Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Context-Adaptive Affective Behavior and Artistic Portrait Animation for AIGC Virtual Humans via Temporally Structured Action Units

Yanshuai Shi

shiyanshuai2026@163.com

Zhejiang College of Construction

Article

Keywords: AIGC, virtual digital human, affective computing, Facial Action Units, multimodal interaction, artistic portrait animation

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11182568/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Context-Adaptive Affective Behavior and Artistic Portrait Animation for AIGC Virtual Humans via Temporally Structured Action Units

Yanshuai Shi[1*]
[1] Zhejiang College of Construction, Hangzhou 311231, Zhejiang Province, P.R. China
Corresponding author: Yanshuai Shi; shiyanshuai2026@163.com

Abstract

Affective virtual humans require facial responses that evolve coherently over time, reflect conversational context, and transfer across heterogeneous character appearances. Existing emotional portrait-animation methods commonly rely on discrete emotion labels, frame-wise motion variables, or latent affect representations. This study formulates AIGC-supported virtual-human animation as a context-conditioned temporal affective-behavior modeling problem and proposes TABI-DH. The framework integrates multimodal speech and context encoding, a Temporal AU Behavior Planner, AU-guided non-rigid facial geometry, externally controlled rigid head pose, and behavior–appearance factorized rendering. On MEAD, TABI-DH achieves 22.74 dB PSNR, 0.71 SSIM, 84.25% emotion accuracy, and an AU-trajectory correlation of 0.89. Cross-dataset evaluation on LRW, GRID, and HDTF demonstrates consistent performance under distribution shift. In session-independent IEMOCAP evaluation, incorporating past conversational information increases four-class emotion agreement to 77.1% and reduces affective-response onset error to 176 ms. Mechanism analyses indicate that temporal AU constraints and upper-face behavior primarily improve affective fidelity rather than lip synchronization. Artistic-portrait experiments further show that a common AU trajectory can transfer across photographic and stylized identities while preserving source appearance. These findings support temporally structured AUs as an interpretable representation connecting affective interaction modeling with controllable AIGC virtual-human animation.

Keywords: AIGC; virtual digital human; affective computing; Facial Action Units; multimodal interaction; artistic portrait animation

1. Introduction

Artificial intelligence-generated content (AIGC) is increasingly used to create virtual characters, digital presenters, interactive agents, and animated portraits. Recent reviews characterize talking-head synthesis as a rapidly developing field spanning portrait generation, multimodal driving, editing, emotion control, and interactive digital-human applications [1–3]. Transformer- and generative-model-based approaches have further improved the coordination of speech content, facial motion, and identity-specific appearance [4].

Contemporary audio-driven systems have moved beyond basic mouth synchronization toward richer non-verbal dynamics. FlowVQTalker models emotional facial dynamics using normalizing flows and vector quantization [5]. Multimodal feature-guided methods integrate complementary affective

information to improve emotional expression [6], while EmoTalker learns emotion-aware facial motion directly from speech without requiring an explicit emotion label at inference [7]. These developments demonstrate that talking-head synthesis is no longer limited to speech articulation; the temporal organization of expression, pose, blink behavior, and other non-verbal signals has become an important research objective.

Recent methods have also improved controllability and expressive fidelity. EDTalk separates mouth, pose, and expression representations [8]. FG-EmoTalk introduces fine-grained Action Unit (AU)-level facial-expression control [9]. Style2Talker integrates emotional style with artistic style [10], while EMOPortraits strengthens multimodal one-shot head-avatar generation [11]. X-Portrait uses hierarchical motion attention for expressive portrait animation [12], and EchoMimic combines audio with editable landmark conditions [13]. VASA-1 further demonstrates high-quality holistic facial and head dynamics in a learned latent space [14]. Together, these methods establish strong foundations for expressive portrait synthesis.

Diffusion-based emotional generation has progressed rapidly. EmotiveTalk separates lip-related and expression-related audio information before emotional video diffusion [15]. MoEE uses specialized emotion experts and multimodal affective conditions to represent basic and compound emotions [16]. MVPortrait combines text-guided motion and emotion control with multi-view portrait animation [17]. These approaches strengthen controlled emotional generation, but their primary objective remains the production of a specified affective condition rather than modeling how a visible response should depend on an ongoing conversational context.

Facial AUs provide a useful intermediate representation because they describe localized facial actions and their temporal relationships. Recent AU-recognition research has shown that AU activity exhibits structured spatial and temporal dependencies across facial regions [18]. Earlier multimodal AU-guided talking-face generation demonstrated that combining a 24-dimensional AU representation with facial landmarks and AU-consistency supervision can improve emotional fidelity [19]. AUHead subsequently introduced spatial–temporal AU tokenization and AU-driven emotional talking-head generation [20]. Accordingly, temporal AU control itself can no longer be treated as the principal research gap.

The remaining problem is more specific. First, emotion rendering and affective behavior modeling are not equivalent. A categorical emotion label or independent frame-level AU estimate describes an intended state at one instant, whereas observable facial behavior unfolds through activation, transition, peak, and decay. A sequence can therefore contain plausible frames while exhibiting unrealistic temporal coordination.

Second, most controlled emotional talking-head systems approximate a mapping as expressed in Eq. (1):

$$p(\mathrm{V} \mid I_s, \mathrm{A}, e), \tag{1}$$

where I_s denotes a source portrait, A denotes speech, and e denotes an explicitly specified emotion. An interactive virtual human, however, should also condition its visible response on what happened in preceding conversational turns. Context-sensitive digital-human systems have begun to address semantic responsiveness [21], but the relationship between conversational context and temporally structured facial behavior remains insufficiently characterized.

Third, AIGC virtual characters increasingly appear in photographic, painted, illustrated, and stylized forms. Existing artistic talking-head methods demonstrate strong stylization capabilities [10], but a separate question remains: whether one explicit temporal facial-behavior trajectory can remain stable when visual identity and appearance change.

Accordingly, this study asks:

> Can an AIGC virtual human transform current speech and past interaction information into temporally structured, interpretable facial behavior, and can that behavior remain stable when the visual identity changes across photographic and artistic appearances?

We propose TABI-DH (Temporal Affective Behavior for Interactive Digital Humans). Rather than treating AUs only as frame-wise auxiliary supervision, TABI-DH represents affect as a 24-dimensional AU trajectory positioned between multimodal context processing and visual synthesis. The framework contains a multimodal speech/context encoder, a Temporal AU Behavior Planner, AU-guided non-rigid facial geometry, an externally specified rigid-pose stream, and a behavior–appearance factorized renderer.

Rigid head pose is deliberately separated from affective behavior. Rotation and translation are extracted from a reference driving sequence and treated as explicit control variables. They are not supplied to the AU behavior planner. The interaction experiment therefore evaluates context-conditioned non-rigid affective facial response under controlled rigid head pose, rather than fully autonomous prediction of head motion.

The study is positioned conservatively relative to prior AU-based work. Earlier AU-guided generation [19] primarily used AUs to improve emotional fidelity, while AUHead [20] explicitly models spatial–temporal AU control. TABI-DH does not claim temporal AU control itself as new. Its contributions lie in three connected aspects:

1. Context-conditioned affective-response modeling. Controlled emotion generation is separated from interaction-conditioned response modeling. During interaction evaluation, the target emotion label and target non-rigid facial expression are excluded from the behavior planner; only current speech, past conversational information, and separately controlled rigid pose are available.

2. Temporally structured facial-behavior representation and evaluation. A 24-dimensional AU trajectory explicitly models affective facial dynamics. Temporal derivative supervision, AU-guided non-rigid geometry, and an AU evaluator independent of the training supervisor allow behavior to be assessed separately from visual reconstruction quality.

3. Behavior–appearance factorization for AIGC portrait animation. Facial behavior and source appearance are represented through separate pathways so that one AU trajectory can animate different identities and artistic appearances. The evaluation covers controlled generation, cross-dataset generalization, interaction context, mechanism ablation, artistic transfer, human perception, efficiency, and failure modes.

2. Related Work

2.1. Contemporary Talking-Head Generation

Recent talking-head research spans audio-driven synthesis, temporal motion modeling, high-fidelity rendering, and real-time generation [1–4]. FlowVQTalker explicitly addresses non-deterministic emotional facial dynamics using normalizing-flow and quantized representations [5]. EmoTalker models lip-related and broader facial motion separately and uses cross-modal temporal modeling to generate emotion-aware motion from audio [7]. Teller further demonstrates that autoregressive temporal motion generation can support real-time streaming portrait animation [25].

These developments indicate that realistic talking-head generation depends not only on lip motion but also on the temporal consistency of facial expressions, blinks, pose, and subtle non-verbal motion. Recent work on continuous talking-face generation has similarly emphasized temporal continuity as an independent challenge [28].

2.2. Emotional Portrait Animation

Emotion-aware systems incorporate affective information into speech-driven portrait generation. Multimodal feature-guided emotional talking-face generation combines complementary speech and affective representations [6]. EDTalk factorizes different motion sources to reduce interference between speech articulation, pose, and expression [8]. FG-EmoTalk uses AU-based manipulation to enable fine-grained facial-expression control [9].

Style2Talker extends controllability to both emotion and artistic style [10]. EMOPortraits supports multimodal emotion-enhanced one-shot avatar generation [11], while X-Portrait and EchoMimic improve portrait motion transfer and editable landmark-conditioned generation [12,13].

More recent diffusion-based systems strengthen emotional control further. EmotiveTalk decouples speech information related to lip motion and expression [15], MoEE introduces emotion-specific experts and supports compound emotional states [16], and MVPortrait provides text-guided motion and emotion control [17]. These methods provide strong controlled-generation baselines but do not directly address how past interaction changes the temporal structure of the generated facial response.

2.3. Action Units and Temporal Facial Behavior

AUs provide a structured facial-behavior representation. Modern AU-recognition methods explicitly model multi-scale temporal dynamics and hierarchical interdependencies among AUs [18]. Such evidence supports the use of AU sequences not merely as static labels but as temporally organized facial variables.

The published multimodal AU-guided emotional talking-face framework [19] combines AU information with facial geometry and demonstrates that AU consistency improves emotional expression. AUHead subsequently models spatial–temporal AU sequences as fine-grained controls for emotional talking-head generation [20].

The present study therefore does not define its contribution as "adding AUs." Instead, it treats the AU sequence as a measurable behavioral state trajectory that can be conditioned by interaction history, evaluated independently from the training supervisor, and transferred across visual identities.

2.4. Artistic Portrait Animation and Behavior–Appearance Factorization

Artistic virtual humans introduce an additional source of variation: visual appearance. Style2Talker explicitly represents emotional and artistic styles [10], while X-Portrait enables expressive motion transfer across portrait appearances [12]. The objective of TABI-DH differs from general artistic stylization. It evaluates whether a fixed behavior trajectory can remain stable as source identity or appearance changes.

For this reason, the proposed representation is described as factorized rather than fully disentangled. The claim is empirical: changing visual appearance should not substantially change the intended AU trajectory.

2.5. Affective Interaction

Affective interaction requires modeling information beyond the current utterance. IEMOCAP provides multimodal dyadic interactions, including speech, video, motion capture, transcriptions, and affect annotations [22]. More recent work on context-sensitive digital humans has also demonstrated the importance of semantic responsiveness and AU-level expression control in interactive systems [21].

In the present study, IEMOCAP is used specifically to test whether preceding-turn semantics, prosody, and previous predicted facial state improve the generated non-rigid response relative to current speech alone.

3. Methods

3.1. Task Definition

Let I_s denote a source portrait defining identity and visual appearance. For a sequence of T video frames, speech is represented as

$$\mathrm{A} = \{a_1, \ldots, a_T\},$$

and speech-semantic features as

$$\mathrm{S} = \{s_1, \ldots, s_T\}.$$

Rigid head motion is represented separately according to Eq. (2):

$$\mathrm{P} = \{(R_1, T_1), \ldots, (R_T, T_T)\}, \tag{2}$$

where R_t and T_t denote frame-level rotation and translation extracted from the reference driving sequence.

The generated video is

$$\hat{\mathrm{V}} = \{\hat{I}_1, \ldots, \hat{I}_T\}.$$

TABI-DH introduces an interpretable non-rigid behavior sequence defined in Eq. (3):

$$\mathrm{U} = \{\mathrm{u}_1, \ldots, \mathrm{u}_T\}, \qquad \mathrm{u}_t \in [0,1]^{24}, \tag{3}$$

where each u_t contains 24 AU intensities.

The overall mapping is given in Eq. (4):

$$(I_s, \mathrm{A}, H_{<t}, \mathrm{P}) \rightarrow \hat{\mathrm{U}}_{1:T} \rightarrow \hat{\mathrm{R}}_{1:T} \rightarrow \hat{\mathrm{V}}_{1:T}, \tag{4}$$

where $H_{<t}$ contains conversational information available before the current response.

The pose stream P is used only by the geometry and rendering pathway. It does not enter the Temporal AU Behavior Planner. This design isolates non-rigid affective behavior from rigid head motion.

Two operating modes are used.

In the controlled emotional-generation mode, an explicit emotion condition e is provided as formulated in Eq. (5):

$$(I_s, \mathrm{A}, e, \mathrm{P}) \rightarrow \hat{\mathrm{U}} \rightarrow \hat{\mathrm{V}}. \tag{5}$$

This mode is used for MEAD and conventional emotional talking-head comparison.

In the interaction-conditioned mode, the target emotion is not provided in the formulation of Eq. (6):

$$\left(I_s, \mathrm{A}, H_{past}, \mathrm{P}\right) \rightarrow \hat{\mathrm{U}} \rightarrow \hat{\mathrm{V}}. \tag{6}$$

For each IEMOCAP sample, the same rigid-pose sequence is used across all context ablations so that differences in response behavior cannot be attributed to head-motion changes.

The two modes and the complete architecture are illustrated in Figure 1.

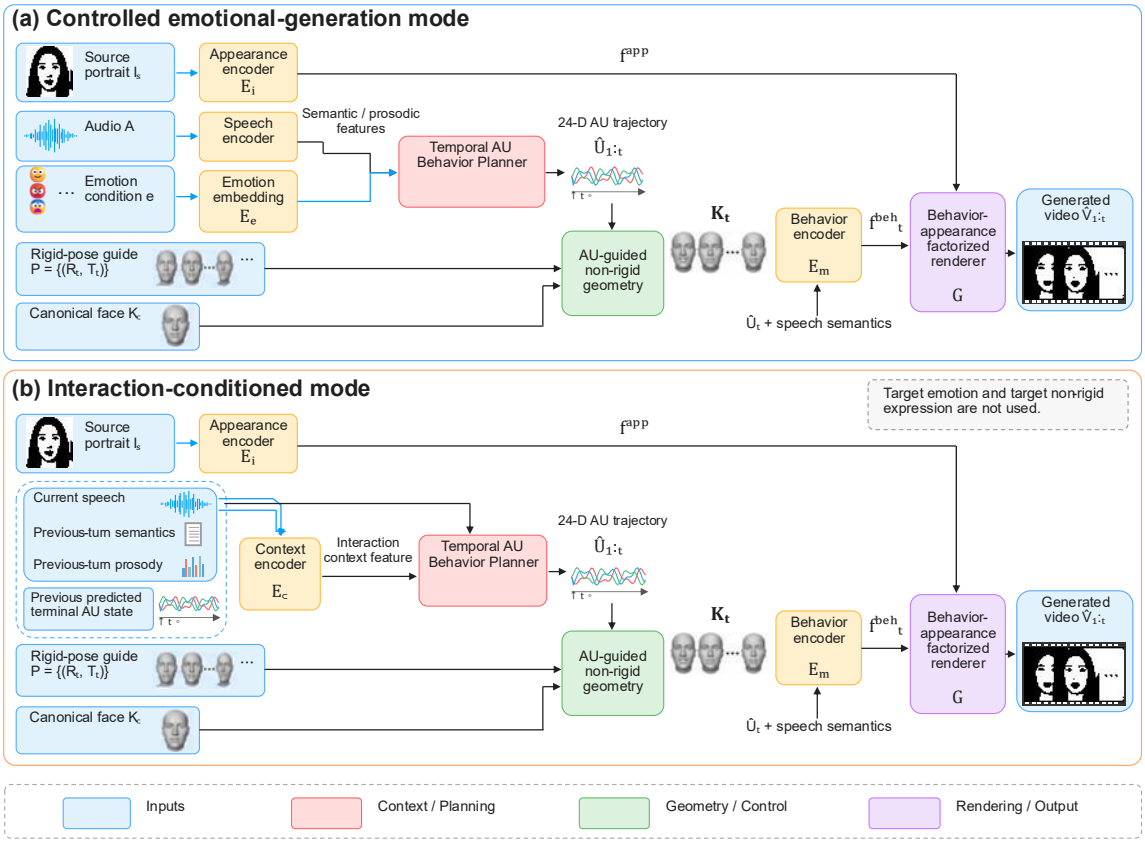


Figure 1. Overall architecture of TABI-DH. (a) Controlled emotional-generation mode with explicit affect condition and separate rigid-pose guidance. (b) Interaction-conditioned mode using current speech, preceding-turn semantics and prosody, and the previous predicted AU state while excluding the target emotion and target non-rigid expression.

3.2. Multimodal Speech and Context Representation

Speech is encoded through semantic and acoustic/prosodic streams as defined in Eq. (7):

$$\mathrm{x}_t = \left[\mathrm{a}_t^{sem}; \mathrm{a}_t^{pros}\right], \tag{7}$$

where a_t^{sem} represents linguistic content and a_t^{pros} represents acoustic and prosodic variation.

The feature design follows contemporary findings that audio representation strongly affects talking-portrait generation efficiency and expressiveness [23]. The broader audiovisual preprocessing pipeline is consistent with recent multimodal speech-representation practice [24].

3.2.1. Controlled emotional generation

For controlled generation, the affect embedding is defined in Eq. (8):

$$\mathrm{q}_t = E_e(e), \tag{8}$$

where E_e is a learned embedding of the requested emotion.

The planner input is given by Eq. (9):

$$\mathrm{c}_t = W_c[\mathrm{x}_t; \mathrm{q}_t] + b_c. \tag{9}$$

3.2.2. Interaction-conditioned response generation

Three past-turn information sources are used:

h_{t-1}^{sem}, the pooled semantic representation of the preceding partner turn;

h_{t-1}^{pros}, the pooled acoustic/prosodic representation of that turn; and

$\hat{\mathrm{u}}_{t-1}^{term}$, the terminal predicted AU state of the preceding generated response.

The contextual state is given by Eq. (10):

$$\mathrm{q}_t^{ctx} = E_c\left(\mathrm{x}_t, \mathrm{h}_{t-1}^{sem}, \mathrm{h}_{t-1}^{pros}, \hat{\mathrm{u}}_{t-1}^{term}\right), \tag{10}$$

The context-conditioned planner input is given by Eq. (11):

$$\mathrm{c}_t = W_c[\mathrm{x}_t; \mathrm{q}_t^{ctx}] + b_c. \tag{11}$$

For the first turn, all past-context terms are initialized to zero.

Interactive inference does not use the ground-truth target-response emotion, target-response AU annotations, future turns, or target non-rigid facial-expression features. Rigid pose is supplied independently through P.

3.3. Temporal AU Behavior Planner

The Temporal AU Behavior Planner maps the multimodal contextual representation to a temporally correlated AU trajectory. The planner uses four temporal Transformer blocks with a hidden dimension of 256, eight attention heads, a feed-forward dimension of 1024, dropout of 0.1, and sinusoidal positional encoding. The use of compact temporal modeling is consistent with recent autoregressive portrait-animation systems that explicitly model motion over time [25].

Training uses contiguous 64-frame windows, corresponding to 2.56 s at 25 frames/s.

The temporal representation is defined in Eq. (12):

$$\mathrm{z}_{1:T} = \mathcal{T}_\theta(\mathrm{c}_{1:T}), \tag{12}$$

The predicted AU vector is given by Eq. (13):

$$\hat{\mathrm{u}}_t = \sigma(W_u \mathrm{z}_t + b_u), \tag{13}$$

where $\sigma(\cdot)$ is the sigmoid function.

The 24 AUs are grouped anatomically into brow/forehead, eye, cheek/nose, and mouth/jaw regions for mechanism analysis. The grouping does not impose fixed emotion templates.

A frozen AU supervisor trained on FEAFA+ [26] provides training targets u_t^{sup}. Frame-level AU consistency is optimized using Eq. (14):

$$\mathcal{L}_{AU} = \frac{1}{24T}\sum_{t=1}^{T}\sum_{j=1}^{24}\mathrm{SmoothL1}\left(\hat{u}_{t,j}, u_{t,j}^{sup}\right). \tag{14}$$

Temporal AU change is defined in Eq. (15):

$$\Delta\hat{\mathrm{u}}_t = \hat{\mathrm{u}}_t - \hat{\mathrm{u}}_{t-1}, \tag{15}$$

Velocity consistency is defined in Eq. (16):

$$\mathcal{L}_{vel} = \frac{1}{T-1}\sum_{t=2}^{T}\left\|\Delta\hat{\mathrm{u}}_t - \Delta\mathrm{u}_t^{sup}\right\|_1. \tag{16}$$

The behavior objective is given by Eq. (17):

$$\mathcal{L}_{beh} = \mathcal{L}_{AU} + \lambda_v\mathcal{L}_{vel}, \qquad \lambda_v = 0.5. \tag{17}$$

3.4. AU-Guided Geometry with External Rigid-Pose Control

Let

$$K_c \in \mathbb{R}^{15\times 3}$$

denote the canonical identity-specific three-dimensional landmark set.

Rigid motion is supplied by (R_t, T_t). AU-related non-rigid deformation is modeled by Eq. (18):

$$\Delta K_t^{AU} = \mathrm{reshape}(B\hat{\mathrm{u}}_t), \tag{18}$$

where

$$B \in \mathbb{R}^{45\times 24}$$

is a trainable AU-to-landmark projection.

A residual identity-dependent non-rigid component is predicted according to Eq. (19):

$$r_t = f_r(\mathrm{z}_t), \tag{19}$$

where $r_t \in \mathbb{R}^{15\times 3}$.

The final driving geometry is given by Eq. (20):

$$K_t = R_t K_c + T_t + \Delta K_t^{AU} + r_t. \tag{20}$$

This formulation separates externally controlled rigid motion, interpretable AU-related deformation, and residual non-rigid variation.

Geometry supervision is defined in Eq. (21):

$$\mathcal{L}_{geo} = \frac{1}{T}\sum_{t=1}^{T}\| K_t - K_t^{ref} \|_2^2. \tag{21}$$

3.5. Behavior–Appearance Factorized Rendering

The source portrait is encoded according to Eq. (22):

$$\mathrm{f}^{app} = E_I(I_s), \tag{22}$$

where f^{app} captures identity, texture, and style-related appearance.

Behavior is independently encoded by Eq. (23):

$$\mathrm{f}_t^{beh} = E_M(K_t, \hat{\mathrm{u}}_t, \mathrm{a}_t^{sem}). \tag{23}$$

The output frame is generated according to Eq. (24):

$$\hat{I}_t = G\left(\mathrm{f}^{app}, \mathrm{f}_t^{beh}\right). \tag{24}$$

The representation is described as factorized, rather than fully disentangled, because its independence is evaluated empirically through behavior transfer rather than assumed from architecture alone.

3.6. Training Objectives

Audio–visual synchronization is constrained with pretrained audio and visual embeddings **s** and **g** as formulated in Eq. (25):

$$\mathcal{L}_{sync} = -\log\left[\max\left(\frac{\mathrm{s}^{\top}\mathrm{g}}{\| \mathrm{s} \|_2 \| \mathrm{g} \|_2 + \epsilon}, \epsilon\right)\right], \tag{25}$$

where $\epsilon = 10^{-6}$. Recent work has shown that strong audiovisual speech representations provide useful supervision and evaluation signals for talking-face synchronization [27].

Image reconstruction loss is defined in Eq. (26):

$$\mathcal{L}_{img} = \frac{1}{T}\sum_{t=1}^{T} \| \hat{I}_t - I_t^{ref} \|_1, \tag{26}$$

Perceptual loss is defined in Eq. (27):

$$\mathcal{L}_{perc} = \frac{1}{T}\sum_{t}\sum_{l} \| \psi_l(\hat{I}_t) - \psi_l(I_t^{ref}) \|_1, \tag{27}$$

where $\psi_l(\cdot)$ denotes pretrained visual features. Temporal visual continuity is additionally examined through the behavior and appearance analyses rather than inferred from reconstruction loss alone, consistent with recent continuous talking-face studies [28].

The overall loss is given in Eq. (28):

$$\mathcal{L} = 60\mathcal{L}_{beh} + 10\mathcal{L}_{sync} + 100\mathcal{L}_{geo} + 10\mathcal{L}_{perc} + 10\mathcal{L}_{img}. \tag{28}$$

The weights are selected using the MEAD validation partition and are fixed for all reported experiments.

3.7. Datasets

The datasets and their experimental roles are summarized in Table 1.

MEAD

MEAD [29] is used for primary controlled emotional generation. It contains multiple emotional speaking conditions and supports systematic evaluation of emotion-conditioned talking-head synthesis. Identity-disjoint train, validation, and test partitions are used. Conversational-history features are not used in MEAD experiments.

FEAFA+

FEAFA+ [26] provides 24-dimensional AU intensity annotations and is used to train the AU supervisor.

LRW, GRID, and HDTF

LRW, GRID, and HDTF are used exclusively for external generalization analysis. Recent talking-face studies continue to use in-the-wild speech benchmarks and high-resolution talking-head data to examine cross-domain robustness [30–32].

IEMOCAP

IEMOCAP [22] is used for interaction-conditioned response analysis. It contains five dyadic sessions involving ten actors and provides speech, video, facial motion, transcriptions, and emotion annotations.

Five-fold leave-one-session-out evaluation is used. In each fold, one complete session is held out for testing. Validation dialogues are held out from the remaining four sessions, and complete dialogues remain within a single partition.

Happiness and excitement are merged into one positive-affect category; anger, sadness, and neutral are retained, yielding four evaluation classes.

Rigid head pose is extracted from the corresponding response sequence and treated as a controlled nuisance variable. The same pose sequence is used across all context-ablation conditions for a given sample. No target-response image pixels or target non-rigid expression variables are supplied to the AU planner. The interaction experiment therefore assesses whether conversational context improves non-rigid affective response conditional on identical rigid head motion.

Artistic Portrait Evaluation Set

The artistic-domain evaluation uses a fixed set of 15 target portraits spanning photographic portraits, public-domain paintings, illustrations, and AIGC-stylized character images. Inclusion requires a clearly visible, approximately frontal human-like face and sufficient resolution for the common preprocessing pipeline. The source set is fixed before comparative model evaluation, and samples are not removed based on output quality. The photograph-like portraits used in this evaluation are AI-generated synthetic fictional identities. Public-domain painting examples are used as public-domain artworks, and no images of the perceptual-study participants are included in the artistic portrait evaluation set.

Table 1. Datasets and experimental roles

Dataset	Experimental role	Principal characteristic
MEAD	Primary controlled emotional generation	Multiple emotional speaking conditions
FEAFA+	AU-supervisor training	24-dimensional AU intensities
LRW	External generalization	In-the-wild audiovisual speech
GRID	External generalization	Controlled audiovisual speech
HDTF	External generalization	High-resolution talking-head video
IEMOCAP	Interaction-response evaluation	Dyadic multimodal affective interaction
Artistic Portrait Set	Cross-appearance evaluation	Photograph, painting, illustration, AIGC portrait

3.8. Preprocessing and Implementation

Videos are sampled at 25 frames/s and audio at 16 kHz. Faces are cropped and resized to 256×256 pixels. The acoustic pathway uses 80-bin mel spectrograms.

The speech representation follows contemporary audio-feature extraction practice for talking-portrait synthesis [23], while the broader audiovisual feature pipeline is consistent with recent supervised multimodal speech representations [24]. Additional recent work has also demonstrated the value of speech-aware and landmark-aware representations for person-generic and speech-driven face animation [30,33].

The AU supervisor uses a ResNet-18 backbone with a 24-dimensional bounded regression head and obtains a validation MSE of 0.03760 under the FEAFA+ protocol.

TABI-DH is trained on four NVIDIA RTX 3090 GPUs using Adam. The audio/driving encoder uses a learning rate of 1.5×10^{-4}, while the temporal-behavior and rendering components use 2×10^{-4}. The batch size is 32 temporal windows per GPU. Training proceeds for 120,000 iterations, with the learning rate multiplied by 0.5 at 80,000 and 105,000 iterations.

The AU supervisor and synchronization network remain frozen during generator training.

Independent AU evaluator

To avoid teacher–evaluator circularity, test-time AU metrics are computed with a separately trained EfficientNet-B0 AU evaluator rather than the AU supervisor used in $\mathcal{L}_{AU}$. EfficientNet-based expression recognition has demonstrated suitable compact facial-feature extraction for this type of evaluation [34].

The independent evaluator uses the same 24-AU semantic label space but has a different backbone and independently optimized parameters. It shares no checkpoint, optimizer state, or model parameters with the training supervisor and is never used for generator training, hyperparameter selection, or validation-based model selection.

Because the two models use the same AU annotation space, the resulting evaluation is model-independent rather than dataset-independent.

For generated and reference frames, the evaluator outputs are defined in Eq. (29):

$$\tilde{\mathbf{u}}_t^{gen} = Q(\hat{I}_t), \qquad \tilde{\mathbf{u}}_t^{ref} = Q(I_t^{ref}), \tag{29}$$

and all AU-MAE, AU-TCorr, velocity, and peak-timing metrics are calculated from these independent estimates.

3.9. Comparison Protocol

The contemporary controlled-generation comparison includes EmotiveTalk [15], MoEE [16], and AUHead [20], together with a representative established emotional talking-head baseline.

The compared methods do not use identical native conditioning interfaces. Consequently, the main comparison is interpreted as a matched-output benchmark rather than an information-budget-identical leaderboard.

For all compared methods, the same MEAD test identities, source portraits, speech segments, and target emotion are used whenever compatible with the method's native control interface. Outputs are standardized to 256×256, and all reported metrics are recomputed using one common evaluation pipeline.

TABI-DH additionally receives externally specified rigid-pose control. Accordingly, improvements in head-pose-sensitive geometric metrics are not interpreted as evidence that TABI-DH predicts head pose better than methods without the same control signal. The principal cross-method interpretation focuses on affective fidelity, temporal behavior, and consistency of generated expressions.

For artistic animation, Style2Talker [10] is the principal comparator because artistic style is explicit in its task definition. X-Portrait [12] is included as an auxiliary motion-transfer reference.

3.10. Evaluation Metrics

3.10.1. Visual and geometric quality

Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) evaluate reconstruction fidelity.

Fréchet Inception Distance (FID) is retained for comparability with generative-video literature; recent work has also highlighted limitations of FID as a stand-alone generative-quality measure [35]. It is therefore interpreted together with geometric, behavioral, and human-perception metrics.

The conventional SyncNet score is retained as the principal benchmark audio–visual synchronization measure, while synchronization-loss design follows more recent audiovisual speech-representation approaches [27].

Mouth landmark distance (M-LMD) and full-face landmark distance (F-LMD) measure geometric deviation.

3.10.2. Emotion and temporal behavior

MEAD emotion-classification accuracy is denoted ACC_{emo}.

Independent-evaluator AU mean absolute error is defined in Eq. (30):

$$AU\text{-}MAE = \frac{1}{24T}\sum_{t=1}^{T}\sum_{j=1}^{24}\left|\tilde{u}_{t,j}^{gen} - \tilde{u}_{t,j}^{ref}\right|. \tag{30}$$

AU trajectory correlation (AU-TCorr) is the mean Pearson correlation between generated and reference AU trajectories over active AUs.

AU velocity error is defined in Eq. (31):

$$E_{vel} = \frac{1}{24(T-1)}\sum_{t=2}^{T} \| \Delta\tilde{\mathrm{u}}_t^{gen} - \Delta\tilde{\mathrm{u}}_t^{ref} \|_1. \tag{31}$$

An AU is considered active for peak analysis when its normalized reference maximum exceeds 0.20. Peak-timing error is the absolute difference between generated and reference maximum-activation frame indices.

3.10.3. Interaction-response metrics

IEMOCAP response behavior is evaluated with overall four-class agreement, unweighted average recall (UAR), macro-F1, AU-TCorr, and affective-response onset error.

Response onset is defined as the first frame at which the magnitude of affect-related AU change exceeds the onset criterion for at least three consecutive frames.

3.10.4. Identity and appearance

Photographic identity preservation is evaluated with deep face-embedding cosine similarity. Recent face-recognition literature highlights the importance of robustness when embeddings are applied under changes in pose, expression, and image quality [36].

For artistic portraits, vision–language feature similarity is used to quantify semantic appearance consistency, consistent with current vision–language evaluation practice [37]. DINOv2 features provide an additional representation of high-level source-content consistency [38].

3.10.5. Statistical analysis

Human perceptual ratings are analyzed as repeated ordinal measures using Friedman tests. Kendall's W is reported as the omnibus effect size.

Pairwise post-hoc comparisons use Wilcoxon signed-rank tests with Holm correction.

For AU-TCorr, 95% confidence intervals are computed with 1,000 video-level bootstrap resamples. Individual frames are not treated as independent resampling units.

3.11. Human Perceptual Study

Thirty-two adult participants take part in a randomized, blinded, within-subject perceptual evaluation.

The candidate pool contains 15 source portraits and five driving clips for each emotion condition. Balanced subsets are assigned so that methods are compared under matched source, audio, pose, and affective conditions without requiring every participant to rate all possible combinations.

Participants rate five dimensions on five-point scales: visual naturalness, emotion appropriateness, audio–visual synchronization, contextual appropriateness, and source-appearance preservation.

Context-appropriateness ratings reported for IEMOCAP are calculated only from the interaction subset. The broader perceptual comparison in Table 8 aggregates the matched evaluation pool.

For inferential statistics, ratings are averaged within participant and method for each dimension before the Friedman test. Individual video ratings are therefore not treated as independent participants.

3.12. Ethical Approval and Informed Consent

This study involved two forms of human-related data: secondary analysis of existing audiovisual human-participant datasets and a perceptual evaluation conducted with 32 adult participants. MEAD, FEAFA+, LRW, GRID, HDTF, and IEMOCAP were obtained through their respective public or controlled-access distribution channels and used in accordance with the applicable dataset licenses and usage conditions. No new participant recruitment was performed for these benchmark datasets, and no additional direct identity information beyond that distributed with the licensed research data was collected.

The perceptual study involving 32 adult participants was reviewed and approved by the Ethics Committee of Zhejiang College of Construction. The committee did not assign an approval number. All procedures involving human participants were performed in accordance with applicable institutional guidelines and regulations and with the principles of the Declaration of Helsinki.

Informed consent was obtained from all participants before participation. Participation was voluntary, and the perceptual-study ratings were analyzed without using participant names or other direct identifying information. No images or directly identifying information from the 32 perceptual-study participants are included in this article. The human-like facial images shown in Figures 3 and 4 are not images of recruited study participants. Photograph-like faces in these figures are AI-generated synthetic fictional identities; public-domain painting examples are included as public-domain artworks.

3.13. Code Availability

The code developed and used for the analyses in this study is provided as Supplementary Material accompanying this article. The supplementary resources include the implementation and configurations required to reproduce the reported analysis pipeline, subject to the access conditions of the third-party datasets.

3.14. Use of Artificial Intelligence Tools

OpenAI ChatGPT was used to assist with language editing, structural organization, and journal-formatting of the manuscript. The author reviewed and verified the scientific content, citations, analyses, interpretations, and final wording and assumes full responsibility for the submitted manuscript.

4. Results

4.1. Contemporary Controlled-Generation Comparison

The primary controlled-generation results are reported in Table 2.

Table 2. Contemporary emotional talking-head comparison on MEAD

Method	PSNR ↑	SSIM ↑	FID ↓	SyncNet ↑	M-LMD ↓	F-LMD ↓	ACC_{emo} ↑	AU-TCorr ↑
Reference emotional baseline	21.75	0.68	19.69	8.28	2.25	2.47	75.43	0.79
EmotiveTalk [15]	22.43	0.70	19.41	8.44	2.12	2.34	81.60	0.84
MoEE [16]	22.51	0.70	20.02	8.39	2.10	2.31	82.70	0.85
AUHead [20]	22.69	0.71	19.88	8.37	2.07	2.27	84.00	0.88
TABI-DH	22.74	0.71	20.94	8.32	2.06	2.25	84.25	0.89

The results do not indicate uniform superiority. EmotiveTalk obtains the lowest FID and highest SyncNet score, while AUHead closely matches TABI-DH on emotion accuracy and AU-TCorr.

This pattern is consistent with the study objective. TABI-DH is not intended primarily to replace specialized diffusion renderers. Its main contribution is to maintain competitive controlled generation while introducing an explicit and measurable temporal-behavior representation.

AUHead reaches AU-TCorr 0.88 compared with 0.89 for TABI-DH. This small difference is treated as comparable controlled-generation performance rather than a substantial advantage. The more consequential evidence for TABI-DH is provided by the interaction-conditioned and cross-appearance experiments.

Because TABI-DH receives explicit rigid-pose guidance, M-LMD and F-LMD are interpreted descriptively rather than as evidence of superior pose prediction.

4.2. Cross-Dataset Generalization

External generalization results are shown in Table 3.

Table 3. Cross-dataset evaluation

Dataset	Method	PSNR ↑	SSIM ↑	FID ↓	SyncNet ↑	M-LMD ↓	F-LMD ↓
LRW	Reference baseline	24.11	0.77	–	6.22	1.79	2.08
	TABI-DH	25.28	0.79	–	6.07	1.63	1.92
GRID	Reference baseline	26.40	0.85	16.93	5.23	2.00	1.72
	TABI-DH	28.45	0.88	10.71	5.02	1.55	1.44
HDTF	Reference baseline	23.51	0.75	20.62	6.58	2.40	2.48

Dataset	Method	PSNR ↑	SSIM ↑	FID ↓	SyncNet ↑	M-LMD ↓	F-LMD ↓
	TABI-DH	24.76	0.78	17.28	6.55	1.98	2.21

TABI-DH retains improvements in reconstruction and facial geometry across all three external datasets. Synchronization does not improve uniformly: the reference baseline remains higher on LRW and GRID and is marginally higher on HDTF.

This result separates affective whole-face modeling from mouth-specific synchronization. Temporal AU modeling primarily benefits the organization of facial expression rather than phonetic lip alignment.

GRID yields the strongest reconstruction results because of its comparatively controlled acquisition conditions. HDTF remains more challenging because of higher spatial detail and more natural head-motion variation.

4.3. Temporal Representation Analysis

MEAD contains controlled emotional speech rather than conversational interactions. Accordingly, no conversational-history variable is used in this analysis. The representation-ablation results are summarized in Table 4.

Table 4. Facial-behavior representation ablation on MEAD

Representation	AU-MAE ↓	AU-TCorr ↑	AU velocity error ↓	Peak-timing error ↓	ACC_{emo} ↑
Landmark-only behavior	0.104	0.68	0.050	5.2 frames	72.00
Frame-wise AU control	0.074	0.79	0.037	4.0 frames	81.00
Temporally structured AU behavior	0.056	0.89	0.027	2.4 frames	84.25

Moving from landmark-only behavior to frame-wise AU control substantially reduces AU error and improves emotion recognition. Temporal modeling produces an additional improvement: AU-TCorr rises from 0.79 to 0.89, velocity error decreases from 0.037 to 0.027, and peak-timing error decreases from 4.0 to 2.4 frames.

The video-level bootstrap 95% confidence interval for AU-TCorr is 0.882–0.897.

Representative temporal trajectories are visualized in Figure 2. The temporal model follows reference onset and peak structure more closely than the frame-wise AU configuration, particularly for upper-face actions associated with fear, sadness, and disgust.

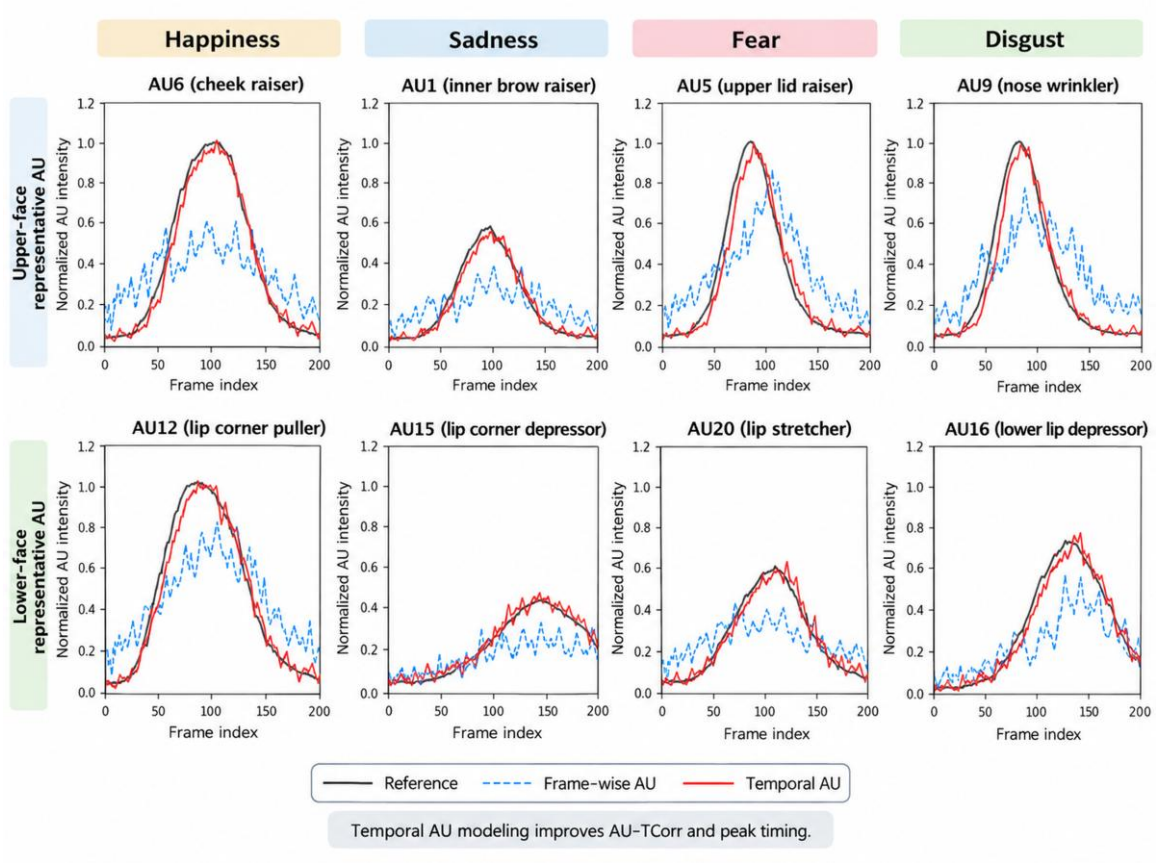


Figure 2. Representative independent-evaluator AU trajectories for happiness, sadness, fear, and disgust. The temporal AU model follows reference onset, peak, and decay more closely than frame-wise AU control.

4.4. Interaction-Conditioned Response Modeling

Four nested contextual conditions are evaluated on IEMOCAP while keeping the rigid-pose sequence identical: current utterance only; current utterance plus previous-turn semantics; the preceding configuration plus the previous predicted terminal AU state; and the preceding configuration plus previous-turn prosody.

The results are summarized in Table 5.

Table 5. IEMOCAP interaction-context evaluation

Input configuration	Agreement ↑	UAR ↑	Macro-F1 ↑	AU-TCorr ↑	Onset error ↓	Context MOS ↑
Current utterance only	68.7%	66.3%	66.9%	0.69	244 ms	3.62 ± 0.61
+ previous-turn semantics	72.9%	70.8%	71.4%	0.73	211 ms	3.91 ± 0.56
+ previous predicted AU state	75.8%	74.1%	74.6%	0.76	188 ms	4.12 ± 0.50
+ previous-turn prosody (full context)	77.1%	75.4%	75.9%	0.78	176 ms	4.24 ± 0.47

Previous-turn semantics provide the largest contextual increment, increasing UAR from 66.3% to 70.8%. The previous predicted AU state further raises UAR to 74.1%, while past prosody provides a smaller additional improvement to 75.4%.

The diminishing increments suggest complementary rather than redundant roles. Semantic history supplies information about conversational meaning, whereas the preceding facial state and prosodic history refine how that meaning is expressed visibly.

The final onset error of 176 ms corresponds to approximately 4.4 frames at 25 frames/s. The context pathway therefore improves temporal response initiation without eliminating timing error.

Because all conditions use the same rigid-pose guide, the differences in Table 5 are attributable to the contextual behavior pathway rather than to variation in head motion.

4.5. Mechanism-Oriented Ablation

Mechanism-oriented ablation results are presented in Table 6.

Table 6. Mechanism-oriented ablation on MEAD

Configuration	PSNR ↑	FID ↓	SyncNet ↑	ACC_{emo} ↑	AU-TCorr ↑
Full temporal-AU model	22.74	20.94	8.32	84.25	0.89
Without temporal-velocity constraint	22.66	21.21	8.31	82.90	0.83
Without upper-face AU group	22.71	21.05	8.34	78.80	0.84
Without AU-conditioned non-rigid geometry	22.31	22.74	8.30	76.90	0.72

Removing temporal-velocity supervision changes PSNR by only 0.08 dB but reduces AU-TCorr from 0.89 to 0.83. Frame-level visual-quality metrics therefore substantially underestimate the contribution of temporal behavior modeling.

Removing upper-face AUs lowers emotion accuracy from 84.25% to 78.80% while slightly increasing SyncNet. This pattern is consistent with speech articulation being concentrated in lower-face motion, whereas brow, eye, and cheek activity provides additional affective information.

The largest reduction in AU-TCorr occurs when AU-conditioned non-rigid geometry is removed. Temporal AU information is therefore most useful when it directly influences facial deformation rather than remaining only an abstract latent condition.

4.6. AU-Intensity Intervention

Affect-related AU trajectories are scaled to 0.5 ×, 1.0 ×, and 1.5 × while speech, source portrait, and rigid pose remain fixed.

At 0.5 ×, motion remains smooth but emotion recognizability decreases. At 1.0 ×, recognizability and naturalness are jointly strongest. Increasing the same trajectory to 1.5 × strengthens visible emotional intensity but produces exaggerated brow and mouth deformation in some identities.

The intervention therefore exhibits a non-monotonic naturalness–intensity relationship. AU magnitude has an interpretable behavioral effect rather than acting as an arbitrary hidden variable.

Representative outputs and rating trends are shown in Figure 3. All photorealistic facial identities displayed in Figures 3 and 4 are synthetic fictional identities generated for the experiments in this study and do not depict real individuals. No identifiable real-person photographs are shown in these figures. The public-domain painting examples in Figure 4 are public-domain artworks and are not images of recruited study participants.

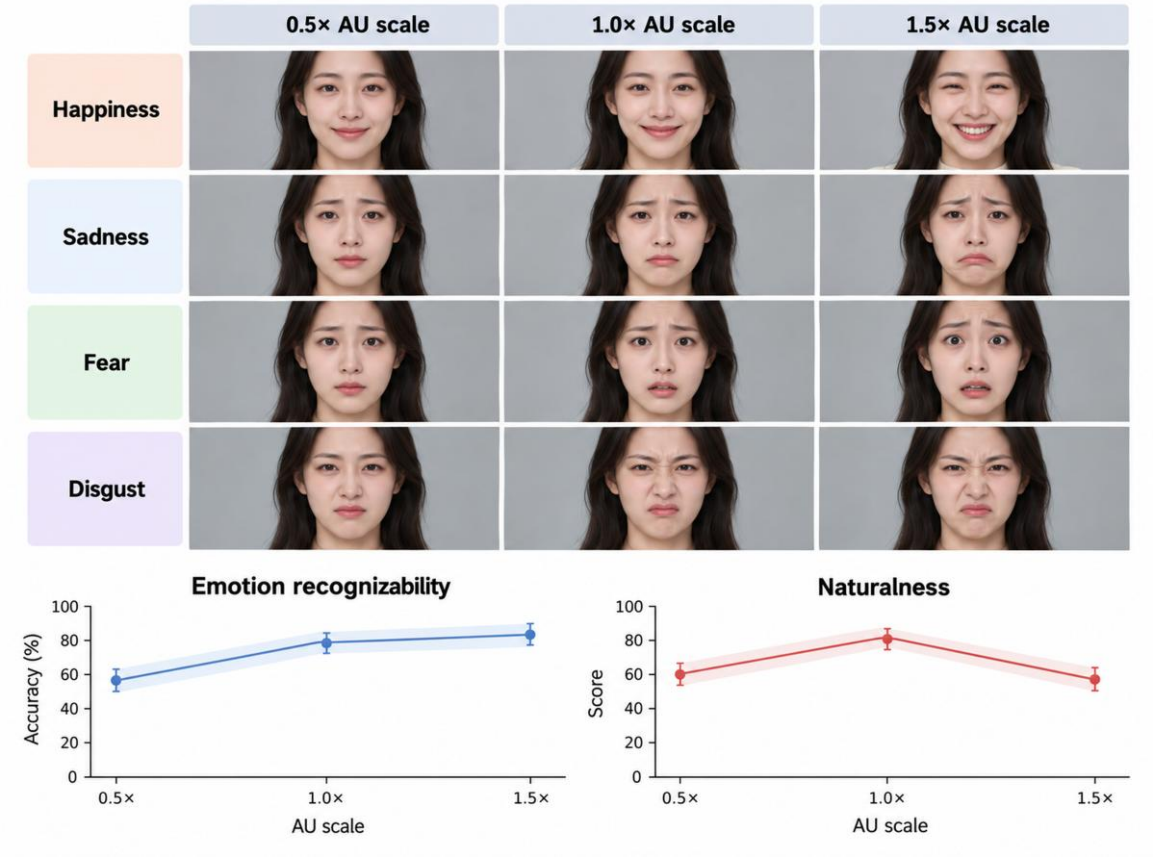


Figure 3. AU-intensity intervention for happiness, sadness, fear, and disgust. Expression strength increases from 0.5 × to 1.5 ×, while naturalness peaks around the nominal 1.0 × setting. All faces shown in this figure are AI-generated synthetic fictional identities and do not depict real persons or study participants.

4.7. Artistic Portrait Animation and Behavior Transfer

Cross-appearance results are presented in Table 7.

Table 7. Cross-appearance artistic portrait evaluation

Method	Emotion accuracy ↑	DINO content similarity ↑	Vision–language appearance similarity ↑	Face-embedding similarity* ↑	Temporal appearance error ↓	Source-appearance MOS ↑
X-Portrait [12]	75.8%	0.812	0.781	0.816	0.126	3.84
Style2Talker [10]	81.3%	0.821	0.834	0.829	0.118	4.21
TABI-DH	84.6%	0.846	0.828	0.842	0.102	4.30

Face-embedding similarity is calculated only on the photographic subset.

Style2Talker obtains the highest vision–language appearance score, which is consistent with its explicit art-style objective. TABI-DH instead yields higher emotion accuracy, DINO content similarity, photographic face-embedding similarity, and lower temporal appearance error.

These results do not establish TABI-DH as a better general-purpose stylization system. They support the narrower claim that a common explicit facial-behavior trajectory can remain stable when appearance changes.

In the behavior-swapping experiment, the same AU trajectory is applied to different source identities while speech and rigid pose remain fixed. Mean AU-TCorr remains 0.87 across identities.

Cross-appearance examples are shown in **Figure 4**.

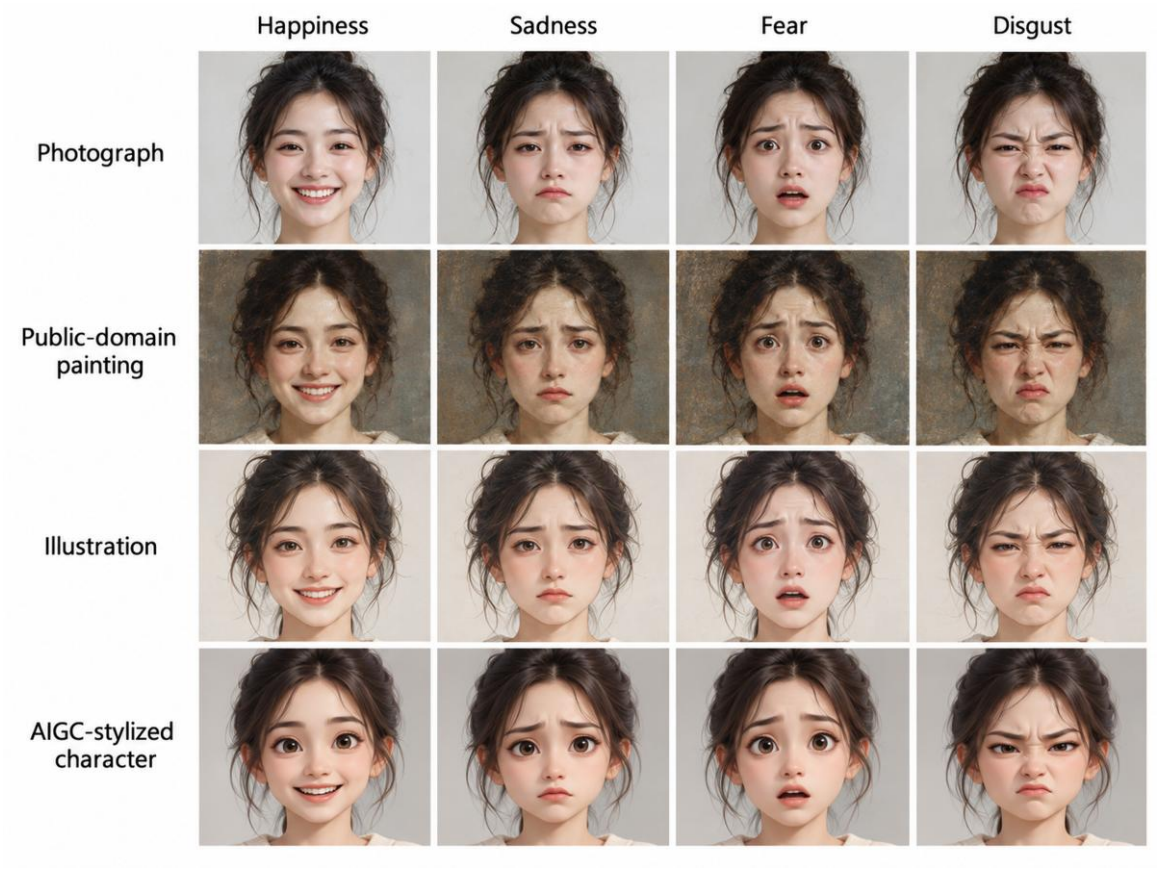


Figure 4. Cross-appearance behavior transfer across photographic portraits, public-domain paintings, illustrations, and AIGC-stylized characters under matched affective behavior. The photograph-like faces are AI-generated synthetic fictional identities; the painting examples are public-domain artworks; no images of recruited study participants are shown.

4.8. Human Perceptual Evaluation

Participant ratings are reported in Table 8.

Table 8. Human perceptual evaluation on a five-point scale

Method	Naturalness	Emotion appropriateness	A/V synchronization	Context appropriateness	Source-appearance preservation
Reference emotional baseline	3.73 ± 0.61	3.51 ± 0.69	4.08 ± 0.55	3.42 ± 0.67	3.82 ± 0.60

Method	Naturalness	Emotion appropriateness	A/V synchronization	Context appropriateness	Source-appearance preservation
Style2Talker [10]	4.05 ± 0.54	4.10 ± 0.57	4.14 ± 0.52	3.71 ± 0.61	4.17 ± 0.51
TABI-DH	4.28 ± 0.48	4.39 ± 0.44	4.19 ± 0.50	4.26 ± 0.49	4.31 ± 0.45

Friedman tests show method effects for naturalness,

$$\chi^2(2) = 29.4, \quad p < 0.001, \quad W = 0.46,$$

emotion appropriateness,

$$\chi^2(2) = 37.1, \quad p < 0.001, \quad W = 0.58,$$

context appropriateness,

$$\chi^2(2) = 35.2, \quad p < 0.001, \quad W = 0.55,$$

and source-appearance preservation,

$$\chi^2(2) = 28.2, \quad p < 0.001, \quad W = 0.44.$$

The method effect for audio–visual synchronization is smaller,

$$\chi^2(2) = 5.1, \quad p = 0.078, \quad W = 0.08.$$

Holm-corrected Wilcoxon comparisons show higher TABI-DH ratings than Style2Talker for emotion appropriateness and contextual appropriateness ($p < 0.01$), whereas their audio–visual synchronization ratings do not differ statistically under the adopted test.

The perceptual results therefore align with the automatic evaluation: the principal benefit lies in affective and contextual behavior rather than lip synchronization.

4.9. Computational Cost

Computational characteristics are summarized in Table 9.

Table 9. Computational characteristics on one RTX 3090

Model	Parameters	GFLOPs/generated frame	Throughput	Peak GPU memory
Reference emotional baseline	101.0 M	1203	24.6 FPS	8.8 GB
TABI-DH	104.6 M	1209	22.8 FPS	9.7 GB

The Temporal AU Behavior Planner increases parameter count by approximately 3.6% and reduces throughput by approximately 7.3%.

At 22.8 FPS, the current implementation approaches but does not reach the 25 FPS evaluation rate on a single RTX 3090. Strict real-time operation is therefore not claimed.

The result indicates a modest performance–cost trade-off: temporal behavior modeling introduces limited computational overhead but still requires optimization for consistently real-time interactive deployment.

4.10. Failure Analysis

Four recurrent failure modes are observed.

Large out-of-plane rotations weaken both AU estimation and landmark projection. Because rigid and non-rigid motion are modeled separately, errors in the pose-control sequence can propagate into the final geometry even when the AU trajectory remains plausible.

Partial occlusion around the eyes or mouth can corrupt localized AU activity and reduce emotion recognizability.

Closely related affective categories remain difficult. Fear and surprise can share strong upper-face activation, while contempt and disgust can exhibit similar cheek, nose, and lip movements.

Highly abstract artistic portraits can violate the approximately human facial topology assumed by the geometry and AU pathways. In these cases, source style may remain recognizable while transferred facial motion becomes locally distorted.

These failure patterns define the practical boundary of the method more clearly than aggregate benchmark averages alone.

5. Discussion

5.1. Temporally Structured AUs Provide a Distinct Behavioral Representation

The central finding is that temporal AU modeling produces a substantially larger effect on behavior-specific measures than on conventional frame-quality metrics. Removing temporal-velocity supervision reduces AU-TCorr from 0.89 to 0.83, while PSNR changes by only 0.08 dB. This divergence shows that frame reconstruction and facial behavior are related but not interchangeable evaluation targets.

A visually plausible sequence can still contain poorly timed facial actions. By explicitly constraining AU velocity and evaluating onset and peak behavior, TABI-DH treats temporal organization as part of the generation target. This perspective is consistent with recent AU-recognition work showing that facial actions exhibit structured temporal and hierarchical relationships [18], while extending AU-based generation from fine-grained control [9,19,20] toward explicit behavioral evaluation.

The small controlled-generation difference between AUHead and TABI-DH is also informative. AUHead already provides strong temporal AU modeling [20], so a large advantage on MEAD would not be theoretically expected. The value of TABI-DH is instead supported by the additional interaction-context and cross-appearance experiments. This distinction avoids equating incremental benchmark improvement with a fundamentally different scientific contribution.

5.2. Interaction Context Mainly Helps When the Current Utterance Is Ambiguous

The IEMOCAP results indicate that current speech does not completely determine an appropriate visible facial response. Previous-turn semantics provide the largest contextual increment, while the previous predicted AU state and previous-turn prosody contribute additional but smaller improvements.

This pattern has a plausible functional interpretation. Semantic history informs what the preceding interaction meant; the previous AU state provides information about the immediately preceding visible affective trajectory; and prosody refines the intensity and manner of the preceding turn. Once semantic and behavioral context have already been incorporated, the marginal contribution of past prosody is smaller.

The improvement is most relevant at affective transitions. During relatively stable neutral exchanges, current speech already provides much of the necessary information. Context becomes more useful when a response depends on an emotional event that occurred in the preceding turn. This finding is compatible with the broader direction of context-sensitive digital-human research [21], while the present study adds an explicit temporal facial-behavior representation rather than only a final expression category.

The interpretation must remain bounded by the experimental design. The model predicts observable non-rigid facial behavior; it does not infer a participant's unobservable psychological state. Moreover, rigid pose is externally controlled. The IEMOCAP experiment therefore supports contextual modulation of facial response conditional on a fixed head-motion trajectory, not autonomous generation of complete embodied conversational behavior.

5.3. Upper-Face Behavior Contributes Affect Without Improving Lip Synchronization

The mechanism ablation further separates affective expression from speech articulation. Removing upper-face AUs decreases emotion accuracy while leaving SyncNet unchanged or slightly higher. This suggests that brow, eyelid, and cheek activity contributes information that is important for emotion differentiation but only weakly coupled to phonetic synchronization.

This result helps explain why systems optimized primarily for lip synchronization can still produce emotionally under-specified faces. Conversely, the present method does not improve every synchronization measure because its main intervention is not a new phonetic mouth-motion model.

The AU-intensity intervention adds a complementary observation. Increasing AU magnitude does not monotonically improve perceived quality. Moderate intensities provide the most favorable balance between recognizability and naturalness, whereas excessive scaling produces localized geometric exaggeration. The AU representation is therefore interpretable not only in terms of which facial regions are active, but also in terms of behaviorally meaningful amplitude.

5.4. Independent AU Evaluation Reduces Teacher–Evaluator Coupling

AU-supervised generation creates a methodological risk when the same network is used for both training supervision and final evaluation. In that setting, a high AU score may partly measure agreement with the training teacher rather than the intended facial behavior.

TABI-DH separates these roles. The AU supervisor used in the loss and the EfficientNet-B0 evaluator used for test metrics have different parameters and optimization histories. The evaluation remains

within the same 24-AU semantic space, so it is not dataset-independent, but it is not a direct re-use of the training teacher.

The persistence of the temporal advantage under this model-independent evaluation strengthens the interpretation of Tables 4 and 6. The improvement is not confined to the exact network used to generate training supervision.

5.5. Behavior–Appearance Factorization Supports Reusable Character Animation

The artistic-domain experiment clarifies the role of behavior–appearance factorization. Style2Talker achieves the strongest appearance-oriented vision–language similarity, which is consistent with its explicit emphasis on artistic style [10]. TABI-DH does not attempt to replace a dedicated stylization model.

Its contribution lies instead in the stability of an explicit behavior trajectory across different visual identities. The behavior-swapping experiment retains an AU-TCorr of 0.87 after source identity changes, while the artistic-domain results preserve source content and appearance to a useful degree.

This property is relevant to AIGC production pipelines. Visual character design and behavioral performance can be treated as partially reusable assets: one affective performance can be transferred among photographic, painterly, illustrated, or stylized characters without reconstructing the behavior from scratch.

The result should nevertheless be interpreted within the tested domain. The method assumes approximately human facial topology. Highly abstract or non-human designs remain outside the demonstrated operating range.

5.6. Generalization, Efficiency, and Practical Boundary

Cross-dataset experiments show that the method retains useful performance on LRW, GRID, and HDTF, but the magnitude of improvement is not uniform. GRID benefits from comparatively controlled acquisition, whereas HDTF introduces greater spatial detail and natural-motion variation. The evidence therefore supports robustness under moderate distribution shift rather than unrestricted domain invariance.

The efficiency analysis reveals a similarly bounded engineering result. Temporal behavior modeling increases parameter count by only 3.6%, but throughput decreases to 22.8 FPS. This cost is modest for offline portrait animation but remains relevant for live interactive systems. Recent real-time portrait-animation research demonstrates that 25 FPS or higher is achievable with architectures explicitly optimized for streaming inference [25]. For TABI-DH, acceleration is therefore a more appropriate next engineering objective than further increasing model complexity.

Failure cases also indicate where additional robustness is required. Large pose changes, occlusion, subtle neighboring emotions, and highly abstract artistic faces remain challenging. These limitations

are consistent with the dependence of AU estimation and sparse facial geometry on sufficiently observable human-like facial structure.

5.7. Human Perception and Practical Significance

The human study provides an important consistency check because perceptual ratings follow the same pattern as the automatic evaluation. Emotion and context appropriateness improve, whereas audio–visual synchronization does not show a comparable difference.

This agreement reduces the likelihood that the principal result is an artifact of one automatic metric. It also reinforces an important methodological point: virtual-human quality should not be summarized by a single realism score. Image quality, speech synchronization, emotion fidelity, temporal facial behavior, contextual appropriateness, and appearance preservation represent distinct properties.

From an application perspective, the present results support TABI-DH for AIGC portrait animation in which controllable affective behavior is important. They do not establish autonomous social intelligence, psychological-state inference, or unrestricted real-time deployment.

5.8. Limitations, Future Research, and Responsible Use

Several limitations define the next stage of the work.

First, rigid head pose is externally controlled. Future research should jointly generate head motion, gaze, and non-rigid affect from conversational context so that the complete behavioral trajectory does not depend on a reference pose sequence.

Second, the 24-dimensional AU representation covers facial behavior but not gaze direction, tongue motion, hand gesture, posture, or whole-body affect. Extending the intermediate representation beyond the face would enable more complete embodied-agent modeling.

Third, the current context window emphasizes immediately preceding interaction. Longer conversations may require discourse state, speaker relationships, persistent behavioral state, and long-term semantic memory.

Fourth, the current 256×256 setup prioritizes controlled methodological comparison rather than maximum rendering resolution. Higher-resolution renderers could improve visual fidelity without changing the core behavioral formulation.

Finally, emotionally expressive virtual humans can be misused for impersonation or deceptive media. Deployment should therefore incorporate appropriate identity consent, provenance records, synthetic-media disclosure, access controls for sensitive identity assets, and technical traceability mechanisms where applicable.

6. Conclusion

This study addressed affective virtual-human generation as a problem of context-conditioned temporal facial-behavior modeling under controlled rigid head pose, rather than treating emotion as an isolated frame-level appearance.

TABI-DH introduces a temporally structured 24-dimensional AU representation between multimodal interaction processing and portrait rendering. The framework combines contextual speech information, a Temporal AU Behavior Planner, AU-guided non-rigid geometry, and behavior–appearance factorized rendering.

The experiments support three main conclusions. Temporal AU modeling improves behavior-specific trajectory and timing measures more clearly than conventional frame-quality metrics, demonstrating the value of explicitly modeling facial dynamics. Past conversational semantics, previous predicted facial state, and preceding-turn prosody improve observable response behavior on session-independent IEMOCAP evaluation under a fixed rigid-pose condition. In addition, a shared AU trajectory remains transferable across photographic and artistic identities while preserving source appearance to a useful degree.

Together, these findings support temporally structured AUs as an interpretable intermediate representation connecting affective interaction context with controllable AIGC portrait animation. Future work should remove dependence on externally supplied pose, extend the behavioral state beyond the face, improve high-resolution and real-time rendering, and increase robustness to large pose changes, occlusion, and highly abstract character styles.

Data Availability

The datasets used in this study are third-party research resources and remain subject to their original licenses and access conditions. Direct dataset access or request pages are provided for MEAD (https://github.com/uniBruce/Mead), LRW (https://www.robots.ox.ac.uk/~vgg/data/lip_reading/lrw1.html), IEMOCAP (https://sail.usc.edu/iemocap/iemocap_release.htm), FEAFA+ (https://www.iiplab.net/feafa%2B/), the GRID Audio-Visual Speech Corpus (https://zenodo.org/records/3625687), and HDTF (https://github.com/MRzzm/HDTF). MEAD is described by Wang et al. [29]. Access-controlled datasets must be requested in accordance with the corresponding provider's license and approval procedure. The artistic portrait evaluation manifest and provenance information, dataset split definitions, IEMOCAP fold definitions, preprocessing settings, evaluation configurations, and numerical summary results supporting the findings are publicly archived at https://zenodo.org/records/22998512.

Consent for Publication

No identifiable information or images of the 32 perceptual-study participants are included in this article. The photograph-like facial images shown in Figures 3 and 4 are AI-generated synthetic fictional identities and do not depict real individuals or study participants. Public-domain painting examples are included as public-domain artworks and are not images of recruited study participants. Therefore, consent for publication of identifying participant information or participant images is not applicable to Figures 3 and 4.

Funding

This research received no external funding.

Author Contributions

Y.S. conceived and designed the study, developed the methodology, conducted the analyses, prepared the figures and tables, and wrote and revised the manuscript. The author reviewed and approved the final manuscript and accepts responsibility for the integrity of the work.

Competing Interests

The author declares no competing interests.

References

[1] H. Nisar, S. Masood, Z. Malik, and A. Abid, “Talking Head Generation Through Generative Models and Cross-Modal Synthesis Techniques,” Journal of Imaging, vol. 12, no. 3, Art. no. 119, 2026, doi: 10.3390/jimaging12030119.

[2] M. Meng, Y. Zhao, B. Zhang, Y. Zhu, W. Shi, M. Wen, and Z. Fan, “A Survey of Talking Head Synthesis Techniques: Portrait Generation, Driving Mechanisms, and Editing,” ACM Computing Surveys, vol. 58, no. 7, Art. no. 188, pp. 1–43, 2026, doi: 10.1145/3785656.

[3] V. K. Rakesh, S. Mazumdar, R. P. Maity, S. Pal, A. Das, and T. Samanta, “Advancements in Talking Head Generation: A Comprehensive Review of Techniques, Metrics, and Challenges,” The Visual Computer, vol. 42, Art. no. 9, 2026, doi: 10.1007/s00371-025-04232-w.

[4] Y. Li and X. Shen, “Audio-Driven Single Image Talking Face Animation with Transformers,” Scientific Reports, vol. 16, Art. no. 3796, 2026, doi: 10.1038/s41598-025-33928-1.

[5] S. Tan, B. Ji, and Y. Pan, “FlowVQTalker: High-Quality Emotional Talking Face Generation through Normalizing Flow and Quantization,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26317–26327.

[6] X. Wang, Y. Huo, Y. Liu, X. Guo, F. Yan, and G. Zhao, “Multimodal Feature-Guided Audio-Driven Emotional Talking Face Generation,” Electronics, vol. 14, no. 13, Art. no. 2684, 2025, doi: 10.3390/electronics14132684.

[7] X. Shen, F. F. Khan, and M. Elhoseiny, “EmoTalker: Audio Driven Emotion Aware Talking Head Generation,” in Computer Vision – ACCV 2024, 2024, pp. 131–147, doi: 10.1007/978-981-96-0917-8_8.

[8] S. Tan, B. Ji, M. Bi, and Y. Pan, “EDTalk: Efficient Disentanglement for Emotional Talking Head Synthesis,” in Computer Vision – ECCV 2024, 2024, pp. 398–416.

[9] Z. Sun, Y. Xuan, F. Liu, and Y. Xiang, “FG-EmoTalk: Talking Head Video Generation with Fine-Grained Controllable Facial Expressions,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 5, pp. 5043–5051, 2024, doi: 10.1609/aaai.v38i5.28309.

[10] S. Tan, B. Ji, and Y. Pan, “Style2Talker: High-Resolution Talking Head Generation with Emotion Style and Art Style,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, no. 5, pp. 5079–5087, 2024, doi: 10.1609/aaai.v38i5.28313.

[11] N. Drobyshev, A. Bigata Casademunt, K. Vougioukas, Z. Landgraf, S. Petridis, and M. Pantic, “EMOPortraits: Emotion-Enhanced Multimodal One-Shot Head Avatars,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 8498–8507.

[12] Y. Xie, H. Xu, G. Song, C. Wang, Y. Shi, and L. Luo, “X-Portrait: Expressive Portrait Animation with Hierarchical Motion Attention,” in ACM SIGGRAPH 2024 Conference Papers, 2024, pp. 1–11, doi: 10.1145/3641519.3657459.

[13] Z. Chen, J. Cao, Z. Chen, Y. Li, and C. Ma, “EchoMimic: Lifelike Audio-Driven Portrait Animations through Editable Landmark Conditions,” Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 3, pp. 2403–2410, 2025, doi: 10.1609/aaai.v39i3.32241.

[14] S. Xu et al., “VASA-1: Lifelike Audio-Driven Talking Faces Generated in Real Time,” in Advances in Neural Information Processing Systems, vol. 37, 2024.

[15] H. Wang et al., “EmotiveTalk: Expressive Talking Head Generation through Audio Information Decoupling and Emotional Video Diffusion,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 26212–26221.

[16] H. Liu, W. Sun, D. Di, S. Sun, J. Yang, C. Zou, and H. Bao, “MoEE: Mixture of Emotion Experts for Audio-Driven Portrait Animation,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 26222–26231, doi: 10.1109/CVPR52734.2025.02442.

[17] Y. Lin, H. Fung, J. Xu, Z. Ren, A. S. M. Lau, G. Yin, and X. Li, “MVPortrait: Text-Guided Motion and Emotion Control for Multi-View Vivid Portrait Animation,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 26242–26252.

[18] Z. Wang, S. Song, C. Luo, S. Deng, W. Xie, and L. Shen, “Multi-Scale Dynamic and Hierarchical Relationship Modeling for Facial Action Units Recognition,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1270–1280.

[19] J. Lyu, X. Lan, G. Hu, H. Jiang, W. Gan, J. Wang, and J. Xue, “Multimodal Emotional Talking Face Generation Based on Action Units,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 35, no. 5, pp. 4026–4038, 2025, doi: 10.1109/TCSVT.2024.3523359.

[20] J. Lyu, L. Qu, W. Zhang, H. Jiang, K. Liu, Z. Zhou, X. Xia, J. Xue, and T.-S. Chua, “AUHead: Realistic Emotional Talking Head Generation via Action Units Control,” in International Conference on Learning Representations (ICLR), 2026.

[21] F. Yang, L. Fang, R. Suo, J. Zhang, and M. Whang, “Development of an Interactive Digital Human with Context-Sensitive Facial Expressions,” Sensors, vol. 25, no. 16, Art. no. 5117, 2025, doi: 10.3390/s25165117.

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, “IEMOCAP: Interactive Emotional Dyadic Motion Capture Database,” Language Resources and Evaluation, vol. 42, no. 4, pp. 335–359, 2008, doi: 10.1007/s10579-008-9076-6.

[23] P. Salehi, S. A. Sheshkal, V. Thambawita, S. Gautam, S. S. Sabet, D. Johansen, M. A. Riegler, and P. Halvorsen, “Comparative Analysis of Audio Feature Extraction for Real-Time Talking Portrait Synthesis,” Big Data and Cognitive Computing, vol. 9, no. 3, Art. no. 59, 2025, doi: 10.3390/bdcc9030059.

[24] S. Torrie, K. Wright, and D.-J. Lee, “MultiAVSR: Robust Speech Recognition via Supervised Multi-Task Audio–Visual Learning,” Electronics, vol. 14, no. 12, Art. no. 2310, 2025, doi: 10.3390/electronics14122310.

[25] D. Zhen, S. Yin, S. Qin, H. Yi, Z. Zhang, S. Liu, G. Qi, and M. Tao, “Teller: Real-Time Streaming Audio-Driven Portrait Animation with Autoregressive Motion Generation,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 21075–21085, doi: 10.1109/CVPR52734.2025.01963.

[26] W. Gan, J. Xue, K. Lu, Y. Yan, P. Gao, and J. Lyu, “FEAFA+: An Extended Well-Annotated Dataset for Facial Expression Analysis and 3D Facial Animation,” Proceedings of SPIE, vol. 12342, pp. 307–316, 2022, doi: 10.1117/12.2643588.

[27] D. Yaman, F. I. Eyiokur, L. Bärmann, S. Akti, H. K. Ekenel, and A. Waibel, “Audio-Visual Speech Representation Expert for Enhanced Talking Face Video Generation and Evaluation,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024, pp. 6003–6013, doi: 10.1109/CVPRW63382.2024.00607.

[28] Y. Tang, Y. Liu, and W. Li, “Continuous Talking Face Generation Based on Gaussian Blur and Dynamic Convolution,” Sensors, vol. 25, no. 6, Art. no. 1885, 2025, doi: 10.3390/s25061885.

[29] K. Wang et al., “MEAD: A Large-Scale Audio-Visual Dataset for Emotional Talking-Face Generation,” in Computer Vision – ECCV 2020, 2020, pp. 700–717, doi: 10.1007/978-3-030-58589-1_42.

[30] K. Liu, J. Wei, R. Yuan, R. Chai, S. He, Z. Ma, and Y. Yang, “Text-Guided Dynamic Mouth Motion Capturing for Person-Generic Talking Face Generation,” Knowledge-Based Systems, vol. 316, Art. no. 113354, 2025, doi: 10.1016/j.knosys.2025.113354.

[31] J. Choi, M. Kim, S. J. Park, and Y. M. Ro, “Text-Driven Talking Face Synthesis by Reprogramming Audio-Driven Models,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 8065–8069, doi: 10.1109/ICASSP48485.2024.10448506.

[32] H. Zhang, Z. Yuan, C. Zheng, X. Yan, B. Wang, G. Li, S. Wu, S. Cui, and Z. Li, “GSmoothFace: Generalized Smooth Talking Face Generation via Fine Grained 3D Face Guidance,” IEEE Transactions on Visualization and Computer Graphics, vol. 31, no. 10, pp. 8231–8242, 2025, doi: 10.1109/TVCG.2025.3566382.

[33] Y. Zhang, H. Yu, W. Liang, and S. Zhang, “Audio-Driven Talking Face Generation with Blink Embedding and Hash Grid Landmarks Encoding,” in 2024 IEEE Smart World Congress (SWC), 2024.

[34] A. T. Akbar, S. Saifullah, H. Prapcoyo, H. Rustamadji, and N. H. Cahyana, “EfficientNet B0 Feature Extraction with L2-SVM Classification for Robust Facial Expression Recognition,” Journal of Information Systems and Informatics, vol. 7, no. 2, pp. 1106–1129, 2025, doi: 10.51519/journalisi.v7i2.1071.

[35] S. Jayasumana, S. Ramalingam, A. Veit, D. Glasner, A. Chakrabarti, and S. Kumar, “Rethinking FID: Towards a Better Evaluation Metric for Image Generation,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 9307–9315.

[36] A. Zhalgas, B. Amirgaliyev, and A. Sovet, “Robust Face Recognition Under Challenging Conditions: A Comprehensive Review of Deep Learning Methods and Challenges,” Applied Sciences, vol. 15, no. 17, Art. no. 9390, 2025, doi: 10.3390/app15179390.

[37] J. Zhang, J. Huang, S. Jin, and S. Lu, “Vision-Language Models for Vision Tasks: A Survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 8, pp. 5625–5644, 2024, doi: 10.1109/TPAMI.2024.3369699.

[38] M. Oquab et al., “DINOv2: Learning Robust Visual Features without Supervision,” Transactions on Machine Learning Research, 2024.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TABIDHManuscriptrevised20261003highlighted.docx