AFFECTATLAS: STRUCTURED NEURO-SYMBOLIC AFFECT MODELING FOR EMOTIONALLY CONSISTENT TALKING FACE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-driven talking-face generation has made substantial progress in visual fidelity, identity preservation, and lip synchronization. However, existing methods often treat emotion as a weak or isolated conditioning signal, leading to facial behaviour that is visually plausible but affectively inconsistent with speech. We propose **AffectAtlas**, a neuro-symbolic affect modeling framework for emotion-consistent talking-face generation. We model affect as a structured system over three complementary representations: expressions, action units, valence-arousal. We first construct a probabilistic affect atlas that captures bidirectional relationships among these affect spaces. We then introduce a teacher-guided atlas adaptation mechanism to adapt the atlas to talking-face data while preserving its structured prior. Finally, we incorporate differentiable neuro-symbolic rules to enforce psychologically meaningful consistency among the three affect spaces, and inject the refined representations into a diffusion-based generator via an emotion-aware A/V module. Extensive quantitative, qualitative, cross-dataset, ablation and user-study evaluations show that AffectAtlas improves emotion fidelity, temporal affect alignment, visual quality, and overall perceptual realism compared to state-of-the-art methods.

1 INTRODUCTION

Audio-driven talking-face generation aims to synthesize realistic facial videos from a reference identity and a driving speech signal. Recent advances, particularly diffusion-based methods such as Hallo (Xu et al., 2024) and VExpress (Wang et al., 2024), have significantly improved visual fidelity, identity preservation and lip synchronization. However, generating a face that merely moves in synchrony with speech is not sufficient for realistic human communication. Facial behaviour also conveys emotion, intention and affective state. When the generated expression is emotionally inconsistent with the speech, the result may appear unnatural even if the lips are well synchronized. Most existing talking-face generation approaches treat emotion in a limited manner. Early works (Chen et al., 2018; Cheng et al., 2022) primarily focused on lip synchronization and head motion, while more recent methods (Gan et al., 2023; Wang et al., 2025; Xu et al., 2023; Tan et al., 2024a; Liu et al., 2025b; Feng et al., 2026) incorporate emotion using discrete labels, reference images or videos, or a limited set of facial action units. While these approaches improve controllability, they rely on incomplete affect representations. A single emotion label cannot capture intensity or temporal variation; reference-based control may introduce identity-specific biases; and AU-based representations focus on local muscle activations without fully capturing semantic emotion or continuous affective intensity. In natural facial behaviour, emotion can be represented through multiple affect spaces, including categorical expressions (Ekman & Friesen, 1978), facial action units (AUs) (Ekman & Friesen, 1978), and continuous valence-arousal (VA) (Russell, 1980). These representations capture complementary aspects of affect and should remain consistent with each other. For example, a "happy" expression should align with positive valence and relevant AUs such as AU6 and AU12, while a given AU configuration should correspond to a plausible expression and VA state.

To address this limitation, we propose **AffectAtlas**, a neuro-symbolic framework for structured affect modeling in talking-face generation. The core idea is to construct a unified probabilistic affect prior that explicitly models the relationships among expressions (EXP), facial action units (AUs),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and valence-arousal (VA) through bidirectional mappings across affect spaces. Since an offline prior may not fully match the distribution of talking-face videos, we further introduce a teacher-guided adaptation mechanism that refines the atlas using outputs from pretrained affect models while preserving its structured relationships. In addition, we introduce a neuro-symbolic consistency framework that injects prior knowledge of facial behaviour into the learning process. Building on recent advances in differentiable logic and neuro-symbolic learning (Daniele et al., 2022), we encode relationships among EXP, AUs, and VA as soft logical rules and enforce them through a differentiable loss. Finally, we integrate the refined affect representation into a diffusion-based talking-face generation model via an emotion-aware audio-visual module, enabling the generated face to follow both the phonetic structure and emotional content of speech while producing temporally coherent facial behaviour. Overall, our contributions are as follows:

1. We introduce a probabilistic affect atlas that unifies categorical expressions, facial action units, and valence-arousal into a structured affect prior.
2. We propose a teacher-guided adaptation mechanism with residual correction and cycle consistency to adapt the affect prior to talking-face data.
3. We develop a neuro-symbolic consistency framework that uses differentiable facial behaviour rules to enforce agreement across affect representations.
4. We integrate the refined affect representation into a diffusion-based talking-face generation model for dynamic, emotion-consistent facial synthesis while preserving visual quality and lip synchronization. Extensive qualitative and quantitative experiments and user studies demonstrate improvements in emotion fidelity, temporal affect alignment, and generation quality over SOTA.

2 Related Work

Audio-driven talking-face generation. Audio-driven talking face generation aims to generate realistic talking videos from audio and a reference image. Early works (Prajwal et al., 2020; Yin et al., 2022; Cheng et al., 2022; Chen et al., 2018; Suwajanakorn et al., 2017) mainly focused on lip synchronization while keeping other facial attributes unchanged. Later works (Hong et al., 2024; Wang et al., 2021; Zhang et al., 2023; Zhou et al., 2020; Qi et al., 2023) incorporated head poses and eye blinks using intermediate representations such as 3D Morphable Models (3DMM) and facial landmarks. However, these methods often generate inaccurate motions from the audio, which limits the realism and expressiveness of the generated videos. In contrast, end-to-end methods (Shen et al., 2023; 2022; Wang et al., 2023b; Zhou et al., 2019; Thies et al., 2020; Chen et al., 2023) combine features from multiple encoders, such as image, identity, and audio, to generate talking face videos. Recently, Hallo (Xu et al., 2024) and VExpress (Wang et al., 2024) used the Stable Diffusion Model (Rombach et al., 2022) to maintain appearance consistency and improved the visual quality of talking face videos. However, these methods do not explicitly consider facial emotion, which is an important factor for generating realistic talking face videos. Our approach aims to address this limitation by integrating emotional information into the talking face generation process.

Emotion-aware talking-face generation. To improve the realism of talking face videos, researchers have attempted to integrate emotion into traditional talking face models in different ways. While some studies (Wang & Fu, 2025; Gan et al., 2023; Tan et al., 2024b; Ji et al., 2021; Pan et al., 2024; Wang et al., 2020) use emotion labels to incorporate emotional information, others (Tan et al., 2024c; Ji et al., 2022; Liang et al., 2022; Wang et al., 2023a; Ma et al., 2023; Wang et al., 2023a) employ reference images or videos to transfer emotion from reference resources to the target. However, these approaches still struggle to preserve expressive emotions aligned with the emotional content of speech. SynchroRaMa (Yee et al., 2026) applied multi-modal emotion embedding based on text, speech emotion recognition, and audio-derived VA signals. Recent methods also explore AUs: EmoSpeaker (Feng et al., 2026) used AU-based contrastive learning. FG-EmoTalk (Sun et al., 2024b) and MoEE (Liu et al., 2025a) fused audio and AU features to control facial expressions. However, relying only on AU may not provide sufficiently faithful emotion embeddings for talking face generation, because AU-based representations mainly focus on local facial muscle movements. Some recent works (Sun et al., 2024a; Liu et al., 2025a) have utilized AU to represent facial expressions. However, these methods use AU labels as text prompts and typically modify only a limited number of AUs, usually one or two. Since facial expressions are usually formed by the combination of multiple AUs, modifying only one or two AUs may not produce realistic expressions. To address

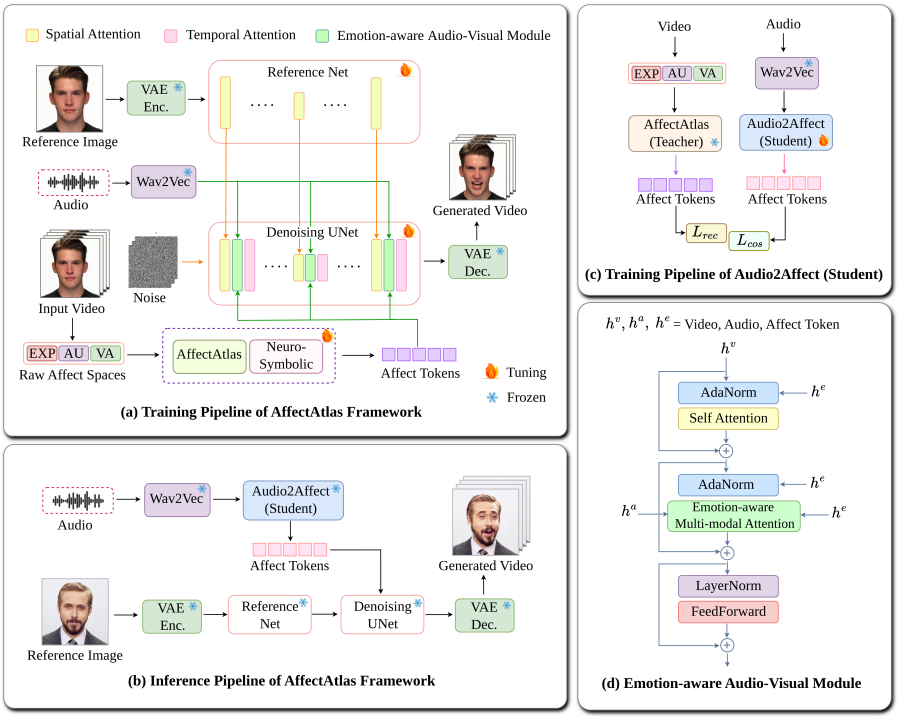


Figure 1: Overview of AffectAtlas. (a) In training, EXP, AU, and VA from the input video are refined by AffectAtlas and neuro-symbolic module, and Diffusion U-Net generates videos conditioned on affect tokens, reference image, and audio. (b) In inference, Audio2Affect predicts affect tokens directly from audio for emotion representation. (c) Audio2Affect learns to predict audio-based affect tokens matching those from the video-based AffectAtlas teacher. (d) Emotion-aware AV module fuses visual, audio, and affect features for accurate lip-sync and emotion-consistent expressions.

these limitations, we propose an emotion embedding that combines AUs, EXP, and VA signals, to provide richer and more comprehensive emotional cues.

3 METHOD

We propose **AffectAtlas**, a neuro-symbolic affect modeling framework for emotion-consistent talking-face generation. Given a reference image and a driving audio, our goal is to generate a realistic talking-face video that preserves identity, accurate lip synchronization, and expresses emotions that are aligned with affective content of speech. Our framework models affect as a structured system composed of 3 complementary representations: EXPs, AUs and VA. The overall framework consists of an offline affect atlas, a teacher-guided atlas adaptation, a neuro-symbolic affect consistency, an Audio2Affect (student) module and an emotion-aware audio-visual generation module. Fig. 1 presents the overall framework, while Fig. 2 provides details of the affect modeling process.

3.1 OFFLINE AFFECT ATLAS: A STRUCTURED AFFECT PRIOR

The proposed method first builds an offline probabilistic affect atlas that captures the relationships among EXP, AU, VA. This atlas provides structured mappings between affect representations.

Let $e \in \{1, \ldots, C\}$ denote a categorical EXP, $\mathbf{au} \in \{0,1\}^K$ denote a binary AU vector, and $\mathbf{va} \in \mathbb{R}^2$ denote a VA state. The atlas models bidirectional relationships among the three spaces:

$$p(\mathbf{au} \mid e), \quad p(e \mid \mathbf{au}), \quad p(\mathbf{va} \mid e), \quad p(e \mid \mathbf{va}), \quad p(\mathbf{va} \mid \mathbf{au}), \quad p(\mathbf{au} \mid \mathbf{va}).$$

The mappings allow information from an affect space to be interpreted and aligned with the others.

EXP-to-VA mapping EXPs correspond to regions rather than fixed points in VA space. Therefore, for each EXP class e, we model $p(\mathbf{va} \mid e)$ using a Gaussian mixture model (GMM): $p(\mathbf{va} \mid e) = \sum_{m=1}^{K_e} \pi_{e,m} \mathcal{N}(\mathbf{va}; \boldsymbol{\mu}_{e,m}, \boldsymbol{\Sigma}_{e,m})$, where the number of components K_e is selected via Bayesian

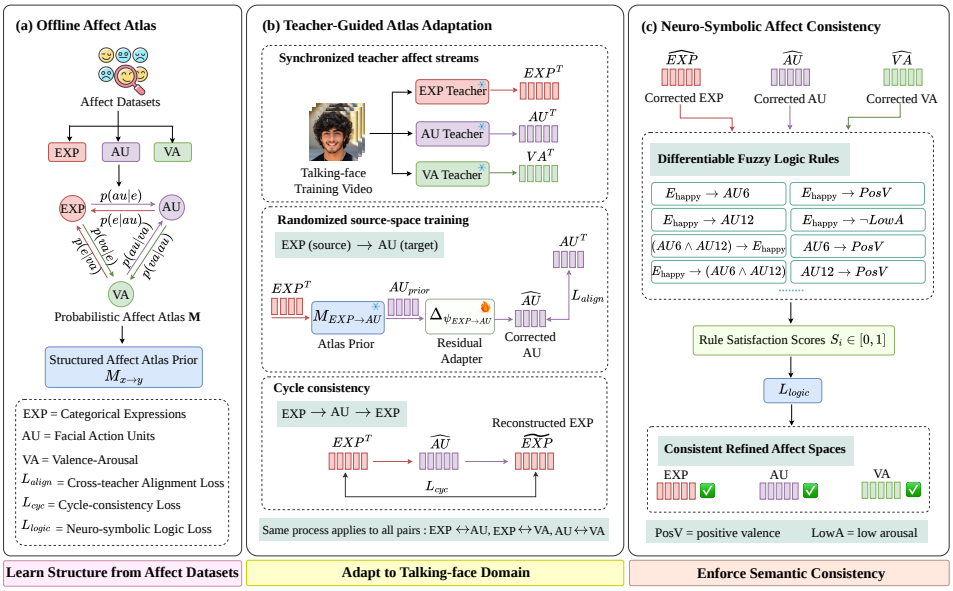


Figure 2: Overview of neuro-symbolic affect modeling framework. (a) Offline affect atlas learns bidirectional probabilistic mappings among EXP, AU, and VA. (b) Teacher-Guided Atlas Adaptation adapts the affect prior to the talking-face domain. One affect space is randomly chosen, mapped to the others through the frozen atlas, and refined with residual corrections using cross-teacher alignment and cycle consistency losses. (c) Neuro-Symbolic Affect Consistency applies differentiable fuzzy logic rules to refine semantically consistent affect representations.

Information Criterion (BIC) and $\pi_{e,m}$, $\boldsymbol{\mu}_{e,m}$, and $\boldsymbol{\Sigma}_{e,m}$ denote the mixture weight, mean, covariance of the m-th component.

VA-to-EXP mapping. The reverse mapping is obtained using Bayes' rule (where $p(e)$ is the frequency of each expression in the data): $p(e \mid \mathbf{va}) = \frac{p(e)\, p(\mathbf{va}|e)}{\sum_{e'} p(e')\, p(\mathbf{va}|e')}$.

EXP-to-AU mapping. An EXP corresponds to a structured configuration of multiple AUs rather than a single facial movement. For example, happiness is commonly associated with AU6 and AU12. However, the same EXP can be realized through different AU combinations, and the same AU may appear in multiple EXPs. Thus, a simple deterministic EXP-to-AU mapping is insufficient. We model $p(\mathbf{au} \mid e)$ using a pairwise binary Markov Random Field (MRF) over the AU vector:

$$p(\mathbf{au} \mid e) = \frac{1}{Z_e} \exp\left(\sum_{k=1}^{K} \theta_k^{(e)} au_k + \sum_{i<j} J_{ij}^{(e)} au_i au_j \right), \tag{1}$$

where $\theta_k^{(e)}$ captures the activation likelihood of AU k under EXP e, $J_{ij}^{(e)}$ captures AU co-activation or inhibition for that EXP, and Z_e normalizes the distribution over all possible AU configurations.

This formulation has two important advantages. First, it models EXP-specific AU activation patterns, allowing each EXP to have its own characteristic AU structure. Second, it captures dependencies among AUs, reflecting the coordinated muscle movements that produce EXP.

AU-to-EXP mapping. The reverse mapping is computed as: $p(e \mid \mathbf{au}) = \frac{p(e)p(\mathbf{au}|e)}{\sum_{e'} p(e')p(\mathbf{au}|e')}$.

AU-to-VA mapping. Mapping AU configurations to VA is challenging because their affective meaning depends on both individual activation and co-occurring AUs. For example, AU12 may indicate positive valence with cheek raising, but different combinations can convey other affective states. Directly learning $p(\mathbf{va} \mid \mathbf{au})$ for all 2^K AU configurations is infeasible, as many are rarely observed. We therefore approximate the full mapping using a mixture of lower-order affect experts.

Given an AU vector $\mathbf{au}$, we first infer an EXP distribution using the atlas posterior $p(e \mid \mathbf{au})$, providing a semantic bridge from AUs to affect: $p(\mathbf{va} \mid \mathbf{au}) = \sum_e p(e \mid \mathbf{au}) p(\mathbf{va} \mid e, \mathbf{au})$, where $p(\mathbf{va} \mid e, \mathbf{au})$ is an EXP-conditioned VA distribution refined by the observed AU pattern.

To model $p(\mathbf{va} \mid e, \mathbf{au})$, we combine 3 complementary sources of affective evidence: (i) EXP-level expert $p(\mathbf{va} \mid e)$ provides a coarse semantic prior over the VA region associated with e; (ii) single-AU experts $p(\mathbf{va} \mid au_k = 1)$ capture the affective tendency of individual facial actions; (iii) AU-pair experts $p(\mathbf{va} \mid au_i = 1, au_j = 1)$ capture compositional facial patterns that cannot be represented by individual AUs alone. We combine them using a Product-of-Experts (PoE) formulation:

$$p(\mathbf{va} \mid e, \mathbf{au}) \propto p(\mathbf{va} \mid e) \prod_{k=1} p(\mathbf{va} \mid au_k = 1) \prod_{(i,j)\in\mathcal{P}(\mathbf{au})} p(\mathbf{va} \mid au_i = 1, au_j = 1), \tag{2}$$

where $\mathcal{P}(\mathbf{au})$ denotes active AU pairs retained in the dictionary $\mathcal{D}$. Single-AU and AU-pair experts are modeled using GMMs. This formulation gives the AU-to-VA mapping three important properties. It is *scalable*, as it avoids estimating a VA density for every AU configuration; *compositional*, as it combines EXP, single-AU, and AU-pair evidence; and *interpretable*, since each contribution can be traced to an EXP prior, individual AU, or meaningful AU co-activation pattern.

VA-to-AU mapping. We approximate reverse mapping from VA to AUs by treating EXP as an intermediate latent variable that explains much of the dependence between VA and AU configuration:

$$p(\mathbf{au} \mid \mathbf{va}) \approx \sum_{e} p(\mathbf{au} \mid e) p(e \mid \mathbf{va}). \tag{3}$$

3.2 Teacher-Guided Atlas Adaptation

The offline AffectAtlas provides a structured prior over EXP, AUs and VA. However, as it is estimated from heterogeneous affect datasets, it may not fully match the talking-face videos that contain domain-specific biases caused by speech articulation, head motion, identity variation, video quality and annotation noise. Direct use of the offline atlas may thus produce affect representations that are principled but not sufficiently aligned with the target generation domain. To address this, we introduce a **teacher-guided atlas adaptation** mechanism. It preserves the structured prior while adapting it in a controlled manner. Frozen affect teachers provide synchronized EXP, AU, and VA streams, while lightweight residual modules correct the atlas mappings. This maintains the atlas structure while correcting domain-specific deviations in talking-face data.

Synchronized teacher affect streams. Given a training video $\mathbf{V} = \{\mathbf{I}_t\}_{t=1}^{T}$, we apply three frozen affect teachers (EXP, AU and VA ones) to process each frame: $\mathbf{q}_t^{e,T} = F_e(\mathbf{I}_t)$, $\mathbf{p}_t^{au,T} = F_{au}(\mathbf{I}_t)$, $\mathbf{v}_t^{va,T} = F_{va}(\mathbf{I}_t)$, where: $\mathbf{q}_t^{e,T} \in \Delta^C$ is EXP distribution, $\mathbf{p}_t^{au,T} \in [0,1]^K$ is AU activation vector, and $\mathbf{v}_t^{va,T} \in \mathbb{R}^2$ is VA prediction. These streams provide the same facial behaviour across three affect spaces. We use them to adapt the atlas mappings between affect spaces.

Randomized source-space training. At each training iteration, we randomly select one affect space as the source space and use the other two as target spaces. For example, if EXP is selected as the source, the atlas is asked to infer the corresponding AU and VA representations. This randomized training strategy prevents the adaptation process from privileging one affect representation and encourages all cross-space mappings to become mutually reliable.

Residual correction of atlas mappings. Although the offline atlas provides a fixed mapping $M_{x\to y}$ from source space x to target space y, its outputs may deviate from the target talking-face domain, since M is estimated from general affect data. We therefore introduce a residual correction:

$$\hat{\mathbf{s}}_t^y = T_{x\to y}(\mathbf{s}_t^{x,T}) = M_{x\to y}(\mathbf{s}_t^{x,T}) + \Delta_{\psi_{x\to y}}\left(M_{x\to y}(\mathbf{s}_t^{x,T})\right). \tag{4}$$

Here, $M_{x\to y}$ is frozen and $\Delta_{\psi_{x\to y}}$ is a lightweight learnable adapter. The adapter is initialized close to zero, so the adapted mapping initially behaves like the offline atlas. During training, it learns only the domain-specific correction needed to align atlas predictions with talking-face teacher outputs.

This design is important: it avoids unconstrained relearning and overfitting to noisy teacher predictions while preserving the psychologically meaningful atlas prior.

Cross-teacher distillation and bidirectional cycle consistency. Adapters are trained using cross-teacher alignment and cycle consistency, ensuring the refined mappings remain both data-consistent and semantically coherent. We compute the cross-teacher alignment loss: $\mathcal{L}_{align}^{x\to y} = \ell_y\left(\hat{\mathbf{s}}_t^y, \mathbf{s}_t^{y,T}\right)$, where ℓ_y depends on the target affect space: categorical cross-entropy (CE) for EXP, binary cross-entropy (BCE) for AUs, Concordance Correlation Coefficient (CCC)-based loss for VA. Cycle-consistency loss is computed as: $\mathcal{L}_{cyc}^{x\leftrightarrow y} = \ell_x\left(\tilde{\mathbf{s}}_t^x, \mathbf{s}_t^{x,T}\right)$ and apply this to all affect-space pairs.

Expression	AUs
Happy	AU6, AU12
Sad	AU1, AU4, AU15, AU17
Anger	AU4, AU5, AU7, AU23
Disgust	AU9, AU10
Fear	AU1, AU2, AU4, AU5, AU20
Surprise	AU1, AU2, AU5, AU26

(a) Relationship between EXP and AUs.

Rule Set	Rule
EXP → AU	$E_{\text{happy}} \rightarrow AU6$
EXP → AU	$E_{\text{happy}} \rightarrow AU12$
EXP → AU	$E_{\text{happy}} \rightarrow (AU6 \wedge AU12)$
AU → EXP	$(AU6 \wedge AU12) \rightarrow E_{\text{happy}}$
EXP → VA	$E_{\text{happy}} \rightarrow PosV$
EXP → VA	$E_{\text{happy}} \rightarrow \neg LowA$
AU → VA	$AU6 \rightarrow PosV$
AU → VA	$AU12 \rightarrow PosV$

(b) Example rules for "Happy" expression.

The overall teacher-guided atlas adaptation loss is: $\mathcal{L}_{adapt} = \sum_{x \neq y} \lambda_{align}^{x \to y} \mathcal{L}_{align}^{x \to y} + \lambda_{cyc} \mathcal{L}_{cyc}$. This encourages the adapted atlas to behave as a coherent affect coordinate system: moving from one representation to another and back should preserve the original affective meaning.

3.3 Neuro-Symbolic Affect Consistency Framework

While AffectAtlas captures statistical relationships across EXP, AU and VA spaces, it remains fundamentally data-driven and may thus inherit inconsistencies from noisy supervision or domain shifts. In particular, independently learned predictors can produce locally plausible but globally inconsistent affect states. For example, a prediction may assign high probability to a *happy* expression while simultaneously producing negative valence or failing to activate key AUs such as AU6 and AU12. Such inconsistencies degrade both realism and interpretability. To address this limitation, we introduce a **neuro-symbolic affect consistency framework** that explicitly enforces structured relationships among affect representations using differentiable logical reasoning.

Unified multi-space reasoning. We model affect jointly across 3 spaces: EXP (e), AU ($\mathbf{au}$), and VA ($\mathbf{va}$), and define cross-space rules that couple these representations: $\mathcal{R} = \mathcal{R}_{e \to au} \cup \mathcal{R}_{au \to e} \cup \mathcal{R}_{e \to va} \cup \mathcal{R}_{au \to va}$. This enables bidirectional reasoning, ensuring high-level semantic predictions (EXP), mid-level structural signals (AUs), and continuous affect (VA) remain mutually consistent.

Differentiable fuzzy logic formulation. Since affect predictions are probabilistic and continuous, hard logical constraints do not fit. We thus use fuzzy logic formulation that allows soft, differentiable reasoning. For $x, y \in [0, 1]$, we define AND ($\wedge$), OR ($\vee$), NOT ($\neg$), and IMPLY ($\rightarrow$) as: $x \wedge y = xy, \quad x \vee y = x + y - xy, \quad \neg x = 1 - x, \quad x \rightarrow y = 1 - x + xy$. This allows each rule to produce a continuous *satisfaction score* rather than a binary outcome.

Rule construction across affect spaces. To construct reliable rules, we first define affect relationships (See Table 1a) based on established facial-behavior knowledge and prior studies (Clark et al., 2020; Tian et al., 2001b; iMotions, n.d.; Zhang et al., 2014a; Hwooi et al., 2022). We then define 69 rules across four groups: EXP → AU, AU → EXP, EXP → VA, and AU → VA. These rules capture relationships across different affect spaces:

(i) EXP → AU These rules enforce that EXPs activate characteristic facial muscle patterns, e.g.: $E_{\text{happy}} \rightarrow (AU6 \wedge AU12), \quad E_{\text{sad}} \rightarrow (AU1 \wedge AU4 \wedge AU15)$.

(ii) AU → EXP These rules provide reverse semantic grounding, e.g.: $(AU6 \wedge AU12) \rightarrow E_{\text{happy}}$.

(iii) EXP → VA These rules link discrete emotions to continuous affect, e.g.: $E_{\text{happy}} \rightarrow \text{PosV}, \quad E_{\text{sad}} \rightarrow \text{NegV} \wedge \text{LowA}$.

(iv) AU → VA These rules encode fine-grained affective cues, e.g.: $AU12 \rightarrow \text{PosV}, \quad AU4 \rightarrow \text{NegV}, \quad AU5 \rightarrow \text{HighA}$. A subset of rules related to "Happy" expression is shown in Table 1b.

Rule satisfaction and logic loss. For each rule r_i, we compute a satisfaction score $s_i \in [0, 1]$. The overall logic loss is : $\mathcal{L}_{logic} = \sum_i w_i(1 - s_i)$, where w_i controls the importance of each rule. This formulation softly penalizes violations while allowing flexibility in ambiguous or uncertain cases.

Integration with atlas learning. A key aspect of our approach is the tight integration of neuro-symbolic module with AffectAtlas. The atlas provides structured probabilistic mappings, while neuro-symbolic layer enforces these mappings consistent with known facial-behaviour constraints. This combination yields a *closed-loop reasoning system*: atlas propagates information across affect spaces; adaptation module aligns predictions with data; neuro-symbolic layer enforces consistency. **Appendix C provides additional details on the affect atlas mapping, teacher-guided atlas adaptation, neuro-symbolic affect consistency framework, and the full set of rules.**



Figure 3: Comparison with SOTA methods on MEAD, RAVDESS, and HDTF datasets. Frame-wise valence (V) and arousal (A) values show how each method captures emotional changes over time.

3.4 AUDIO2AFFECT (STUDENT) MODULE.

During inference, target video and visual affect models are not available. Therefore, we train an Audio2Affect student to predict the same refined affect tokens used during the generator training. As shown in Fig. 1 (c), Wav2Vec2 encodes the driving audio, and its contextual features are passed to Audio2Affect, which consists of a 3-layer MLP with ReLU. For each paired audio-video sample, the frozen teacher provides video-derived affect tokens as supervision. The student is trained with an L_1 reconstruction loss to minimize element-wise differences between audio-predicted and video-derived affect tokens, and a cosine-similarity loss to align their directions in the embedding space. This forms a cross-modal distillation process: the teacher provides structured, cross-space-consistent visual affect representations and the student predicts the corresponding affect tokens from audio.

3.5 EMOTION-AWARE AUDIO-VISUAL GENERATION

The emotion-aware A/V fusion module integrates visual, speech, and affect information (Fig. 1(d)). Given visual, audio, refined affect tokens ($\mathbf{h}^v$, $\mathbf{h}^a$, $\mathbf{h}^e$), we apply multi-modal attention: $\mathrm{MMAttn}(\mathbf{h}^v, \mathbf{h}^a, \mathbf{h}^e) = \mathrm{Softmax}\left(QK^\top/\sqrt{d}\right)V;\ \ Q = \mathbf{h}^v W_Q, \quad K = [\mathbf{h}^v; \mathbf{h}^a; \mathbf{h}^e]W_K, \quad V = [\mathbf{h}^v; \mathbf{h}^a; \mathbf{h}^e]W_V$. This jointly conditions visual features on speech and affect cues. Affect is further injected via emotion-adaptive normalization: $\mathrm{AdaNorm}(\mathbf{h}^v, \mathbf{h}^e) = \gamma(\mathbf{h}^e)\frac{\mathbf{h}^v - \mu(\mathbf{h}^v)}{\sigma(\mathbf{h}^v)} + \beta(\mathbf{h}^e)$, where $\gamma(\cdot)$ and $\beta(\cdot)$ are learned from the affect embedding, enabling emotion-adaptive visual synthesis.

4 EXPERIMENTS

Datasets For the atlas, we use large-scale affect datasets: Aff-Wild (Kollias et al., 2019; Zafeiriou et al., 2017), AffectNet (Mollahosseini et al., 2017), RAF-DB (Liu et al., 2023), EmotionNet (Fabian Benitez-Quiroz et al., 2016), DISFA (Mavadati et al., 2013), BP4D (Zhang et al., 2014b). For the video generation model, we use MEAD (Wang et al., 2020), CREMA-D (Cao et al., 2014), HDTF

Table 2: Quantitative results with SOTA approaches on the MEAD, HDTF, and RAVDESS datasets. Best results are shown in **bold**, and second-best results are underlined.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CSIM ↑	FVD ↓	SyncC ↑	SyncD ↓	EFID ↓	F1 ↑	EmoAcc ↑	CCCV ↑	CCCA ↑
SadTalker (2023)	28.06	0.65	0.46	171.73	0.92	158.41	6.68	8.89	0.92	0.66	46.33	0.48	0.44
EAT (2023)	15.94	0.68	0.46	163.34	0.87	237.94	**7.15**	**7.65**	1.51	0.58	32.65	0.57	0.51
EDTalk (2024)	16.32	0.56	0.49	182.34	0.92	246.13	5.12	9.16	0.78	0.67	56.67	0.56	0.53
MEMO (2024)	28.12	0.64	0.45	153.14	0.92	**94.81**	5.67	9.12	0.58	0.65	58.76	0.60	0.57
Hallo (2024)	28.00	0.63	0.44	161.26	0.92	113.40	6.10	8.69	0.38	0.68	61.58	0.52	0.54
DiceTalk (2025)	27.45	0.61	0.47	153.29	0.89	342.78	6.49	8.84	1.71	0.64	50.85	0.55	0.54
EchoMimic (2025)	28.17	0.69	0.44	153.69	0.93	162.84	5.43	9.25	0.49	0.68	54.24	0.51	0.49
KeyFace (2025)	28.72	0.62	0.45	171.73	0.93	155.89	4.99	9.59	0.56	0.69	57.21	0.55	0.50
Ours	**31.72**	**0.76**	**0.38**	**149.13**	**0.94**	107.26	6.98	7.86	**0.21**	**0.79**	**74.37**	**0.72**	**0.70**
SadTalker (2023)	28.46	0.50	0.46	75.32	0.92	149.89	7.85	7.37	0.18	0.63	61.25	0.39	0.36
EAT (2023)	14.44	0.51	0.54	124.82	0.85	285.81	7.82	7.59	1.25	0.53	41.88	0.40	0.41
EDTalk (2024)	15.26	0.49	0.56	146.71	0.83	301.16	5.67	9.01	0.20	0.67	67.45	0.41	0.41
MEMO (2024)	30.80	0.69	0.21	28.64	0.90	106.77	8.17	7.21	0.15	0.61	68.12	0.39	0.38
Hallo (2024)	30.32	0.69	0.19	29.12	0.92	100.52	7.63	7.71	0.19	0.66	68.75	0.41	0.42
DiceTalk (2025)	30.56	0.69	0.20	30.46	0.91	125.93	**8.57**	7.12	0.25	0.65	61.87	0.44	0.42
EchoMimic (2025)	28.99	0.51	0.40	55.46	0.92	**93.89**	6.13	8.97	0.17	0.64	68.12	0.43	0.43
KeyFace (2025)	30.82	0.71	0.27	29.26	0.91	104.22	7.23	8.06	0.15	0.66	70.62	0.56	0.51
Ours	**32.88**	**0.76**	**0.15**	**23.12**	**0.93**	101.24	8.15	**7.09**	**0.13**	**0.73**	**79.81**	**0.61**	**0.56**
SadTalker (2023)	27.90	0.62	0.37	53.91	0.85	237.20	5.53	7.83	1.45	0.61	45.70	0.63	0.66
EAT (2023)	11.77	0.62	0.44	97.60	0.78	252.92	5.81	7.98	1.94	0.61	49.34	0.66	0.65
EDTalk (2024)	13.46	0.58	0.48	104.30	0.75	276.44	4.19	8.98	1.41	0.65	61.31	0.61	0.63
MEMO (2024)	31.73	0.72	0.19	30.57	0.85	166.28	6.01	7.60	0.45	0.65	64.47	0.65	0.69
Hallo (2024)	31.24	0.71	0.21	35.10	0.85	**143.10**	4.88	8.06	0.89	0.66	59.87	0.55	0.63
DiceTalk (2025)	31.41	0.71	0.21	41.36	0.80	302.10	5.19	8.67	2.11	0.63	45.39	0.62	0.68
EchoMimic (2025)	28.93	0.62	0.40	54.11	0.85	168.55	4.38	8.51	1.20	0.63	51.97	0.59	0.64
KeyFace (2025)	17.13	0.71	0.20	30.57	0.86	217.05	4.29	8.44	1.04	0.66	62.25	0.59	0.59
Ours	**32.86**	**0.77**	**0.18**	**28.35**	**0.87**	174.03	**6.09**	**7.34**	**0.40**	**0.73**	**69.71**	**0.72**	**0.73**

(Zhang et al., 2021), CelebVHQ (Zhu et al., 2022), and in-the-wild data from Hallo3 Cui et al. (2025). We use 150 videos each from MEAD and HDTF for evaluation, and 150 videos each from RAVDESS (Livingstone & Russo, 2018) and VFHQ (Xie et al., 2022) for cross-dataset evaluation.

Evaluation metrics We evaluate image quality with PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), FID, and FVD; lip-sync using Sync-C and Sync-D from SyncNet (Prajwal et al., 2020). For identity preservation, we extract identity features using ArcFace (Deng et al., 2019a) and compute cosine similarity (CSIM). For AU preservation, we compute Macro-F1. For emotion, we compute 3 metrics between generated and ground truth videos: (i) EXP-FID (E-FID) on EXP parameters extracted by (Deng et al., 2019b); (ii) EmoAcc (Wang & Fu, 2025) on predicted emotions by DeepFace (Serengil, 2020); (iii) CCC on VA extracted by EmoNet (Toisoul et al., 2021).

Qualitative Results. Fig. 3 presents qualitative comparisons with SOTA on MEAD, RAVDESS, HDTF. As shown in the figure, SadTalker, EDTalk and EchoMimic exhibit artifacts around lip regions. EAT produces low-quality results with inconsistent motion. DiceTalk generates exaggerated expressions and blurry lips in some cases. MEMO preserves emotional cues but suffers from artifacts such as missing teeth and unrealistic eye gaze. In contrast, AffectAtlas generates more expressive and natural results. This is also evident on Appendix E, which provides additional results on VFHQ dataset (Fig. 5) and out-of-distribution images (Fig. 6).

Quantitative Comparison. We compare AffectAtlas across MEAD, HDTF, and RAVDESS against 8 SOTA methods: 5 emotion-aware methods (EAT (Gan et al., 2023), EDTalk (Tan et al., 2024a), MEMO (Zheng et al., 2024), DiceTalk (Tan et al., 2025), KeyFace (Bigata et al., 2025)) and 3 widely used methods (SadTalker (Zhang et al., 2023), Hallo (Xu et al., 2024), EchoMimic (Chen et al., 2025)). Table 2 shows our method performing best on most metrics, including image quality (PSNR, SSIM, LPIPS, FID), identity preservation (CSIM), and emotion-related metrics (E-FID, AU F1, EmoAcc, VA CCC). Additionally, to evaluate cross-dataset robustness, we conduct experiments on the VFHQ dataset. As shown in Table 12 of Appendix E, AffectAtlas demonstrates strong generalization to unseen data while preserving visual quality, identity, synchronization, and emotional fidelity.

User Studies. First, we perform a user study comparing AffectAtlas to 8 SOTA in terms of: (1) emotional consistency with speech (Emo); (2) realism of emotional expressions over time (Realism); (3) lip synchronization (Sync); (4) overall video quality (Vid-Q). We randomly select 45 clips (4-7 secs each), from MEAD and RAVDESS datasets, covering 5 emotion classes. For each clip, participants view videos from all 9 methods side-by-side in random order to reduce positional and ordering bias, with method identities hidden. All 40 participants rate the same set of clips. Each

video is rated from 1 to 5 on the four aspects. Scores are averaged over all participants and clips. As shown in Table 6, our method outperforms all SOTA across all aspects. To assess whether the observed differences in user ratings are statistically reliable, we also perform statistical significance analysis. We show in Appendix E.3 that AffectAtlas significantly outperforms all SOTA in emotion, realism and overall video quality; for lip synchronization, AffectAtlas significantly outperforms all SOTA except MEMO, for which the difference is not statistically significant. Finally, we conduct an additional user study comparing AffectAtlas to 8 SOTA in terms of temporal realism. As shown in Appendix (Table 13), AffectAtlas achieves the highest score across all three temporal criteria, with statistically significant improvements over the SOTA.

Ablation Studies. We conduct ablation studies on the RAVDESS dataset. First, we analyze the role of affect space configurations, including single-space (EXP, AU, VA) and multi-space combinations (EXP+AU, EXP+VA, AU+VA, and EXP+AU+VA). For each configuration, AffectAtlas is trained using only the selected signals. Both qualitative (Fig. 8) and quantitative (Table 5) results show that the full EXP+AU+VA configuration achieves the best results, demonstrating that the 3 representations are complementary. Next, we evaluate whether the proposed AffectAtlas effectively learns meaningful affect cues by comparing 4 settings: (1) w/o affect atlas (all affect cues are removed); (2) raw affect streams (EXP, AU, VA are directly concatenated without mapping); (3) only affect atlas; (4) whole AffectAtlas. Both qualitative (Fig. 8) and quantitative (Table 4) results show that without affect atlas, the model produces weak emotional expressions; using raw affect streams or only affect atlas, the method struggles to represent some emotions well; the full method generates more expressive and accurate emotions. This shows that affect atlas is important for learning better emotion representations. Next, to test whether the structured affect atlas is necessary to preserve emotion, we compare it with simple learned fusion of EXP, AU, and VA. We implement MLP Fusion which concatenates EXP, AU, VA predictions from teacher models and projects them in a shared embedding using 2-layer MLP. We also implement a stronger Transformer Fusion which learns interactions among the 3 affect signals via self-attention. We also evaluate direct audio-to-VA and audio-to-EXP conditioning, where affect is predicted directly from audio and provided to video generator without affect atlas. Both qualitative (Fig. 7) and quantitative (Table 3) results show that these alternatives exhibit less stable eye behavior, weaker expressions and more limited facial action dynamics than AffectAtlas. Appendix E.5 provides ablations of fusion and affect-conditioning mechanisms, of the neuro-symbolic consistency framework and of the emotion-aware audio-visual module. Appendix E.4 provides response delay comparison results, where our method achieves the best results.

Table 3: Ablation on mapping.

Modality	EFID ↓	EmoAcc ↑	CCCV ↑	CCCA ↑	F1 ↑
MLP Fusion	0.75	59.68	0.63	0.61	0.61
Transf Fusion	0.61	62.43	0.62	0.64	0.63
Audio-to-VA	0.88	55.14	0.58	0.60	0.59
Audio-to-EXP	0.80	57.36	0.61	0.63	0.59
Ours	**0.40**	**69.71**	**0.72**	**0.73**	**0.73**

Table 4: Ablation on configurations.

Modality	EFID ↓	EmoAcc ↑	CCCV ↑	CCCA ↑	F1 ↑
w/o affect atlas	0.93	46.58	0.58	0.56	0.53
concat	0.67	61.79	0.65	0.68	0.65
only affect atlas	0.51	64.82	0.69	0.68	0.68
Ours (Whole)	**0.40**	**69.71**	**0.72**	**0.73**	**0.73**

Table 5: Ablation on affect spaces.

Modality	EFID ↓	EmoAcc ↑	CCCV ↑	CCCA ↑	F1 ↑
VA	0.92	58.82	0.63	0.64	0.61
EXP	0.98	59.47	0.65	0.67	0.63
AU	0.67	60.13	0.69	0.68	0.65
AU + VA	0.62	61.84	0.68	0.66	0.65
EXP + AU	0.50	63.72	0.70	0.68	0.67
VA + EXP	0.57	61.93	0.66	0.69	0.66
Ours	**0.40**	**69.71**	**0.72**	**0.73**	**0.73**

Table 6: User Study Results.

Method	Emo ↑	Realism ↑	Sync ↑	Vid-Q ↑
SadTalker	3.00	3.03	3.71	3.29
EAT	2.74	2.84	3.97	3.16
EDTalk	3.15	2.47	3.11	2.51
MEMO	3.63	3.11	4.34	3.58
Hallo	2.13	1.89	2.92	2.45
DiceTalk	3.18	2.55	3.82	3.21
EchoMimic	2.24	2.50	3.16	3.05
KeyFace	3.45	3.45	4.16	3.66
Ours	**4.47**	**4.03**	**4.39**	**4.29**

5 Conclusion

We propose a structured neuro-symbolic framework for emotionally consistent talking-face generation. Extensive experiments demonstrate improvements over SOTA methods in emotion fidelity, visual quality, and lip synchronization.

AI USE STATEMENT

In this work, we use generative AI tools to assist with language editing and to improve the clarity and readability of the manuscript. We have not used generative AI tools for research methodology, experiments, or analysis; these are developed by the authors. We have reviewed all AI-assisted text for accuracy and consistency. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work aims to improve emotionally consistent talking face generation for research and beneficial applications such as human-computer interaction, digital media, and virtual avatars. We do not intend the method for misleading or deceptive use. We encourage users to apply it responsibly and with careful consideration of ethical implications.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, training and inference procedures, and implementation details in the main paper and appendix to support reproducibility. We will release the source code and model weights upon acceptance.

REFERENCES

Tadas Baltrušaitis, Peter Robinson, Louis-Philippe Morency, et al. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, volume 110. Lake Placid, NY, 2016.

Antoni Bigata, Michał Stypułkowski, Rodrigo Mira, Stella Bounareli, Konstantinos Vougioukas, Zoe Landgraf, Nikita Drobyshev, Maciej Zieba, Stavros Petridis, and Maja Pantic. Keyface: Expressive audio-driven facial animation for long sequences via keyframe interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5477–5488, 2025.

Breakthrough. Pyscenedetect: Scene cut detection & video splitting tool. `https://github.com/Breakthrough/PySceneDetect`, 2026. Accessed: 2026-04-13.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 520–535, 2018.

Liyang Chen, Zhiyong Wu, Runnan Li, Weihong Bao, Jun Ling, Xu Tan, and Sheng Zhao. Vast: Vivify your talking avatar via zero-shot expressive facial style transfer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2977–2987, 2023.

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.

Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022.

Elizabeth A Clark, J'Nai Kessinger, Susan E Duncan, Martha Ann Bell, Jacob Lahne, Daniel L Gallagher, and Sean F O'Keefe. The facial action coding system for characterization of human affective response to consumer product-based stimuli: a systematic review. *Frontiers in psychology*, 11:920, 2020.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21086–21095, 2025.

Alessandro Daniele, Tommaso Campari, Sagar Malhotra, and Luciano Serafini. Deep symbolic learning: Discovering symbols and rules from perceptions. *arXiv preprint arXiv:2208.11561*, 2022.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019a.

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*, pp. 0–0, 2019b.

Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

C Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5562–5570, 2016.

Guanwen Feng, Haoran Cheng, Yunan Li, Zhiyuan Ma, Chaoneng Li, Zhihao Qian, Qiguang Miao, and Chi-Man Pun. Emospeaker: One-shot fine-grained emotion-controlled talking face generation. *IEEE Transactions on Multimedia*, 2026.

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.

Fa-Ting Hong, Yunfei Liu, Yu Li, Changyin Zhou, Fei Yu, and Dan Xu. Dreamhead: Learning spatial-temporal correspondence via hierarchical diffusion for audio-driven talking head synthesis. *arXiv preprint arXiv:2409.10281*, 2024.

Stephen Khor Wen Hwooi, Alice Othmani, and Aznul Qalid Md Sabri. Deep learning-based approach for continuous affect prediction from facial expression images in valence-arousal space. *IEEE Access*, 10:96053–96065, 2022. doi: 10.1109/ACCESS.2022.3205018.

iMotions. Facial action coding system (facs). `https://imotions.com/blog/learning/research-fundamentals/facial-action-coding-system/`, n.d. Accessed: 2026-04-06.

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.

Dimitrios Kollias, Panagiotis Tzirakis, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, Björn Schuller, Irene Kotsia, and Stefanos Zafeiriou. Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *International Journal of Computer Vision*, 127(6):907–929, 2019.

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. doi: 10.1109/CVPR52688.2022.00338.

Chuanhe Liu, Xinjie Zhang, Xiaolong Liu, Tenggan Zhang, Liyu Meng, Yuchen Liu, Yuanyuan Deng, and Wenqiang Jiang. Facial expression recognition based on multi-modal features for videos in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5872–5879, 2023.

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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26222–26231, 2025a.

Tao Liu, Ziyang Ma, Qi Chen, Feilong Chen, Shuai Fan, Xie Chen, and Kai Yu. Vqtalker: Towards multilingual talking avatars through facial motion tokenization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5586–5594, 2025b. doi: 10.1609/aaai.v39i6.32595.

Steven R Livingstone and Frank A 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.

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, pp. 1896–1904, 2023.

Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, Aibin Huang, and Yigang Wang. Poster++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, 157:110951, 2025.

S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE transactions on affective computing*, 10(1):18–31, 2017.

Ye Pan, Shuai Tan, Shengran Cheng, Qunfen Lin, Zijiao Zeng, and Kenny Mitchell. Expressive talking avatars. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2538–2548, 2024.

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.

Zipeng Qi, Xulong Zhang, Ning Cheng, Jing Xiao, and Jianzong Wang. Difftalker: Co-driven audio-image diffusion for talking faces via intermediate landmarks. *arXiv preprint arXiv:2309.07509*, 2023.

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*, pp. 10684–10695, 2022.

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

Sefik Serengil. Deepface: A lightweight face recognition framework for python, February 2020. URL `https://github.com/serengil/deepface`.

Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European conference on computer vision*, pp. 666–682. Springer, 2022.

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*, pp. 1982–1991, 2023.

Zhaoxu Sun, Yuze Xuan, Fang Liu, and Yang Xiang. Fg-emotalk: talking head video generation with fine-grained controllable facial expressions. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024a. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i5.28309. URL https://doi.org/10.1609/aaai.v38i5.28309.

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, pp. 5043–5051, 2024b.

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.

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. 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*, pp. 26317–26327, 2024b.

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, 2024c.

Weipeng Tan, Chuming Lin, Chengming Xu, FeiFan Xu, Xiaobin Hu, Xiaozhong Ji, Junwei Zhu, Chengjie Wang, and Yanwei Fu. Disentangle identity, cooperate emotion: Correlation-aware emotional talking portrait generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9987–9995, 2025.

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*, pp. 716–731. Springer, 2020.

Y.-I. Tian, T. Kanade, and J.F. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001a. doi: 10.1109/34.908962.

Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001b.

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. doi: 10.1038/s42256-020-00280-0.

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

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, 2025.

Jiayu Wang, Kang Zhao, Shiwei Zhang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Lipformer: High-fidelity and generalizable talking face generation with a pre-learned facial codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13844–13853, 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 *European conference on computer vision*, pp. 700–717. Springer, 2020.

Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2head: Audio-driven one-shot talking-head generation with natural head motion. *arXiv preprint arXiv:2107.09293*, 2021.

Wenqing Wang and Yun Fu. Realtalk: Realistic emotion-aware lifelike talking-head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1481–1490, 2025. doi: 10.1109/ICCVW69036.2025.00159.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 657–666, 2022.

Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, Chengjie Wang, Zhifeng Xie, and Yong Liu. High-fidelity generalized emotional talking face generation with multi-modal emotion space learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6609–6619, 2023.

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

Phyo Thet Yee, Dimitrios Kollias, Sudeepta Mishra, and Abhinav Dhall. Synchrorama: Lip-synchronized and emotion-aware talking face generation via multi-modal emotion embedding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4546–4555, 2026. doi: 10.1109/WACV61042.2026.00442.

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*, pp. 85–101. Springer, 2022.

Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal'in-the-wild'challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 34–41, 2017.

Ligang Zhang, Dian Tjondronegoro, and Vinod Chandran. Representation of facial expression categories in continuous arousal–valence space: feature and correlation. *Image and Vision Computing*, 32(12):1067–1079, 2014a.

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, 2023.

Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014b.

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, 2024. URL https://arxiv.org/abs/2412.04448.

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.

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.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. Celebv-hq: A large-scale video facial attributes dataset. In *European conference on computer vision*, pp. 650–667. Springer, 2022.

A Preliminaries to Expression, AU, and Valence-Arousal

Facial Action Unit (AU) is a simple and interpretable way to describe facial expressions based on facial muscle movements. AUs are defined based on the Facial Action Coding System (Ekman & Friesen, 1978), where each unit represents the activation of specific facial muscles (e.g., AU6 for cheek raiser, AU12 for lip corner puller). Unlike categorical emotion labels (e.g., happy or sad), AUs describe expressions in a more detailed way by showing which muscles are active. Multiple AUs can appear together to form different expressions. In contrast, valence-arousal (VA) is a continuous representation used to describe human emotion. Valence represents how positive or negative an emotion is, while arousal represents the intensity or activation level of that emotion. For example, happy emotion has high valence and medium-to-high arousal, while sad emotion has low valence and low arousal. Figure 4 illustrates examples of facial action units and the valence-arousal space.

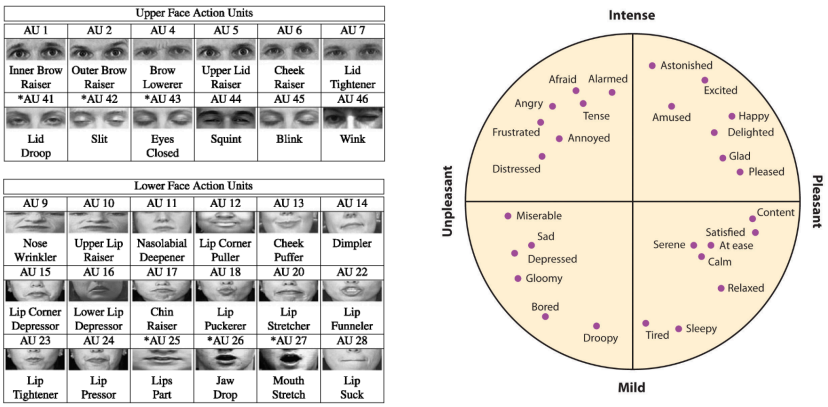


Figure 4: Examples of Facial Action Units and Valence-Arousal space, reproduced from Tian et al. (2001a) and Russell (1980), respectively.

B Additional Related Work

Affect representations. Facial affect can be represented in multiple complementary forms. Categorical expression labels provide high-level semantic descriptions such as happiness, sadness, anger, and surprise. Facial action units, defined by the Facial Action Coding System (FACS) (Ekman & Friesen, 1978), describe local facial muscle activations and provide a more fine-grained representation of facial behaviour. Continuous valence-arousal models describe affective states in terms of emotional polarity and intensity, and have been widely used in affective computing benchmarks such as Aff-Wild and Aff-Wild2. Although these representations are strongly related, most generative talking-face models use only one of them. This can lead to incomplete or inconsistent affect

modeling. Our work addresses this limitation by constructing an AffectAtlas: a unified probabilistic prior that explicitly models the relationships among expressions, AUs, and VA.

Structured and consistent affect modeling. Multi-task affect analysis has shown that expressions, AUs, and VA provide complementary information and can benefit from joint modeling. However, in talking-face generation, these affect spaces are usually treated as independent supervision signals or simple conditioning features. Few works explicitly model their cross-space relationships. This is important because affective consistency imposes natural constraints: for example, a happy expression should be compatible with positive valence and the activation of AUs such as AU6 and AU12. Our proposed Affect Atlas differs from prior emotion-conditioning approaches by explicitly learning probabilistic mappings between affect spaces and refining them through teacher-guided adaptation and cycle consistency. This allows the model to preserve semantic agreement across representations while adapting to the target talking-face domain.

Neuro-symbolic learning for visual generation. Neuro-symbolic learning combines neural networks with symbolic knowledge, enabling models to benefit from both data-driven learning and explicit rules (Daniele et al., 2022). Such methods are useful when domain knowledge can be expressed as logical constraints, especially in settings where purely learned correlations may produce inconsistent predictions. In facial affect modeling, there exist well-established relationships between expressions, AUs, and VA, derived from psychological and facial-behaviour studies (Ekman & Friesen, 1978; Clark et al., 2020). However, these relationships have rarely been incorporated into talking-face generation. We introduce a differentiable neuro-symbolic consistency framework that encodes facial-behaviour rules as soft logical constraints. This encourages the generated affect representation to be not only visually plausible but also psychologically meaningful and internally consistent.

We clarify that the absence of other methods from our experimental tables was due to either unavailable executable releases (code, pretrained models) or differences in the supported inference protocol. For MoEE and EmotiveTalk, no usable official implementation or pretrained checkpoints are publicly available. MoEE has no official code repository and the code link on the EmotiveTalk project page currently points to a 'page not found' website. AUHead has an official repository, but it currently contains only a README. RealTalk does not provide any pretrained weights. EMOPortraits' repository supports only video-driven pipeline (no released audio-driven mode) so it cannot be fairly compared to our reference image plus audio setting. These approaches primarily differ from AffectAtlas in where they introduce emotion modeling. EDTalk disentangles facial motion into separate mouth, pose and expression latent spaces; MoEE uses emotion-specific experts and an emotion-to-latent module for basic and compound emotion control; EmotiveTalk separates lip- and expression-related audio representations and introduces an emotion-conditioned video-diffusion architecture; RealTalk deforms audio-derived facial landmarks using categorical emotion embeddings before NeRF rendering; AUHead predicts AU sequences from speech and uses them as fine-grained diffusion controls; EMOPortraits improves the transfer of intense and asymmetric expressions through its avatar architecture and training procedure.

In contrast, AffectAtlas does not primarily propose another emotion-specific rendering architecture or treat affect as a single control label, AU sequence, or unconstrained latent embedding. Its main contribution is a structured affect representation that explicitly connects three complementary descriptions of facial affect: categorical expressions provide semantic emotion identity, AUs describe local facial-muscle activations and valence–arousal represents continuous affective polarity and intensity. AffectAtlas models bidirectional probabilistic mappings among these spaces, including coordinated AU-pair evidence and adapts the mappings to the talking-face domain through lightweight residual corrections while keeping the underlying structured prior fixed. Cycle-consistency and neuro-symbolic constraints further penalize contradictory predictions across the three spaces. The resulting cross-space-consistent representation is then used to condition the generator. Thus, the contribution is complementary to prior advances in motion disentanglement, expert routing, audio-feature separation, landmark deformation, AU control and rendering: it addresses the missing structured relationship among heterogeneous affect representations before they are injected into the generation model.

Table 7: Summary of the main notation

Notation	Description
e, au, va	expression, facial action unit, valence-arousal, respectively
$\mathcal{M}$	Probabilistic Affect Atlas containing cross-space affect mappings
$M_{x\to y}$	Offline atlas mapping from affect space x to affect space y
$p(\mathbf{au} \mid e), p(e \mid \mathbf{au})$	EXP-AU bidirectional mappings
$p(\mathbf{va} \mid e), p(e \mid \mathbf{va})$	EXP-VA bidirectional mappings
$p(\mathbf{va} \mid \mathbf{au}), p(\mathbf{au} \mid \mathbf{va})$	AU-VA bidirectional mappings
K_e	Number of GMM components for expression class e
$\pi_{e,m}$	Mixture weight of the m-th GMM component for expression e
$\boldsymbol{\mu}_{e,m}$	Mean vector of the m-th GMM component for expression e
$\boldsymbol{\Sigma}_{e,m}$	Covariance matrix of the m-th GMM component for expression e
$\theta_k^{(e)}$	Expression-specific bias controlling the activation tendency of AU k
$J_{ij}^{(e)}$	Interaction parameter between AUs i and j under expression e
Z_e	Partition function of the expression-conditioned MRF
$n_{e,k}$	Number of samples with expression e in which AU k is active
n_e	Total number of samples belonging to expression e
α, β	Beta-Bernoulli smoothing parameters
$\mathcal{D}$	Dictionary of retained informative AU pairs
$\mathcal{P}(\mathbf{au})$	Set of active AU pairs in $\mathbf{au}$ that belong to $\mathcal{D}$
$\mathrm{PMI}(i,j)$	Pointwise mutual information between AUs i and j
$\mathrm{cPMI}_e(i,j)$	Expression-specific contrastive PMI score for AU pair (i,j)
$F_e,\ F_{\mathrm{au}},\ F_{\mathrm{va}}$	Frozen EXP, AU, and VA teacher models
$s_t^{x,T}$	Teacher prediction in source affect space x
$s_t^{y,T}$	Teacher prediction in target affect space y
$\Delta\psi_{x\to y}$	Learnable residual adapter for mapping $x \to y$
$\hat{s}_t^y$	Corrected prediction in target affect space y
$\tilde{s}_t^x$	Reconstructed source-space representation after bidirectional mapping
$\mathcal{L}_{\mathrm{align}}^{x\to y}$	Cross-teacher alignment loss for mapping $x \to y$
$\mathcal{L}_{\mathrm{cyc}}^{x\leftrightarrow y}$	Cycle-consistency loss for the affect-space pair (x,y)
$\mathcal{L}_{\mathrm{cyc}}$	Total cycle-consistency loss over all affect-space pairs
$\mathcal{L}_{\mathrm{adapt}}$	Overall teacher-guided atlas adaptation loss
$\mathcal{R}$	Complete set of neuro-symbolic affect consistency rules
r_i	The i-th neuro-symbolic rule
s_i	Fuzzy satisfaction score of rule r_i
$\mathcal{L}_{\mathrm{logic}}$	Neuro-symbolic logic-consistency loss
h^v, h^a, h^e	Visual, audio, and refined affect tokens, respectively

C Additional Method Details

The main notation used throughout the method is summarized in Table 7.

C.1 Offline AffectAtlas: A structured affect prior

C.1.1 EXP-to-AU mapping.

Expressions are typically realized through coordinated patterns of facial muscle activations. Modeling only marginal AU probabilities would therefore ignore expression-specific co-activation struc-

ture that characterizes realistic facial behaviour. We thus seek a structured distribution that captures both which AUs are likely under expression e and how they interact with one another.

Joint model For each expression, we define a pairwise binary Markov Random Field (MRF) over au:

$$p(\mathbf{au} \mid e) = \frac{1}{Z^{(e)}} \exp\left(\sum_{k=1}^{17} \theta_k^{(e)} au_k + \sum_{1 \le i < j \le 17} J_{ij}^{(e)} au_i au_j \right), \tag{5}$$

where $Z^{(e)}$ is the partition function:

$$Z^{(e)} = \sum_{\mathbf{au}' \in \{0,1\}^{17}} \exp\left(\sum_{k=1}^{17} \theta_k^{(e)} au'_k + \sum_{1 \le i < j \le 17} J_{ij}^{(e)} au'_i au'_j \right). \tag{6}$$

This formulation assigns a probability to every possible AU vector under expression e.

The model is parameterized by two complementary components. The bias terms $\theta_k^{(e)}$ capture the tendency of AU k to be active under expression e, while the interaction terms $J_{ij}^{(e)}$ capture whether pairs of AUs tend to co-activate or inhibit one another. Together, these parameters encode both first-order AU-expression associations and higher-order dependency structure.

As a robust first-order statistic, we also estimate the marginal probability of each AU under expression e using Beta–Bernoulli smoothing:

$$p(au_k = 1 \mid e) = \frac{n_{ek} + \alpha}{n_e + \alpha + \beta}, \tag{7}$$

where: n_{ek} is the number of samples of expression e in which AU k is active; n_e is the total number of samples of expression e; $\alpha, \beta > 0$ are smoothing parameters. These smoothed marginals provide stable per-AU statistics and are used to initialize the AU bias structure. However, they do not determine the full joint distribution $p(\mathbf{au} \mid e)$, as they do not capture dependencies among AUs.

Parameter estimation via pseudo-likelihood Direct maximum-likelihood estimation of the joint model is generally inconvenient because it requires repeated computation of $Z^{(e)}$ during optimization. We thus estimate the model parameters using pseudo-likelihood, which decomposes the joint estimation problem into node-wise conditional models:

$$\mathcal{L}_{\mathrm{PL}}^{(e)} = \sum_{n \in \mathcal{D}_e} \sum_{k=1}^{17} \log p\Big(au_k^{(n)} \mid \mathbf{au}_{-k}^{(n)}, e \Big), \tag{8}$$

where: $\mathcal{D}_e$ denotes the subset of training samples with expression e; $\mathbf{au}_{-k}$ denotes all AUs except au_k. Under the pairwise MRF, each node-wise conditional takes the auto-logistic MRF form:

$$p(au_k = 1 \mid \mathbf{au}_{-k}, e) = \sigma\left(\theta_k^{(e)} + \sum_{j \ne k} J_{kj}^{(e)} au_j \right), \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function. In practice, this corresponds to fitting 17 logistic regressions per expression, one for each AU, using the remaining AUs as predictors. The bias term $\theta_k^{(e)}$ captures the intrinsic tendency of AU k to activate under expression e, while the interaction terms modulate this tendency according to the state of the other AUs. To promote sparsity and improve interpretability, we apply an ℓ_1 penalty to the interaction parameters.

As the node-wise conditional models are estimated independently, the learned pairwise interaction parameters are not guaranteed to be symmetric ($J_{ij}^{(e)} \ne J_{ji}^{(e)}$). However, a valid pairwise MRF requires symmetric interactions. We thus enforce consistency by symmetrizing the learned parameters:

$$J_{ij}^{(e)} \leftarrow \frac{1}{2}\left(J_{ij}^{(e)} + J_{ji}^{(e)} \right). \tag{10}$$

C.1.2 AU-TO-VA MAPPING.

Directly estimating this conditional for every possible AU vector is impractical, since many full AU configurations are rare and would not support reliable density estimation. We thus approximate $p(\mathbf{va} \mid \mathbf{au})$ by combining lower-order affect experts that capture complementary evidence about VA at different levels of granularity.

Specifically, we use: (i) $p(\mathbf{va} \mid e)$ (from eq. 3.1), which provides a coarse semantic prior; (ii) single-AU VA experts $p(\mathbf{va} \mid au_k = 1)$, which capture the affective tendency of individual facial actions; and (iii) selected AU-pair VA experts $p(\mathbf{va} \mid au_i = 1, au_j = 1)$, which capture higher-order compositional facial patterns. The full AU-to-VA mapping is then written as:

$$p(\mathbf{va} \mid \mathbf{au}) = \sum_e p(\mathbf{va} \mid e, \mathbf{au})\, p(e \mid \mathbf{au}) \tag{11}$$

where: $p(\mathbf{va} \mid e, \mathbf{au})$ is a refined VA distribution obtained by fusing the relevant lower-order experts for the observed AU pattern. We define $p(\mathbf{va} \mid e, \mathbf{au})$ using a product-of-experts (PoE) formulation that combines the expression expert, the active single-AU experts and the active AU-pair experts:

$$p(\mathbf{va} \mid e, \mathbf{au}) \propto p(\mathbf{va} \mid e) \prod_{k:au_k=1} p(\mathbf{va} \mid au_k = 1) \prod_{(i,j):au_i=au_j=1} p(\mathbf{va} \mid au_i = 1, au_j = 1) \tag{12}$$

Thus, the lower-order AU-conditioned VA models are not themselves the final target, but rather act as reusable experts whose evidence is fused to approximate the VA distribution of the full AU vector.

Single-AU VA experts Individual AUs encode localized facial movements that often correlate with valence or arousal. To capture these fine-grained cues, for each AU k we fit a GMM:

$$p(\mathbf{va} \mid au_k = 1) = \sum_{m=1}^{K_k} \pi_{k,m}\, \mathcal{N}(\mathbf{va}; \boldsymbol{\mu}_{k,m}, \boldsymbol{\Sigma}_{k,m}). \tag{13}$$

These single-AU experts provide local affect evidence that complements the broader expression-level semantics.

Selected AU-pair VA experts While single AUs provide useful local cues, affective meaning is often expressed through coordinated facial muscle activations rather than isolated actions. AU pairs can therefore capture richer and more specific compositional patterns that are not recoverable from single AUs alone. We accordingly augment the atlas with VA experts conditioned on informative AU pairs.

However, not all AU pairs are equally useful. A naive strategy that models all possible pairs is undesirable because: (i) many pairs are too rare to support reliable density estimation in VA space; (ii) some pairs co-occur frequently simply because their constituent AUs are individually common, rather than because they form a meaningful affective unit; and (iii) many top-ranked pairs are redundant variants of the same broader facial bundle. We therefore construct a compact AU-pair dictionary through a multi-stage selection procedure.

Let $\mathbf{au}^{(n)} \in \{0,1\}^{17}$ denote the AU vector of sample n, and let N be the total number of samples. First, for each AU pair (i, j) with $i < j$, we compute its co-occurrence count (i.e., the number of samples in which both AUs are active; we do this because meaningful AU-pair modeling requires observing the pair sufficiently often in the data):

$$c(i,j) = \sum_{n=1}^{N} \mathbf{1}[au_i^{(n)} = 1 \wedge au_j^{(n)} = 1], \tag{14}$$

and its empirical support (i.e., how frequently the pair appears in the dataset): $\mathrm{supp}(i,j) = \frac{c(i,j)}{N}$. Pairs with support below a minimum threshold are discarded, since very rare pairs do not provide stable estimates. However, support alone is insufficient, because a frequent pair may simply reflect two common AUs. To measure association beyond chance, we compute pointwise mutual information (PMI); we use it because we want the atlas to encode meaningful co-activation structure rather than merely frequent individual AUs):

$$\mathrm{PMI}(i,j) = \log \frac{p(au_i = 1, au_j = 1)}{p(au_i = 1)\, p(au_j = 1)} \tag{15}$$

where the probabilities are estimated empirically:

$$p(au_i = 1, au_j = 1) = \frac{c(i,j)}{N};\ p(au_i = 1) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}[au_i^{(n)} = 1];\ p(au_j = 1) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}[au_j^{(n)} = 1]. \tag{16}$$

To balance frequency and association strength, we define a global score:

$$S_{\text{global}}(i,j) = \text{supp}(i,j) \cdot \text{PMI}(i,j) \tag{17}$$

and rank all candidate pairs and retain the top-M globally informative pairs. This preserves AU pairs that capture stable affective co-activation structure across the full dataset. However, global ranking alone is biased toward dominant expressions (e.g. Happy or Neutral) and may suppress characteristic pairs from rarer expressions (e.g. Fear or Disgust). To ensure semantic diversity, we additionally perform expression-specific selection. For each expression subset $\mathcal{D}_e$, we compute its support ($\text{supp}_e(i,j)$) and expression-specific PMI ($\text{PMI}_e(i,j)$). To identify pairs that are specifically informative for a given expression, we define the expression-contrastive PMI which compares how strongly the pair is associated within expression e relative to all competing expressions:

$$\text{cPMI}_e(i,j) = \text{PMI}_e(i,j) - \max_{e' \neq e} \text{PMI}_{e'}(i,j) \tag{18}$$

We retain a pair for expression e only if $\text{cPMI}_e(i,j) > 0$, (i.e., the pair is more strongly associated with expression e than with any other expression). This step reduces the selection of misleading pairs that are common but not expression-specific. We then rank the remaining candidates using:

$$S_e(i,j) = \text{supp}_e(i,j) \cdot \text{cPMI}_e(i,j). \tag{19}$$

The top-N pairs per expression are retained. In this way, the final AU-pair dictionary contains both globally informative pairs and expression-discriminative pairs that might otherwise be hidden by class imbalance.

Even after global and expression-specific selection, many retained pairs remain redundant. For example, several pairs may belong to the same broader facial bundle and therefore contribute largely overlapping information. Keeping all such pairs would unnecessarily increase atlas complexity and may lead to over-representation of a small number of facial patterns. To address this, we build an undirected AU graph in which each AU is a node and the edge weight between nodes i and j is given by the pair association strength. We then apply Louvain community detection to partition the graph into communities of tightly connected AUs. Within each community, we retain only the highest-scoring pair(s), thereby preserving representative pairs from different co-activation clusters while removing near-duplicate variants. This graph-based pruning makes the AU-pair component more compact, diverse and interpretable. Finally, for each retained AU pair (i,j), we fit a GMM:

$$p(\mathbf{va} \mid au_i = 1, au_j = 1) = \sum_{m=1}^{K_{ij}} \pi_{ij,m}\, \mathcal{N}(\mathbf{va}; \boldsymbol{\mu}_{ij,m}, \boldsymbol{\Sigma}_{ij,m}). \tag{20}$$

The number of components K_{ij} is selected using the same model-selection strategy used for the expression- and AU-conditioned VA models.

Overall, the AU-pair pipeline is designed to ensure that the selected pairs are: (i) sufficiently frequent to support reliable estimation, (ii) statistically associated beyond chance, (iii) both globally informative and expression-discriminative, and (iv) non-redundant.

C.2 Teacher-Guided Atlas Adaptation

Given a training video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, we apply 3 pretrained frozen teacher models frame by frame: (i) a **VA teacher**, producing $\mathbf{va}_t^T \in \mathbb{R}^2$; (ii) an **Expression teacher**, producing $\mathbf{q}_t^{T,e} \in \Delta^7$, and (iii) an **AU teacher**, producing $\mathbf{p}_t^{T,au} \in [0,1]^{17}$. Thus, each video is converted into three synchronized affect streams, one in each representation space. At each training iteration, we randomly choose one of the three spaces as the *source* space. The corresponding teacher output is used as input to the atlas, while the teacher outputs in the remaining spaces act as supervision targets. This randomized source selection prevents the training procedure from privileging one affect representation and encourages all cross-space mappings to be learned under a common framework.

C.2.1 RESIDUAL CORRECTION OF ATLAS MAPPINGS

The offline atlas already captures meaningful cross-space structure. Re-learning all mappings directly from noisy teacher outputs would discard this prior and may lead to overfitting. We therefore keep the atlas mappings frozen and learn only small residual corrections on top of them. This ensures that the refined mappings remain anchored to the atlas while adapting to the empirical distributions induced by the teachers on real videos.

Let $x, y \in \{\mathrm{EXP}, \mathrm{AU}, \mathrm{VA}\}$ with $x \neq y$. For each directed mapping $x \to y$, the offline atlas provides a fixed base mapping: $M_{x\to y}$. Given a source state s_t^x, the corrected mapping is defined as:

$$T_{x\to y}(s_t^x) = M_{x\to y}(s_t^x) + \Delta_{\psi_{x\to y}}(M_{x\to y}(s_t^x)), \tag{21}$$

where $\Delta_{\psi_{x\to y}}$ is a lightweight residual adapter. Each adapter is initialized to 0, so that at the start of training the corrected mapping coincides exactly with the offline atlas. The correction is applied in the natural parameter space of the target variable: (i) **Expression/AU targets** (in logit space followed by softmax/sigmoid); and (ii) **VA targets** (correction of the mean of the base VA distribution).

C.2.2 ATLAS-GUIDED CROSS-TEACHER DISTILLATION

A correction adapter should not simply perturb the atlas output arbitrarily. Instead, it should refine the atlas prediction toward what is observed by the teacher in the target space. This gives rise to a natural training principle: the atlas provides a structured prior, while the teacher in the target space provides a data-driven refinement signal. We refer to this as *atlas-guided cross-teacher distillation*. For a source-target pair (x, y), let $s_t^{x,T}$ denote the teacher output in the source space and $s_t^{y,T}$ the teacher output in the target space. The fixed atlas first produces a base prediction: $s_{t,\mathrm{base}}^y = M_{x\to y}(s_t^{x,T})$, which is then refined by the correction adapter: $\hat{s}_t^y = T_{x\to y}(s_t^{x,T})$. We directly align the corrected target prediction with the teacher output in the target space through:

$$\mathcal{L}_{\mathrm{align}}^{x\to y} = \ell_y\Big(\hat{s}_t^y,\, s_t^{y,T}\Big), \tag{22}$$

where ℓ_y is CE for expression, BCE for AUs and CCC-based loss for VA.

The offline affect atlas provides a structured and interpretable prior over the relationships between EXP, AU and VA spaces. However, since it is estimated offline from aggregate statistics, it cannot fully capture dataset-specific biases, teacher noise, or domain-specific deviations present in real videos. Our goal is thus to preserve the atlas as a fixed prior while allowing lightweight, data-driven refinements. To this end, we extract synchronized affect streams from frozen teachers, pass them through the fixed atlas mappings, and learn small residual correction adapters on top of them. These adapters are trained through a combination of *cross-teacher alignment* and *cycle consistency*, so that the refined mappings remain both data-consistent and semantically coherent.

C.2.3 BIDIRECTIONAL CYCLE CONSISTENCY

Target-space teacher alignment ensures that an individual mapping matches the target teacher, but it does not guarantee that different mappings are mutually consistent. For example, a corrected expression-to-AU map may align well with the AU teacher while still producing AU states that, when mapped back, no longer correspond to the original expression. We thus impose cycle consistency so that the refined mappings preserve the semantics of the source affect state. For any bidirectional pair (x, y), let: $\hat{s}_t^y = T_{x\to y}(s_t^{x,T})$, & $\tilde{s}_t^x = T_{y\to x}(\hat{s}_t^y)$. We then enforce reconstruction of source state:

$$\mathcal{L}_{\mathrm{cyc}}^{x\leftrightarrow y} = \ell_x\Big(\tilde{s}_t^x,\, s_t^{x,T}\Big). \tag{23}$$

The same principle is applied to all three bidirectional affect pairs:

$$\mathcal{L}_{\mathrm{cyc}} = \mathcal{L}_{\mathrm{cyc}}^{e\leftrightarrow au} + \mathcal{L}_{\mathrm{cyc}}^{e\leftrightarrow va} + \mathcal{L}_{\mathrm{cyc}}^{au\leftrightarrow va}. \tag{24}$$

C.2.4 OVERALL TRAINING OBJECTIVE

The final objective is: $\mathcal{L} = \sum_{x\neq y} \lambda_{\mathrm{align}}^{x\to y}\, \mathcal{L}_{\mathrm{align}}^{x\to y} + \lambda_{\mathrm{cyc}}\, \mathcal{L}_{\mathrm{cyc}}$, where $\lambda_{\mathrm{align}}^{x\to y}$ and λ_{cyc} balance the alignment and cycle-consistency terms.

The adapted atlas therefore learns to satisfy three requirements simultaneously: it remains anchored to the offline probabilistic prior, it aligns with real talking-face data through teacher supervision, and it preserves semantic consistency across affect spaces through bidirectional cycles.

Discussion This adaptation mechanism is central to our framework. A purely offline atlas may be interpretable but too rigid, while a fully learned mapping may be flexible but unstructured and vulnerable to teacher noise. Our approach combines the strengths of both: the offline atlas provides a psychologically meaningful structure, and the teacher-guided residual adapters provide controlled domain adaptation. As a result, the final affect representation is both principled and data-aligned, enabling more reliable emotion guidance for talking-face generation.

C.3 Neuro-symbolic Affect Consistency Framework

AffectAtlas learns meaningful mappings between EXP, AUs, and VA in a data-driven manner. However, it mainly learns correlations from data, and does not explicitly enforce the predicted affect representations to be consistent with known relationships among these components. As a result, the predictions may be inconsistent across EXP, AUs, and VA in some cases. For example, the model may predict a "Happy" expression without enforcing the activation of its corresponding AUs (e.g., AU6 and AU12), even when they are missing. To address this limitation, we introduce a neuro-symbolic consistency framework that integrates data-driven learning with knowledge-based constraints. In this framework, the neural component (AffectAtlas) generates affect representations, while the symbolic component encodes prior affective knowledge in the form of logical rules that describe relationships among these affect spaces. Instead of enforcing these rules as strict constraints, we adopt a differentiable formulation that enables their integration into the learning process.

The proposed framework works as follows. First, frame-wise predictions of affect signals (EXP, AUs, and VA) are obtained from the AffectAtlas. These predictions are then passed to a neuro-symbolic module, which evaluates their consistency using a predefined set of rules under a differentiable fuzzy logic formulation. Each rule computes a satisfaction score that measures how well the prediction satisfies the corresponding affective relationships. These scores are then used to define a logic loss, which penalizes rule violations and optimizes the affect atlas to refine its predictions.

Fuzzy Logic Specification. To integrate symbolic rules with neural predictions, we adopt a fuzzy logic formulation. Fuzzy logic allows variables to take continuous truth values in the range $[0, 1]$, instead of binary values (True or False) (Daniele et al., 2022). Since the outputs of the affect atlas are continuous (e.g., probabilities and normalized VA values), we employ fuzzy logic to represent logical relationships in a differentiable form. We use standard product fuzzy logic operators, AND ($\wedge$), OR ($\vee$), NOT ($\neg$), and IMPLY ($\rightarrow$). Specifically, $x \wedge y = xy, \quad x \vee y = x + y - xy, \quad \neg x = 1 - x, \quad x \rightarrow y = 1 - x + xy$.

Neuro-Symbolic Rules. To construct reliable rules, we leverage prior studies (Clark et al., 2020; Tian et al., 2001b; iMotions, n.d.; Zhang et al., 2014a; Hwooi et al., 2022) on the relationships among EXP, AUs, and VA to build affect relations (Table 1a. Based on these, we define 69 rules in four groups: EXP $\rightarrow$ AU, AU $\rightarrow$ EXP, EXP $\rightarrow$ VA, and AU $\rightarrow$ VA, capturing bidirectional dependencies among affect representations.

EXP $\rightarrow$ AU To ensure that the predicted expression activates its corresponding AUs, we define EXP-to-AU rules. These include both single AU rules and pairwise co-activation rules. Single AU rules ensure the presence of key AUs associated with an expression, while pairwise AU rules capture AUs that typically occur together. For example, the "Happy" expression is associated with AU6 and AU12, and their co-activation is a strong indicator of happiness. Accordingly, rules such as $E_{\text{happy}} \rightarrow AU6$, $E_{\text{happy}} \rightarrow AU12$, and $E_{\text{happy}} \rightarrow (AU6 \wedge AU12)$ are defined.

AU $\rightarrow$ EXP These rules ensure reverse consistency by aligning AU patterns with predicted expressions. They map AU combinations back to their corresponding expressions, keeping both consistent. For example, co-activation of AU6 and AU12 indicates "Happy", while AU1 with AU15 or AU17 corresponds to "Sad". Based on this, we define rules such as $(AU6 \wedge AU12) \rightarrow E_{\text{happy}}$ and $(AU1 \wedge (AU15 \vee AU17)) \rightarrow E_{\text{sad}}$.

EXP $\rightarrow$ VA These rules capture the relationship between discrete expressions and continuous affective dimensions. For example, "Happy" is associated with positive valence and medium-to-high

arousal, while "Sad" corresponds to negative valence and low arousal. Therefore, we define rules such as $E_{\text{happy}} \rightarrow PosV$, $E_{\text{happy}} \rightarrow \neg LowA$, $E_{\text{sad}} \rightarrow LowV$, and $E_{\text{sad}} \rightarrow LowA$. These rules ensure that the predicted VA values are consistent with the corresponding expression.

AU $\rightarrow$ VA These rules describe the relationship between facial muscle activations and affective dimensions. We include only AUs strongly correlated with valence or arousal, excluding weak or ambiguous ones to maintain stability. For example, AU12 is associated with positive valence, while AU4 is associated with negative valence. Similarly, AU5 and AU7 are associated with high arousal. Accordingly, rules such as $AU_{12} \rightarrow PosV$, $AU_4 \rightarrow NegV$, $AU_5 \rightarrow HighA$, and $AU_7 \rightarrow HighA$ are defined. These rules ensure predicted VA values align with activated AUs.

By combining these rules, consistency is further enhanced across EXP, AUs, and VA representations.

Rule Satisfaction Score. Given the defined rule sets, we evaluate their consistency by computing a satisfaction score for each rule. For an implication rule of the form $x \rightarrow y$, where x and y represent the antecedent and consequent respectively, the satisfaction score is computed as:

$$s = 1 - x + x \cdot y, \tag{25}$$

where $s \in [0, 1]$ denotes the satisfaction of the rule.

Logic Loss. We define a logic loss that penalizes violations of the rules and guides the model toward consistent predictions. The loss encourages each rule to achieve a high satisfaction score and is defined as:

$$L_{\text{logic}} = \sum_i w_i(1 - s_i), \tag{26}$$

where s_i is the satisfaction score of the i-th rule and w_i is the corresponding weight. Notably, the neuro-symbolic framework does not directly modify the predictions. Instead, it refines them by encouraging rule consistency during training, resulting in more coherent affect representations across EXP, AUs, and VA.

Discussion Compared to prior emotion-aware generation methods, which typically treat affect as a weak conditioning signal, our framework introduces an explicit reasoning layer that operates over multiple affect representations. This provides three key advantages: (i) improved robustness to noisy or conflicting supervision, (ii) enhanced interpretability through rule-based reasoning, and (iii) stronger cross-representation coherence. As a result, the generated facial behaviour is not only visually plausible but also semantically and psychologically consistent.

C.3.1 Neuro-Symbolic Rules

We define four sets of rules: $EXP \rightarrow AU$, $AU \rightarrow EXP$, $EXP \rightarrow VA$, and $AU \rightarrow VA$. Each rule set is presented in Tables 8-11.

Table 8: EXP to AU (single + pairwise) rules.

Expression	Rule	Expression	Rule
Happy	$E_{happy} \rightarrow AU12$	Surprise	$E_{surprise} \rightarrow AU2$
Happy	$E_{happy} \rightarrow AU6$	Surprise	$E_{surprise} \rightarrow AU5$
Happy	$E_{happy} \rightarrow (AU6 \wedge AU12)$	Surprise	$E_{surprise} \rightarrow AU26$
Anger	$E_{anger} \rightarrow AU4$	Surprise	$E_{surprise} \rightarrow (AU1 \wedge AU2)$
Anger	$E_{anger} \rightarrow AU5$	Surprise	$E_{surprise} \rightarrow (AU5 \wedge AU26)$
Anger	$E_{anger} \rightarrow AU7$	Surprise	$E_{surprise} \rightarrow ((AU1 \vee AU2) \wedge AU5)$
Anger	$E_{anger} \rightarrow AU23$	Fear	$E_{fear} \rightarrow AU1$
Anger	$E_{anger} \rightarrow (AU4 \wedge AU7)$	Fear	$E_{fear} \rightarrow AU2$
Anger	$E_{anger} \rightarrow ((AU4 \vee AU7) \wedge (AU23 \vee AU24))$	Fear	$E_{fear} \rightarrow AU4$
Sad	$E_{sad} \rightarrow AU1$	Fear	$E_{fear} \rightarrow AU5$
Sad	$E_{sad} \rightarrow AU4$	Fear	$E_{fear} \rightarrow AU20$
Sad	$E_{sad} \rightarrow AU15$	Fear	$E_{fear} \rightarrow (AU1 \wedge AU2)$
Sad	$E_{sad} \rightarrow AU17$	Fear	$E_{fear} \rightarrow (AU1 \wedge AU5)$
Sad	$E_{sad} \rightarrow (AU1 \wedge (AU15 \vee AU17))$	Neutral	$E_{neutral} \rightarrow \neg AU4$
Disgust	$E_{disgust} \rightarrow AU9$	Neutral	$E_{neutral} \rightarrow \neg AU9$
Disgust	$E_{disgust} \rightarrow AU10$	Neutral	$E_{neutral} \rightarrow \neg AU10$
Surprise	$E_{surprise} \rightarrow AU1$	Neutral	$E_{neutral} \rightarrow \neg AU12$

Table 9: AU to EXP rules.

Expression	Rule
Happy	$(AU6 \wedge AU12) \rightarrow E_{happy}$
Anger	$(AU4 \wedge AU7) \rightarrow E_{anger}$
Anger	$((AU4 \vee AU7) \wedge (AU23 \vee AU24)) \rightarrow E_{anger}$
Sad	$(AU1 \wedge (AU15 \vee AU17)) \rightarrow E_{sad}$
Surprise	$(AU1 \wedge AU2 \wedge AU5) \rightarrow E_{surprise}$
Surprise	$(AU5 \wedge AU26) \rightarrow E_{surprise}$
Fear	$(AU1 \wedge AU2 \wedge AU4) \rightarrow E_{fear}$
Disgust	$(AU9 \vee AU10) \rightarrow E_{disgust}$
$\neg Act$	$(\neg Act) \rightarrow E_{neutral}$

No AU pair rule is defined for the "Disgust" expression, as $AU9$ and $AU10$ are already the primary indicators of disgust expression.

We define the "Neutral" expression as the absence of AUs, except for AU25 (lip part) and AU26 (jaw drop), as these are speech-related AUs that are often activated during speech.

Table 10: EXP to VA rules.

Expression	Rule	Expression	Rule
Happy	$E_{happy} \rightarrow PosV$	Happy	$E_{happy} \rightarrow \neg LowA$
Sad	$E_{sad} \rightarrow NegV$	Sad	$E_{sad} \rightarrow LowA$
Anger	$E_{anger} \rightarrow NegV$	Anger	$E_{anger} \rightarrow HighA$
Disgust	$E_{disgust} \rightarrow NegV$	Disgust	$E_{disgust} \rightarrow \neg LowA$
Fear	$E_{fear} \rightarrow NegV$	Fear	$E_{fear} \rightarrow HighA$
Neutral	$E_{neutral} \rightarrow NeuV$	Neutral	$E_{neutral} \rightarrow LowA$
		Surprise	$E_{surprise} \rightarrow HighA$

Table 11: AU to VA rules.

Rule	Rule
$AU12 \rightarrow PosV$	$AU24 \rightarrow NegV$
$AU6 \rightarrow PosV$	$AU5 \rightarrow HighA$
$AU4 \rightarrow NegV$	$AU7 \rightarrow HighA$
$AU15 \rightarrow NegV$	$AU15 \rightarrow LowA$
$AU9 \rightarrow NegV$	$AU20 \rightarrow HighA$
$AU10 \rightarrow NegV$	$AU26 \rightarrow HighA$
$AU23 \rightarrow NegV$	

In the expression-to-valence relation, we exclude the "Surprise" expression from the rule, as it can correspond to either positive or negative valence. Therefore, it is assigned only to arousal.

Since we use a fuzzy logic framework, we convert the predicted VA values from the range $[-1, 1]$ to truth values in $[0, 1]$ using sigmoid-based mapping. Specifically, we define: $PosV = \sigma(\alpha v)$, $NegV = \sigma(-\alpha v)$, $HighA = \sigma(\beta a)$, $LowA = \sigma(-\beta a)$, and $NeuV = 1 - |v|$, where α, $\beta = 5$ control the sharpness of sigmoid mapping, determining how strongly VA values approach binary truth values.

In $AU \rightarrow VA$ rules, we only include AUs that have strong psychological correlation with valence or arousal, and avoid weak or ambiguous ones to keep the logic stable.

C.4 Training and Inference Pipelines

Training Pipeline. We first construct an offline AffectAtlas from affect datasets to model the bidirectional relationships among EXP, AU, and VA. For each talking-face training video, three frozen affect teachers extract synchronized EXP, AU, and VA representations. The base Atlas mappings remain frozen, while lightweight residual adapters are trained to adapt these mappings to the talking-face domain. The neuro-symbolic objective enforces cross-space consistency during training. The resulting refined affect representation is then injected into the Diffusion U-Net through the emotion-

aware audio-visual module together with the reference image and driving audio. Additionally, the Audio2Affect student is trained to reproduce the same refined affect tokens from audio.

Inference Pipeline. During inference, neither the target video nor the visual affect models are available. Therefore, the pretrained Audio2Affect student predicts the refined affect tokens from the contextual audio features extracted by Wav2Vec2. These affect tokens are injected into the Diffusion UNet through emotion-aware audio-visual module. The generator then uses the reference image, driving audio, and the predicted affect tokens to generate the emotion-aware talking-face video.

D Implementation Details

We train the mapping network, AffectAtlas and the video generation pipeline separately. AffectAtlas is trained in two stages. In the first stage, we construct probabilistic mappings between EXP, AUs, and VA by extracting affect annotations from large-scale emotion datasets and estimating their statistical relationships. In the second stage, we train the mapping network using the same datasets of the video generation pipeline, including MEAD, CREMA-D, HDTF, CelebVHQ, and in-the-wild data from Hallo3. This ensures that the mapping network aligns with the distribution of talking face videos, improving consistency between affect predictions and the video generation process. In addition, we incorporate the neuro-symbolic consistency framework to enforce consistency between EXP, AUs, and VA based on predefined rules. We train the model on an NVIDIA A100 GPU for 30 epochs with a batch size of 256, Adam optimizer with a learning rate of $1e^{-5}$. The training process takes about 12 hours and requires approximately 2GB of memory.

For the video generation model, videos are resized to a resolution of 512×512 at 25 fps, and the audio is resampled to 16kHz. To ensure temporal consistency across frames, we detect scene changes using PySceneDetect (Breakthrough, 2026). The resulting clips are 2-10 seconds long. Each training sample consists of 14 frames. For affect signal extraction, we use three teacher models: PosterV2(Mao et al., 2025) to extract expression (EXP), OpenFace (Baltrušaitis et al., 2016) to extract action units (AUs), and EmoNet (Toisoul et al., 2021) to extract valence-arousal (VA). During training, we initialize the ReferenceNet and Diffusion UNet with weights from SD1.5 (Rombach et al., 2022). The model is trained on an NVIDIA A100 GPU for 20,000 iterations with a batch size of 1. We use a learning rate of $1e^{-5}$. The training process requires approximately 30 GB of memory and takes around 24 hours to complete.

E Experiment Results

E.1 Cross-dataset Evaluation

Quantitative Results. To evaluate cross-dataset robustness of our method, we conduct a cross-dataset evaluation on VFHQ dataset. As shown in Table 12, AffectAtlas achieves the best results on most metrics, including PSNR, SSIM, FID, CSIM, FVD, E-FID, F1, EmoAcc, CCCV, and CCCA. It also achieves the second-best LPIPS and Sync-C scores, while comparable on Sync-D. These results show that AffectAtlas generalizes well to unseen data while preserving visual quality, identity, synchronization, and emotional fidelity.

Qualitative Results. Figure 5 presents qualitative comparisons with SOTA methods (Zhang et al., 2023; Tan et al., 2024a; 2025; Zheng et al., 2024; Gan et al., 2023; Xu et al., 2024; Chen et al., 2025; Bigata et al., 2025) on VFHQ dataset (Xie et al., 2022). Compared with the competing methods, AffectAtlas generates more expressive and temporally consistent facial emotions while preserving identity, visual quality, and lip synchronization. These qualitative results are consistent with the quantitative improvements reported in Table 12.

E.2 Results on Out-of-Distribution Images.

Figure 6 illustrates the results of our method on out-of-distribution images, although the model is not trained on such data. As shown in the figure, our method generates expressive videos in which emotions vary across frames according to the emotional content of the speech.

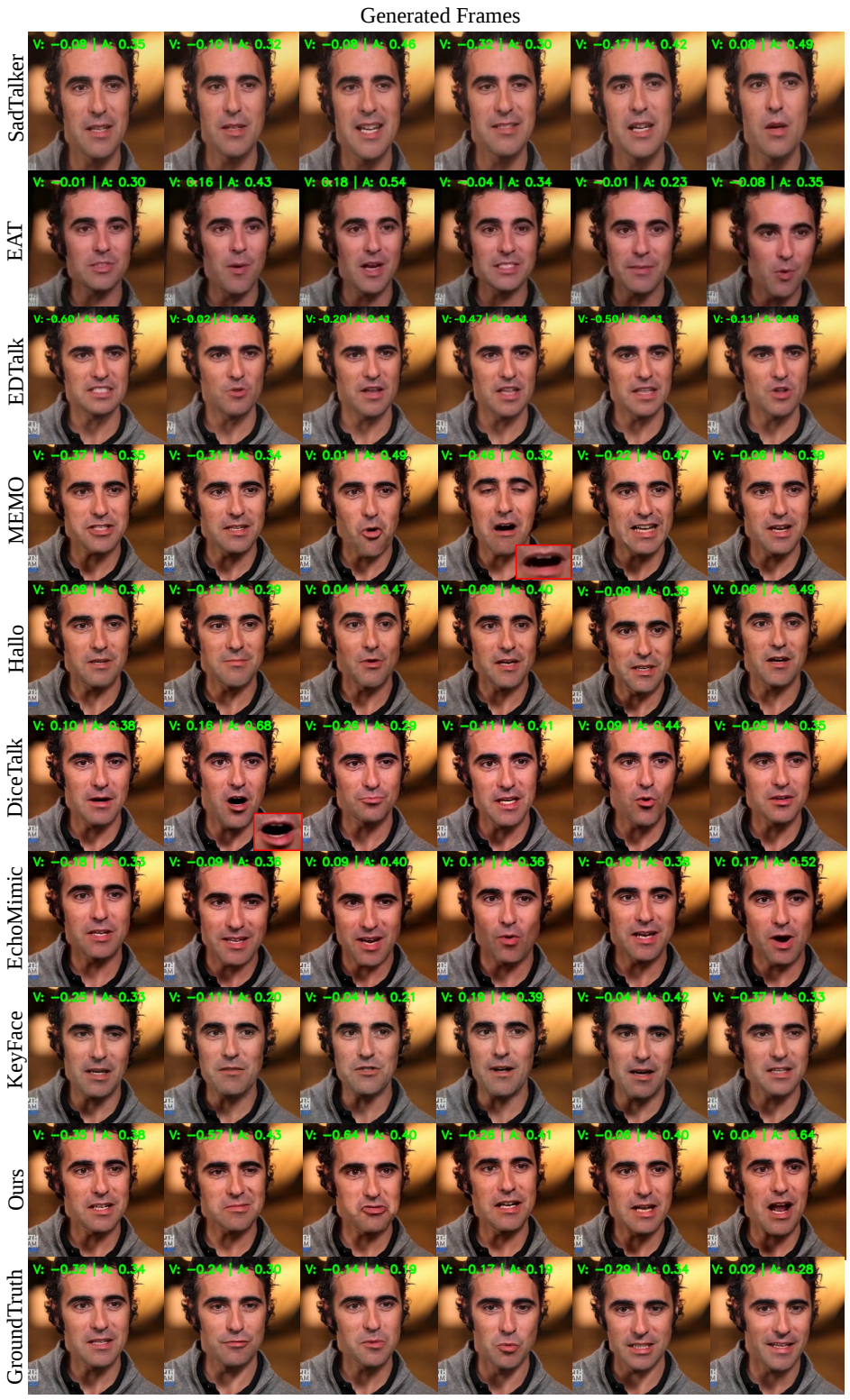


Figure 5: Comparison results with SOTA approaches on VFHQ dataset. We present frame-wise valence (V) and arousal (A) values to compare how different methods capture emotional changes over time.



Figure 6: Generation with out-of-distribution images. AffectAtlas generates dynamic facial expressions across frames that are aligned with the affective content of the speech. See the supplementary videos for a clear illustration.

Table 12: Quantitative results with SOTA approaches on VFHQ dataset. Best results are shown in **bold**, and second-best results are underlined.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CSIM ↑	FVD ↓	SyncC ↑	SyncD ↓	EFID ↓	F1 ↑	EmoAcc ↑	CCCV ↑	CCCA ↑
SadTalker (2023)	28.51	0.53	0.42	68.85	0.88	162.50	6.58	7.79	0.21	0.63	64.67	0.47	0.46
EAT (2023)	13.65	0.51	0.54	127.98	0.82	237.99	6.87	7.79	0.90	0.52	49.33	0.52	0.49
EDTalk (2024)	14.32	0.47	0.53	150.14	0.80	254.47	5.26	9.46	0.25	0.65	66.90	0.50	0.47
MEMO (2024)	30.26	0.62	0.26	36.74	0.88	158.28	7.04	7.56	0.27	0.59	70.00	0.49	0.49
Hallo (2024)	29.90	0.62	0.27	40.78	0.88	131.12	6.47	8.12	0.22	0.65	68.00	0.52	0.52
DiceTalk (2025)	20.87	0.61	0.27	38.41	0.89	143.28	**7.28**	**7.48**	0.26	0.66	59.33	0.54	0.48
Echomimic (2025)	29.19	0.54	0.37	50.86	0.87	121.90	4.96	9.35	0.23	0.62	71.33	0.49	0.49
KeyFace (2025)	30.23	0.63	**0.25**	41.70	0.88	154.09	5.18	9.08	0.21	0.65	68.00	0.51	0.55
Ours	**32.76**	**0.72**	0.26	**33.36**	**0.90**	**117.11**	7.06	7.76	**0.17**	**0.72**	**74.18**	**0.59**	**0.61**

E.3 Additional User Study Analyses

Statistical Analysis. To further analyze the perceptual ratings from the user study reported in Table 6, we perform statistical significance tests using the same ratings collected from 40 participants over 45 clips and 9 evaluated methods. For each evaluation criterion, we first average the clip-level ratings for each participant and method. We then apply the Friedman test to assess whether ratings differ significantly among the evaluated methods. When the Friedman test indicates a significant overall difference, we perform paired Wilcoxon signed-rank tests between AffectAtlas and each baseline. Holm correction is applied to the resulting p-values within each criterion to account for multiple comparisons. The Friedman test shows significant overall differences among the methods for all four evaluation criteria ($p < 0.001$). Holm-corrected Wilcoxon tests show that AffectAtlas significantly outperforms all baselines in emotional consistency, dynamic emotional realism, and overall video quality ($p_{adj} < 0.05$). For lip synchronization, AffectAtlas significantly outperforms all baselines except MEMO, for which the difference is not statistically significant. These results indicate that the perceptual improvements are strongest in the emotion-related dimensions targeted by AffectAtlas, while lip synchronization remains competitive with the strongest baseline.

Temporal Realism. We conduct an additional user study focused specifically on temporal realism. We follow the same video selection and presentation protocol as the main user study. We randomly select 45 clips from MEAD and RAVDESS, covering five emotion classes. Each clip is 4-7 secs long. For each clip, the generated videos from all nine methods are presented side-by-side with audio. Method identities are hidden and their display positions are randomized to reduce positional and ordering bias. We recruit 22 new participants, all of whom evaluate the same completed set of clips and methods. Participants rate each video on a five-point scale according to: (i) temporal coherence & stability, (ii) naturalness of facial motion, and (iii) emotional dynamics & transitions. For each method and criterion, we report the mean score across participants and evaluated clips.

As shown in Table 13, AffectAtlas achieves the highest score across all three temporal criteria. The largest improvement is observed for emotional dynamics and transitions, which directly reflects the objective of AffectAtlas to model temporally evolving and affectively consistent facial behaviour. These results are consistent with our original user study and perceptual ablation study, in which AffectAtlas obtained the highest ratings for emotional appropriateness, dynamic emotional realism and overall video quality.

For statistical analysis, we first average the 45 clip-level ratings for each participant and method within each criterion. We then apply a Friedman test to assess overall differences among the nine methods. Following each significant Friedman test, we perform paired Wilcoxon signed-rank tests between AffectAtlas and each of the eight baselines, with Holm correction to the p-values within that evaluation criterion. The Friedman test shows significant overall differences for all three temporal criteria ($p < 0.001$). Holm-corrected Wilcoxon tests further show that AffectAtlas significantly outperforms all eight baselines across all three criteria ($p_{adj} < 0.05$).

Ablation Perceptual Evidence To provide complementary perceptual evidence, we additionally conducted a user study examining whether the gains arise merely from supplying continuous EXP, AU, and VA signals, or from explicitly modeling their relationships through affect atlas and NeSy. The complete model achieves the highest rating across all four criteria. When affect atlas is re-

Table 13: User Study on Temporal Realism. Higher is better.

Method	Temporal Coherence & Stability ↑	Naturalness of Facial Motion ↑	Emotional Dynamics & Transitions ↑
SadTalker	3.10	3.05	2.95
EAT	2.85	2.80	2.75
EDTalk	2.95	3.01	2.98
MEMO	<u>3.78</u>	3.40	3.35
Hallo	2.30	2.15	2.05
DiceTalk	2.95	2.85	2.95
EchoMimic	2.65	2.60	2.40
KeyFace	3.62	<u>3.58</u>	<u>3.52</u>
Ours	**4.28**	**4.32**	**4.51**

placed by Transformer fusion while NeSy is retained, emotional appropriateness and realism decrease from 4.54/4.43 to 3.12/3.12. When NeSy is removed while affect atlas is retained, they decrease to 4.00/4.11. Moreover, generic concatenation, MLP fusion, and Transformer fusion remain substantially below the complete framework. These findings indicate that the gains do not arise merely from supplying continuous AU magnitudes or increasing fusion capacity; explicitly modeling and reconciling EXP, AU, and VA through the structured atlas and neuro-symbolic consistency constraints also improves perceived emotional realism.

This perceptual ablation, together with the continuous-AU affect-space ablation, shows that: (i) continuous AUs alone do not make VA redundant; and (ii) generic fusion of continuous affect signals does not reproduce the benefit of the proposed structured multi-space modeling.

Table 14: User study results comparing different model configurations. Higher is better.

Method	Emo↑	Realism↑	Sync↑	Vid-Q↑
Concat	1.53	1.52	1.55	2.60
MLP Fusion	2.70	2.35	2.90	2.78
Transf Fusion	2.89	2.90	3.03	3.05
Transf + NeSy (w/o affect atlas)	3.12	3.12	3.76	3.43
Only affect atlas (w/o NeSy)	4.00	4.11	4.08	4.00
Full	**4.54**	**4.43**	**4.30**	**4.27**

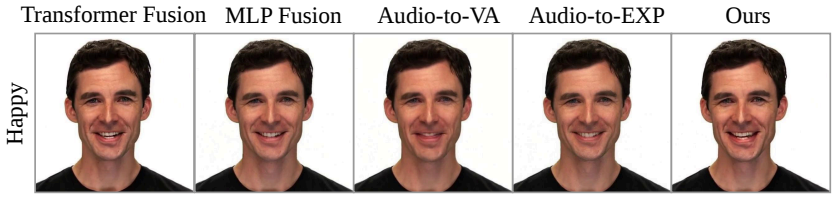


Figure 7: Ablation study on AffectAtlas mapping.

E.4 Evaluation of response delay (lag)

To evaluate whether the generated face changes at the correct time compared to the ground-truth video, we compute the response delay (lag) between the ground-truth and generated videos. Specifically, we use valence-arousal curves extracted from both videos. For each video, the generated valence-arousal signal is shifted within ± 20 frames, and the shift with the highest correlation to the ground-truth signal is selected as the lag. We compute this separately for valence and arousal, and calculate the average absolute lag across all videos. A lower value indicates better temporal alignment between generated and ground-truth facial expressions. As presented in Table 15, our method achieves the best results on MEAD, RAVDESS, and VFHQ, and second-best result on HDTF.

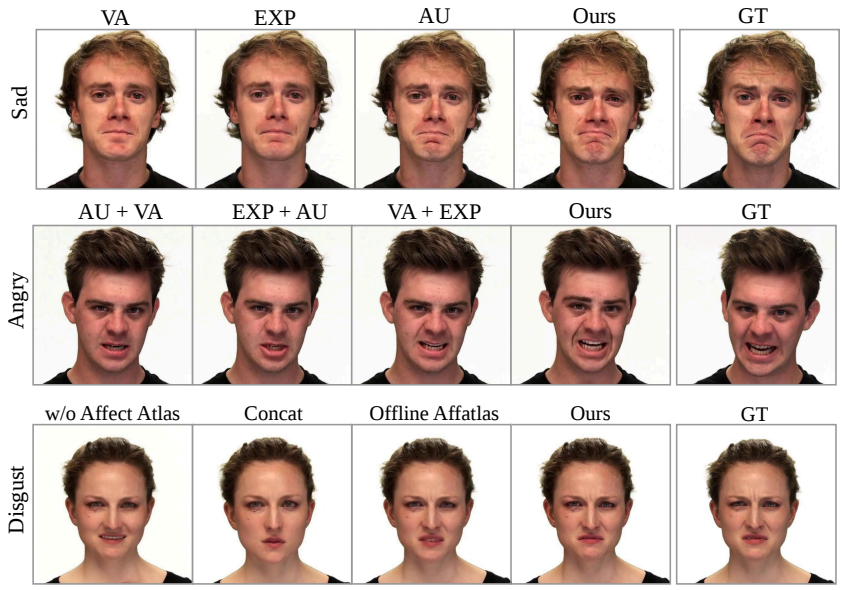


Figure 8: Ablation study on different affect spaces (1^{st} and 2^{nd} row) and Affect Atlas configurations (3^{rd} row).

Table 15: Response delay across datasets (frames / seconds). Lower is better.

Method	MEAD	HDTF	RAVDESS	VFHQ
SadTalker (2023)	9.38 / 0.38s	6.75 / 0.26s	10.39 / 0.41s	9.37 / 0.31s
EAT (2023)	7.19 / 0.29s	7.07 / 0.28s	7.95 / 0.32s	9.38 / 0.31s
EDTalk (2024)	9.79 / 0.26s	8.46 / 0.32s	11.03 / 0.41s	11.14 / 0.32s
MEMO (2024)	11.19 / 0.45s	11.88 / 0.47s	10.63 / 0.43s	12.26 / 0.49s
Hallo (2024)	7.13 / 0.29s	6.45 / 0.26s	10.27 / 0.42s	8.42 / 0.28s
DiceTalk (2025)	6.45 / 0.26s	**5.55 / 0.22s**	8.92 / 0.36s	8.09 / 0.27s
EchoMimic (2025)	8.17 / 0.33s	8.81 / 0.35s	9.55 / 0.39s	9.40 / 0.31s
KeyFace (2025)	9.82 / 0.39s	8.15 / 0.33s	11.31 / 0.45s	10.17 / 0.34s
Ours	**6.02 / 0.24s**	6.29 / 0.25s	**7.17 / 0.28s**	**6.69 / 0.27s**

E.5 Ablation Study

Ablation of Neuro-Symbolic Consistency Framework. To evaluate the consistency of predicted EXP, AUs, and VA with predefined symbolic rules, we compute rule satisfaction as $s(x \rightarrow y) = 1 - x + x \cdot y$ and violation as $v(x \rightarrow y) = 1 - s(x \rightarrow y)$. For this analysis, we train separate models using individual rule sets, including EXP $\rightarrow$ AU, AU $\rightarrow$ EXP, EXP $\rightarrow$ VA, and AU $\rightarrow$ VA. The full rule set achieves the highest rule satisfaction, while EXP $\rightarrow$ AU rules performs better than AU $\rightarrow$ EXP and EXP $\rightarrow$ VA rules. The model without rules gives the lowest satisfaction score. These results indicate that the proposed neuro-symbolic consistency framework effectively enforces structured relationships among EXP, AUs, and VA. Table 16 reports the quantitative results.

Table 16: Ablation study on Neuro-Symbolic rules.

Modality	Rule sat. ↑	Rule vio. ↓
EXP $\rightarrow$ AU rules	0.86	0.14
AU $\rightarrow$ VA rules	0.84	0.16
AU $\rightarrow$ EXP rules	0.82	0.18
EXP $\rightarrow$ VA rules	0.82	0.18
No rules	0.81	0.19
Ours	**0.90**	**0.10**

Ablation of fusion and affect-conditioning mechanisms To directly test whether comparable performance can be obtained using simpler affect-conditioning mechanisms, we conducted four additional controlled experiments.

First, we implemented an MLP fusion baseline that concatenates the EXP, AU and VA teacher outputs and learns a shared affect embedding through a two-layer MLP. Second, we implemented a

stronger Transformer fusion baseline, which uses self-attention to learn interactions among the three affect signals. We additionally evaluated direct audio-to-VA and direct audio-to-EXP conditioning, where the respective affect prediction is directly supplied to the generator without the Atlas (audio-to-VA conditioning directly uses the VA predicted from the driving audio as the emotion condition; audio-to-EXP conditioning directly uses the expression distribution predicted from the driving audio as the emotion condition). All variants use the same reference encoder, audio encoder, diffusion generator, training data and optimization protocol; only the construction of the affect condition is changed. As shown in Table 17, the full method consistently outperforms all four simpler alternatives. Compared with the strongest alternative for each affect-related metric, AffectAtlas reduces EFID from 0.31 to 0.22, improves AU F1 from 0.72 to 0.75, improves EmoAcc from 62.43 to 72.15, and increases valence/arousal CCC from 0.59/0.58 to 0.67/0.66. It also improves visual quality, temporal quality, synchronization and identity preservation. Qualitatively, the simpler alternatives exhibit less stable eye behavior, weaker expression intensity and more limited facial muscle movement.

These experiments isolate two aspects of the design. First, the MLP and Transformer baselines receive the same three affect signals as our method, but replace the probabilistic Atlas with generic learned fusion. Their lower performance indicates that the improvement cannot be attributed solely to combining multiple affect detectors. Second, the audio-to-VA and audio-to-EXP baselines show that conditioning on a single audio-derived affect space does not capture the complementary semantic, structural and continuous affect information provided jointly by EXP, AUs and VA.

Teacher-guided adaptation serves a separate purpose and is not a second redundant affect model. The offline AffectAtlas is constructed once from large-scale affect datasets and provides a general structured prior over the relationships among EXP, AUs and VA. Because these datasets differ from the talking-face domain, we keep the Atlas fixed and learn only lightweight, zero-initialized residual adapters guided by the three frozen teachers. This preserves the structured prior while correcting domain-specific deviations, rather than relearning the complete representation from noisy teacher outputs. The existing Table 4 further shows that teacher-guided adaptation improves over the offline Atlas alone.

Overall, the new fusion and direct-conditioning baselines, together with the existing affect-space and adaptation ablations, show that the gains arise not merely from using multiple pretrained detectors, but from the structured cross-space modeling and its controlled adaptation to talking-face data.

Table 17: Comparison of different fusion and conditioning strategies.

MEAD	PSNR↑	SSIM↑	LPIPS↓	FID↓	CSIM↑	FVD↓	SyncC↑	SyncD↓	EFID↓	F1 (AU)↑	EmoAcc↑	CCC_V ↑	CCC_A ↑
MLP Fusion	28.93	0.64	0.46	152.54	0.93	123.46	6.22	8.74	0.31	0.72	60.28	0.57	0.53
Transf Fusion	28.76	0.62	0.47	160.03	0.92	136.72	6.31	8.62	0.42	0.72	62.43	0.59	0.58
Audio-to-VA	28.91	0.63	0.45	156.62	0.93	142.31	6.15	9.01	0.44	0.71	55.14	0.51	0.53
Audio-to-EXP	28.90	0.65	0.46	156.66	0.92	142.47	6.17	9.12	0.41	0.71	57.36	0.50	0.49
Ours	**30.89**	**0.72**	**0.42**	**152.22**	**0.94**	**111.34**	**6.78**	**8.59**	**0.22**	**0.75**	**72.15**	**0.67**	**0.66**

Ablation of Emotion-aware Audio-Visual Module. To evaluate the effectiveness of the emotion-aware audio-visual module, we inject audio and affect features directly into the Diffusion U-Net through cross-attention. We observe that cross-attention provides limited interaction between audio and affect cues, resulting in weaker expressions and lip synchronization. In contrast, our module jointly fuses visual, audio, and affect tokens through multimodal attention and further injects affect information using emotion-adaptive normalization. This produces more expressive facial motions aligned with the audio emotion while maintaining accurate lip synchronization. Quantitative results with and without the emotion-aware audio-visual module are presented in Table 18.

Table 18: Ablation study on emotion-aware audio-visual module.

Modality	Syn-C ↑	Sync-D ↓	EmoAcc ↑
w/o audio module	5.45	9.12	60.72
Ours	**6.98**	**7.86**	**74.37**

F LIMITATIONS

AffectAtlas is constructed from EXP, AU, and VA signals detected from the training video, instead of relying on predefined emotion labels or reference inputs. Since video-based affect signals are

unavailable at inference, we use an Audio2Affect student that learns from the video-based AffectAtlas teacher to predict affect tokens from driving audio. Therefore, the realism of the generated expressions depends on the emotional content of the driving audio. If the audio contains weak or unnatural emotional cues, such as non-natural or TTS-generated speech, the model may fail to produce accurate facial expressions. Therefore, realistic emotional content in the driving audio is important for generating realistic expressions. As future work, we plan to extend our method to full-body generation, with hand gestures aligned with the expressed emotion.