Bridging the Generative Trilemma: An Efficient Framework for Real-Time Affective Talking Head Synthesis

Anonymous authors
Paper under double-blind review

Abstract

Audio-driven talking head synthesis faces a "Generative Trilemma", where existing methods struggle to balance precise lip-synchronization, expressive naturalness, and real-time performance. We propose AffecTalk, an efficient unified framework that achieves a strong balance across these three objectives. AffecTalk introduces a novel differentiable 3DMM bridge, which provides robust geometric guidance to ensure accurate lip-sync while avoiding the error propagation inherent in cascaded systems. To achieve nuanced expressiveness without manual emotion labels for training, we devise an annotation-free emotion learning strategy via cross-modal knowledge distillation and a cross-synthesis consistency loss. These geometric and emotional representations are hierarchically injected into a single-pass GAN generator to minimize feature interference. Extensive experiments show that AffecTalk achieves a strong overall balance among state-of-the-art methods, generating highly expressive, high-fidelity video at speeds exceeding 55 FPS.

1 Introduction

Audio-driven portrait animation (Zhou et al., 2019; Pataranutaporn et al., 2021; Liu et al., 2024a), the synthesis of photorealistic talking head videos from a single static image and a driving audio signal, is a rapidly emerging field. Its wide-ranging applications in virtual avatars, film production, and online education highlight its immense practical value (Wang et al., 2022c; Yu et al., 2023).

A practical SOTA system must satisfy three pillars: (1) **Lip-Synchronization**: This pillar (Prajwal et al., 2020; Sun et al., 2022; Wang et al., 2023b; Yu et al., 2020) represents the fundamental requirement of audio-visual coherence. It demands precise temporal alignment between the synthesized mouth movements and the phonemes of the driving audio. (2) **Expressive Naturalness:** Beyond accurate lip-sync, the animation must exhibit emotional congruence with the audio input (Gururani et al., 2023; Tian et al., 2024; Shi et al., 2025; Lyu et al., 2024). This requires translating cues in the audio—such as prosody, pitch, and cadence—into corresponding and realistic facial expressions (Peng et al., 2023; Kim et al., 2025). Without this emotional fidelity, the resulting animation appears unnatural and fails to convey the presence of a living subject. (3) **Real-Time Performance:** For real-world viability (Guo et al., 2024; Ji et al., 2024; Wang et al., 2025b; Ye et al., 2024), particularly in interactive applications, the system must operate with low latency and minimal computational overhead. Fast inference is therefore essential for smooth real-time performance and broader adoption.

Unfortunately, current methods often exhibit a generative trilemma in practice: improving one of these three pillars can weaken another. As illustrated by the performance radar in Figure 1, recent approaches can be broadly categorized into three groups, each reflecting a different compromise. (1) *Multi-Stage Geometric Pipelines*, such as SadTalker (Zhang et al., 2023b), rely on intermediate geometric representations, including 3DMM coefficients. While this approach effectively decouples motion from appearance, it is highly susceptible to error accumulation (Tan et al., 2024a). Inaccuracies introduced in the initial audio-to-geometry prediction stage are propagated and often amplified during final rendering, which degrades visual quality and yields sub-optimal lip-sync accuracy, struggling to fully satisfy Pillar 1. (2) *Label-Dependent Emotional Models*, including EAMM (Ji et al., 2022) and EAT (Gan et al., 2023), rely on external inputs like reference videos

or discrete emotion labels to guide expression synthesis. While some achieve high emotion accuracy, their reliance on heavy architectural modules or discrete labels overlooks continuous audio cues and severely limits real-time viability, compromising Pillar 3. (3) *End-to-End Diffusion Models*, such as KDTalker (Yang et al., 2025), significantly advance photorealistic fidelity and expressive naturalness. However, their iterative sampling process incurs prohibitive computational costs and high inference latency, making them impractical for real-time use and falling short on Pillar 3. Consequently, navigating these trade-offs to achieve a truly holistic balance remains a critical gap.

To mitigate this trilemma, we propose **AffecTalk**, an efficient framework for **Affec**tive **Talk**ing head animation within a single end-to-end architecture. Rather than maximizing a single metric at the expense of others, AffecTalk is designed to achieve a favorable trade-off across lip-synchronization, expressive naturalness, and real-time performance. Our key design choices lead to strong overall performance across these aspects: (1) **Pillar 1 (Lip-Synchronization)**: We introduce a 3DMM as a differentiable bridge within our end-to-end framework. This avoids the error accumulation of multi-stage pipelines, achieving a superior lip-sync confidence score (6.21 Sync_c on HDTF) that approaches ground-truth levels. (2) **Pillar 2 (Expressive Naturalness)**: Our framework operates entirely without emotion labels. We distill nuanced emotional cues from audio via cross-modal knowledge transfer, while a novel cross-synthesis consistency loss ensures the learned emotion is identity-agnostic. This yields photorealism on par with leading diffusion models (14.80 FID) while rendering highly accurate expressions (62.35% ACC_e). (3) **Pillar 3 (Real-Time Performance)**: AffecTalk is engineered for efficiency, using a lightweight geometric prior and a single-pass GAN generator. This results in true real-time inference at over 55 FPS, achieving a 2.5–3.6× speedup over strong competitors and running orders of magnitude faster than iterative diffusion models. Overall, the motion and emotion representations are integrated via a novel Hierarchical Injection strategy, which introduces the geometric prior to early layers and emotion features to later layers, preventing interference. Fidelity is further enhanced by a Temporal Variance Masking (TVM) module that focuses reconstruction on dynamic facial regions.

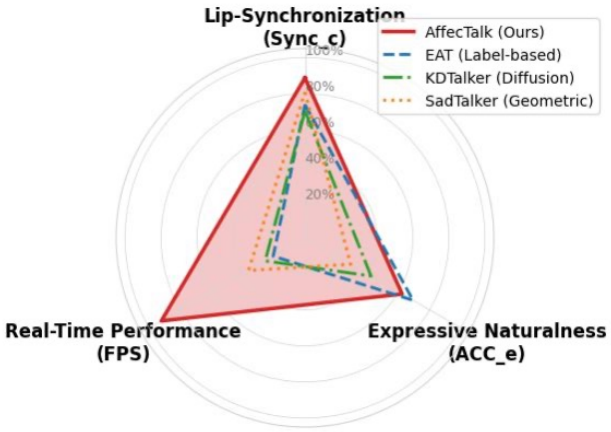


Figure 1: Comparison of our AffecTalk with prior SOTA methods. AffecTalk substantially alleviates this generative trilemma, achieving a strong overall balance across lip-synchronization, expressive naturalness, and real-time performance.

In summary, our key contributions are:

- We propose AffecTalk, a unified end-to-end framework for affective talking head synthesis that jointly considers lip-synchronization, expressive naturalness, and real-time efficiency in a single architecture.
- We introduce a novel use of a 3DMM as a differentiable bridge, which provides strong geometric guidance for robust lip-sync while avoiding the error accumulation common in multi-stage pipelines.
- We present a label-free strategy to learn continuous emotion representations from audio via cross-modal distillation, coupled with a Hierarchical Injection mechanism that strategically integrates motion and emotion signals into different generator layers to prevent interference between the coarse-grained geometric signals (for structure) and the fine-grained emotion signals (for detail).
- Extensive experiments on the MEAD and HDTF benchmarks show that AffecTalk achieves competitive overall performance across visual quality, lip-synchronization, and emotional expressiveness, while maintaining real-time generation speed.

This paper is structured as follows. Section 2 reviews related work. Section 3 introduces our AffecTalk framework, covering feature disentanglement and hierarchical generation. Section 4 presents comprehensive experiments and ablation studies, and we conclude in Section 5.

2 Related Work

2.1 Geometric Prior-based Methods

A dominant strategy in talking head generation is to employ a two-stage pipeline reliant on geometric priors (Ye et al., 2024). These priors evolve from 2D landmarks (Chen et al., 2019) to more structured 3D Morphable Models (3DMMs) used in methods like SadTalker (Zhang et al., 2023b), GMTalker (Xia et al., 2025), and LES-Talker (Feng et al., 2025). More recently, SyncTalk (Peng et al., 2024) leverages a NeRF-based pipeline with a 3D facial blendshape model to enhance expression control and pose stability. These methods first predict 3DMM coefficients from audio to drive a separate renderer, a design that suffers from compounding errors. In contrast, AffecTalk's unified framework uses the 3DMM as a differentiable regularizer—not a supervised target—learning an implicit geometric prior via backpropagation from the final video to avoid error accumulation.

2.2 Emotional Control Methods

Emotional control has been a central challenge, with methods evolving from explicit supervision to more sophisticated, audio-driven techniques (Zhang et al., 2023a; Sinha et al., 2022; Sheng et al., 2023). Early approaches, such as the GCN-based method of (Sinha et al., 2022), relied on discrete, one-hot emotion vectors, which limited expressiveness to predefined categories. A second line of work achieved more dynamic control by using external guidance, either by transferring style from a reference video, as in EAMM (Ji et al., 2022) and DREAM-Talk (Zhang et al., 2023a), or by using text prompts, as seen in Style2Talker (Tan et al., 2024b) and EAT (Gan et al., 2023). While flexible, these methods all depend on auxiliary, non-audio inputs. A more advanced paradigm aims to derive emotion directly from the rich cues within the audio signal itself, typically by disentangling it from other motion factors. PC-AVS (Zhou et al., 2021), for example, focuses on separating head pose from lip motion. More directly targeting emotion, EVP (Ji et al., 2021) uses unimodal audio cross-reconstruction, GC-AVT (Liang et al., 2022) employs input-space masking of facial regions, and PD-FGC (Wang et al., 2023a) utilizes statistical decorrelation. Although these audio-driven strategies are self-contained, they often define emotion indirectly through structural or statistical separation. AffecTalk advances this by introducing a more direct learning signal: without relying on external supervision or discrete labels, our framework autonomously learns to extract nuanced emotional cues from the input audio's acoustic properties. This is achieved via cross-modal distillation from a visual expert, which transfers knowledge without requiring explicit emotion labels.

2.3 Diffusion-based Methods

The advent of diffusion models (Ho et al., 2020) extends to talking head generation (Du et al., 2023; Ma et al., 2023; Zheng et al., 2026), where two main strategies emerge. One line of work uses diffusion models as renderers to translate geometric priors into photorealistic frames, as seen in AniPortrait (Wei et al., 2024), EchoMimic (Chen et al., 2025). A second approach, demonstrated by AniTalker (Liu et al., 2024b) and KDTalker (Yang et al., 2025), uses diffusion as a motion generator to create dynamic keypoint sequences from audio. More recently, EmotiveTalk (Wang et al., 2025a) proposed a video diffusion framework that uses a Vision-guided Audio Information Decoupling (V-AID) approach to generate separate representations for lips and expression. While these techniques achieve SOTA visual quality, their high computational cost and slow inference speeds make them unsuitable for real-time applications. In contrast, AffecTalk's single-pass GAN architecture is designed for efficiency, enabling real-time performance.

3 Method

As illustrated in Fig. 2, our proposed framework, AffecTalk, synthesizes a photorealistic talking-head sequence $\hat{\mathbf{I}}_{1:T}$ from a single identity image $\boldsymbol{I}^i$ and a corresponding audio clip $\mathbf{A}_{1:T}$. Our methodology is built on the principle of strategically orchestrating the knowledge encapsulated in powerful, pre-trained models to solve higher-level challenges. This allows us to focus our contributions on the primary bottleneck: the generation of novel, disentangled control signals and their effective integration. Specifically, we leverage a pre-trained

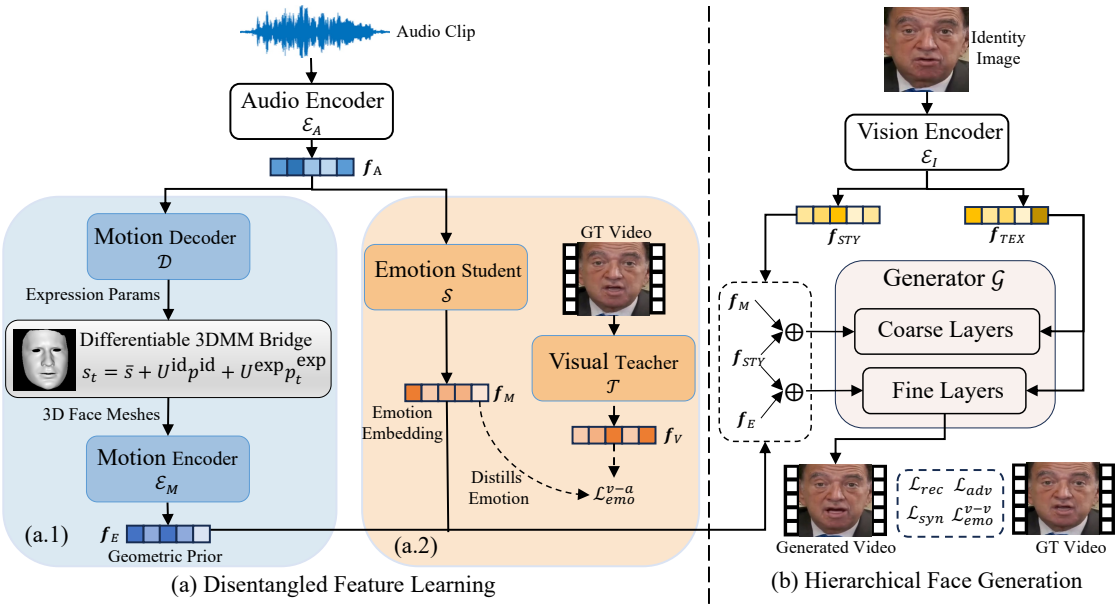


Figure 2: AffecTalk overview: The main pipeline generates a talking head video from a single image and audio clip. It first performs Disentangled Feature Learning (a) to extract a geometric prior (a.1) and an emotion representation (a.2). These features are then integrated into the Hierarchical Face Generation stage (b) using a core Hierarchical Injection mechanism.

Wav2Vec 2.0 (Baevski et al., 2020) $\mathcal{E}_A$ for audio encoding, a ViT as a visual "teacher" $\mathcal{T}$, and SyncNet for lip-sync supervision, while adapting the generator ($\mathcal{G}$) and image encoder ($\mathcal{E}_I$) as a high-fidelity synthesis backbone. The core of our innovation lies in generating and integrating two novel audio-driven signals, a geometric prior $\boldsymbol{f}_M$ and an emotion embedding $\boldsymbol{f}_E$, to control this backbone. The overall generation process consists of two main stages:

- **Complementary Feature Learning** (Fig. 2(a)): An audio encoder $\mathcal{E}_A$ first extracts features $\boldsymbol{f}_A = \mathcal{E}_A(\mathbf{A}_{1:T})$ from the input audio. These features are then fed into two parallel branches: a **Geometric Prior Learning** branch, which predicts an implicit motion prior $\boldsymbol{f}_M$, and an **Emotion Representation Learning** branch, which extracts a dynamic emotion embedding $\boldsymbol{f}_E$. While speech geometry and emotional expression are often correlated, our framework is designed to functionally decouple these aspects into two complementary streams. This approach encourages each branch to specialize: the geometric branch focuses on predictable, speech-driven motion for robust lip-sync, while the emotion branch models the nuanced expressions tied to audio prosody. This functional separation is crucial for resolving the ambiguity in audio-to-video synthesis by separating predictable, speech-correlated geometric motion from more stochastic, nuanced emotional expressions.

- **Hierarchical Face Generation** (Fig. 2(b)): The synthesis process is orchestrated by a GAN-based generator $\mathcal{G}$, which hierarchically integrates the learned features. The process begins by encoding the source identity image $\boldsymbol{I}^i$ using a vision encoder $\mathcal{E}_I$. This yields two distinct representations: a global, non-spatial style code $\boldsymbol{f}_{STY}$ encapsulating the holistic appearance, and a pyramid of multi-scale spatial features $\boldsymbol{f}_{TEX}$ that retain rich, high-fidelity texture details. The generator is first conditioned on $\boldsymbol{f}_{STY}$ to set the subject's foundational appearance. Adopting a coarse-to-fine approach, the geometric prior $\boldsymbol{f}_M$ is injected into the early layers of $\mathcal{G}$ to dictate the overall facial structure and head motion. In parallel, the emotion embedding $\boldsymbol{f}_E$ modulates the later, fine-grained layers to synthesize nuanced expressions. To guarantee robust identity preservation, the identity feature $\boldsymbol{f}_{TEX} = \mathcal{E}_I(\boldsymbol{I}^i)$ is also fed into the generator along with these control signals.

The framework is trained end-to-end using a self-reconstruction objective, where a composite loss supervises generated video $\hat{\mathbf{I}}_t$ against the ground-truth video $\mathbf{I}_t^{\text{GT}}$. A key part of our training is the *Emotion Disentan-*

glement Strategy (Fig. 3), which uses a cross-synthesis consistency loss to learn the emotion embedding $\boldsymbol{f}_E$ without any labels. The following sections detail these components.

3.1 Functionally Disentangled Feature Learning

As highlighted by Zhou et al. (2020), a primary challenge in talking head generation is the ambiguity of direct audio-to-pixel mapping, which can lead to inaccurate lip movements and unnatural motion. This arises from the inherent one-to-many relationship, as a single phoneme can correspond to diverse articulatory and emotional facial movements. To address this, our framework approaches the problem through functional disentanglement, mapping the audio signal to two complementary and more predictable streams: geometric motion and emotion representation. Though these two aspects may be correlated in the input audio (e.g., a smile affects both geometry and emotion), our architecture and targeted loss functions encourage each branch to specialize in a distinct function. Specifically, we employ the pre-trained Wav2vec 2.0 model (Baevski et al., 2020) as our audio encoder, $\mathcal{E}_A$, to extract rich features, $\boldsymbol{f}_A = \mathcal{E}_A(\mathbf{A}_{1:T})$, from the raw audio. These features then serve as the common input for the two parallel branches described in the following sections.

3.1.1 Geometric Prior Learning

We model the geometric motion stream using a 3D Morphable Model (3DMM), which transforms the ill-posed audio-to-pixel problem into a constrained task of predicting low-dimensional facial parameters. This parametric representation inherently disentangles the strong audio-lip correlation from weaker correlations with head pose or blinks, mitigating the one-to-many ambiguity (Xia et al., 2025).

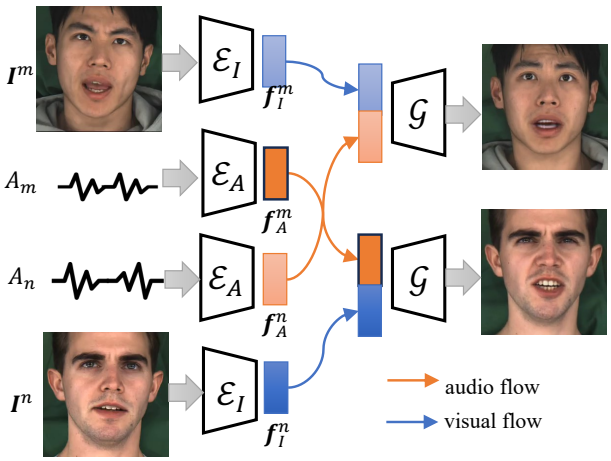


Figure 3: The overview of cross-synthesis strategy. The emotion representation is learned without labels via this strategy.

Crucially, instead of treating 3DMM coefficients as pre-computed supervision targets, we use the 3DMM parameterization as a differentiable geometric regularizer within an end-to-end framework. This avoids the error accumulation common in previous multi-stage methods (Zhang et al., 2023b; Feng et al., 2025), where early prediction errors degrade the final rendering. Moreover, our "differentiable bridge" design offers two key advantages. First, it constrains the optimization: instead of mapping audio to an unconstrained, high-dimensional pixel space, the model predicts a few low-dimensional 3DMM parameters, shrinking the search space, providing strong geometric guidance, and steering optimization toward a reasonable, anatomically plausible optimum—thereby addressing the one-to-many ambiguity by focusing on predictable, speech-driven facial motion (Pillar 1). Second, it improves inference speed (Pillar 3), as predicting a small set of parameters requires only a lightweight network, unlike estimating computationally expensive dense motion fields.

As illustrated in Fig. 2(a.1), audio features $\boldsymbol{f}_A$ given by $\mathcal{E}_A$ are fed into a motion decoder $\mathcal{D}$ (a Temporal Convolutional Network, TCN), which predicts a sequence of 3DMM expression parameters $\boldsymbol{p}_t^{\text{exp}}$. These parameters, in turn, define a sequence of 3D facial meshes $\boldsymbol{S}_{1:T} = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_T\}$ via the standard 3DMM formulation $\boldsymbol{s}_t = \bar{\boldsymbol{s}} + \boldsymbol{U}^{\text{id}}\boldsymbol{p}^{\text{id}} + \boldsymbol{U}^{\text{exp}}\boldsymbol{p}_t^{\text{exp}}$, where $\bar{\boldsymbol{s}}$ is the mean face shape, and $\boldsymbol{U}^{\text{id/exp}}$ are the 3DMM bases. A key aspect of our design is the strategic handling of identity and expression parameters within the 3DMM. We make the following choices:

- Fixed Identity Prior: We do not learn the identity parameter $\boldsymbol{p}^{\text{id}}$. Instead, following (Ren et al., 2021), we fix it to a prior value representing a generic mean face. The rationale is that the low-dimensional $\boldsymbol{p}^{\text{id}}$ only captures coarse, low-frequency aspects of facial identity (e.g., bone structure) and cannot preserve the high-frequency details (e.g., skin texture, eye color, and fine wrinkles). We therefore delegate this task to

the vision encoder $\mathcal{E}_I$, which extracts rich, high-dimensional appearance features ($\boldsymbol{f}_{STY}$, $\boldsymbol{f}_{TEX}$) directly from the source image, which will be discussed in Subsection 3.2.

- Expression for Geometry: While $\boldsymbol{p}_t^{\text{exp}}$ is learned, we use it primarily to guide speech-related geometry (e.g., lip and jaw movements). The 3DMM's expression space is too coarse for subtle emotions, so we use a separate, dedicated emotion feature $\boldsymbol{f}_E$ to synthesize fine-grained expressions.

The resulting mesh sequence $\boldsymbol{S}_{1:T}$ is then encoded by a lightweight CNN-based motion encoder $\mathcal{E}_M$ to produce the final motion feature $\boldsymbol{f}_M$. Notably, neither $\boldsymbol{p}_t^{\text{exp}}$ nor the intermediate meshes are directly supervised; they serve as a differentiable geometric bridge from audio to synthesis. Supervision comes only from the rendered output, letting $\mathcal{D}$ learn an optimal geometric prior via end-to-end backpropagation so the geometry is optimized for final image generation rather than for matching a potentially noisy intermediate representation.

3.1.2 Emotion Representation Learning

As shown in Fig. 2(a.2), to learn a rich, visually-grounded emotion representation from audio without labels, we propose a cross-modal knowledge distillation strategy, as inspired by (Jin et al., 2023; Sun et al., 2024). The approach employs a visual "teacher" $\mathcal{T}$, which is a Vision Transformer (ViT) pre-trained on a large-scale facial expression dataset and acts as an expert emotion extractor, producing an embedding $\boldsymbol{f}_V = \mathcal{T}(\mathbf{I}^{GT})$ from each ground-truth video frame. Concurrently, our trainable "student" $\mathcal{S}$ (a TCN-based emotion decoder) processes audio features to generate a corresponding embedding $\boldsymbol{f}_E = \mathcal{S}(\boldsymbol{f}_A)$. The student is trained to mimic the teacher by minimizing the cosine distance between their embeddings: $\mathcal{L}_{emo}^{v-a} = 1 - \cos(\boldsymbol{f}_E, \text{sg}(\boldsymbol{f}_V))$, where $\cos(\cdot,\cdot)$ is the cosine similarity and $\text{sg}(\cdot)$ is the stop-gradient operator. This process effectively transfers nuanced emotional understanding from the visual domain to our audio network and bypasses the need for manual annotations.

3.2 Hierarchical Face Generation

The final synthesis stage targets both high-fidelity rendering and real-time performance using a conditional GAN generator $\mathcal{G}$, whose architecture builds upon established warping-based paradigms (Wang et al., 2022b; Tan et al., 2024a). To ensure training stability and leverage robust facial-animation priors, we initialize the core backbone modules ($\mathcal{G}$ and its encoder $\mathcal{E}_I$) using publicly available pre-trained weights from standard motion-transfer models.

The core challenge here is to effectively integrate the diverse control signals—identity, motion, and emotion—into warping without interference. To this end, we first encode the source identity image $\boldsymbol{I}^i$ with the encoder $\mathcal{E}_I$ into two complementary representations. A non-spatial **global style code**, $\boldsymbol{f}_{STY}$, is extracted from the final layer of $\mathcal{E}_I$ to encapsulate the holistic appearance identity. A pyramid of multi-scale **spatial texture features**, $\boldsymbol{f}_{TEX}$, is extracted from $\mathcal{E}_I$'s intermediate layers. These features retain the rich, spatially-aligned textures and structural details that serve as the high-fidelity "content canvas" to be animated. Together, $\boldsymbol{f}_{STY}$ and $\boldsymbol{f}_{TEX}$ ensure that the subject's identity remains stable and detailed across the entire generated sequence. The generator $\mathcal{G}$ then renders the target frame $\hat{\mathbf{I}}_t$ by synergizing them with the audio-driven motion prior $\boldsymbol{f}_M$ and the emotion embedding $\boldsymbol{f}_E$.

3.2.1 Analysis of the Generation Process

As shown in Fig. 4, we analyze the generator's synthesis process using Principal Component Analysis (PCA) on its intermediate features. By projecting features from different layers onto their top principal components and visualizing them as RGB channels, we observe a distinct coarse-to-fine progression. Early-stage features establish global structure and layout, while later-stage features progressively add intricate local details and textures. This finding motivates our Hierarchical Injection strategy.

3.2.2 Hierarchical Injection

Informed by our analysis, we introduce a Hierarchical Injection strategy to align our control signals with the generator's synthesis stages. This design maximizes controllability by separating the "what to animate" content from the "how to animate" instructions. The "how to animate" instructions are delivered through

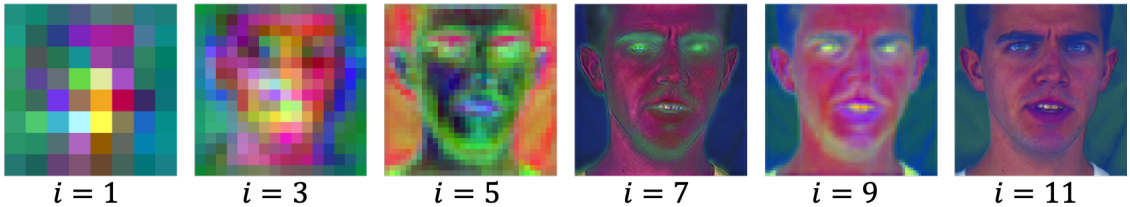


Figure 4: Visualization of the coarse-to-fine synthesis process within the generator. Each image displays the output from a different generator layer, where i denotes the output of the i-th layer.

a composite control signal, $\boldsymbol{f}_{CTRL}$, which modulates the generator's behavior at different layers. For the early, coarse-grained layers that control global structure, the control signal $\boldsymbol{f}_{CTRL} = \boldsymbol{f}_{STY} + \boldsymbol{f}_M$, ensuring robust structural animation and precise lip-sync (Pillar 1). In contrast, for the latter, fine-grained layers which refine local details, $\boldsymbol{f}_{CTRL} = \boldsymbol{f}_{STY} + \boldsymbol{f}_E$. This allows for the synthesis of nuanced, high-frequency emotional details (Pillar 2) without disrupting the underlying facial geometry. The hierarchical control is implemented via Adaptive Instance Normalization (AdaIN) (Huang & Belongie, 2017). Concretely, $\boldsymbol{f}_{CTRL}$ is first mapped through a learned affine transformation AF to produce style parameters (α, β), which then stylize the generator's intermediate feature map $\boldsymbol{f}^j_{INT}$ for a specific layer j:

$$\begin{aligned} \alpha, \beta &= \mathrm{AF}(\boldsymbol{f}_{\mathrm{CTRL}}), \\ \mathrm{AdaIN} &= \alpha \left(\frac{\boldsymbol{f}^j_{\mathrm{INT}} - \mu(\boldsymbol{f}^j_{\mathrm{INT}})}{\sigma(\boldsymbol{f}^j_{\mathrm{INT}})} \right) + \beta, \end{aligned} \tag{1}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the channel-wise mean and standard deviation.

This creates two synergistic pathways within the generator. The control pathway uses the AdaIN layers to translate the control codes ($\boldsymbol{f}_{STY}$, $\boldsymbol{f}_M$, $\boldsymbol{f}_E$) into the precise instructions needed to generate the optical flow field. The content pathway, in parallel, supplies the spatial texture features $\boldsymbol{f}_{TEX}$ via skip connections. The generator then uses the predicted flow field to warp the detailed "content" from $\boldsymbol{f}_{TEX}$, effectively animating the source identity according to the audio-driven commands.

We note that the high-level principle of staged generation is also explored in diffusion models (Peng et al., 2025; Liang et al., 2025), which may use early denoising steps for motion and later steps for detail. Our Hierarchical Injection is architecturally distinct: it operates across the spatial layers of a single-pass GAN to resolve interference between explicitly disentangled motion ($\boldsymbol{f}_M$) and emotion ($\boldsymbol{f}_E$) signals. This structural choice is crucial for achieving real-time performance and, as confirmed by our ablation study (Table 3), is essential for preventing the feature clash that would otherwise degrade both lip-sync and expression.

3.3 Self-reconstruction Training Strategy

We train the framework end-to-end using a self-reconstruction strategy, where the generated video is supervised against the ground-truth. The composite loss is:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{syn}\mathcal{L}_{syn} + \lambda_{emo}(\mathcal{L}^{v-v}_{emo} + \mathcal{L}^{v-a}_{emo}) \tag{2}$$

These terms correspond to a reconstruction loss, an adversarial loss, a lip-sync loss, and an emotion perceptual loss, with weights set following prior work (Tan et al., 2024a). Since both $\mathcal{L}^{v-v}_{emo}$ and $\mathcal{L}^{v-a}_{emo}$ share the same objective of aligning emotional representations, they are assigned an equal weight. In addition, we introduce an Emotion Disentanglement Strategy that decouples emotion by applying losses across different videos. Below we elaborate on each component.

Reconstruction Loss. Following prior work (Gan et al., 2023; Tan et al., 2024a), we begin with a composite reconstruction loss, $\mathcal{L}_{rec}$, that combines a standard L1 loss for pixel-level accuracy with a VGG-19 based perceptual loss (Simonyan & Zisserman, 2014) to preserve high-frequency details. This baseline loss is formulated as: $\mathcal{L}_{rec} = \lambda_{L1}||\hat{y} - y||_1 + \lambda_{perc}||\phi(\hat{y}) - \phi(y)||_1$. However, this objective treats all pixels equally. To specifically enhance the reconstruction quality of dynamic regions, we augment this with an additional

masked L1 term, guided by a dynamically generated motion mask $\mathbf{M}_{motion}$. This mask is produced by our proposed **Temporal Variance Masking (TVM)**, a lightweight, parameter-free algorithm. TVM operates on the principle that spatial regions exhibiting high variance along the temporal dimension correspond to dynamic content (e.g., mouth movements), whereas low-variance regions represent the static background. The mask is thus constructed by isolating these high-variance areas (see Appendix B.2 for details). Our final reconstruction loss integrates this dynamic component, adaptively concentrating the model's capacity on these challenging moving regions:

$$\mathcal{L}_{rec} = \lambda_{L1}||\hat{y} - y||_1 + \lambda_{perc}||\phi(\hat{y}) - \phi(y)||_1 + \lambda_{motion}||(\hat{y} - y) \odot \mathbf{M}_{motion}||_1 \tag{3}$$

where λ_{motion} weights the motion-focused reconstruction term.

3.3.1 Adversarial Loss

To enhance realism (Zhou et al., 2021; Zhang et al., 2023b), we employ an adversarial loss. Our generator $\mathcal{G}$ is trained to minimize a relativistic average objective, while the discriminator $\mathcal{D}_{adv}$ is trained on a symmetric loss:

$$\mathcal{L}_{adv} = \mathbb{E}_{\hat{y},y}\left[\text{softplus}(\mathcal{D}_{adv}(y) - \mathcal{D}_{adv}(\hat{y}))\right] \tag{4}$$

where $\hat{y}$ and y are generated and real samples, respectively.

3.3.2 Syncnet Loss

To ensure precise temporal alignment between the generated lip movements and the input audio, we incorporate a lip-sync loss $\mathcal{L}_{syn}$. We leverage a pre-trained SyncNet (Chung & Zisserman, 2016a) model, a widely-used expert network for audio-visual synchronization (Gan et al., 2023). The loss is formulated to maximize the cosine similarity between these two embeddings:

$$\mathcal{L}_{syn} = 1 - \cos(\boldsymbol{v}, \boldsymbol{s}) = 1 - \frac{\boldsymbol{v} \cdot \boldsymbol{s}}{\max(\|\boldsymbol{v}\|_2 \|\boldsymbol{s}\|_2, \epsilon)} \tag{5}$$

where $\boldsymbol{v}$ and $\boldsymbol{s}$ are the embeddings extracted by the SyncNet's visual and audio encoders, respectively. The term ϵ is a small constant added for numerical stability.

3.3.3 Emotion Perceptual Loss

To ensure the emotional fidelity of the final rendered images, we employ a perceptual loss that operates in the feature space of the pre-trained ViT model. Repurposing the ViT teacher model $\mathcal{T}$ as a fixed perceptual supervisor, we extract emotion logits from the generated image (z_{gen}) and its ground-truth (GT) frame (z_{gt}). The perceptual emotion loss, $\mathcal{L}_{emo}^{v-v}$, is then defined as the KL divergence (Goodfellow et al., 2016) between their Softmax probability distributions:

$$\mathcal{L}_{emo}^{v-v} = D_{\text{KL}}\left(\mathcal{P}(\text{sg}(z_{\text{gt}})) \parallel \mathcal{P}(z_{\text{gen}})\right) \tag{6}$$

where the stop-gradient $sg(\cdot)$ on z_{gt} emphasizes that the GT distribution is a fixed target. Minimizing $\mathcal{L}_{emo}^{v-v}$ guides the generator to produce images that are perceptually aligned with the true emotion.

3.3.4 Emotion Disentanglement Strategy

To ensure the learned emotion embedding $\boldsymbol{f}_E$ is truly disentangled from identity-specific attributes, we introduce a cross-synthesis consistency loss. This strategy enables self-supervision by forcing the model to recombine identity and emotion features from different source videos. For any two training samples, $(\mathbf{I}^m, \boldsymbol{A}_m)$ and $(\mathbf{I}^n, \boldsymbol{A}_n)$, we extract their respective identity features $(\boldsymbol{f}_I^m, \boldsymbol{f}_I^n)$ and audio-derived emotion embeddings $(\boldsymbol{f}_E^m, \boldsymbol{f}_E^n)$. We then perform a feature swap to generate a cross-synthesized image, $\hat{\mathbf{I}}^{mn} = \mathcal{G}(\boldsymbol{f}_I^m, \boldsymbol{f}_E^n)$, by feeding the generator the identity of sample m but the emotion of sample n. If disentanglement is successful, $\hat{\mathbf{I}}^{mn}$ should retain the identity of $\mathbf{I}^m$ but exhibit the emotional expression of $\mathbf{I}_n$. Consequently, the emotional ground truth for $\hat{\mathbf{I}}^{mn}$ is the original frame $\mathbf{I}^n$. We enforce this consistency by applying $\mathcal{L}_{emo}^{v-v}$ in Eq. (6) between the cross-synthesized image $\hat{\mathbf{I}}^{mn}$ and the emotion source's ground truth $\mathbf{I}^n$.

Table 1: Quantitative comparisons with SOTA methods. ↑ denotes higher the better, ↓ denotes lower the better. The best result is in **bold**, and the second-best is underlined. * denotes diffusion methods.

Method	Inference (FPS)	MEAD					HDTF			
		PSNR ↑	SSIM ↑	FID ↓	Sync_c ↑	ACC_e ↑	PSNR ↑	SSIM ↑	FID ↓	Sync_c ↑
EAMM	16.87	17.21	0.60	33.27	1.98	32.90	19.17	0.62	40.36	2.33
EAT	12.23	21.33	0.70	20.48	5.17	**69.35**	22.15	0.70	27.58	4.32
SadTalker	21.74	21.39	0.68	25.84	5.36	29.33	21.20	0.60	15.33	5.66
RealPortrait	3.77	20.55	0.62	21.22	5.67	43.55	21.41	0.69	27.27	5.32
AniTalker*	3.57	20.90	0.70	24.17	4.53	33.49	21.82	0.68	18.01	4.69
Aniportrait*	3.84	21.23	0.69	23.17	2.35	48.22	21.69	0.61	15.02	3.18
EchoMimic*	0.23	21.89	0.75	19.11	4.24	39.21	22.33	0.73	17.45	4.48
KDTalker*	15.34	**22.32**	**0.79**	**16.59**	4.22	42.40	**23.07**	**0.78**	16.52	4.93
AffecTalk	**55.32**	22.19	0.77	17.09	**5.84**	62.35	22.82	0.75	**14.80**	**6.21**
GT	-	-	-	-	6.35	77.37	-	-	-	**6.49**

4 Experimental Results

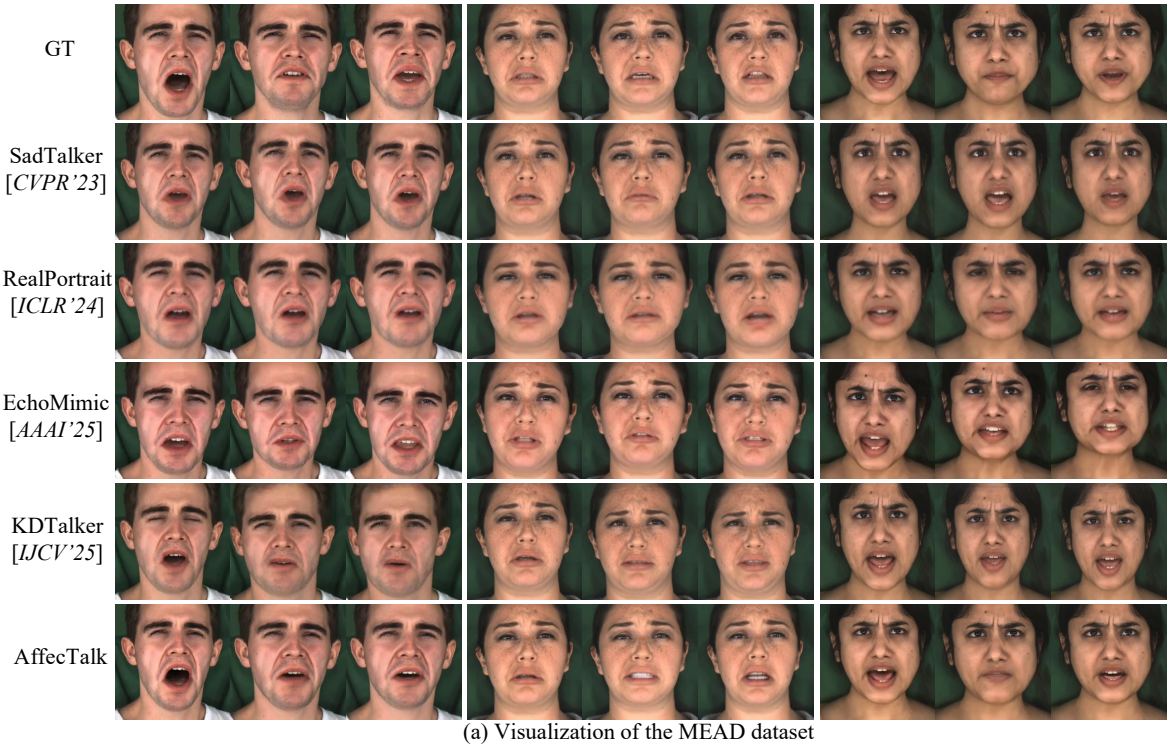
We train and evaluate our model on the MEAD (Wang et al., 2020) and HDTF (Zhang et al., 2021) datasets, covering both controlled and in-the-wild scenarios with video frames resized to 256×256 (Danecek et al., 2022). We benchmark AffecTalk against SOTA methods categorized by their core strategy: emotion label-dependent (EAMM (Ji et al., 2022), EAT (Gan et al., 2023)), geometric prior-based (SadTalker (Zhang et al., 2023b), RealPortrait (Ye et al., 2024)), and diffusion-based (Aniportrait (Wei et al., 2024), AniTalker (Liu et al., 2024b), EchoMimic (Chen et al., 2025), KDTalker (Yang et al., 2025)). To rigorously evaluate performance against our three pillars, we employ a holistic suite of metrics: for **Image Quality and Realism**, we use PSNR, SSIM, and the Fréchet Inception Distance (**FID**) (Heusel et al., 2017); for **Lip-Synchronization (Pillar 1)**, we report the SyncNet confidence score (Sync_c) (Chung & Zisserman, 2016b); for **Expressive Naturalness (Pillar 2)**, we measure emotion accuracy (**ACC_e**) (Meng et al., 2019); and for **Real-Time Performance (Pillar 3)**, we report Frames Per Second (**FPS**). More details regarding the datasets, the baselines, the evaluation metrics, and the implementation details are provided in Appendix C.

4.1 Main Results

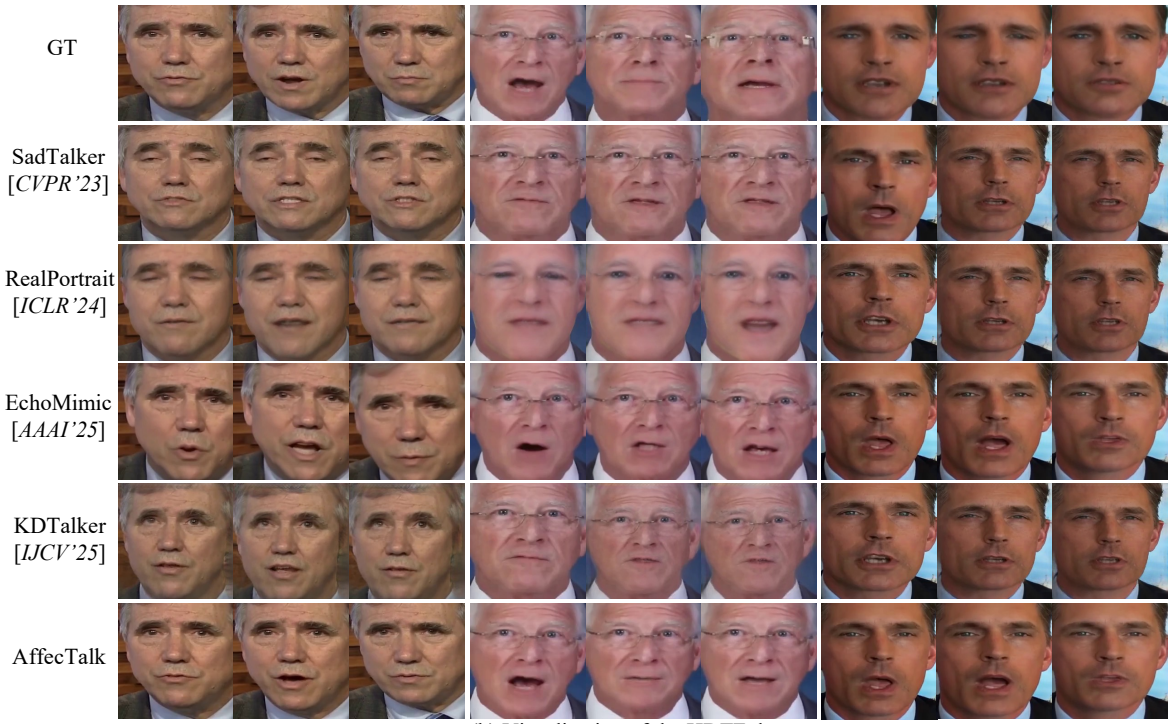
As shown in Table 1, AffecTalk achieves strong overall performance across lip-synchronization, expressive naturalness, and real-time inference. Rather than optimizing a single metric, our method provides a favorable balance among these aspects within a unified framework.

Pillar 1: Strong Lip-Synchronization. AffecTalk achieves the highest Sync_c score on both datasets (e.g., 6.21 on HDTF), surpassing all baselines and approaching the ground-truth level (6.49). This validates our end-to-end framework, where the 3DMM acts as a differentiable bridge. Unlike multi-stage methods like SadTalker, our design avoids the accumulation of geometric prediction errors, resulting in more precise lip articulation. While diffusion models like KDTalker and AniTalker show strong FID scores, their lower Sync_c scores highlight a key weakness: averaging results blurs fine-grained lip motion, degrading audio-visual coherence. AffecTalk's GAN-based architecture, by contrast, preserves these critical dynamic details.

Pillar 2: Expressive Naturalness without Labels. AffecTalk excels in generating both photorealistic and emotionally congruent animations. Regarding **photorealism**, our model achieves a top-tier FID score (14.80 on HDTF), proving it can generate images with quality and diversity comparable to leading diffusion models, but without their computational overhead. While some methods like KDTalker report slightly higher PSNR/SSIM values, this points to the classic perception-distortion trade-off (Zhang et al., 2018). These methods over-optimize for pixel-wise similarity, often producing overly smooth, "averaged" images that lack realism, as seen qualitatively in Fig. 5. AffecTalk, by contrast, prioritizes perceptual quality, which is confirmed by its strong FID and user study results. On the other hand, in terms of **expressiveness**, AffecTalk achieves a high ACC_e score (62.35% on MEAD), demonstrating its ability to render accurate



(a) Visualization of the MEAD dataset



(b) Visualization of the HDTF dataset

Figure 5: Qualitative comparison of our AffecTalk with SadTalker, RealPortrait, EchoMimic, and KDTalker.

emotional expressions. Crucially, it does so in a fully label-free manner, a significant practical advantage over methods like EAT, which requires explicit emotion labels to reach its slightly higher score (69.35%). This result validates our cross-modal distillation strategy for learning nuanced emotion directly from audio.

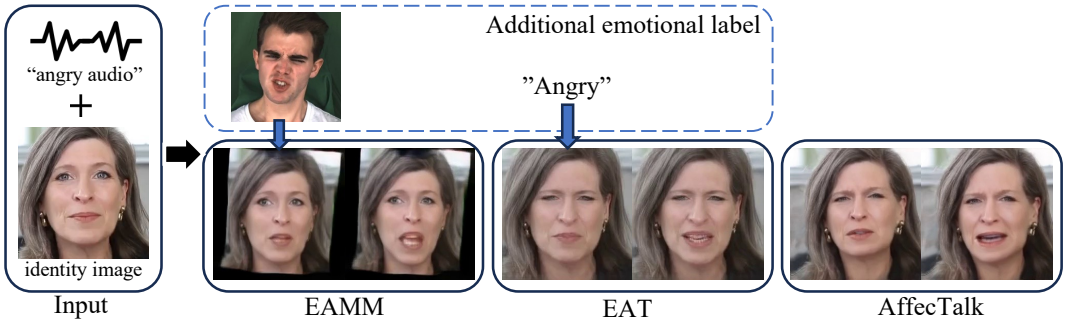


Figure 6: Comparison of emotional expressiveness across generation methods.

To validate our claims, we compare AffecTalk against emotion-driven methods like EAT and EMMN in Fig. 6. These methods face two critical limitations that AffecTalk overcomes. First, regarding expressiveness, their reliance on discrete labels results in static, "canned" emotions. In contrast, our label-free approach learns expressions from audio prosody. Second, regarding real-time performance, they sacrifice speed for control. EAT (12.23 FPS) and EAMM (16.87 FPS) are too slow for interactive use. AffecTalk, however, delivers superior emotional fidelity at over 55 FPS (a 3.6× speedup). Thus, AffecTalk offers a favorable balance in the core trade-off between expressiveness and real-time performance.

Table 2: User study results. Higher is better.

Metric	Lip-sync	Realness	Naturalness
RealPortrait	0.53	0.62	0.55
Aniportrait	0.35	0.33	0.32
KDTalker	0.49	0.65	0.65
Ours	0.66	0.71	0.75

Pillar 3: Real-Time Performance. Our framework is designed for efficiency, achieving an inference speed of 55.32 FPS on a single A100 GPU. This is a 2.5× to 3.6× speedup over strong competitors like SadTalker (21.74 FPS) and KDTalker (15.34 FPS), and orders of magnitude faster than diffusion models. This performance, which meets real-time requirements, stems from our single-pass GAN generator and the lightweight geometric prior, avoiding the high latency of iterative sampling in diffusion models.

Qualitative and Human Evaluation. The quantitative results are strongly supported by human perception. The qualitative comparisons in Fig. 5 offer a clear visual narrative that corroborates our metric-based findings. While diffusion-based methods like KDTalker and Aniportrait produce smooth outputs, they often fail to capture the precise emotional nuances of the ground truth, resulting in a generalized or muted expression. In contrast, Fig. 5 demonstrates that AffecTalk excels at replicating the specifics of the ground-truth expression and emotion. This expressive fidelity, combined with visibly more accurate lip shapes, directly showcases our model's strengths in both Expressive Naturalness (Pillar 2) and Lip-Synchronization (Pillar 1). Additional visualizations are available at HTTP link. Furthermore, a user study (Table 2) confirms AffecTalk's superiority, as participants rated it highest in lip-synchronization, photorealism, and overall naturalness. This convergence of metric-based and human-centric evaluation confirms AffecTalk as a robust and practical solution.

Table 3: Ablation study for main components in our method. GPL: Geometric Prior Learning; ERL: Emotion Representation Learning; HI: Hierarchical Injection; EDS: Emotion Disentanglement Strategy.

Index	Method	PSNR	SSIM	Sync_c	ACC_e
-	Full Model	22.19	0.77	5.84	62.35
A.1	w/o GPL	21.21	0.68	1.85	61.72
A.2	w/o ERL	21.37	0.70	5.65	41.23
A.3	w/o 3DMM	21.05	0.66	3.23	57.41
B.1	w/o HI	21.47	0.71	4.91	52.64
B.2	w/o $\boldsymbol{f}_{STY}$	21.63	0.70	5.75	60.15
C.1	w/o EDS	21.85	0.73	5.82	45.79
C.2	w/o TVM	21.93	0.75	5.77	62.10

Additional evaluations regarding 3D mesh alignment, algorithmic computational costs, and out-of-distribution robustness are detailed in Appendix D.

Table 4: Ablation Study of Loss Function.

Model	PSNR	SSIM	Sync_c	Acc_e
Full Model	22.19	0.77	5.84	62.35
w/o $\mathcal{L}_{syn}$	22.10	0.77	3.12	58.35
w/o $\mathcal{L}_{emo}^{v-v}$	22.13	0.78	5.43	51.42
w/o $\mathcal{L}_{emo}^{v-a}$	22.01	0.76	5.63	55.76

Table 5: Sensitivity Analysis for Loss Weights.

λ_{motion}	λ_{L1}	λ_{emo}	PSNR	SSIM	Acc_e
1	1	1	21.89	0.70	60.07
2	2	2	22.10	0.73	60.72
1	2	1	22.15	0.76	61.21
1	2	2	22.19	0.77	62.35

4.2 Ablation Study

Disentangled Feature Learning. Experiments A.1 and A.2 confirm the efficacy of our feature disentanglement. Removing the Geometric Prior Learning (GPL) branch (A.1) cripples lip-sync performance, causing the Sync_c score to plummet from 5.84 to a mere 1.85. Conversely, removing the Emotion Representation Learning (ERL) branch (A.2) devastates emotional expressiveness, with ACC_e dropping from 62.35% to 41.23%. This orthogonal impact provides clear proof that we have successfully isolated the drivers for Pillar 1 (Lip-Sync) and Pillar 2 (Expressiveness) into independent, non-interfering pathways. Furthermore, replacing our 3DMM "differentiable bridge" with an unstructured decoder (A.3) also severely degrades Sync_c (to 3.23), confirming that structured, anatomical guidance is critical for high-fidelity lip motion.

Hierarchical Generation and Identity Preservation. Our analysis validates the principled design of the generator. Disabling Hierarchical Injection (HI) and naively concatenating all control signals (B.1) leads to a "feature clash" that harms both lip-sync (Sync_c 4.91) and expression (ACC_e 52.64). This confirms our coarse-to-fine injection strategy is crucial for preventing interference between motion and emotion signals. Similarly, removing the global style anchor $\boldsymbol{f}_{STY}$ (B.2) degrades performance by removing the stable identity context needed for high-fidelity rendering. The spatial texture features, $\boldsymbol{f}_{TEX}$, are structurally integral to the warping-based synthesis, and their removal impacts training stability and causes a runtime error.

Emotion and Detail Enhancement. We isolate the impact of our enhancement modules. Disabling the Emotion Disentanglement Strategy (EDS) (C.1) causes a major drop in ACC_e to 45.79%, proving our cross-synthesis consistency loss is essential for learning an identity-agnostic emotion embedding, free from the "pollution" of identity-specific attributes. Concurrently, removing the Temporal Variance Masking (TVM) (C.2) results in a noticeable dip in PSNR and SSIM, demonstrating its role in focusing the model's capacity on dynamic regions to enhance fine-grained reconstruction quality.

Ablation Study of Loss Function. Table 4 demonstrates the contribution of each loss term. Removing the sync loss ($\mathcal{L}_{syn}$) drops the Sync_c score from 5.84 to 3.12, underscoring its critical role in audio-lip alignment. Similarly, disabling the emotion losses ($\mathcal{L}_{emo}^{v-v}$ and $\mathcal{L}_{emo}^{v-a}$) degrades emotion accuracy, confirming their necessity for visual emotion consistency. Ultimately, the full model achieves top performance across all metrics, validating our loss design.

Sensitivity Analysis for Loss Weights. We investigated the sensitivity of our model to the primary loss weights, as shown in Table 5. We found that the L1 reconstruction weight (λ_{L1}) is important for visual quality; increasing it to 2 significantly boosts both PSNR and SSIM. Concurrently, the emotion loss weight (λ_{emo}) directly enhances expressive accuracy. The best trade-off was achieved with the weights set to ($\lambda_{motion} = 1$, $\lambda_{L1} = 2$, $\lambda_{emo} = 2$), as this configuration delivered the highest performance across all evaluation metrics. Therefore, we use these values for our final model.

5 Conclusion

We propose AffecTalk, a unified end-to-end framework for affective talking head synthesis. By integrating a differentiable 3DMM bridge and a label-free audio-driven expressive strategy, AffecTalk simultaneously improves lip-synchronization, emotion fidelity, and inference speed. Extensive experiments confirm it achieves competitive real-time performance, effectively balancing generation quality, expressiveness, and efficiency.

References

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.

V BLANA. A morphable model for the synthesis of 3d faces. In *SIGGRAPH'99 Conference Proceedings*, pp. 187–194, 1999.

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*, pp. 7832–7841, 2019.

Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2403–2410, 2025.

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

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016b.

Radek Danecek, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20311–20322, 2022.

Chenpeng Du, Qi Chen, Tianyu He, Xu Tan, Xie Chen, Kai Yu, Sheng Zhao, and Jiang Bian. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4281–4289, 2023.

Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. Unitalker: Scaling up audio-driven 3d facial animation through a unified model. In *European Conference on Computer Vision*, pp. 204–221. Springer, 2024.

Guanwen Feng, Zhihao Qian, Yunan Li, Siyu Jin, Qiguang Miao, and Chi-Man Pun. Les-talker: Fine-grained emotion editing for talking head generation in linear emotion space, 2025. URL `https://arxiv.org/abs/2411.09268`.

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*, pp. 22634–22645, 2023.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

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

Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, and Ming-Yu Liu. Space: Speech-driven portrait animation with controllable expression. In *Proceedings of the ieee/cvf international conference on computer vision*, pp. 20914–20923, 2023.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL `https://arxiv.org/abs/2006.11239`.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

Xiaozhong Ji, Chuming Lin, Zhonggan Ding, Ying Tai, Junwei Zhu, Xiaobin Hu, Donghao Luo, Yanhao Ge, and Chengjie Wang. Realtalk: Real-time and realistic audio-driven face generation with 3d facial prior-guided identity alignment network. *arXiv preprint arXiv:2406.18284*, 2024.

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*, pp. 14080–14089, 2021.

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*, pp. 1–10, 2022.

Yufeng Jin, Guosheng Hu, Haonan Chen, Duoqian Miao, Liang Hu, and Cairong Zhao. Cross-modal distillation for speaker recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12977–12985, 2023.

Jisoo Kim, Jungbin Cho, Joonho Park, Soonmin Hwang, Da Eun Kim, Geon Kim, and Youngjae Yu. Deeptalk: Dynamic emotion embedding for probabilistic speech-driven 3d face animation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 4275–4283, 2025.

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*, pp. 3387–3396, 2022.

Chao Liang, Jianwen Jiang, Wang Liao, Jiaqi Yang, Weihong Zeng, Han Liang, et al. Alignhuman: Improving motion and fidelity via timestep-segment preference optimization for audio-driven human animation. *arXiv preprint arXiv:2506.11144*, 2025.

Jin Liu, Xi Wang, Xiaomeng Fu, Yesheng Chai, Cai Yu, Jiao Dai, and Jizhong Han. Osm-net: One-to-many one-shot talking head generation with spontaneous head motions. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6888–6900, 2024a.

Tao Liu, Feilong Chen, Shuai Fan, Chenpeng Du, Qi Chen, Xie Chen, and Kai Yu. Anitalker: animate vivid and diverse talking faces through identity-decoupled facial motion encoding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6696–6705, 2024b.

Jiayi Lyu, Xing Lan, Guohong Hu, Hanyu Jiang, Wei Gan, Jinbao Wang, and Jian Xue. Multimodal emotional talking face generation based on action units. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

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

Debin Meng, Xiaojiang Peng, Kai Wang, and Yu Qiao. Frame attention networks for facial expression recognition in videos. In *2019 IEEE international conference on image processing (ICIP)*, pp. 3866–3870. IEEE, 2019.

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.

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*, pp. 20687–20697, 2023.

Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 666–676, 2024.

Ziqiao Peng, Jiwen Liu, Haoxian Zhang, Xiaoqiang Liu, Songlin Tang, Pengfei Wan, Di Zhang, Hongyan Liu, and Jun He. Omnisync: Towards universal lip synchronization via diffusion transformers. *arXiv preprint arXiv:2505.21448*, 2025.

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*, pp. 484–492, 2020.

Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13759–13768, 2021.

Zhicheng Sheng, Liqiang Nie, Min Zhang, Xiaojun Chang, and Yan Yan. Stochastic latent talking face generation toward emotional expressions and head poses. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2734–2748, 2023.

Wuzhen Shi, Zibang Xue, and Yang Wen. Keypoints and action units jointly drive talking head generation for video conferencing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Sanjana Sinha, Sandika Biswas, Ravindra Yadav, and Brojeshwar Bhowmick. Emotion-controllable generalized talking face generation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022.

Teng Sun, Yinwei Wei, Juntong Ni, Zixin Liu, Xuemeng Song, Yaowei Wang, and Liqiang Nie. Muti-modal emotion recognition via hierarchical knowledge distillation. *IEEE Transactions on Multimedia*, 26: 9036–9046, 2024.

Yasheng Sun, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Zhibin Hong, Jingtuo Liu, Errui Ding, Jingdong Wang, Ziwei Liu, and Koike Hideki. Masked lip-sync prediction by audio-visual contextual exploitation in transformers. In *SIGGRAPH Asia 2022 conference papers*, pp. 1–9, 2022.

Shuai Tan, Bin Ji, Mengxiao Bi, and Ye Pan. Edtalk: Efficient disentanglement for emotional talking head synthesis. In *European Conference on Computer Vision*, pp. 398–416. Springer, 2024a.

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, pp. 5079–5087, 2024b.

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*, pp. 244–260. Springer, 2024.

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*, pp. 17979–17989, 2023a.

Haotian Wang, Yuzhe Weng, Yueyan Li, Zilu Guo, Jun Du, Shutong Niu, Jiefeng Ma, Shan He, Xiaoyan Wu, Qiming Hu, et al. Emotivetalk: Expressive talking head generation through audio information decoupling and emotional video diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26212–26221, 2025a.

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

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 *ECCV*, 2020.

Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20333–20342, 2022a.

Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. In *International Conference on Learning Representations*, 2022b.

Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022c.

Zhongjian Wang, Peng Zhang, Jinwei Qi, Guangyuan Wang, Chaonan Ji, Sheng Xu, Bang Zhang, and Liefeng Bo. Omnitalker: One-shot real-time text-driven talking audio-video generation with multimodal style mimicking. *arXiv preprint arXiv:2504.02433*, 2025b.

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

Yibo Xia, Lizhen Wang, Xiang Deng, Xiaoyan Luo, Yunhong Wang, and Yebin Liu. Gmtalker: Gaussian mixture-based audio-driven emotional talking video portraits, 2025. URL `https://arxiv.org/abs/2312.07669`.

Chaolong Yang, Kai Yao, Yuyao Yan, Chenru Jiang, Weiguang Zhao, Jie Sun, Guangliang Cheng, Yifei Zhang, Bin Dong, and Kaizhu Huang. Unlock pose diversity: Accurate and efficient implicit keypoint-based spatiotemporal diffusion for audio-driven talking portrait. *arXiv preprint arXiv:2503.12963*, 2025.

Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. *arXiv preprint arXiv:2401.08503*, 2024.

Lingyun Yu, Jun Yu, Mengyan Li, and Qiang Ling. Multimodal inputs driven talking face generation with spatial–temporal dependency. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1): 203–216, 2020.

Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, and Baoyuan Wang. Talking head generation with probabilistic audio-to-visual diffusion priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7645–7655, 2023.

Chenxu Zhang, Chao Wang, Jianfeng Zhang, Hongyi Xu, Guoxian Song, You Xie, Linjie Luo, Yapeng Tian, Xiaohu Guo, and Jiashi Feng. Dream-talk: Diffusion-based realistic emotional audio-driven method for single image talking face generation, 2023a. URL `https://arxiv.org/abs/2312.13578`.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

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*, pp. 8652–8661, 2023b.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3661–3670, 2021.

Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng YAN. MEMO: Memory-guided diffusion for expressive talking video generation. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=uBcHcM7Kzi`. J2C Certification.

Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 9299–9306, 2019.

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*, pp. 4176–4186, 2021.

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.

A Main Notation Introduction

Table 6: Meanings of Notations

Notation	Type	Meaning
$\boldsymbol{I}^i$	$B \times 3 \times H \times W$	The identity image
$\mathbf{A}_{1:T}$	$B \times T \times N$	The audio clip
$\mathcal{E}_I$	Module	The image encoder
$\mathcal{E}_A$	Module	The audio encoder
$\mathcal{E}_M$	Module	The motion encoder
$\mathcal{D}$	Module	The motion decoder
$\mathcal{S}$	Module	The student module
$\mathcal{T}$	Module	The teacher module
$\mathcal{G}$	Module	The generator
$\boldsymbol{f}_E$	$B \times 512$	The feature of emotion
$\boldsymbol{f}_M$	$B \times 512$	The feature of facial motion
$\boldsymbol{f}_{TEX}$	$B \times 512$	The spatial texture feature
$\boldsymbol{f}_{STY}$	$B \times 512$	The global style feature
$\boldsymbol{f}_{CTRL}$	$B \times 512$	The feature of composite control
$\boldsymbol{f}_{INT}$	$B \times C \times H \times W$	The intermediate feature of generator
$\mathcal{L}_{rec}$	Loss	The reconstruction loss function
$\mathcal{L}_{L1}$	Loss	The L1 loss function
$\mathcal{L}_{perc}$	Loss	The perceptual loss function
$\mathcal{L}_{adv}$	Loss	The adversarial loss function
$\mathcal{L}_{syn}$	Loss	The syncnet loss function
$\mathcal{L}_{emo}^{v-v}$	Loss	The visual-visual emotion perceptual loss function
$\mathcal{L}_{emo}^{v-a}$	Loss	The visual-audio emotion perceptual loss function
λ_i	Constant	The loss weights

B Network Architecture

B.1 An Introduction to 3DMM

A 3D Morphable Model (3DMM) is a powerful statistical model of 3D facial geometry and texture, first introduced by BLANA (1999). It is constructed by applying Principal Component Analysis (PCA) to a large dataset of registered 3D face scans. The core principle of a 3DMM is that any specific face can be parametrically represented as a linear combination of a mean face and a set of basis vectors (the principal components), which capture the primary modes of variation across the training population.

The 3D shape $\boldsymbol{s}_t$ of a face at time t, represented as a mesh of vertices, can be mathematically formulated using the notation from Section 3.1 as:

$$\boldsymbol{s}_t = \bar{\boldsymbol{s}} + \boldsymbol{U}^{\text{id}}\boldsymbol{p}^{\text{id}} + \boldsymbol{U}^{\text{exp}}\boldsymbol{p}_t^{\text{exp}} \tag{7}$$

where:

- $\bar{\boldsymbol{s}}$ is the average face shape computed from the training dataset.
- $\boldsymbol{U}^{\text{id}}$ and $\boldsymbol{U}^{\text{exp}}$ are the orthonormal bases for identity and expression, respectively. The identity basis, $\boldsymbol{U}^{\text{id}}$, captures variations in facial structure that are unique to an individual (e.g., bone structure, nose shape), while the expression basis, $\boldsymbol{U}^{\text{exp}}$, captures non-rigid deformations corresponding to facial expressions (e.g., smiling, frowning).
- $\boldsymbol{p}^{\text{id}} \in \mathbb{R}^m$ is the static, low-dimensional coefficient vector that controls the identity of the face.

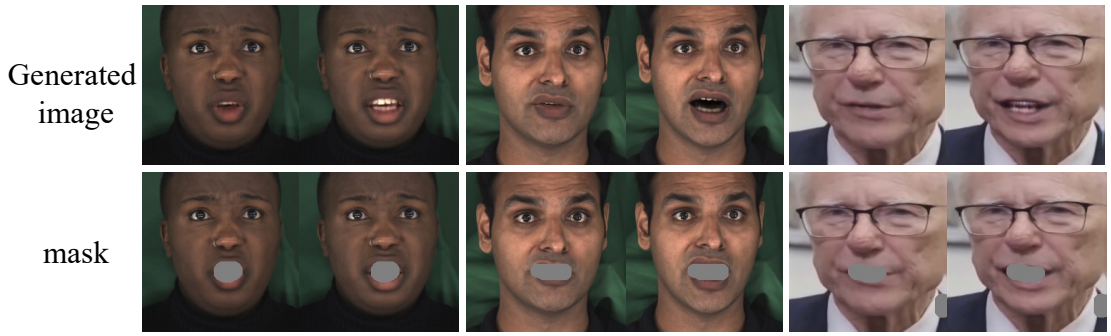


Figure 7: Visualization of the dynamic mask generated by Temporal Variance Masking module.

- $\boldsymbol{p}_t^{\text{exp}} \in \mathbb{R}^n$ is the dynamic, low-dimensional coefficient vector that controls the facial expression at time t. By manipulating these vectors, one can generate novel and realistic facial geometries.

For the purpose of talking head animation, this model is augmented to include rigid transformations, namely head pose. The pose is typically described by a rotation matrix $\boldsymbol{r} \in SO(3)$ (i.e., $\boldsymbol{r}$ is a 3x3 matrix that describes a valid physical rotation of the head in 3D space, without any stretching or mirror-flipping) and a translation vector $\boldsymbol{t} \in \mathbb{R}^3$. Thus, the complete set of parameters to define a specific talking head instance at a moment in time is a combination of identity, expression, and pose coefficients: $(\boldsymbol{p}^{\text{id}}, \boldsymbol{p}_t^{\text{exp}}, \boldsymbol{r}, \boldsymbol{t})$. More advanced 3DMMs, such as FaceVerse (Wang et al., 2022a) used in GMTalker (Xia et al., 2025), may include additional parameters for finer control over eye blinks and gaze.

The primary advantage of using a 3DMM in audio-driven talking head generation is its ability to **disentangle** the complex factors of facial appearance and motion. As demonstrated in works like SadTalker (Zhang et al., 2023b) and LES-Talker (Feng et al., 2025), this disentangled, parametric representation transforms the ill-posed, one-to-many problem of mapping audio directly to pixels into a more constrained and manageable task of predicting low-dimensional coefficients. Models can learn the strong correlation between audio and lip-related expression coefficients ($\boldsymbol{p}_t^{\text{exp}}$) separately from the weaker, more varied correlation with head pose $(\boldsymbol{r}, \boldsymbol{t})$, thereby mitigating ambiguity and enabling the generation of diverse and realistic animations. As used in RealPortrait (Ye et al., 2024), 3DMM parameters can also be used to derive appearance-agnostic motion representations (e.g., PNCC) to animate other 3D structures. In our work, we leverage this property by using the 3DMM as a differentiable regularizer to guide the learning of speech-related geometric motion.

B.2 Temporal Variance Masking.

We introduce Temporal Variance Masking (TVM), a lightweight, plug-and-play module designed to automatically identify dynamic regions within a sequence of feature maps. The core principle of TVM is that spatial regions exhibiting high variance along the temporal dimension correspond to dynamic content (e.g., a speaking mouth), while low-variance regions represent the static background.

The mask generation process, detailed in Algorithm 1, follows a multi-stage refinement pipeline. First, a variance map is generated by computing the pixel-wise variance of the feature maps $\boldsymbol{f}$ across the temporal dimension, followed by an averaging operation across the channel dimension to produce a 2D map $\mathbf{V}$. This map is then normalized to the range $[0, 1]$ and binarized using an initial threshold τ_{init} to form a coarse mask $\mathbf{M}_{coarse}$. To enhance robustness, this mask is refined via a morphological opening operation, which effectively eliminates spurious pixel clusters not associated with coherent motion. Subsequently, a Gaussian blur is applied to smooth the contours of the refined binary mask. The final motion mask, $\mathbf{M}_{motion}$, is obtained by resizing the smoothed map to the target dimensions and performing a final binarization with a threshold τ_{final}. As illustrated in Fig. 7, the mask generated by TVM accurately isolates the primary motion areas, such as the mouth region during speech.

Algorithm 1 Temporal Variance Masking

Require:
Sequence of feature maps $\boldsymbol{f} \in \mathbb{R}^{T \times C \times H \times W}$
Initial threshold $\tau_{\text{init}} \in [0, 1]$
Morphological opening kernel size k
Gaussian blur standard deviation $\sigma_g > 0$
Final threshold $\tau_{\text{final}} \in [0, 1]$
Target output size $(H_{\text{out}}, W_{\text{out}})$
Ensure:
Motion mask $\mathbf{M}_{motion} \in \{0, 1\}^{H_{\text{out}} \times W_{\text{out}}}$

▷ Compute a 2D variance map from temporal dynamics
$\mathbf{V} \leftarrow \text{Mean}_C(\text{Var}_T(\boldsymbol{f}))$
$\mathbf{M}_{norm} \leftarrow (\mathbf{V} - \min(\mathbf{V}))/(\max(\mathbf{V}) - \min(\mathbf{V}))$
$\mathbf{M}_{coarse} \leftarrow \mathbf{1}[\mathbf{M}_{norm} > \tau_{\text{init}}]$
▷ Refine the binary mask
$\mathbf{M}_{refined} \leftarrow \text{Opening}(\mathbf{M}_{coarse}, k)$
$\mathbf{M}_{smooth} \leftarrow \text{GaussianBlur}(\mathbf{M}_{refined}, \sigma_g)$
▷ Resize and re-binarize
$\mathbf{M}_{resized} \leftarrow \text{Resize}_{\text{bilinear}}(\mathbf{M}_{smooth}, H_{\text{out}}, W_{\text{out}})$
$\mathbf{M}_{motion} \leftarrow \mathbf{1}[\mathbf{M}_{resized} > \tau_{\text{final}}]$
return $\mathbf{M}_{motion}$

C Experimental Settings

C.1 Datasets

MEAD. This dataset is a large-scale, high-quality laboratory dataset meticulously designed for research in audio-visual emotion recognition. It contains recordings of 60 actors (43 of whom are publicly accessible) portraying eight different emotions: anger, contempt, disgust, fear, happiness, sadness, surprise, and neutral.

HDTF. This dataset is a large-scale "in-the-wild" dataset curated for the task of talking head generation and audio-driven facial synthesis. Sourced from high-definition YouTube videos, it is renowned for its quality and features over 300 unique identities across a wide range of ages, genders, and ethnicities.

C.2 Implementation Details

For video preprocessing, we follow the methodology of EMOCA (Danecek et al., 2022) by cropping the faces and resizing the resulting videos to a 256 × 256 resolution for both training and testing. The dimensionality of the style code ($\boldsymbol{f}_{STY}$) and texture code ($\boldsymbol{f}_{TEX}$) is set to 512. For the loss functions, the weights for λ_{L1} and λ_{emo} are set to 2, while all remaining weights are set to 1. We initialize the image encoder and generator with pre-trained weights from Tan et al. (2024a), which is based on LIA (Wang et al., 2022b). Our audio encoder is a pre-trained Wav2Vec 2.0 model (Baevski et al., 2020). All audio is down-sampled to 16 kHz, and the learning rate is set to 0.001. The model was trained on 8 NVIDIA A100 GPUs, each with 80GB of memory.

C.3 Baselines

The eight baselines we compared to are summarized below:

- **EAMM.** The core of this method is decoupling and superposition. It first breaks the task into two steps: 1) generating a "base animation" that is lip-synced to the audio but has a neutral expression; and 2) extracting pure "emotional dynamics" from a separate source video and adding them as a "displacement" on top of the base animation.

- **EAT.** The core of this method is an efficient, plugin-based adaptation strategy. It freezes a pre-trained, emotion-agnostic talking-head model and inserts lightweight "emotional plugins" at various stages. These plugins alone are trained to add expressive, controllable emotions, thus avoiding costly, full-model retraining.
- **SadTalker.** The core of this method involves decoupling motion. It separately predicts 3D expression coefficients from audio and generates diverse head poses. It then bypasses 3D rendering by translating these motion parameters into 2D keypoints, which drive a 2D animation engine to synthesize the final video.
- **AniTalker.** The core of this method is to learn a universal, identity-decoupled motion space using self-supervised disentanglement. A diffusion model then generates these motion codes from audio to drive a renderer, synthesizing the final video.
- **RealPortrait.** This method uses a "reconstruct-then-animate" pipeline. It first reconstructs a high-quality 3D head (tri-plane) from a single image. A lightweight motion adapter then computes a "motion diff-plane" from the driving signal, which is added to the head. Finally, a specialized synthesizer generates the video.
- **KDTalker.** The core of KDTalker is that it's the first to use a Diffusion Model for predicting keypoint motion. It does not generate images directly; instead, it utilizes a custom spatiotemporal diffusion model to generate a series of dynamic changes for implicit 3D keypoints (including deformation, head pose, etc.) based on the audio. These keypoints are then fed into a separate renderer to create the final video.
- **EchoMimic.** The core of EchoMimic is a strategy of hybrid driving and collaborative training. It builds a unified diffusion model framework that simultaneously learns to be driven by both audio and facial landmarks. Through a novel training technique (e.g., randomly dropping parts of the driving signal), the model learns to generate high-quality videos whether driven by audio alone, landmarks alone, or a combination of audio and selected landmarks.
- **AniPortrait.** The core of AniPortrait is a two-stage process: first, it converts the audio signal into a sequence of 2D facial landmarks as an intermediate motion representation; then, this landmark sequence is used as a "pose controller" (similar to ControlNet) to guide a powerful pre-trained image diffusion model (like Stable Diffusion) to generate the final, high-quality animated video frame by frame.

D Additional Experimental Results

D.1 3D Mesh Alignment and Quality.

As shown in Fig. 8, the 3D mesh generated by our method maintains high fidelity with the final rendered mouth shape, validating its effectiveness as a geometric prior for the synthesis process. Notably, this quality is achieved without requiring any explicit 3D supervision. Instead, our end-to-end framework directly optimizes the 3DMM parameters from 2D video frames, thereby bypassing the error accumulation inherent in pipelines that rely on intermediate mesh supervision. Consequently, when benchmarked against the recent SOTA method UniTalker (Fan et al., 2024), our approach yields more accurate and realistic mouth shape predictions.

D.2 Computational Cost Analysis

To evaluate the practical applicability of AffecTalk in real-world scenarios, this section provides a detailed analysis of its computational efficiency. The model was trained for approximately 20 hours on 8 NVIDIA A100 GPUs. All subsequent inference benchmarks were conducted on a single NVIDIA A100 GPU to ensure a fair comparison. Key performance metrics are summarized in Table 7.

- **Compared to non-diffusion-based methods:** AffecTalk is also substantially more efficient than prior non-diffusion approaches. Compared with cascaded pipelines such as SadTalker and EAMM, our unified architecture avoids multiple sequential forward passes through separate modules for audio-to-motion prediction, intermediate geometric rendering, and final image synthesis. Instead, AffecTalk uses

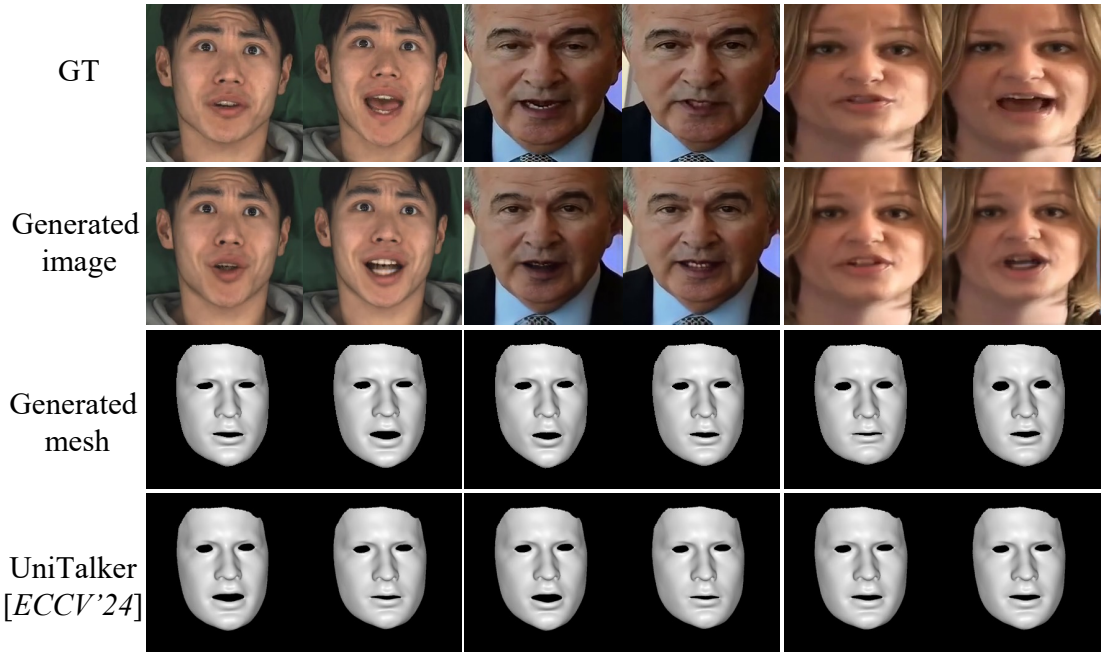


Figure 8: Visualization of Talking Head Generation and 3D Mesh Alignment.

Table 7: Computational profile of AffecTalk. All inference metrics are benchmarked on a single NVIDIA A100 GPU.

Model	VRAM (GB)	Inference Speed (FPS)	Inference Latency (ms)
EAT	5.54	12.23	81.77
SadTalker	4.39	21.74	46.00
Aniportrait	15.84	3.84	260.42
RealPortrait	7.18	3.77	265.25
KDTalker	18.85	15.34	65.19
AffecTalk	6.48	55.32	18.08

a differentiable 3DMM bridge to inject both geometric and emotional priors directly into a single-pass GAN generator, reducing both inference latency and computational overhead. Compared with end-to-end models such as EAT, the efficiency advantage arises from different architectural choices. EAT relies on Transformer-based modules to predict intermediate 3D representations, where self-attention becomes increasingly expensive for longer sequences. In contrast, AffecTalk employs lightweight temporal convolutional networks (TCNs) for temporal feature extraction together with an efficient convolutional generator. This design provides strong synthesis quality while remaining computationally efficient.

- **Compared to diffusion-based methods:** While diffusion-based methods, such as KDTalker, have set a high bar for photorealism and expressive naturalness, they inherently struggle to meet the strict latency requirements of real-time applications. This limitation stems from their mathematical formulation, specifically the iterative sampling process, which requires dozens of sequential denoising steps to generate a single frame. This multi-step inference fundamentally incurs prohibitive computational costs and high latency, making them orders of magnitude slower than what interactive scenarios demand. In stark contrast, AffecTalk is architecturally designed for efficiency. As a single-pass GAN-based framework, our model circumvents iterative denoising by generating each frame in a single, efficient forward pass. This algorithmic advantage allows AffecTalk to achieve inference speeds exceeding 55 FPS, effectively navigating the critical trade-off between generative quality and real-time viability.

Figure 9: Qualitative comparisons on VFHQ dataset.

Table 8: Generalization performance comparison on VFHQ.

Method	PSNR	SSIM	Sync_c
RealPortrait	18.86	0.51	4.27
AffecTalk (ours)	20.43	0.65	5.63

D.3 Evaluation on Generalization and Robustness

To rigorously assess the real-world applicability of our framework, we evaluated AffecTalk's zero-shot generalization and robustness on the VFHQ dataset. This "in-the-wild" video benchmark, entirely excluded from our training set, presents a challenging out-of-distribution scenario with its diverse unseen subjects and complex, noisy backgrounds.

As illustrated in Table 8, AffecTalk achieves strong performance, quantitatively outperforming the recent GAN-based method, RealPortrait. Qualitatively, Fig. 9 further illustrates that AffecTalk maintains superior lip consistency and identity preservation even against cluttered backgrounds. This performance confirms that AffecTalk not only animates novel identities with high fidelity but also operates robustly in unconstrained environments. This achievement on a completely unseen domain provides compelling evidence that AffecTalk is a generalizable and robust framework well-suited for real-world deployment.

E Ablation Study Setting.

To validate the contribution of each key component, we conduct a comprehensive ablation study. The complete quantitative results are presented in Table 3, with a detailed description and analysis of each setting provided below.

- **w/o GPL:** Removes the Geometric Prior Learning (GPL) branch from the Disentanglement Feature Learning module.
- **w/o ERL:** Removes the Emotion Representation Learning (ERL) branch from the Disentanglement Feature Learning module.
- **w/o 3DMM:** Replaces the 3DMM-based geometric prior with a direct prediction from a simple decoder.
- **w/o HI:** Disables hierarchical injection, instead concatenating all features ($\boldsymbol{f}_{STY}$, $\boldsymbol{f}_M$, and $\boldsymbol{f}_E$) at the input layer of generator.
- **w/o $\boldsymbol{f}_{STY}$:** Excludes the style feature $\boldsymbol{f}_{STY}$ from the generator's input.
- **w/o EDS:** Disables the Emotion Disentanglement Strategy (EDS).

- **w/o TVM:** Removes the Temporal Variance Masking (TVM) from the reconstruction loss calculation.

F Limitations and Future Work

Limitations. While AffecTalk excels in synthesizing highly expressive and synchronized facial animations, the current framework primarily focuses on the head and neck regions. It does not explicitly model the coupled movements of the upper torso and shoulders. Consequently, when the synthesized head exhibits highly energetic emotional responses, the lack of corresponding body language can occasionally make the overall portrait appear slightly constrained.

Future Work. To address this, our future work will focus on extending the framework to explicitly model torso dynamics and natural shoulder movements. By incorporating holistic body language priors, we aim to provide a more immersive and comprehensive virtual avatar experience.