Emotional Talking Face Generation via Spatial-Temporal Decoupled Training and Continuous Emotion Guidance

Yuhai Deng
School of Computer Science
Nankai University
yuhai_deng@126.com

Abstract

Diffusion-based models have recently advanced talking face generation by producing high-resolution videos with accurate audio-lip synchronization. However, generating vivid and fine-grained facial expressions remains challenging due to two key limitations. First, the lack of fine-grained emotional representations limits expressive control and hinders the modeling of subtle emotional nuances. Second, the temporal module in diffusion models enforces smoothness constraints during training, which induces a coupling between temporal modeling and expression learning. This coupling limits the model's ability to learn fine-grained and dynamic facial expressions. To address these issues, we introduce a continuous Valence-Arousal (VA) emotion representation and a spatial-temporal decoupled training strategy. The VA representation enables precise and interpretable emotional control, while the decoupled strategy mitigates the training-time coupling. Specifically, we use VA signals to guide expression generation and first train the model without the temporal module to learn precise emotion-to-expression mappings, then introduce the temporal module to ensure temporal consistency. This design enhances emotional control and improves the alignment between facial expressions and audio-driven emotions.

1 Introduction

Talking face generation has gained widespread attention due to its applications in digital communication [1], virtual assistants [2], and AI-driven content creation [3]. While early studies primarily focused on accurate audio-lip synchronization, achieving truly natural and engaging talking videos requires the synthesis of vivid and expressive facial behaviors. In particular, generating expressive talking videos with vivid and fine-grained facial expressions is crucial for conveying emotions and intentions [4, 5], which in turn enables more engaging and realistic human-computer interactions.

Recent studies [6, 7, 8, 9, 10, 11] in talking face generation have achieved impressive results in audio-lip synchronization. Notably, with the advancement of video diffusion models, diffusion-based methods [12, 13, 14, 15, 16, 17] enhance visual quality through high-resolution and temporally coherent video synthesis. However, these methods provide limited control over facial emotions. To mitigate this constraint, several recent works [9, 10] have explored emotion guidance by conditioning the generation process on discrete emotion categories [18, 10] or global affective vectors [9]. For example, MoEE [18] employs discrete emotion labels and introduces emotion-specific experts to decouple six basic emotions, enabling the synthesis of both base and compound expressions. Despite their progress, their use of coarse emotion representations restricts the generation of fine-grained expressions, such as gradual changes in emotional intensity and smooth transitions between similar affective states.

Preprint. Under review.

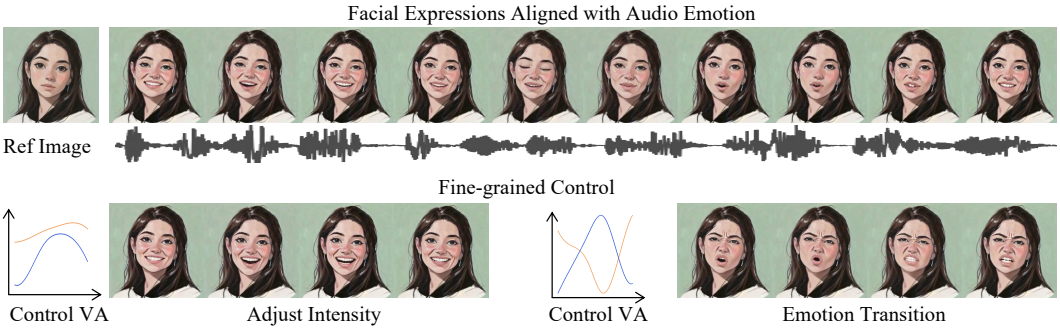


Figure 1: Our method enables expressive talking face generation with fine-grained emotional control. (Top) Given a reference image and an audio clip, our model generates a talking face video with expressions that align with the affective state of the speech. (Bottom left) By adjusting a specified Valence-Arousal (VA) sequence, our approach allows for fine-grained control over emotion intensity, enabling nuanced expression manipulation. (Bottom right) Our method supports emotion transition, enabling dynamic expression changes over time.

A key limitation of current methods lies in the entanglement between temporal consistency modeling and the learning of fine-grained expressions. Temporal attention mechanisms are commonly introduced to ensure smooth transitions across frames. However, such mechanisms often lead to overly smooth facial motions, suppressing subtle variations across frames. Consequently, the joint optimization of spatial and temporal objectives within a single-stage training pipeline limits the model's ability to capture expressive, dynamic facial changes. Another limitation lies in the use of discrete emotion categories (e.g., 'happy', 'sad'). Although these categories provide high-level emotional control, they fail to capture nuanced intra-class variations. For instance, a gentle smile and an exuberant grin are both labeled as 'happy' but differ significantly in expression intensity and visual dynamics. Such coarse labels restrict fine-grained control over facial behavior, ultimately limiting the expressiveness and diversity of the generated videos.

To address these issues, we propose a framework that integrates continuous emotion representation with a spatial-temporal decoupled training strategy. Specifically, we adopt the Valence-Arousal (VA) space [4] as a continuous affective representation, where valence reflects emotional polarity, and arousal denotes the level of activation. This allows for more precise and flexible control of facial expressions across a broad emotional spectrum. Furthermore, to alleviate the coupling between spatial expression learning and temporal smoothness during training, we introduce a two-stage spatial-temporal decoupled training strategy. In the first stage, we train the model on image-level data without temporal modules, encouraging accurate and fine-grained expression synthesis from VA signals without being constrained by temporal smoothness. In the second stage, we reintroduce temporal modules and fine-tune the model on video sequences to ensure temporal coherence. This progressive strategy not only alleviates the over-smoothing effect caused by temporal constraints but also enhances data efficiency by reducing reliance on large-scale annotated videos. As illustrated in Figure 1, our approach enables expressive talking face generation with fine-grained emotional control, producing facial behaviors that align with the underlying emotional cues in the audio and supporting continuous emotion editing.

In summary, our primary contribution is enabling fine-grained control of facial expressions in talking face generation. To this end, we make the following contributions:

- We adopt the Valence-Arousal (VA) emotion space as a continuous and fine-grained representation of affective states, enabling more nuanced and flexible emotion control compared to discrete emotion categories.
- We propose a spatial-temporal decoupled training strategy, which first trains the model on image-level data for precise expression learning and then fine-tunes it on video sequences to ensure temporal coherence. This efficient two-stage process facilitates fine-grained expression control with limited training data.
- Through extensive experiments, we demonstrate that our method outperforms existing approaches in expression controllability and realism, delivering expressive talking face videos with dynamic and vivid facial movements.

2 Related Work

Audio-driven Talking Face Generation. Audio-driven talking face generation, which aims to synthesize a realistic talking video from a reference image and an audio clip, has been widely explored. Early methods [19, 20, 21] focused on achieving lip synchronization by directly mapping audio features to lip movements, while maintaining other facial attributes unchanged. Subsequent works [22, 23, 24, 25, 26] employed GAN-based models to enhance visual realism and improve the quality of synthesized talking faces. To further improve temporal coherence and controllability, intermediate representations such as facial landmarks [7, 27, 24], 3D head poses [28, 29], and 3DMM parameters [30] have been introduced. Recently, diffusion models [31, 32] have emerged as a powerful framework for generative modeling, driving significant advancements in high-fidelity image and video generation [33]. Many models [13, 12, 14, 15, 34] built on the foundation of pre-trained diffusion models [32], achieving high-fidelity talking face video generation. We also adopt a diffusion-based generative framework to ensure high-quality video synthesis.To facilitate more effective learning of expressive facial behaviors, we employ a spatial-temporal decoupled training strategy. Specifically, the model is first trained on single frames to learn accurate spatial expression mappings, and then temporally fine-tuned to capture dynamic consistency across frames.

Facial Expression Generation in Talking Face Video. Facial expression generation in talking face video enhances realism by modeling emotional variations beyond lip motion. Most prior methods adopt coarse-grained emotion control, using discrete emotion labels [10, 35, 36, 37, 38, 39] (e.g., happy, sad, angry) or emotion style images [8, 40, 41, 42] to guide expression generation. These approaches help produce visually plausible expressions but lack flexibility and fail to capture subtle, dynamic facial changes. Recent work explores fine-grained control of facial expressions. Some methods [43, 44] extract emotion patterns from expression reference videos, such as 2D landmarks, allowing one-shot generation for unseen identities. While these methods offer finer control, they often rely on auxiliary driving videos or manually defined expression codes, increasing complexity and limiting generalizability. Some approaches [45, 36] extend discrete labels with emotion intensity values, which allows partial refinement but still lacks continuous control. Another line of work leverages facial Action Units (AUs) [46] for fine-grained control. For example, FG-EMOTalk [47] uses a predefined set of AUs to modulate expressions, enabling localized control over facial muscle movements. However, AU-based methods rely on manually designed codes and lack a compact, interpretable representation of overall emotion. In contrast, we leverage the continuous Valence-Arousal (VA) space [4], which provides a compact and interpretable representation of affective states. Compared to discrete labels and AUs, VA enables smooth and scalable emotion control, allowing the model to generate expressive behaviors that better reflect subtle emotional variations. Our approach utilizes Valence-Arousal guidance to achieve fine-grained expression control and generate talking face videos with expressions consistent with the emotional state of the speech.

3 Method

Our framework aims to generate expressive talking face videos that are both lip-synchronized to the input speech and controllable in facial emotion. As illustrated in Figure 2, the inputs include a speech audio segment, a reference image of the target speaker, and an optional emotion vector represented in the Valence-Arousal (VA) space. The emotion condition can be either explicitly provided or inferred from the audio, enabling both standard talking face generation and fine-grained expression editing. The core of the framework is a diffusion-based generator built upon a 3D U-Net backbone [33], which incorporates a spatial-temporal separable attention mechanism [48]. The spatial module integrates reference, audio, and emotion features via cross-attention to model speaker identity, lip movements, and emotional cues, respectively. The temporal module then applies self-attention across frames to ensure smooth and coherent facial dynamics. To facilitate effective learning, we first train the spatial module independently for accurate expression modeling, and then jointly train it with the temporal module to ensure temporal coherence.

3.1 Preliminaries

Latent diffusion models (LDMs) [32] are a class of diffusion models that perform denoising in a compressed latent space, significantly reducing computational cost while preserving generation quality. A pretrained autoencoder $\mathcal{E}, \mathcal{D}$ is used to map an image $\mathbf{x}$ into a latent representation $\mathbf{z} =$

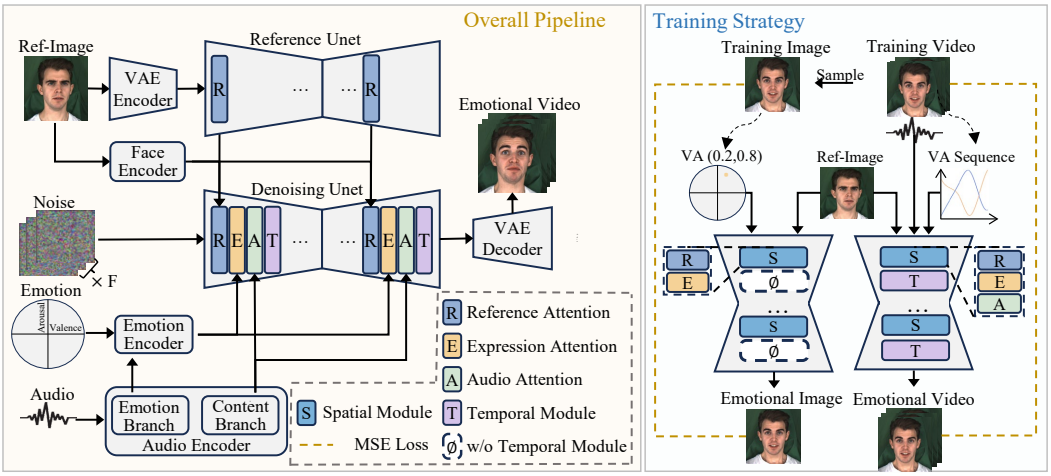


Figure 2: Overall framework and training strategy. The model generates expressive talking face videos conditioned on audio, reference image, and emotion signals in the Valence-Arousal space. The training follows a two-stage decoupled strategy, starting with static expression learning from images and proceeding to dynamic modeling on videos for temporal consistency.

$\mathcal{E}(\mathbf{x})$, where the diffusion process is performed. After denoising in the latent space, the final result $\hat{\mathbf{z}}_0$ is decoded back into the image domain via the fixed decoder $\mathcal{D}$. The model learns to reverse a noise process by predicting either the clean data $\mathbf{x}_0$ or the added noise $\boldsymbol{\epsilon}$ at each timestep.

Conditioning via Cross-Attention. To enable controllable generation, LDMs incorporate conditioning signals (e.g., text, image, audio, emotion) through cross-attention mechanisms [49] in the denoising network. At each timestep t, the model predicts noise via a conditional denoising function $\boldsymbol{\epsilon}_\theta(\mathrm{z}_t, t, \mathrm{c})$, where c represents the encoded condition. Cross-attention is used to inject c into the latent feature stream:

$$\mathrm{Attn}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \mathrm{softmax}\left(\mathrm{QK}^\top / \sqrt{d_k}\right)\mathrm{V}, \tag{1}$$

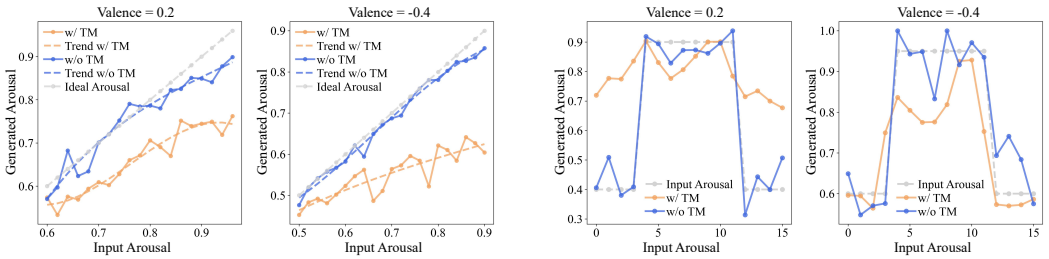
where the latent feature z_t is linearly projected to queries Q, and the condition c is encoded into keys K and values V. This allows the model to dynamically align the generation process with the conditioning signal. In our framework, we leverage this mechanism to inject continuous Valence-Arousal emotion representations into the denoising trajectory, enabling fine-grained and temporally consistent expression control in talking face generation.

Temporal Module. In talking face generation, temporal consistency is typically modeled using self-attention across frames [48]. The input feature tensor z with shape $\mathbb{R}^{b\times c\times f\times h\times w}$, where b is the batch size, c is the feature dimension, f is the number of frames, and h and w are the spatial dimensions, is reshaped into $(b\times h\times w)\times f\times c$ to apply self-attention along the temporal dimension f. Temporal layers are incorporated into the Backbone Network at each resolution level to capture the dynamic content of the video. These temporal modules offer advantages in ensuring smooth transitions and maintaining temporal coherence. However, they often result in overly smooth expressions, which can limit the diversity of emotional variations.

3.2 Continuous Emotion Representation

In affective computing and psychology [50, 51], the Valence-Arousal (VA) model is a widely adopted dimensional emotion representation [52] that characterizes emotional states within a continuous two-dimensional space. Specifically, Valence quantifies the degree of emotional pleasantness, ranging from negative (e.g., sadness, anger) to positive (e.g., happiness, satisfaction), while Arousal measures the level of activation, spanning from low (e.g., calmness, fatigue) to high (e.g., excitement, anxiety). This continuous representation provides greater flexibility compared to discrete emotion categories, enabling more precise control over emotional expressions (see Appendix A for a detailed discussion).

Our approach leverages the VA model as a continuous emotion condition to guide facial expression generation. During training, we employ Emonet [53], a pre-trained emotion regressor on large-scale



(a) Spatial suppression of expression intensity by temporal modules.

(b) Temporal smoothing behavior induced by temporal attention modules.

Figure 3: Impact of temporal modules on expression controllability. (a) At the frame level, temporal modules suppress expression intensity, resulting in flatter arousal responses. (b) In terms of temporal dynamics, they introduce delayed reactions to changing arousal inputs, leading to overly smooth transitions across frames.

facial expression datasets, to extract frame-level VA values. These VA values serve as conditional inputs for the generation model. At inference time, the emotion condition can either be inferred from the input audio using a speech emotion recognition model [54] or specified by the user. This enables both standard talking face generation and fine-grained expression editing.

To integrate VA condition into the generation pipeline, each (Valence, Arousal) pair is processed by a lightweight multi-layer perceptron (MLP), referred to as the Emotion-Encoder, which transforms the low-dimensional emotion vector into a compact latent representation. This representation is then used in the VA-Attention module, where emotional information is dynamically injected into the generation process. Specifically, inspired by recent advances in cross-attention for semantic guidance in diffusion models [55, 32, 56], the VA-Attention treats the VA embeddings as keys and values, while the intermediate UNet features serve as queries. This design allows the model to selectively attend to emotion signals at each denoising step, aligning local spatial details with global affective context.

3.3 Spatial-Temporal Decoupled Expression-Progressive Training

To better capture fine-grained facial expressions, we propose a spatial-temporal decoupled training strategy. Instead of jointly optimizing spatial and temporal modules from the beginning, we adopt a two-stage approach. We first train the model without temporal modules to accurately capture frame-level emotional expressions. Then, we fine-tune the model with temporal modules to learn coherent motion patterns. This design mitigates the over-smoothing effects caused by early temporal regularization and helps balance expression fidelity and temporal consistency. The following sections present an analysis of the temporal constraints, along with detailed descriptions of the spatial-temporal decoupled training strategy.

3.3.1 Analysis of Temporal Smoothness Constraints in Temporal Modules.

A key challenge in expressive talking face generation lies in balancing fine-grained emotional expressiveness with temporal coherence across video frames. When temporal modules are incorporated from the early stages of training, they may impose excessive temporal regularization that suppresses the model's capacity to learn expressive and emotionally accurate facial details. As a result, the model tends to prioritize temporal continuity at the expense of accurate and vivid facial expression modeling. Specifically, we observe two forms of constraint: suppressed expression intensity at the frame level, and delayed responses to changes in emotional input over time.

To analyze these effects, we compare two models, one trained with temporal modules from the beginning, and one trained without them. Our first analysis focuses on frame-level emotional controllability. We construct a series of inputs with varying arousal levels while fixing the valence value, and use a pre-trained emotion regressor to estimate the emotional content of the generated faces. As shown in Figure 3a, the model trained without temporal modules produces outputs that closely follow the target arousal trajectory, indicating precise control over fine-grained expression

strength. In contrast, the model trained with temporal modules exhibits notably flatter responses, suggesting that temporal constraints hinder sensitivity to input-driven expression variations.

We further examine the impact of temporal modules on dynamic emotional responsiveness. In this experiment, we introduce abrupt transitions in the arousal input to assess the model's ability to adapt expressions over time. As illustrated in Figure 3b, the model without temporal modules adjusts facial expressions in a timely and direct manner. Conversely, the model with temporal modules exhibits gradual transitions through intermediate expression states, resulting in a noticeable temporal response lag. This smoothing effect indicates a temporal rigidity that limits the model's capacity to reflect sudden emotional shifts.

These observations suggest that although temporal modules enhance inter-frame consistency, they also constrain expression fidelity and responsiveness when applied prematurely. To mitigate these effects, we adopt a decoupled training strategy. We first optimize the model without temporal modules to capture accurate static emotional representations, and then fine-tune it with temporal modules to enhance temporal consistency. This progressive approach preserves fine-grained emotion control while ensuring temporally coherent talking face generation.

3.3.2 Static Expression Learning

In the first stage, we train the model without temporal modules to focus purely on learning the mapping from emotional signals to facial expressions at the frame level. Each training sample consists of a reference image and a target valence-arousal pair, and the model is tasked with generating an expressive face that reflects the specified emotion while preserving the identity of the reference. By excluding temporal components, we avoid premature regularization and allow the spatial module to develop precise, emotion-controllable expression generation.

To construct the training pairs, we randomly sample two images of the same speaker, where one is the reference image and the other shows a different emotional expression as the target image. We then extract the valence-arousal values from the target image as the conditioning signal. This setup enables the model to learn how to edit the emotional expression of the reference image to match the target emotional state while preserving the speaker's identity.

To ensure that emotional conditioning is focused on the expressive regions of the face, we apply a face mask to modulate the cross-attention response between the spatial features z_s and the emotion condition c_e. The updated representation is computed as:

$$z'_s = z_s + \mathrm{CrossAttn}(Q(z_s), K(c_e), V(c_e)) \cdot M, \tag{2}$$

where M is a binary mask that highlights the facial region, $Q(z_s)$ denotes the query projection of the latent z_s, and $K(c_e), V(c_e)$ are the key and value projections of the emotion condition c_e. This spatially-aware modulation improves the model's ability to generate accurate expressions by restricting emotional influence to the relevant facial areas.

This static training phase provides a strong initialization for expressive generation. Once the model has learned to produce emotionally aligned frames, we introduce temporal modules in the next stage to refine motion continuity without degrading expression fidelity.

3.3.3 Temporal Adaptation for Expression Dynamics

Following the static expression learning stage, we incorporate the temporal attention module in the model to enhance inter-frame coherence. At this stage, the model takes raw audio, a reference identity, and the corresponding valence-arousal sequence as input, aiming to generate talking face videos with both expressive accuracy and temporal consistency. We also introduce an audio branch to estimate valence-arousal signals directly from speech. This branch is initialized from a pretrained Transformer-based emotion encoder proposed in [54]. During fine-tuning, we update its emotion prediction head and final Transformer layer to better align audio-derived emotional cues with the visual expression synthesis. Through this progressive fine-tuning process, the model integrates static expression controllability, temporal coherence, and speech-driven emotional guidance, resulting in expressive and temporally fluid talking face generation.

Table 1: Quantitative comparisons with state-of-the-art methods.

	Method	MEAD [38]					RAVDESS [57]				
		FID ↓	FVD ↓	E-FID ↓	Acc_{emo} ↑	Sync-C ↑	FID ↓	FVD ↓	E-FID ↓	Acc_{emo} ↑	Sync-C ↑
GAN-based	Wav2Lip [21]	99.78	856.82	5.20	16.24	**8.84**	34.01	549.21	4.63	7.69	**6.40**
	SadTalker [25]	102.34	678.64	5.32	16.24	5.32	38.21	731.43	5.16	7.69	3.51
	EAT [10]	73.01	592.18	1.82	64.37	7.71	27.58	538.56	3.09	34.62	5.16
	EDTalk [11]	84.36	515.36	2.55	68.83	7.76	36.66	359.12	4.98	54.81	5.34
Diff.-based	AniPotrait [14]	52.57	605.59	3.40	16.55	3.41	26.34	340.66	3.13	7.69	3.86
	Echomimic [12]	59.86	496.54	3.85	16.24	5.58	37.02	502.78	3.11	7.69	4.03
	Sonic [15]	46.03	359.93	2.60	15.64	8.41	21.17	329.20	2.62	7.69	6.36
	Hallo2 [13]	56.53	450.95	2.72	16.24	6.62	26.37	376.32	3.20	7.69	5.19
	Ours	**44.79**	**346.42**	**1.64**	**71.17**	6.45	**20.26**	**317.73**	**2.42**	**60.58**	5.03

4 Experiments

4.1 Experimental Setups

Datasets. We conduct experiments on two emotional talking face datasets: MEAD [38] and RAVDESS [57]. MEAD contains high-quality talking face videos from 47 actors, annotated with eight distinct emotion categories and three intensity levels (low, medium, and high). RAVDESS consists of audiovisual recordings from 24 actors performing both speech and singing, each portraying eight different emotions. In the MEAD dataset, videos of four actors are reserved for testing, and the remaining data are used for training. The RAVDESS dataset is employed exclusively during evaluation to examine the model's ability to generalize across different datasets.

Evaluation Metrics. We evaluate the generated expressive talking face videos in terms of generation quality, audio-visual synchronization, and expressiveness of the facial expressions. Fréchet Inception Distance (FID) [58] and Fréchet Video Distance (FVD) [59] measure realism at the frame and video levels, where lower values indicate greater similarity to real data. For lip-sync accuracy, we use Sync Confidence (Sync-C) computed from a pre-trained SyncNet [60], with higher Sync-C indicating better synchronization. To assess expressiveness of the facial expressions, we compute the expression accuracy (Acc_{emo}) using a pre-trained emotion recognition model, following EDTalk [11]. Additionally, we use the Expression-FID (E-FID) metric, introduced in EMO [61], which calculates FID on facial expression parameters extracted with a 3D face reconstruction model [62]. This metric quantifies the expression divergence between the synthesized and ground truth videos.

Baselines. We compared our proposed method with publicly available implementations of non-diffusion-based methods, such as Wav2lip [21], SadTalker [25], EAT [10], and EDTalk [11], as well as diffusion-based methods including AniPotrait [14], Echomimic [12], Sonic [15], and Hallo2 [13]. Among these, EAT and EDTalk are emotion-aware methods that utilize explicit emotion modeling. The others do not model emotion and are thus regarded as emotion-agnostic.

4.2 Quantitative Results

As shown in Table 1, our method achieves significant improvements over previous state-of-the-art approaches, particularly in expression-related metrics. Specifically, our model achieves the highest expression accuracy (Acc_{emo}) of 71.17 on the MEAD dataset and 60.58 on the RAVDESS dataset, surpassing both GAN-based and diffusion-based baselines. It also outperforms all methods in terms of Expression-FID (E-FID), achieving 1.64 on MEAD and 2.42 on RAVDESS, which reflects its ability to generate more precise and expressive facial animations. Additionally, our approach excels in both FID and FVD scores, with 346.42 and 317.73 respectively, demonstrating the superior capability of diffusion models in producing high-fidelity and temporally coherent videos. While the model does not have the highest Sync-C, it maintains competitive values (6.45 on MEAD and 5.03 on RAVDESS), balancing both high expression accuracy and audiovisual synchronization effectively.

4.3 Qualitative Results

Emotional Talking Face Generation. We present qualitative comparisons with recent state-of-the-art methods as shown in Figure 4. The results can be divided into two settings: emotion-

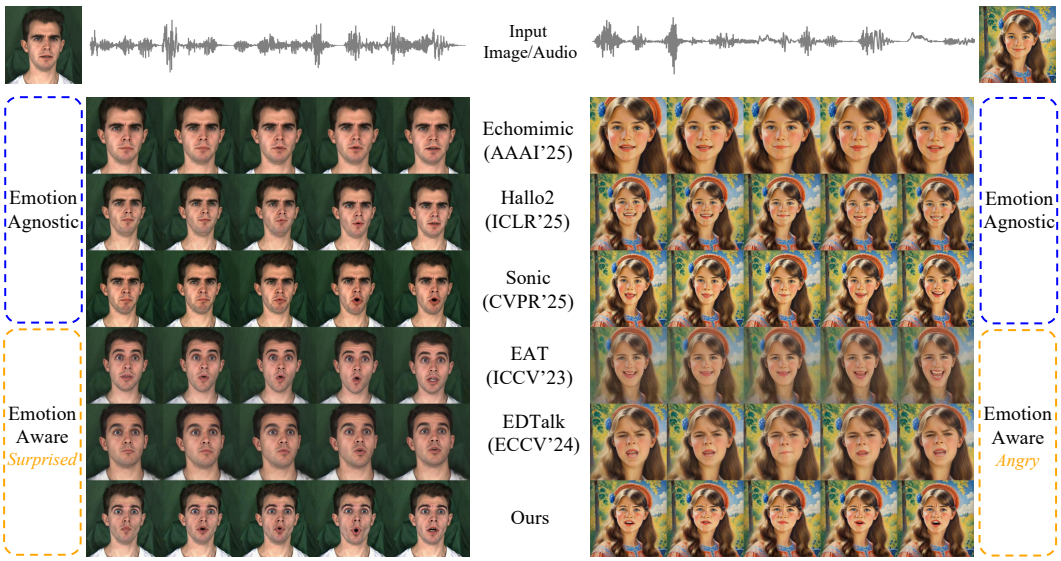


Figure 4: Qualitative comparison with state-of-the-art methods. Given an input image and audio, we compare emotion-agnostic methods (top) and emotion-aware methods (bottom) across different models. Our method generates expressive talking faces that accurately reflect the target emotion (e.g., Surprised and Angry), achieving more vivid and natural emotional expressions compared to previous approaches.

agnostic and emotion-aware. In the emotion-agnostic setting, existing methods mainly focus on lip synchronization while often producing neutral or limited facial expressions. In contrast, under the emotion-aware setting, where explicit emotion inputs are provided, our method generates talking faces with significantly more vivid, dynamic, and emotionally aligned expressions. Compared to previous works, our approach better captures the subtle variations in facial muscles and overall expression intensity, leading to more realistic and engaging talking faces. These results demonstrate the effectiveness of our continuous emotion guidance and spatial-temporal decoupled training strategy in enhancing expressive video generation. Moreover, as shown in the right half of Figure 4, previous emotion-aware methods struggle to drive facial images outside the training set, whereas our method maintains strong identity preservation. Furthermore, as demonstrated in Figure 7, our approach successfully drives diverse types of faces.

Results of Expression Manipulation. To verify the model's ability for fine-grained expression control, we present qualitative results in Figure 5. The first row illustrates the gradual modulation of expression intensity for 'Angry' and 'Surprised', where small changes in the control signal lead to correspondingly subtle variations in facial expressions. This demonstrates the model's precision in capturing nuanced emotional shifts. Additionally, the second row shows a continuous morphing between 'Surprised' and 'Happy', highlighting the model's capability to smoothly transition across different expression types without abrupt changes. These results confirm that our method enables fine-grained and continuous control over facial expressions. Additional results are provided in the supplementary material.

4.4 Ablation Study

Effectiveness of Decoupled Training. To evaluate the impact of Spatial-Temporal Decoupled Training, we compare the full model with a jointly trained one-stage variant. As shown in Table 2, removing decoupled training leads to the lowest Acc_{emo} (42.39%), revealing a clear drop in emotional discriminability. We attribute this to temporal smoothness constraints in joint training, which suppress subtle expression differences and result in similar facial behaviors across emotions. Decoupled training mitigates this by first learning accurate emotion-to-expression mappings in the static domain, avoiding many-to-one mappings and enhancing expression clarity in videos.

Effectiveness of VA-Guidance. To assess the role of continuous emotion conditioning, we replace VA-Guidance with discrete emotion labels. This results in the worst E-FID score (1.89), indicating degraded expression quality. Although Acc_{emo} is higher than the w/o decoupled training setting,

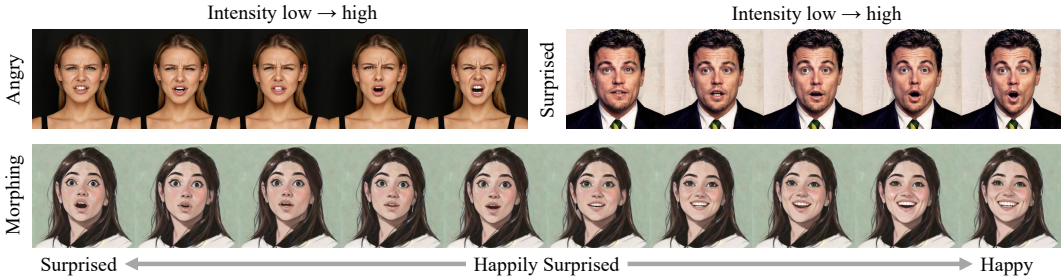


Figure 5: Examples of expression manipulation. The first row demonstrates the control of expression intensity for 'Angry' and 'Surprised' with intensity progressively increasing from left to right. The second row illustrates a continuous morph between 'Surprised' and 'Happy' showcasing the model's ability to smoothly interpolate across different expressions.

Table 2: Ablation study by removing key components from the full framework.

Setting	FID ↓	FVD ↓	E-FID ↓	Acc_{emo} ↑	Sync-C ↑
Full model	**47.79**	**416.24**	**1.48**	**71.73**	**6.32**
w/o Decoupled Training	50.95	441.61	1.89	56.52	6.27
w/o VA-Guidance	49.85	471.52	1.65	42.39	6.11

Table 3: Ablation studies on the use of Valence-Arousal (VA) information. (a) comparison of conditioning mechanisms. (b) comparison of VA-Attention module placement.

(a) Comparison of three conditioning methods for incorporating Valence-Arousal (VA): (A) AdaIN, (B) sinusoidal VA in timestep, and (C) cross-attention.

Conditioning Mechanism	FID ↓	FVD ↓	E-FID ↓	Acc_{emo} ↑	Sync-C ↑
(A)	53.74	485.29	2.55	41.30	6.03
(B)	51.92	472.85	2.17	48.65	6.08
(C)	**47.79**	**416.24**	**1.48**	**71.73**	**6.32**

(b) Comparison of three placements for the VA-Attention module: (A) before the ref-attn, (B) between ref-attn and audio-attn, and (C) after audio-attn.

VA-Attention Location	FID ↓	FVD ↓	E-FID ↓	Acc_{emo} ↑	Sync-C ↑
(A)	51.18	469.97	1.58	68.48	6.27
(B)	**47.97**	**416.21**	**1.48**	**71.73**	**6.32**
(C)	61.53	520.89	2.02	42.39	5.13

the generated expressions are coarser and less dynamic due to the limited granularity of discrete categories. In contrast, VA-Guidance captures fine-grained valence-arousal variations, enabling more nuanced and expressive facial behaviors through flexible one-to-many mappings from emotion to expression.

Effectiveness of Different Conditioning Mechanisms. We compare three conditioning strategies for incorporating VA information: AdaIN, sinusoidal timestep-level embedding, and cross-attention. As shown in Table 3a, cross-attention achieves the best results across all metrics, demonstrating its superiority in achieving expressive and fine-grained emotion control. Therefore, we adopt the cross-attention as our conditioning mechanism.

Effectiveness of VA-Attention Placement. As shown in Table 3b, we evaluate different placements of the VA-Attention module and find that inserting it between the reference and audio branches achieves the best overall performance. This configuration balances emotion control and visual quality, so we adopt this placement as the default setting in our final model.

5 Conclusion

In this work, we proposed a novel framework for expressive talking face generation by integrating continuous Valence-Arousal (VA) emotion guidance with a spatial-temporal decoupled training strategy. By decoupling expression learning from temporal modeling, our two-stage approach enables the expression module to capture fine-grained emotional variations without being constrained by temporal smoothness. The VA-Attention mechanism further allows precise and interpretable control over facial expressions using continuous affective representations. Extensive experiments demonstrate that our method achieves superior emotional expressiveness and realism compared to existing approaches.

References

[1] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021.

[2] Aras Bozkurt, Xiao Junhong, Sarah Lambert, Angelica Pazurek, Helen Crompton, Suzan Koseoglu, Robert Farrow, Melissa Bond, Chrissi Nerantzi, Sarah Honeychurch, et al. Speculative futures on chatgpt and generative artificial intelligence (ai): A collective reflection from the educational landscape. *Asian Journal of Distance Education*, 18(1):53–130, 2023.

[3] Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022, 2021.

[4] Paul Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press, 1971.

[5] Eva G Krumhuber, Lina I Skora, Harold CH Hill, and Karen Lander. The role of facial movements in emotion recognition. *Nature Reviews Psychology*, 2(5):283–296, 2023.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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, 2024.

[12] 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.

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

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

[15] 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.

[16] Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1982–1991, 2023.

[17] Fei Shen, Cong Wang, Junyao Gao, Qin Guo, Jisheng Dang, Jinhui Tang, and Tat-Seng Chua. Long-term talkingface generation via motion-prior conditional diffusion model. *arXiv preprint arXiv:2502.09533*, 2025.

[18] Huaize Liu, Wenzhang Sun, Donglin Di, Shibo Sun, Jiahui Yang, Changqing Zou, and Hujun Bao. Moee: Mixture of emotion experts for audio-driven portrait animation. *arXiv preprint arXiv:2501.01808*, 2025.

[19] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.

[20] Egor Zakharov, Aliaksandra Shysheya, Egor Burkov, and Victor Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9459–9468, 2019.

[21] 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.

[22] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with gans. *International Journal of Computer Vision*, 128(5):1398–1413, 2020.

[23] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-head generation with rhythmic head motion. In *European conference on computer vision*, pages 35–51. Springer, 2020.

[24] Dipanjan Das, Sandika Biswas, Sanjana Sinha, and Brojeshwar Bhowmick. Speech-driven facial animation using cascaded gans for learning of motion and texture. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 408–424. Springer, 2020.

[25] 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.

[26] 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.

[27] 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.

[28] 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*, volume 36, pages 2531–2539, 2022.

[29] Kewei Yang, Kang Chen, Daoliang Guo, Song-Hai Zhang, Yuan-Chen Guo, and Weidong Zhang. Face2face ρ: Real-time high-resolution one-shot face reenactment. In *European conference on computer vision*, pages 55–71. Springer, 2022.

[30] 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.

[31] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[33] 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.

[34] Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024.

[35] Ye Pan, Shuai Tan, Shengran Cheng, Qunfen Lin, Zijiao Zeng, and Kenny Mitchell. Expressive talking avatars. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[36] Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. *arXiv preprint arXiv:2205.01155*, 2022.

[37] 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.

[38] 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 *European Conference on Computer Vision*, pages 700–717. Springer, 2020.

[39] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Wayne Wu, Chen Change Loy, Xun Cao, and Feng Xu. Audio-driven emotional video portraits. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14080–14089, 2021.

[40] 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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 1896–1904, 2023.

[41] Shuai Tan, Bin Ji, Yu Ding, and Ye Pan. Say anything with any style. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5088–5096, 2024.

[42] Shuai Tan, Bin Ji, and Ye Pan. Style2talker: High-resolution talking head generation with emotion style and art style. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5079–5087, 2024.

[43] Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3387–3396, 2022.

[44] 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.

[45] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20687–20697, 2023.

[46] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

[47] Zhaoxu Sun, Yuze Xuan, Fang Liu, and Yang Xiang. Fg-emotalk: Talking head video generation with fine-grained controllable facial expressions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5043–5051, 2024.

[48] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[50] Sicheng Zhao, Xingxu Yao, Jufeng Yang, Guoli Jia, Guiguang Ding, Tat-Seng Chua, Bjoern W Schuller, and Kurt Keutzer. Affective image content analysis: Two decades review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6729–6751, 2021.

[51] James Z Wang, Sicheng Zhao, Chenyan Wu, Reginald B Adams, Michelle G Newman, Tal Shafir, and Rachelle Tsachor. Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion. *Proceedings of the IEEE*, 111(10):1236–1286, 2023.

[52] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[53] Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 3(1):42–50, 2021.

[54] Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759, 2023.

[55] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[57] SR Livingstone and FA Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLoS ONE*, 13(5):e0196391, 2018.

[58] 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. *Advances in neural information processing systems*, 30, 2017.

[59] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

[60] 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.

[61] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2024.

[62] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[63] Mao Li, Bo Yang, Joshua Levy, Andreas Stolcke, Viktor Rozgic, Spyros Matsoukas, Constantinos Papayiannis, Daniel Bone, and Chao Wang. Contrastive unsupervised learning for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6329–6333. IEEE, 2021.

[64] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

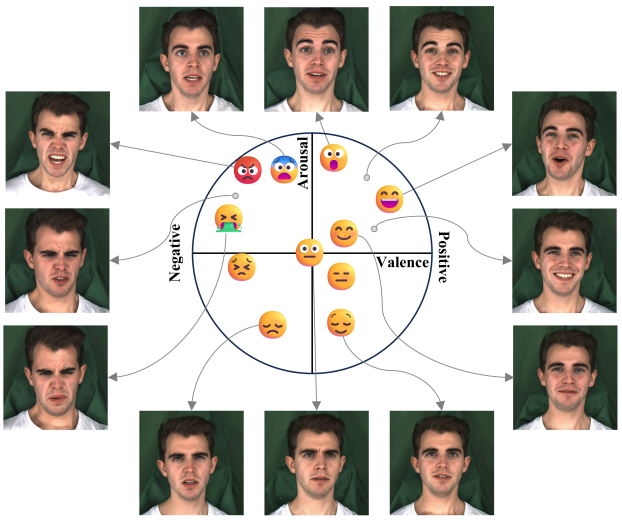


Figure 6: Visualization of the Valence-Arousal (VA) space annotated with real facial expressions. The surrounding facial images illustrate the fine-grained expressiveness of the VA space, distinguishing expressions of varying intensity within the same emotional category and enabling smooth interpolation across emotion types.

A Valence-Arousal Model

As shown in Figure 6, the Valence-Arousal (VA) model [52] provides a continuous two-dimensional emotion space. Valence refers to the pleasantness or unpleasantness of an emotion, ranging from positive to negative. High valence indicates positive emotions (e.g., happiness, satisfaction), while low valence corresponds to negative emotions (e.g., sadness, anger). Arousal reflects the intensity or activation level of an emotion, ranging from calm to excited. High arousal represents more intense emotional states (e.g., excitement, anxiety), while low arousal indicates subdued emotions (e.g., relaxation, boredom).

Compared to discrete emotion categories, the V-A space offers greater flexibility and expressiveness. Figure 6 illustrates the effectiveness of the Valence-Arousal (VA) space in modeling fine-grained facial expressions. Expressions with similar emotion categories but different intensities are located at distinct positions along the arousal axis, enabling continuous control. Moreover, interpolating between typical VA coordinates yields expressions that go beyond discrete emotion classes, capturing subtle or compound emotional states. This continuous structure allows smooth transitions between affective states and captures subtle variations in emotional intensity, making it particularly suitable for generation tasks requiring fine-grained emotional control. Moreover, the V-A model is widely adopted in speech emotion recognition[54, 63], where valence and arousal can be reliably estimated from audio signals. This makes it a natural choice for bridging audio and visual modalities, enabling emotionally consistent talking face generation.

B Implementation Details

B.1 Network Details

Audio Encoder. Our audio encoder includes two branches, the content branch and the emotion branch, both of which are based on the pretrained wav2vec model [64]. The content branch aims to capture linguistic information that drives accurate lip motion. It extracts the hidden representations from all 12 transformer layers of wav2vec. These layer-wise features are concatenated to obtain a comprehensive frame-level audio embedding. To align the high-dimensional features with the downstream talking-face generation task, we apply three linear projection layers to map the concatenated embeddings to a fixed-dimensional representation $\{c_{\text{audio}}^{(s)}\}_{s=1}^{S}$, where $c_{\text{audio}}^{(s)} \in \mathbb{R}^{D_a}$ is the audio embedding for the s frame. The Emotion Branch is also built upon the wav2vec backbone but is specialized to estimate the emotional state in terms of valence-arousal (VA) dimensions. Following the final transformer layer of wav2vec, we perform temporal average pooling over the hidden states and pass the result through a

lightweight prediction head composed of a hidden layer and an output layer. This branch is initialized with pretrained parameters from [54] to leverage emotionally discriminative representations and is further fine-tuned during the second training stage. Specifically, we update the final transformer layer and the emotion prediction head to better adapt to the emotional dynamics of talking face generation.

VA Attention. To enable fine-grained emotional control in facial generation, we propose a Valence-Arousal Attention (VA-Attention) module that injects continuous valence-arousal (VA) signals into visual features through a cross-attention mechanism. Specifically, given the input hidden states, the module first applies Layer Normalization, followed by a cross-attention layer where the VA embeddings serve as keys and values. This structure facilitates effective alignment between spatial representations and the intended emotional cues. To ensure stable training and allow for a smooth incorporation of emotional signals, we adopt a residual connection with a zero-initialized output branch, following the design choice in ControlNet [56]. This initialization encourages the module to initially behave like an identity function and gradually adapt to encode emotion.

VA Encoder. To transform the 2D valence-arousal (VA) signal into a high-dimensional representation suitable for cross-modal fusion, we design a lightweight yet expressive VAEncoder module. This module first encodes the input VA vector through a two-layer MLP with a LeakyReLU activation, projecting it into a hidden space that captures non-linear affective variations. The resulting hidden representation is then concatenated with the original VA input to retain the original signal structure. To ensure compatibility with downstream components, especially the VA-Attention module, the concatenated vector is further processed by a feed-forward network that maps it to a fixed embedding dimension. A zero-padding strategy is employed before this mapping to align the input dimensionality with the target size. Finally, a LayerNorm is applied to normalize the output embedding, improving training stability. The encoded VA embedding is reshaped into a token-like format, enabling its use as key and value vectors in subsequent cross-attention computations.

B.2 Data Details

We conduct experiments on MEAD and RAVDESS, two widely used emotional talking face datasets. MEAD contains video recordings from 60 speakers, with 47 publicly available, each speaking 30 sentences under eight emotions and three intensity levels in a controlled studio environment. Following common practice in prior work [11], we use four speakers (M003, M030, W009, W015) for testing and the remaining for training. For RAVDESS, we select actor01 and actor24 for evaluation. During evaluation, we select a frame labeled with Neutral emotion from the same speaker but a different video to ensure a neutral identity image as the generation source.

B.3 Training and Inference Details

Our model is implemented in PyTorch and trained on NVIDIA A40 GPUs. All face images and video frames are aligned and resized to 512×512 resolution. We adopt a two-stage training strategy. In the first stage (Static Expression Learning), we train the spatial components of the model using still images and corresponding valence-arousal values. The batch size is set to 48, and the model is optimized using AdamW with a learning rate of 1×10^{-5}. No temporal modules or audio inputs are involved in this stage. In the second stage (Dynamic Expression Modeling), we incorporate temporal modules and fine-tune the full model using short video clips. The batch size is reduced to 4 due to higher memory demands. Video clips are processed at a resolution of 512×512 pixels.

During inference, the model generates expressive talking face videos conditioned on a source image and driving audio. The target emotion representation can be either explicitly provided as valence-arousal values derived from video frames or implicitly inferred from the input audio using a pretrained emotion encoder. We adopt the DDIM sampling strategy with 40 denoising steps, and synthesize videos at 25 frames per second (FPS).

B.4 Details of Evaluation Metrics

To comprehensively evaluate the performance of expressive talking face video generation, we adopt a set of quantitative metrics that assess different aspects of the generated videos, including visual realism, audio-visual synchronization, and expression accuracy. Below, we provide detailed descriptions of each metric used in our experiments.

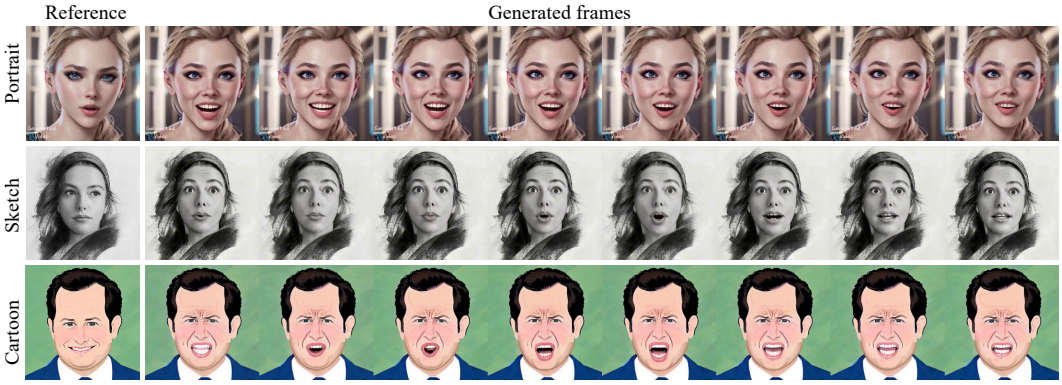


Figure 7: Qualitative results on generalization to different styles of reference images. Our method generates expressive and temporally coherent talking face videos from reference images of diverse visual styles, including realistic portraits, pencil sketches, and cartoons.

Fréchet Inception Distance (FID) [58] is used to measure the realism of individual video frames. It computes the Fréchet distance between the feature distributions of real and generated images extracted from a pre-trained Inception network. Lower FID values indicate that the generated frames are more similar to real images in terms of high-level semantics and visual quality.

Fréchet Video Distance (FVD) [59] extends the FID metric to the video domain by capturing both spatial and temporal consistency. It compares the distribution of features extracted from entire video sequences using a 3D convolutional network (I3D), thus providing a more holistic assessment of video realism. Lower FVD values correspond to higher fidelity and better temporal coherence.

Sync Confidence (Sync-C) measures the audio-visual synchronization quality between the generated lip movements and the input speech. Following prior works, we use a pre-trained SyncNet [60] to extract embeddings from audio and corresponding video frames, and compute the confidence score based on their similarity. A higher Sync-C score indicates better lip-sync performance.

Expression Accuracy ($\mathbf{Acc_{emo}}$) quantifies how well the generated facial expressions align with the target emotional content. We follow the evaluation protocol introduced in EDTalk [11], which uses a pre-trained facial emotion classifier to predict the emotion category of each generated frame. Acc_{emo} is computed as the classification accuracy against the ground truth emotion labels.

Expression Fréchet Inception Distance (E-FID) is used to evaluate the distributional similarity of facial expressions between generated and real videos. This metric was proposed in EMO [61] and is computed by first extracting 3D expression parameters using a 3D face reconstruction model [62], and then calculating the FID between the parameter distributions. Lower E-FID values indicate that the generated expressions are more faithful to those in the ground truth.

C Additional Experimental Results

C.1 Qualitative Comparisons

Generalization to Different Styles of Reference Images. To evaluate the generalization capability of our method across various visual domains, we test the model on reference images with three distinct styles: realistic portrait, sketch, and cartoon. As shown in Figure 7, our approach can synthesize temporally consistent talking face videos even when the input reference image deviates significantly from natural imagery. For portrait-style references, the generated frames maintain fine facial details and expressiveness. In the sketch setting, the model preserves the monochromatic texture while still producing clear and vivid expressions. For cartoons, our method adapts to the exaggerated features and generates sharp and consistent motions. These results demonstrate the robustness of our framework in handling different visual appearances and its potential applicability to stylized talking face generation. **Fine-grained Emotion Transitions.** Figure 8 qualitatively demonstrates the fine-grained emotion control achieved through our method. Each row visualizes a smooth transition between two basic emotions such as Fear and Surprise, Surprise and Happy, or Disgust and Angry. These transitions are generated by linearly varying the target Valence-Arousal values while keeping

Figure 8: Fine-grained emotion transitions achieved by interpolating target Valence-Arousal values. Each row shows a continuous facial expression trajectory between two basic emotions. Intermediate frames reflect compound facial expressions (e.g., Fear-Surprise), demonstrating smooth and controllable emotion editing.

Table 4: Detailed Accuracy of Emotion Classification of MEAD

	Happy	Angry	Disgusted	Fear	Sad	Neutral	Surprised	Contempt	Average
Wav2Lip [21]	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	16.24
SadTalker [25]	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	16.24
AniPotrait [14]	0.00	4.27	0.00	0.00	0.00	97.5	1.71	0.00	16.55
Echomimic [12]	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	16.24
Sonic [15]	0.00	25.64	0.00	0.00	0.00	76.25	0.85	0.84	15.64
Hallo2 [13]	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00	16.24
EAT [10]	92.50	93.16	35.89	4.35	44.17	100.00	100.00	31.09	64.37
EDTalk [11]	100.00	73.50	66.67	11.30	54.17	100.00	100.00	32.77	68.83
Ours	100.00	79.58	67.52	22.61	63.33	100.00	100.00	25.21	71.17

the reference image fixed. The resulting facial expressions exhibit continuous and coherent changes, capturing plausible intermediate states like Fear–Surprise or Surprise–Happy. This highlights the model's ability to produce nuanced emotional expressions beyond discrete emotion categories.

C.2 Detailed Accuracy of Emotion Classification (Acc_{emo})

We report emotion-wise classification accuracy on the MEAD dataset in Table 4. Most baseline models struggle to convey clear emotional cues, often resulting in outputs classified predominantly as *Neutral*. Our method achieves the highest average accuracy (71.17%) and consistently outperforms previous approaches across nearly all emotion categories. Notably, it shows substantial improvements in challenging expressions such as *Fear*, *Sad*, highlighting the benefit of our emotion-aware generation strategy. We further evaluate generalization by reporting emotion-wise classification accuracy on the RAVDESS dataset, which is not used during training (Table 5). Similar to the results on MEAD, most baseline models tend to generate emotionally neutral faces, leading to low accuracy across non-neutral categories. Our method achieves the highest average accuracy (60.58%) and demonstrates consistent improvements across multiple emotion classes, including *Angry*, *Disgusted*, and *Sad*. These results confirm the robustness of our emotion modeling and its ability to transfer to unseen data.

D Discussion

D.1 Limitation.

In our current framework, the background is explicitly kept static throughout the generated video. This design simplifies the learning objective and allows the model to focus on precise facial expression

Table 5: Detailed Accuracy of Emotion Classification of RAVDESS

	Happy	Angry	Disgusted	Fear	Sad	Neutral	Surprised	Average
Wav2Lip [21]	0.00	0.00	0.00	0.00	0.00	100.00	0.00	7.69
SadTalker [25]	0.00	0.00	0.00	0.00	0.00	100.00	0.00	7.69
AniPotrait [14]	0.00	4.27	0.00	0.00	0.00	97.5	1.71	7.69
Echomimic [12]	0.00	0.00	0.00	0.00	0.00	100.00	0.00	7.69
Sonic [15]	0.00	0.00	0.00	0.00	0.00	100.00	0.00	7.69
Hallo2 [13]	0.00	0.00	0.00	0.00	0.00	100.00	0.00	7.69
EAT [10]	100.00	0.00	0.00	0.00	0.00	50.00	100.00	34.62
EDTalk [11]	100.00	6.25	0.00	0.00	100.00	100.00	100.00	54.81
Ours	100.00	43.75	50.00	0.00	50.00	100.00	100.00	60.58

control. However, in practical applications, the visual scene may involve temporal variations in background content, lighting conditions, or camera viewpoints. Since our model does not account for such variations, its applicability to dynamic or scene-rich scenarios remains limited. Future work could explore extending the generation framework to handle both facial dynamics and scene-level changes in a unified manner.

D.2 Broader Impacts and Safeguards.

This work advances the generation of expressive talking face videos by leveraging continuous emotion guidance and a two-stage training strategy. The proposed method has the potential to benefit applications in human-computer interaction, virtual avatars, and assistive communication technologies for individuals with speech or hearing impairments. By enabling more natural and emotionally aligned facial synthesis, it contributes to enhancing the realism and usability of digital agents in educational, healthcare, and entertainment settings.

However, like many generative models involving facial synthesis, this technology may also pose risks of misuse. In particular, it could be employed to generate deceptive or manipulated content, such as emotionally convincing deepfakes, which could contribute to disinformation, impersonation, or violations of privacy. These potential negative societal impacts underscore the need for responsible development and deployment practices.

To mitigate these risks, we will release our pretrained models under a controlled protocol. The release will include a usage license that explicitly prohibits malicious applications such as impersonation or content manipulation. We also plan to provide access through an API-based interface to monitor and limit how the model is used in practice. All training data used in this study are derived from publicly available datasets with appropriate usage permissions, and no private or personally identifiable information has been used. We believe these safeguards strike a balance between fostering research progress and promoting ethical use.