Expressively Controllable Talking Face Generation via Identity-Aware AU Conditioning and Neutral-Reference Debiasing

Tri Ton [1] **Chang D. Yoo** [1]

Abstract

Precise control over expressive dynamics in audio-driven talking face generation remains challenging due to a lack of granular control signals. Existing methods typically rely on coarse categorical emotion labels applied across an entire video, resulting in rigid, unchanging facial poses that lack temporal diversity. To address this, we propose IUFace, a framework that utilizes Action Units (AUs) as a continuous, multimodal conditioning signal for anatomically grounded, frame-level expression control. To ensure these movements translate naturally across diverse subjects, we introduce Identity-Aware AU Conditioning to personalize AU activations based on specific facial structures. Furthermore, our Neutral-Reference Debiasing paradigm explicitly decouples static identity from dynamic expression during training, effectively eliminating expression leakage and reference-pose tethering. Extensive evaluations demonstrate that IUFace achieves state-of-the-art performance in fine-grained expression alignment, temporal coherence, and reliable generation.

1. Introduction

Talking face generation requires the joint synthesis of phonetic synchronization and naturalistic facial affect. While diffusion-based frameworks achieve high-fidelity lip synchronization, controllable synthesis of expressive dynamics remains a bottleneck. Current architectures restrict control mechanisms to coarse emotion labels or text prompts, resulting in either a total absence of affect or a frozen, unchanging smile (Figure 1a). Leading categorical models like DICE-Talk (Tan et al., 2025) exhibit a static expression bias, while non-expressive frameworks like Hallo3 (Cui et al., 2025) fail to reach natural expressive peaks (Figure 1b).

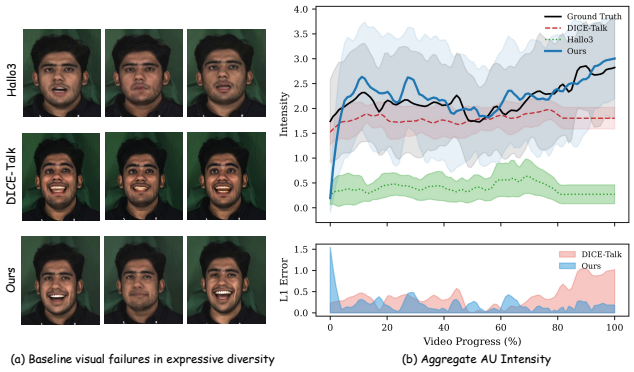

Figure 1. **Aggregate AU intensity analysis.** (a) Qualitative comparison of expressive diversity. (b) AU trajectories and temporal L1 error. Our method recovers natural rise-and-fall dynamics, overcoming the static bias and the absence of expression in baselines.

To transition from rigid, short-term clips to controllable video generation, models require richer, compositional control signals. We propose utilizing Facial Action Units (AUs) as a structured, anatomically grounded multimodal control signal. Unlike text or keyframes, AUs provide a continuous temporal manifold for frame-level expression authoring, allowing creators to explicitly steer subtle micro-expressions and the natural rise-and-fall of human facial dynamics. Realizing this control, however, is challenged by identity-expression entanglement. As shown in Figure 2, existing models over-rely on the reference identity; when provided with a neutral reference, they produce attenuated results that fail to reach the target expressive peaks.

To solve this, we introduce IUFace, a framework for fine-grained, identity-consistent expressive control. We propose *Identity-Aware AU Conditioning* to residually modulate AU activations with identity-specific features, ensuring that the same AU control signal adapts naturally to individual facial geometries. Furthermore, we introduce a *Neutral-Reference Debiasing* paradigm during training to explicitly decouple static identity from dynamic motion, effectively eliminating reference-pose tethering and expression leakage.

In summary, our work makes two main contributions. First, we formulate a continuous, Identity-Aware AU control mechanism that positions Action Units as a powerful multimodal signal for personalized, frame-level expressive video

[1]Korea Advanced Institute of Science and Technology (KAIST). Correspondence to: Chang D. Yoo <cd_yoo@kaist.ac.kr>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

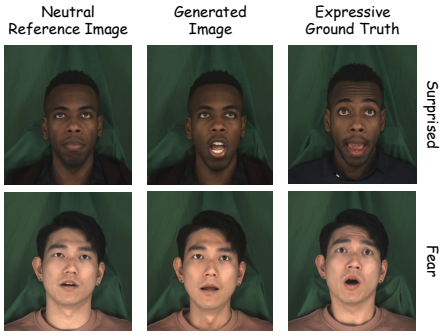

Figure 2. **Failures in Expression Disentanglement.** Prior methods over-rely on reference expressions. Given a neutral reference (left), baselines (middle) fail to capture the high-intensity dynamics of the ground truth (right).

directing. Second, we introduce a Neutral-Reference Debiasing paradigm that explicitly eliminates identity-expression entanglement. Together, these advancements achieve state-of-the-art temporal coherence and visual fidelity without dataset-induced reference leakage.

2. Methodology: IUFace

Figure 3 overviews the IUFace framework. To support reliable, long-horizon talking face generation, audio is processed via wav2vec (Baevski et al., 2020) and refined through a timestep-aware vocal adapter (Tu et al., 2025), mitigating latent distribution drift over extended sequences. For fine-grained, multimodal control, we utilize continuous Action Unit (AU) trajectories. These AU features are embedded, modulated with identity-specific priors, and injected into the generative process. Concurrently, we introduce a neutral-reference debiasing strategy during training to strictly disentangle the subject's static identity from dynamic expressive biases.

Network Architecture. To ensure high-fidelity motion synthesis, we build upon the Wan2.1 1.3B Video Diffusion Transformer (Wan et al., 2025). Rather than relying on coarse text prompts, we adapt the transformer blocks to accept rich, multimodal conditioning by introducing dedicated cross-attention layers for our continuous AU signals alongside the audio embeddings. By isolating the control pathways for audio (driving phonetic sync) and AUs (driving expressive facial dynamics), the architecture provides precise, decoupled control over the final video. Full architectural specifications are provided in the *Suppl* Sec. B.

Identity-Aware AU Conditioning. To enable fine-grained, multimodal control over facial dynamics, we utilize AU trajectories. Each AU vector $a \in \mathbb{R}^{35}$ encodes the intensity of anatomically defined muscle groups (Ekman & Friesen, 1978). Unlike discrete emotion labels, AUs provide a continuous, composable representation of movement. These features are embedded via a lightweight adapter: $a' = \phi_1(a)$, where ϕ_1 is a two-layer MLP.

While AU conditioning effectively drives motion, it is inherently identity-invariant. To ensure anatomically plausible generation, the same AU control signal must produce distinct morphological results based on the subject's unique facial geometry. We achieve this via identity-aware modulation. Rather than relying on standard face recognition embeddings optimized for discrimination, we leverage CLIP features (Radford et al., 2021). We posit that these rich semantic priors provide a superior structural guide for generative models to reconstruct facial textures during dynamic motion. Let $f_{\text{id}} = \text{MeanPool}(\text{CLIP}(I_{\text{ref}})) \in \mathbb{R}^{2048}$ denote the global identity embedding, projected as: $f'\text{id} = \phi_2(f\text{id})$, with ϕ_2 as a two-layer MLP. We then modulate the AU embedding residually:

$$c_a = a' + f'\text{id}. \tag{1}$$

This fusion grounds the AU-driven dynamics in the specific identity (e.g., producing person-specific smile curvatures for a given AU12 intensity). The refined control signal is injected into the video latents via cross-attention: $x_{\text{AU}} = \text{Attention}(x, c_a)$. While we benchmark using video-extracted AUs, the system supports manual expression authoring via predefined 1D templates (e.g., a smile-intensity curve mapped to AU12), allowing creators to steer facial dynamics without a source video. To ensure stable early training, these attention projections are zero-initialized. Finally, the modulated AU features are integrated with the spatial and phonetic pathways:

$$x' = x + x_{\text{img}} + x_{\text{vocal}} + x_{\text{AU}}, \tag{2}$$

yielding expressive, temporally synchronized, and identity-consistent facial generation.

Expression Debiasing via Neutral-Reference Training. When training on expressive datasets like MEAD (Wang et al., 2020), models frequently entangle AU-specific facial priors with the underlying identity representation. Consequently, conditioning on an expressive reference frame causes the model to overfit, leaking residual expressions (e.g., persistent smiling) even during neutral speech. To explicitly decouple static identity from dynamic expression, we introduce a *Neutral-Reference Debiasing* strategy.

During training, rather than using an arbitrary frame from the target sequence, we randomly sample a neutral-expression frame of the same subject to serve as the identity anchor. Formally, given a neutral reference $I_{\text{ref}}^{\text{neutral}}$ and an expressive target video, the model learns the mapping:

$$\hat{V} = \mathcal{F}(I_{\text{ref}}^{\text{neutral}}, \mathbf{C}_{\text{AU}}, \mathbf{C}_{\text{vocal}}), \tag{3}$$

where $\mathbf{C}_{\text{AU}}$ and $\mathbf{C}_{\text{vocal}}$ are the target AU and audio trajectories. This forces the network to derive all facial motion exclusively from the multimodal control signals rather than

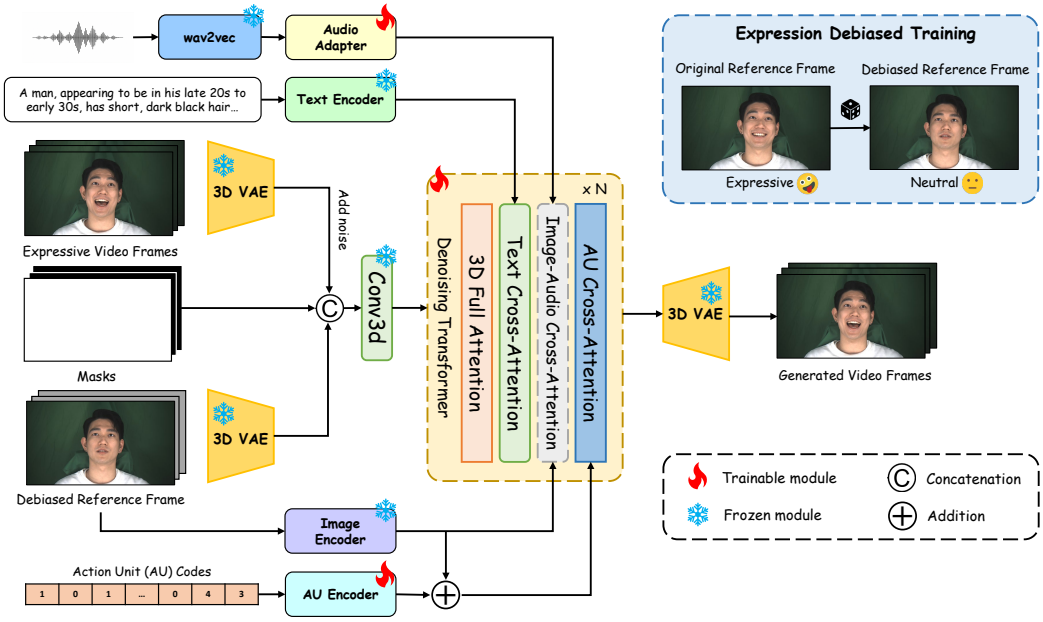

Figure 3. **Overview of IUFace. Training:** A debiased neutral reference anchors the identity, while audio and AU trajectories from an expressive video drive the dynamics. AU embeddings are fused with CLIP identity priors and injected through cross-attention.

the spatial reference, effectively eliminating dataset-induced expression biases and reference-pose tethering.

Classifier-Free Guidance. To provide creators with granular, decoupled control over emotional intensity and audio synchronization at inference time, we adopt a dual-scale classifier-free guidance (CFG) strategy. Let $\mathbf{v}_\theta^{\text{uncond}}$, $\mathbf{v}_\theta^{\text{au}}$, and $\mathbf{v}_\theta^{\text{cond}}$ denote the network's predictions under unconditional, AU-only, and fully conditioned (audio and AU) settings, respectively. The final extrapolated prediction is:

$$\mathbf{v}_\theta = \mathbf{v}_\theta^{\text{uncond}} + w_{au}(\mathbf{v}_\theta^{\text{au}} - \mathbf{v}_\theta^{\text{uncond}}) + w_a(\mathbf{v}_\theta^{\text{cond}} - \mathbf{v}_\theta^{\text{au}}), \quad (4)$$

where w_{au} and w_a dictate the respective guidance strengths. This formulation allows users to independently scale structural expression (via w_{au}) and phonetic precision (via w_a). Empirically, $w_{au} = 3.0$ and $w_a = 5.0$ at 50 sampling steps provide optimal anatomical coherence and sharp identity preservation without artifacts.

3. Experiments

Experimental Setup. We train our model on 16,161 high-quality videos from the MEAD dataset (Wang et al., 2020), reserving 182 videos for testing to ensure zero identity overlap. We fine-tune the Wan2.1 1.3B DiT backbone (Wan et al., 2025) on 4 NVIDIA A100 GPUs for 40,000 iterations. We evaluate generation quality using PSNR and SSIM, distributional similarity via FVD and FID, emotional fidelity via E-FID and AU-F1, and lip-sync via SyncNet. We benchmark against leading audio-driven and categorical models: Hallo3 (Cui et al., 2025), Sonic (Ji et al., 2025), StableAvatar (Tu et al., 2025), and DICE-Talk (Tan et al., 2025). Extended implementation details and cross-dataset evaluations are provided in the *Suppl* Sec. B and Sec. D.

Quantitative Results. To rigorously assess expressive controllability without reference-pose tethering, we use the **Disentanglement Benchmark**: a neutral reference image is paired with audio and AU trajectories from a different expressive video of the same subject. As shown in Table 1, IUFace achieves state-of-the-art performance across nearly all metrics. We obtain the lowest FVD and FID, indicating superior temporal coherence. Critically, our method leads in E-FID and AU-F1, demonstrating unmatched precision in reproducing fine-grained expression dynamics via AU control without leaking identity biases. Results on the HDTF *Generalization Benchmark*, which confirm our framework's robust transferability to in-the-wild subjects and unconstrained backgrounds, are detailed in the *Suppl* Sec. D.

Qualitative Results. Visualizations in Figure 4 compare methods in the challenging neutral-reference setting. Baselines like Hallo3 and StableAvatar exhibit blurry textures, while Sonic and DICE-Talk suffer from head jitter and asymmetric expressions. Conversely, IUFace captures subtle micro-expressions and anatomically plausible deformations, such as symmetric cheek raising (AU6) and lip corner pulling (AU12). Our framework maintains sharp identity details and coherent temporal flow across intense expressions.

Ablation Studies. We evaluate core components in Table 2 and Figure 5. Removing Identity-Aware AU Conditioning (w/o IDA) reverts the system to an audio-driven model, causing weak, generic expressions; its inclusion significantly improves E-FID and F1-AU. Removing Emotion Debias-

Method	FVD ↓	FID ↓	E-FID ↓	PSNR ↑	SSIM ↑	F1-AU ↑	SynC ↑
Hallo3 (CVPR'25) (Cui et al., 2025)	277.79	64.53	4.5369	21.86	0.8292	0.4549	2.5879
Sonic (CVPR'25) (Ji et al., 2025)	192.60	35.02	2.2718	21.14	0.8227	0.4869	2.1125
StableAvatar (arXiv) (Tu et al., 2025)	221.89	37.95	2.5441	22.52	0.8513	0.4694	2.9808
DICE-Talk (ACM MM'25) (Tan et al., 2025)	192.87	37.30	2.3439	21.65	0.8318	0.5289	2.9063
IUFace (Ours)	190.62	34.11	2.0381	24.03	0.8690	0.5895	2.9132

Table 1. **Quantitative comparisons in the *Disentanglement* Benchmark.** We evaluate our framework on MEAD samples. We use colors to denote the first, second respectively.

Figure 4. **Qualitative results in the *Disentanglement* Benchmark.** A neutral reference image from a subject's neutral video is paired with audio and AU trajectories from a different expressive video.

IDA	DBE	FVD ↓	E-FID ↓	F1-AU ↑	SynC ↑
	✓	278.14	3.6821	0.4923	2.1876
✓		224.73	2.4195	0.5618	2.5876
✓	✓	190.62	2.0381	0.5895	2.9132

Table 2. **Ablation study on core components on MEAD test set.**

ing (w/o DBE) means training with expressive references, which leads to expression leakage and identity drift. The full configuration achieves the best balance across all metrics, confirming that AU control and neutral-reference anchoring are essential for high-fidelity, disentangled synthesis. Further ablations on the Identity-Aware module and multi-level intensity training are provided in the *Suppl* Sec. E.

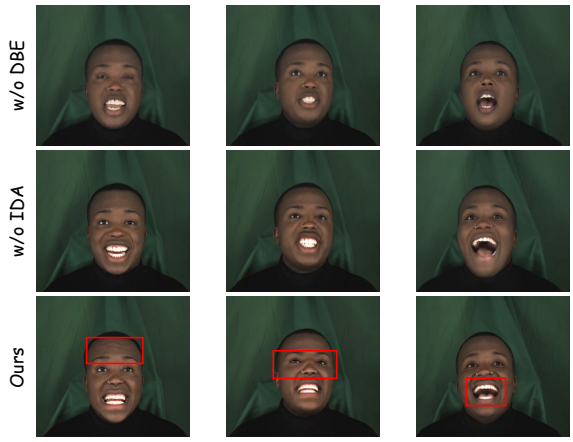

Figure 5. **Ablation on IDA and debiasing DBE. Middle:** w/o IDA → weak, generic expressions. **Top:** w/o DBE → expression leakage, identity drift. **IUFace:** sharp, personalized, debiased.

4. Conclusion

We presented IUFace, a novel framework for expressive talking face generation that achieves fine-grained control over facial dynamics while preserving strong identity consistency. By leveraging Action Units (AUs) as continuous, multimodal control signals and introducing identity-aware AU conditioning, our method enables personalized expression synthesis, overcoming the limitations of coarse categorical labels and text prompts. Furthermore, our neutral-reference debiasing strategy effectively disentangles static identity from dynamic expression, eliminating dataset-induced biases such as persistent smiling. Extensive experiments demonstrate that IUFace achieves state-of-the-art temporal coherence, expressive fidelity, and robust generalization.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 2020.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Chang, D., Yin, Y., Li, Z., Tran, M., and Soleymani, M. Libreface: An open-source toolkit for deep facial expression analysis. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 8205–8215, 2024.

Chen, Z., Cao, J., Chen, Z., Li, Y., and Ma, C. 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.

Cui, J., Li, H., Yao, Y., Zhu, H., Shang, H., Cheng, K., Zhou, H., Zhu, S., and Wang, J. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024.

Cui, J., Li, H., Zhan, Y., Shang, H., Cheng, K., Ma, Y., Mu, S., Zhou, H., Wang, J., and Zhu, S. Hallo3: Highly dynamic and realistic portrait image animation with video diffusion transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Ekman, P. and Friesen, W. V. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Ji, X., Hu, X., Xu, Z., Zhu, J., Lin, C., He, Q., Zhang, J., Luo, D., Chen, Y., Lin, Q., et al. Sonic: Shifting focus to global audio perception in portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. *ICCV*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 2021.

Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Tan, W., Lin, C., Xu, C., Xu, F., Hu, X., Ji, X., Zhu, J., Wang, C., and Fu, Y. Disentangle identity, cooperate emotion: Correlation-aware emotional talking portrait generation. *arXiv preprint arXiv:2504.18087*, 2025.

Team, G., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Tu, S., Pan, Y., Huang, Y., Han, X., Xing, Z., Dai, Q., Luo, C., Wu, Z., and Jiang, Y.-G. Stableavatar: Infinite-length audio-driven avatar video generation. *arXiv preprint arXiv:2508.08248*, 2025.

Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.-W., Chen, D., Yu, F., Zhao, H., Yang, J., et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Wang, K., Wu, Q., Song, L., Yang, Z., Wu, W., Qian, C., He, R., Qiao, Y., and Loy, C. C. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European conference on computer vision*, 2020.

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

Xu, M., Li, H., Su, Q., Shang, H., Zhang, L., Liu, C., Wang, J., Yao, Y., and Zhu, S. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024.

Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

Zhang, Z., Li, L., Ding, Y., and Fan, C. 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*, 2021.

This supplementary material provides comprehensive technical details, an extended literature review, and expanded experimental results to support the main paper. Specifically, Section A provides a broader discussion of related works. Section B outlines our network architecture and implementation details, while Section C offers a comparative analysis of model parameters and computational efficiency. Further empirical evidence is detailed in subsequent sections: Section D presents cross-dataset generalization results on the HDTF dataset, Section E provides additional granular ablation studies, and Section F includes extended standard-setting evaluations on the MEAD dataset. Finally, we discuss current framework limitations in Section G and address ethical considerations regarding the responsible deployment of expressive generative models in Section H.

A. Related Works

Diffusion-based Video Generation. Diffusion models have transformed video synthesis by scaling image priors to temporal domains. Early works like Make-A-Video (Singer et al., 2022) and AnimateDiff (Guo et al., 2023) adapt diffusion via spatial-temporal factorization and motion modules, enabling zero-shot animation from unlabeled videos. Latent diffusion further improves efficiency. Stable Video Diffusion (Blattmann et al., 2023) uses staged pretraining with curated data to achieve state-of-the-art text-to-video and image-to-video quality. Recent large-scale Diffusion Transformers (DiTs) (Peebles & Xie, 2023; Yang et al., 2024; Wan et al., 2025; Kong et al., 2024) push boundaries with billion-parameter models and 3D causal VAEs to enable high-fidelity, controllable generation. These advancements in scalable motion modeling directly inspire modern talking head synthesis, where DiT backbones are conditioned on audio and fine-grained signals to produce expressive, temporally coherent facial dynamics.

Talking Face Generation. Audio-driven talking face generation has progressed from fine-tuning image-based diffusion models with temporal modules to leveraging pretrained video backbones for enhanced realism and scalability. (Wei et al., 2024; Xu et al., 2024; Cui et al., 2024; Chen et al., 2025) adapt AnimateDiff (Guo et al., 2023) by inserting temporal modules and editable landmark conditioning, enabling synchronized lip and expression animation without intermediate representations. More recent methods exploit large video diffusion transformers: Hallo3 (Cui et al., 2025) introduces identity reference networks and causal 3D VAEs for dynamic, non-frontal portrait animation. Sonic (Ji et al., 2025) enhances global audio perception via intra/inter-clip disentanglement and context-aware learning. StableAvatar (Tu et al., 2025) enables infinite-length generation through timestep-aware audio adapters and sliding-window fusion to prevent latent drift. To boost expressiveness, (Cui et al., 2024; Tan et al., 2025) incorporate prompts and emotion labels, but suffer from coarse control and identity leakage. In contrast, our work achieves fine-grained, identity-aware emotional controllability via anatomically precise AU conditioning and neutral-reference debiasing.

B. Implementation details

The Denoising Transformer Network consists of a 30-layer DiT backbone adapted from Wan2.1-Fun-V1.1-1.3B (Wan et al., 2025). We increase the input channel dimension to 36 to accommodate the concatenation of masked latents and binary mask channels. The transformer uses hidden dimension 1536, 12 attention heads, intermediate FFN size 8960, and 3D patching of size $1\times2\times2$. Training is performed with Flow-Matching Euler scheduler, mixed-precision bf16, effective batch size 4 (1 per GPU) on 4×A100-80GB, learning rate 2×10^{-5}, and AdamW optimizer.

During training, we employ a reconstruction loss in latent space. To better supervise the most challenging facial regions, we derive two binary masks from MediaPipe (Lugaresi et al., 2019): a precise inner-lip mask and a broader full-face mask. At each training iteration, with probability 40% we apply no regional weighting, with 10% probability we weight only the full-face region, and with 50% probability we weight only the lip region by element-wise multiplying the per-pixel squared error with the corresponding mask before global averaging. This stochastic region-focused supervision substantially improves lip synchronization accuracy and upper-face expressiveness.

Text prompts, generated by Gemini (Team et al., 2023), are strictly limited to describing static appearance (e.g., hair color, lighting) to ensure that expressive dynamics are driven exclusively by AU and audio conditioning without confounding linguistic cues.

Method	**Params (B)**	**GPU Mem (GB)**	**Speed (s)**	**F1-AU↑**
Hallo3 (Cui et al., 2025)	14.50	49.77	967.00	0.4549
Sonic (Ji et al., 2025)	1.58	19.99	118.26	0.4869
StableAvatar (Tu et al., 2025)	1.75	30.53	578.10	0.4694
DICE-Talk (Tu et al., 2025)	1.63	22.18	176.94	0.5289
IUFace (Ours)	1.89	31.44	816.59	0.5895

Table 3. Model size and inference efficiency on an A100-80GB GPU (81 frames).

C. Parameters and Runtime Comparison

Among methods with fewer than 2B parameters, IUFace achieves the highest F1-AU score by a clear margin while keeping both parameter count and memory footprint highly competitive. The increased inference time primarily stems from the deliberate design choice of using full-attention 3D transformer blocks with explicit AU cross-attention and identity-aware conditioning, which are essential for the fine-grained emotional control and identity preservation demonstrated throughout our results in the main paper and supplementary videos. This represents a favorable trade-off between expressive quality and efficiency for applications that prioritize realism and controllability over performance.

D. HDTF Generalization Results

To evaluate the generalizability of our framework to in-the-wild subjects and unconstrained backgrounds, this section presents our cross-dataset evaluation on the HDTF dataset (Zhang et al., 2021).

Method	**FVD ↓**	**FID ↓**	**E-FID ↓**	**PSNR ↑**	**SSIM ↑**	**F1-AU ↑**	**SynC ↑**
Hallo3 (CVPR'25) (Cui et al., 2025)	153.39	13.96	0.2789	20.85	0.7735	0.6342	3.9839
Sonic (CVPR'25) (Ji et al., 2025)	163.18	12.16	0.2981	21.14	0.7795	0.6385	5.0812
StableAvatar (arXiv) (Tu et al., 2025)	167.40	10.59	0.2378	20.11	0.7727	0.5788	5.0727
DICE-Talk (ACM MM'25) (Tan et al., 2025)	178.50	12.15	0.2415	21.28	0.7848	0.6420	5.3501
IUFace (Ours)	147.10	10.40	0.2254	21.73	0.7841	0.6500	5.1650

Table 4. **Quantitative comparisons in the *Generalization* Benchmark.** We evaluate our framework on the full HDTF dataset.

Quantitative Results. Quantitative results on HDTF in Table 4 demonstrate the robust transferability of our framework to in-the-wild identities and unconstrained backgrounds. We achieve the best temporal coherence and visual quality with the lowest FVD and FID scores. Notably, our method sets a new state of the art in E-FID and F1-AU, indicating that identity-aware AU conditioning effectively preserves expression fidelity during motion. While DICE-Talk (Tan et al., 2025) and Sonic (Ji et al., 2025) maintain slight leads in SSIM and SynC, our framework achieves the highest PSNR and competitive performance across all auxiliary metrics. These results confirm that leveraging CLIP-based semantic structural priors successfully resolves the trade-off between phonetic synchronization and anatomical consistency. These metrics remain stable when evaluated on the full HDTF dataset, confirming the statistical reliability of our generalization claims.

Qualitative Results. Figure 6 illustrates qualitative comparisons on the HDTF dataset, where the reference image and driving signals are sampled from the same expressive sequence. While existing methods often struggle with unconstrained backgrounds, IUFace maintains superior structural integrity and temporal coherence. Specifically, Hallo3 (Cui et al., 2025) and StableAvatar (Tu et al., 2025) exhibit noticeable blurring in the perioral region and tend to produce rigid facial dynamics. Although Sonic (Ji et al., 2025) and DICE-Talk (Tan et al., 2025) capture motion more effectively, they occasionally suffer from asymmetric mouth deformations and unnatural head orientations during high-intensity speech. In contrast, our framework employs identity-aware AU conditioning to preserve subject-specific traits, such as unique smile curvatures and eye geometry, while mitigating identity drift.

E. Detailed Ablation Studies

This section extends the core ablation study presented in the main paper by providing two additional analyses on our finer-grained design choices.

Figure 6. **Qualitative results in the *Generalization* Benchmark.** A reference image is paired with audio and AU trajectories from the same expressive video.

	FVD ↓	E-FID ↓	F1-AU ↑	SynC ↑
w/o IA	267.38	3.5914	0.5012	2.3127
w IA	190.62	2.0381	0.5895	2.9132
Level 3 Only	231.19	2.8472	0.5421	2.6014
All Levels	190.62	2.0381	0.5895	2.9132

Table 5. **Quantitative results of ablation study on detailed modules on MEAD test set.**

E.1. Identity-Aware Modulation

Table 5 presents the impact of the Identity-Aware (IA) module. Without IA, the model utilizes AU conditioning but lacks subject-specific modulation, effectively removing the identity features from the expression control logic. This configuration leads to identity drift and visibly degraded facial dynamics. Introducing IA markedly improves identity preservation and expression fidelity, demonstrating that modeling subject-specific facial priors is essential for high-quality talking head synthesis.

E.2. Training Intensity Diversity

The MEAD dataset categorizes emotion intensities from level 1 (mild) to level 3 (strong). Training exclusively on extreme expressions (Level 3 Only) restricts the model to high muscle activations, causing identity entanglement and poor subtle dynamics. Transitioning to All Levels enables a continuous mapping of AU intensities, which significantly improves E-FID and F1-AU as shown in Table 5. This multi-level strategy allows the model to decouple static identity from emotional variance and capture subtle facial movements.

E.3. Independent AU Evaluation

To verify that the observed performance gains are not artifacts of utilizing OpenFace for both training and evaluation, we assess the generated expressions using an independent AU estimator. As shown in Table 6, IUFace maintains a decisive margin over all baselines under LibreFace (Chang et al., 2024), confirming true anatomical accuracy independent of OpenFace's algorithmic biases.

Method	LibreFace F1-AU	OpenFace F1-AU
Sonic (CVPR'25)	0.4512	0.4869
DICE-Talk (ACM MM'25)	0.4934	0.5289
IUFace (Ours)	0.5608	0.5895

Table 6. LibreFace vs. OpenFace AU-F1 on MEAD test set. IUFace maintains a decisive margin under an independent estimator.

E.4. Identity Encoder Ablation

To evaluate our design choices, we compare the use of CLIP versus ArcFace as the identity embedding within the Identity-Aware AU Conditioning module.

Identity Embedding	FVD ↓	PSNR ↑	SSIM ↑
ArcFace (ResNet-50)	215.42	22.80	0.831
CLIP ViT-L/14 (Ours)	190.62	24.03	0.869

Table 7. CLIP vs. ArcFace identity embeddings on MEAD Disentanglement Benchmark.

CLIP significantly outperforms ArcFace because high-fidelity generative reconstruction requires rich semantic texture priors rather than purely discriminative features. Because ArcFace is trained specifically to discard these intra-class variations, relying on it leads to flattened, less detailed facial textures during generation.

E.5. Identity Preservation Under Debiasing

We measure ArcFace cosine similarity between generated frames and the identity reference. Under the neutral-reference setting, similarity is 0.82; under the expressive-reference setting, it is 0.81. This minimal 0.01 variance confirms that Neutral-Reference Debiasing successfully locks identity representation independent of the driving expression intensity.

F. More MEAD Results

As shown in Table 8, our framework achieves state-of-the-art performance across all metrics on MEAD under the standard setting, where the first frame of the target video serves as the reference image. We obtain the lowest FVD and E-FID, confirming superior temporal dynamics and expression fidelity. Our method also leads in PSNR, SSIM, and AU-F1, demonstrating unmatched structural quality and fine-grained expression control. While FID is competitive, the overall gains, particularly in LPIPS and Sync, validate the robustness of our identity-aware AU conditioning even when expressive references are available. These results confirm general applicability across both debiased and standard training paradigms.

Figure 7 presents qualitative comparisons under the standard setting, where the first frame of the target video serves as the reference image. Our method continues to outperform baselines, generating sharper facial details, more natural mouth deformations, and smoother expression transitions even when the reference contains expressive cues. Hallo3 (Cui et al., 2025), and StableAvatar (Tu et al., 2025) produce overly rigid expressions and blurred lip regions, while Sonic (Ji et al., 2025) and DICE-Talk (Tan et al., 2025) frequently exhibit asymmetric mouth deformations and unnatural head jitter when attempting to transition from the initial expressive pose. These comparisons confirm that IUFace consistently outperforms all evaluated baselines in maintaining high-fidelity, anatomically coherent facial dynamics under the standard setting.

G. Limitations

Our framework faces challenges in generating multiple humans in reference images, as the identity-aware module currently processes one subject at a time. Extending this to multi-person scenes requires robust identity disentanglement and spatial attention mechanisms, which we leave for future exploration. The use of the diffusion model results in lower operational efficiency, with slower inference times due to the computational complexity of the denoising process. We plan to investigate latent consistency models, distillation into flow-based generators to achieve sub-second inference while preserving fidelity.

Method	FVD ↓	FID ↓	E-FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓	F1-AU ↑	SynC ↑
Hallo3 (CVPR'25) (Cui et al., 2025)	137.57	51.01	1.9575	24.23	0.8533	0.2529	0.6627	3.1806
Sonic (CVPR'25) (Ji et al., 2025)	109.71	18.72	1.2161	24.07	0.8912	0.1055	0.6703	2.7603
StableAvatar (arXiv) (Tu et al., 2025)	156.90	22.78	1.4898	26.12	0.8946	0.1455	0.6262	3.1075
DICE-Talk (ACM MM'25) (Tan et al., 2025)	139.27	24.12	1.8593	25.12	0.8968	0.1105	0.6680	3.2597
IUFace (Ours)	97.73	22.40	1.0927	28.79	0.9210	0.1205	0.6918	3.1092

Table 8. **Quantitative comparisons in the second setting.** We evaluate our framework on MEAD samples.

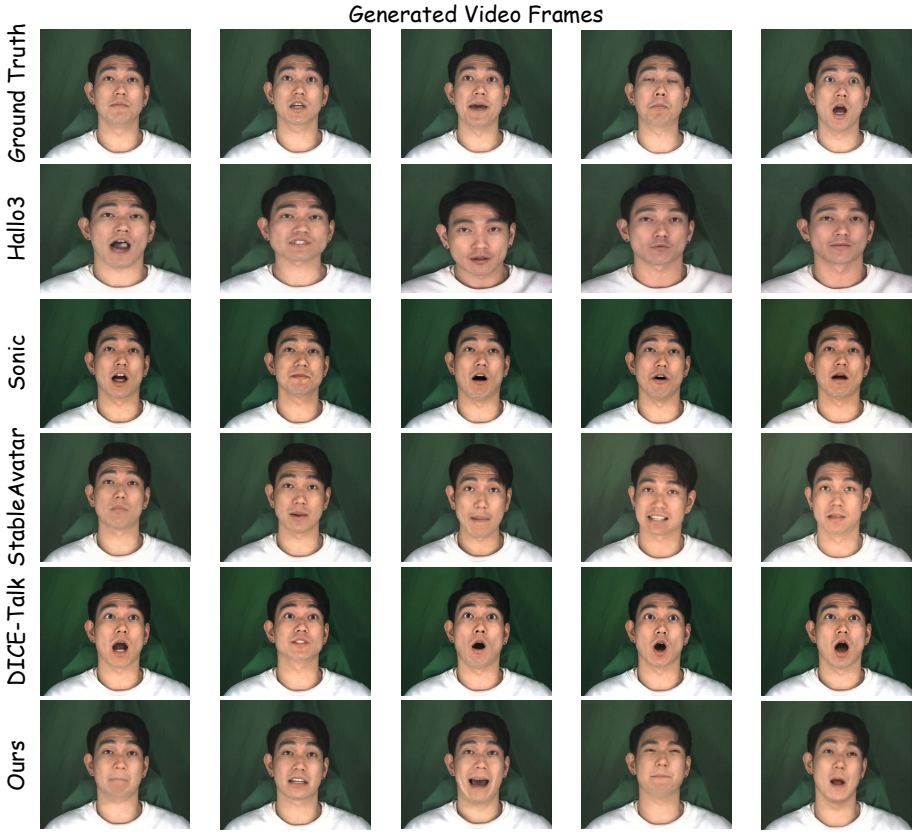

Figure 7. **Qualitative results in the MEAD dataset** A neutral reference image is paired with audio and AU trajectories from the same expressive video.

H. Ethical Concern

Our method synthesizes highly realistic talking-face videos from a single reference image using arbitrary driving audio and fine-grained Action Unit trajectories. This capability supports valuable applications, including virtual avatars, automated dubbing, film post-production, and assistive technologies for speech-impaired individuals. At the same time, it introduces the risk of misuse for creating non-consensual videos of real people, impersonation, or spreading misinformation through synthetic media. To minimize potential societal harm, we strongly recommend that any practical deployment incorporate robust deepfake detection models specifically trained to recognize signatures of AU-conditioned synthesis, as well as imperceptible authentication watermarks embedded in the generated videos.