 / 9.55	0.42 / 0.43
Ground Truth	A+V	-	-	8.63 / 7.30	6.75 / 8.31	-

Table 8. Comparisons of our models with different resolutions on HDTF and MEAD. "A" denotes audio-only driven and "A+V" denotes audio-video driven. "↑" indicates better performance with higher values, while "↓" indicates better performance with lower values.

The results are presented in Fig. 12. The results demonstrate that without utilizing the EDI module (w/o EDI), it is challenging to achieve a transition from happy to angry in the emotion of generated video, even when provided with video information depicting anger. The direct injection approach (DI) fails to effectively eliminate the influence of the happy emotion from the reference image, resulting in the residual coupling of happiness in the generated images and weakening the expression of anger. In contrast, the conditional injection method using the EDI module effectively removes the residual coupling of happiness, enabling a more expressive and accurate transformation from the happy reference image to the generated video portraying anger.

Furthermore, to evaluate the performance of emotional state transfer, we compare our proposed method with other state-of-the-art emotion control methods to compare the results. Fig. 13 shows the qualitative results on emotion control generation by emotion reference video, where "ours" denotes our model with EDI. Results show that StyleTalk and DreamTalk fail to preserve speaker identity. They inadvertently reveal extraneous positional information about the reference video, as the spatial positioning and head size in the generated video are aligned with the emotion reference video rather than maintaining consistency with the reference image. PD-FGC faces challenges in low-definition issues in the lip region. Compared to all the methods, our method with EDI achieves the most expressive emotion control results while preserving speaker identity, resulting in the best performance among all the methods.

7.6. Comparison on Different Resolutions

To quantitatively compare the impact of different resolution training strategies on EmotiveTalk, we have trained two models at resolutions of 512 and 1024 using the configurations described in Sec. 6.4. The same quantitative evaluation metrics as outlined in the main paper Sec. 4 are used to validate performance on the evaluation set of HDTF and MEAD. The results are shown in Tab. 8.

The results demonstrate that models trained at different resolutions exhibit distinct strengths across various metrics on both datasets. The 1024-resolution model significantly outperforms the 512-resolution model in terms of the FID metric, highlighting the superior ability of high-resolution training to better preserve image details. Conversely, the 512-resolution model achieves notably better performance on lip-sync metrics, Sync-C and Sync-D in Tab. 8, underscoring the positive impact of long-time training on generating lip movements that are more consistent with the audio.

Methods	Params (↓)	TFLOPs (↓)	Time (↓)
Hallo [45]	2.17G	4388.32	501.97
Ours	**1.58G**	**3668.16**	**435.18**

Table 9. Comparisons of network efficiency of denoising backbone with diffusion-based methods. "Params" denotes the network parameters of the backbone, "TFLOPs" denotes the computation cost, and "Time" denotes the time cost. "↓" indicates better performance with lower values.

7.7. Comparison on Network Efficiency

To further investigate the advantages of our model's efficient design compared to other diffusion-based talking head generation models, we conducted a comprehensive evaluation from two perspectives: model parameters and inference computational cost. We include a representative method Hallo [45] that is similar to our method, which also employs a 3D U-Net architecture and utilizes pre-trained models of Stable Diffusion [28]. For a fair comparison, we test the total parameters of the backbone network and computational flops of generating the same length of talking head video

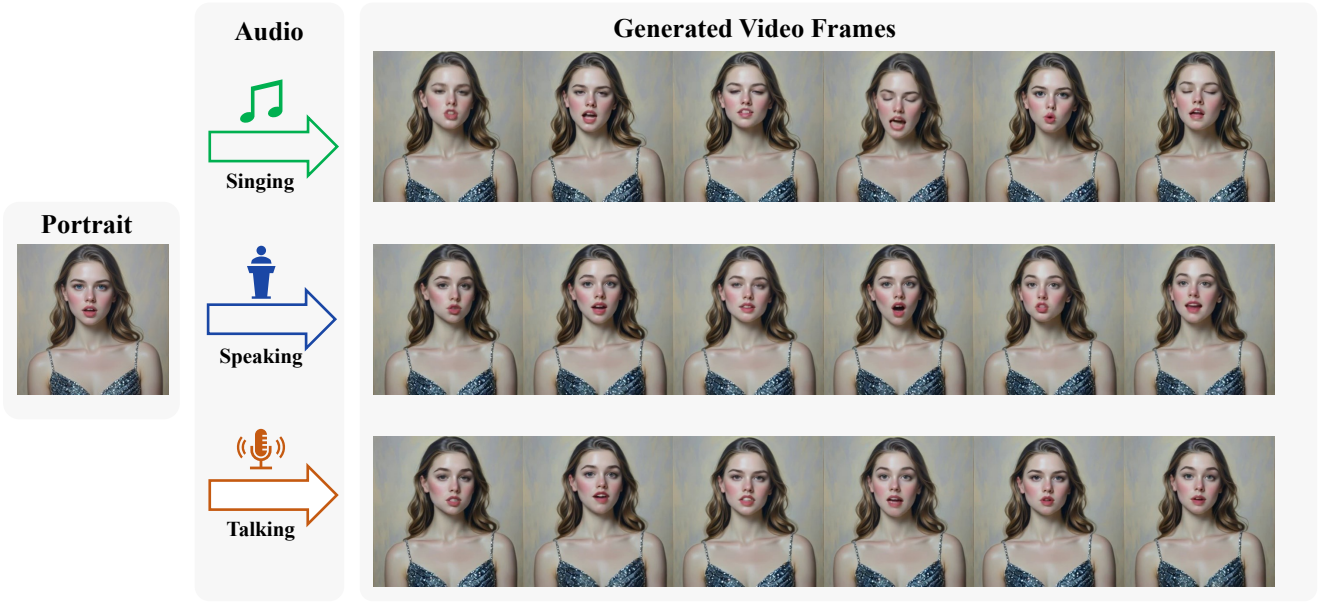

Figure 14. Video generation results of EmotiveTalk given different audios of variable styles.

(4.8s for each model exactly) with the inference settings of each model, respectively. The comparison results are summarized in Tab. 9.

From the metrics, it is evident that our method demonstrates clear advantages over the mainstream reference-denoising UNet architecture, represented by Hallo [45], in terms of parameter count, computational cost, and generation time. Furthermore, as discussed in Tab. 1 of the main paper, our approach also outperforms Hallo in terms of performance on most of the evaluated metrics. This highlights the efficiency of our network design, achieving a balance between performance and efficiency while maintaining superior performance.

7.8. Additional Generation Results

Results on Multiple Audio Styles. We conduct experiments to compare the driving effects of different audio styles on the same reference portrait by generating videos using varying audio inputs. Three distinct audio styles are employed:

- Singing: Driven by English song audio;
- Speaking: Driven by English speech recordings;
- Talking: Driven by English daily talking audio.

The results are shown in Fig. 14. In the visualized results, we observe that using singing audio introduces more frequent blinking, expressive facial changes, and rhythmic head movements. Using speaking driven achieves a vivid and dynamic result, characterized by rich facial expressions, pronounced mouth movements, and noticeable head movements. On the other hand, the result driven by daily talking audio achieves realistic and natural driving effects. These visualized results demonstrate the effectiveness of our proposed method in handling diverse audio styles for driving animations.

Results on Multiple Portrait Styles. To further validate the generalization capability of our proposed method across different portrait styles, we use the same speech audio to drive three distinct styles of portraits: photorealistic, cartoon, and sketch, and generate speech videos corresponding to each portrait. To compare the results, we visualize frames captured at the same time stamps from the generated videos, as shown in Fig. 15. The results demonstrate that our method successfully produces realistic speaking videos for all three styles. Moreover, the lip movements across the three videos remain highly consistent at corresponding time points, confirming the excellent generalization performance of our method across diverse portrait styles.

Multiple Languages. Supplementary video shows that our method generates satisfactory results with speech in French, Chinese, English, and even minority language like Cantonese. This is primarily attributed to the pre-trained wav2vec encoder's strong generalization capability across different languages, which enhances the versatility of the EmotiveTalk framework in generating talking head videos across diverse linguistic contexts

Results on Multiple Emotion Generation. To further evaluate the expressive capability of our proposed method in controlling and generating different emotional states, we applied the emotion control pipeline introduced in Sec. 3.5. Using the same portrait and audio input, we controlled the output by specifying different emotional states via text-based control condition. Four common emotional states,

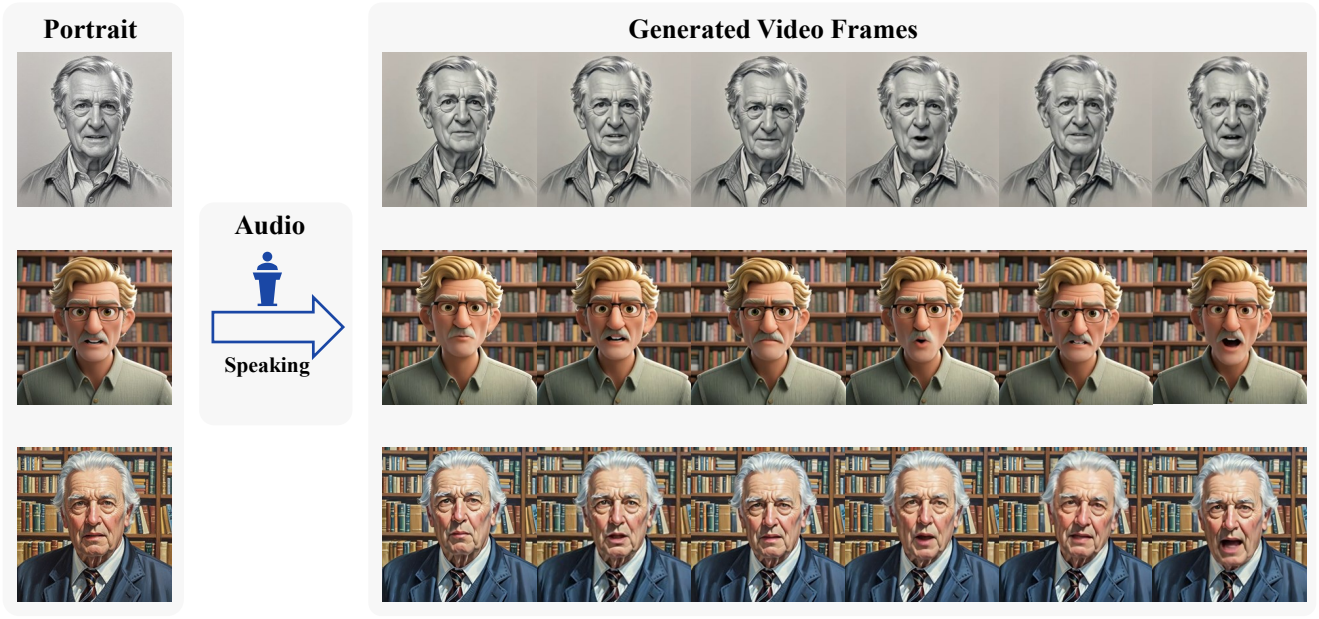

Figure 15. Video generation results of EmotiveTalk given different portraits and same speech audio signal.

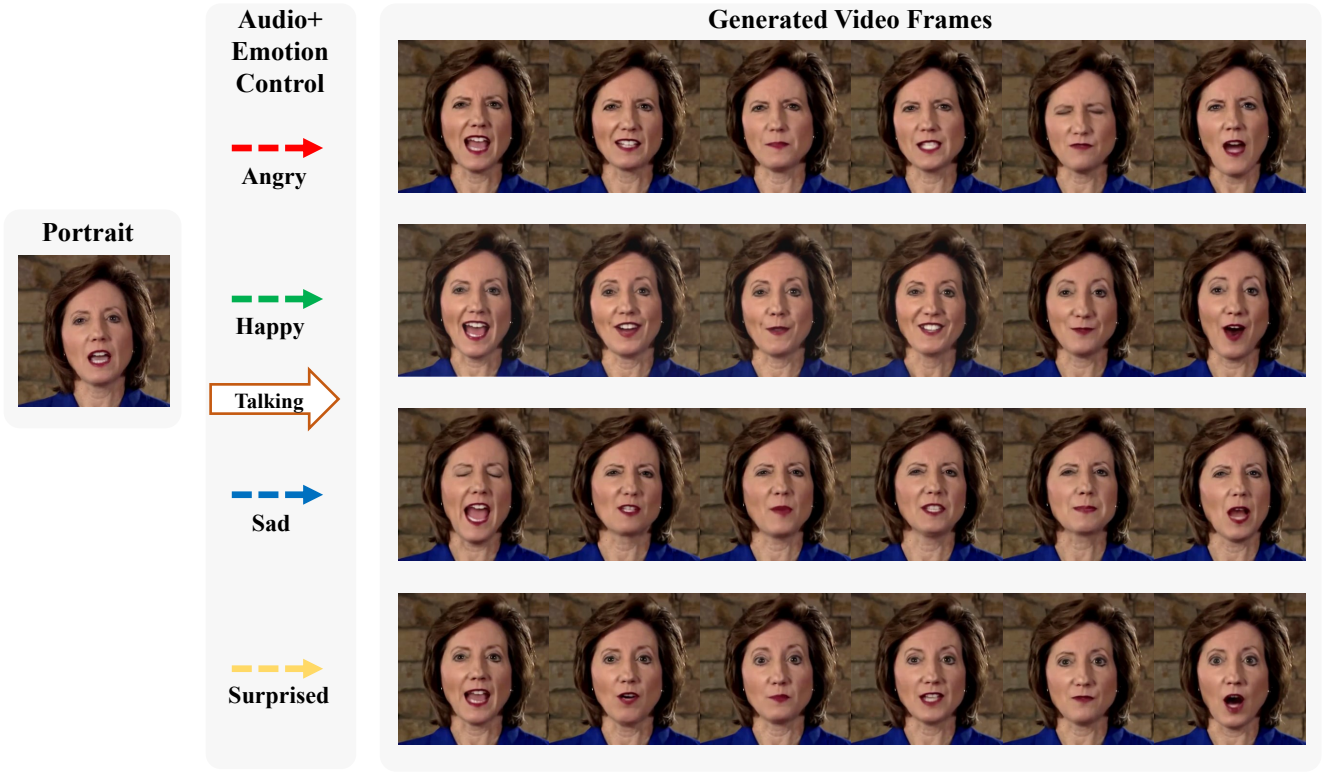

Figure 16. Video generation results of EmotiveTalk given same portraits and same speech audio signal but different emotion control conditions to generate different emotions.

including angry, happy, sad, and surprised, are generated in our experiment. Frames captured at the same time stamps from these videos are visualized for comparison, as shown in Fig. 16.

The results demonstrate that our method can produce vivid, natural, and realistic emotional expressions. Additionally, the lip movements across the videos remain highly consistent at corresponding time points under varying emotional conditions. This confirms that our method effectively decouples the influence of facial expressions and lip synchronization, allowing for accurate lip movements synchronized with speech while transferring emotional states.

8. Limitations and Future Work

Despite EmotiveTalk's promising advancements in expressive talking head generation and emotion control ability, it still encounters several challenges that open the way for future research.

First, EmotiveTalk occasionally experiences motion blur during significant body movements or dramatic facial expression changes, which reduces the resolution of the generated video. This issue is associated with motion blur in the training data during periods of intense movement. A potential solution could involve applying motion blur detection and video processing techniques to the original training videos to eliminate motion blur.

Second, as there are currently no publicly available talking head video datasets with textual annotations, EmotiveTalk currently supports a limited range of emotion categories controlled via textual input, focusing on discrete emotional states. Future work could explore temporally annotating fine-grained facial expressions and emotion states in the training data using Multimodal Large Language Models (MLLMs) to build a talking head dataset with fine-grained textual annotations to support the research on text-guided finer-grained emotional control.

Lastly, EmotiveTalk presently focuses exclusively on controlling emotion states in talking head video generation and does not incorporate explicit control over head movements. In the generated videos, head motion primarily arises from the sampling of the diffusion model. Future enhancements could include explicit control signals for head movement, enabling precise manipulation of desired head motion patterns.

Despite these challenges, EmotiveTalk demonstrates exceptional performance and application potential in generating stable videos with expressive and controllable facial expressions. It holds significant academic and practical value, providing a foundation for future research in the field of talking head generation.

9. Ethical Consideration

EmotiveTalk is capable of generating highly realistic talking head videos that are difficult to distinguish from genuine footage, endowing it with extensive practical value. While EmotiveTalk holds significant positive implications for social development and technological advancement, assisting professionals in practical domains such as human-computer interaction, remote education, and caregiving companionship, the potential misuse of EmotiveTalk could lead to the spread of misinformation. For instance, it could be exploited to create fake videos using the portraits of celebrities, produce videos containing sexual innuendo or violent content, or generate counterfeit videos for purposes of extortion. Such misuses may result in substantial negative impacts, which contravene our fundamental intent of leveraging artificial intelligence to enhance human creativity, drive technological progress, and improve our society. These are outcomes we find absolutely intolerable.

We have taken the risk of potential misuse into careful consideration throughout the development of EmotiveTalk. During the training phase, we meticulously curated the training data to rigorously exclude any undesirable content involving violence, sexual implications, or horror themes. Furthermore, we have imposed strict limitations on the use of EmotiveTalk. The version deployed for academic research purposes is under the supervision of our risk assessment team. All images and audio inputs used to generate talking head videos undergo stringent evaluation and review to ensure that EmotiveTalk is not utilized to produce inappropriate information or content. For potential future releases of EmotiveTalk models intended for engineering applications, we also plan to implement a stringent review and assessment process to guarantee that generated content remains free of harmful materials. Moreover, we advocate for research on advanced forgery detection techniques, which can identify synthetic fake images and videos, thereby helping to mitigate illegal use. We remain resolute in our commitment to preventing the generation of harmful content and mitigating adverse societal impacts, and we will fully address and prevent the various potential misuses of EmotiveTalk.