

# HM-Talker: Hybrid Motion Modeling for High-Fidelity Talking Head Synthesis

## Supplementary Material

### Overview

This supplementary document provides additional materials that complement the main paper. We first extend the discussion of **Related Work** (Sec. 1), offering deeper context on talking head synthesis and explicit motion representations. We then provide detailed **Implementation Information**, including the preprocessing pipeline (Sec. 2.1) and network architectures (Sec. 2.2). Subsequently, we present **Additional Experiments**, featuring comprehensive ablation studies (Sec. 3.1) that further validate our model design. This section also demonstrates the application of our method to a **few-shot personalization framework** (Sec. 3.2) and includes qualitative results on diverse identities (Sec. 3.3). Finally, we offer further **Discussions** (Sec. 4) on the benefits of our personalized motion modeling and different inference modes.

## 1. Related Work

### 1.1. Audio-driven Talking Head Synthesis

Talking head synthesis aims to model the intricate relationship between speech and facial motion, ensuring that generated faces articulate speech with realistic expressiveness and temporal coherence. Early 2D-based methods [18, 19, 27, 28] directly mapped acoustic features to lip textures through generative networks, yet lacked geometric reasoning, often producing rigid or implausible expressions.

The emergence of 3D-aware representations such as Neural Radiance Fields (NeRF) [16] and 3D Gaussian Splatting (3DGS) [9] has significantly improved realism and view consistency in talking head generation [11]. AD-NeRF [7] first projected audio features into dynamic radiance fields to drive head and torso motion, while later works optimized efficiency through multi-resolution hashing [21] and structured motion encoding [12]. However, their fully implicit audio-to-motion mappings limit interpretability and local articulation control.

To improve speech-articulation alignment, recent works have incorporated explicit motion supervision. SyncTalk [17] learns audio-visual synchronization through a cross-modal encoder, while TalkingGaussian [13] decomposes facial regions, controlling upper-face motion via Action Units (AUs) and modeling lips implicitly from audio. PointTalk [23] further captures fine-grained articulation through point clouds but remains implicit for lower-face motion. These studies reveal a growing trend toward hybrid strategies that combine explicit control with data-driven flexibility.

Explicit motion representations are also widely adopted as intermediate priors in generalizable talking head frameworks [8, 10, 15, 20, 22, 24, 29]. For instance, SadTalker [25] maps audio to 3DMM coefficients for disentangled 3D rendering, while GAGAvatar [3] and MG-GTalk [6] employ 3DMM-guided reconstruction to generalize across identities. Yet, their reliance on video-driven or low-accuracy audio-predicted 3DMM limits fine-grained expressiveness and personalized motion synthesis.

### 1.2. Explicit Motion Representations

Explicit motion features provide interpretable and anatomically grounded representations of facial dynamics, enabling precise motion control and improved cross-modal supervision.

**3D Morphable Model (3DMM).** 3DMM [1] encodes facial geometry via low-dimensional shape and expression bases derived from PCA over 3D scans. These coefficients offer a compact and continuous representation widely adopted for motion control [3, 6, 25], but are often limited in capturing subtle, non-linear dynamics such as muscle contractions and micro-expressions.

**Action Unit (AU).** Based on the Facial Action Coding System [4], AUs represent muscle activations through semantically meaningful, anatomically localized units. Each AU corresponds to a distinct movement (e.g., AU12 for lip corner pulling), providing a physiologically interpretable bridge between visual cues and speech articulation. This property makes AUs especially effective for synchronizing audio-driven facial motion [2, 5].

In summary, prior research has progressed from purely implicit neural mappings to explicit or hybrid motion representations. However, balancing interpretability and generalization remains challenging. Building upon this evolution, our framework introduces a **hybrid motion modeling paradigm** that integrates prosody-aware implicit audio features with anatomically grounded explicit representations. By leveraging Action Units as a cross-modal alignment signal, our method enables fine-grained, identity-agnostic control of articulation while preserving the natural expressiveness of 3D Gaussian avatars.

## 2. Implementation Details

### 2.1. Video Preprocess Pipeline

Our preprocessing pipeline follows the procedures of SyncTalk and TalkingGaussian to ensure consistency and reproducibility. Specifically, we adopt the audio feature extraction and head pose normalization process from

SyncTalk. A pre-trained Audio-Visual Encoder is used to extract rich audio representations, which serve as the driving signal for subsequent motion synthesis. For camera parameter estimation, we utilize a Head-Motion Tracker together with a Head-Points Tracker to analyze each video sequence, allowing us to normalize the head pose and stabilize the subject in a canonical coordinate space. For facial and intra-oral region segmentation, we follow the preprocessing setup of TalkingGaussian. We first apply BiSeNet for semantic portrait parsing on each frame to obtain a coarse facial mask. The result is then refined using EasyPortrait to generate a detailed teeth mask. This two-stage process produces a high-fidelity facial mask that separates the head from the background and provides an accurate representation of the mouth interior, which is essential for high-quality composition and rendering.

## 2.2. Network Architecture

The audio features  $\mathbf{a}$  extracted from Audio-Visual Encoder are first concatenated over 8 consecutive frames to capture short-term temporal context, and then passed into **AudioNet** for dimensionality compression. AudioNet comprises three fully connected layers, each followed by a LeakyReLU activation, progressively halving the feature dimension to obtain an intermediate representation of shape  $(8, 32)$ . This representation is then input to **AudioAttNet**, which applies temporal attention to emphasize speech-relevant segments. Specifically, AudioAttNet processes the intermediate features through several 1D convolutional layers, followed by a SoftMax activation along the temporal dimension (of size 8) to compute attention weights. These weights are applied to the intermediate features and aggregated via weighted summation across time, producing the final implicit motion feature  $\mathbf{c}_{a,t}^i \in \mathbb{R}^{1 \times 32}$ .

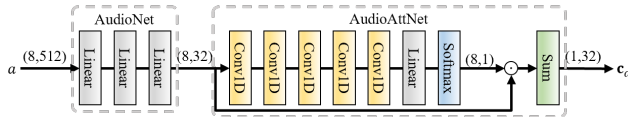


Figure 1. Architecture of AudioNet and AudioAttNet.

## 3. More Experiments

### 3.1. Ablation Study

**Hybrid Motion Modeling Module.** We adopt a three-path training strategy with a fixed ratio of  $\mathcal{P}_{\text{gen}} : \mathcal{P}_{\text{robust}} : \mathcal{P}_{\text{style}} = 4:4:2$ . Ablation studies reveal the influence of different path configurations on performance. When  $\mathcal{P}_{\text{gen}}$  is fixed and varying the ratio between  $\mathcal{P}_{\text{style}}$  and  $\mathcal{P}_{\text{robust}}$ , all evaluation metrics consistently drop (Fig. 2 group 1), highlighting the importance of personal style learning. We attribute this to the role of  $\mathcal{P}_{\text{style}}$  in capturing fine-grained, identity-specific articulatory dynamics that cannot be recov-

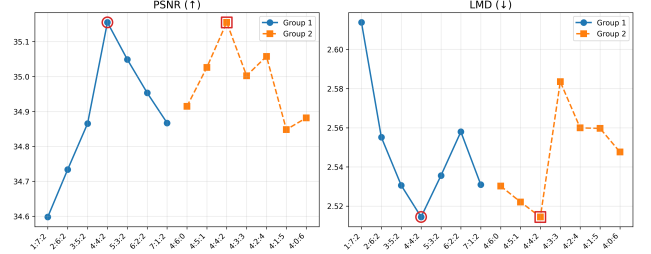


Figure 2. Ablation study of path proportion in Hybrid Motion Modeling Module.

ered from the robustness-oriented or generation paths alone. Furthermore, by fixing  $\mathcal{P}_{\text{style}}$  and varying the ratio between  $\mathcal{P}_{\text{gen}}$  and  $\mathcal{P}_{\text{robust}}$  (Fig. 2 group 2), we observe that performance deteriorates when  $\mathcal{P}_{\text{robust}}$  is either too high (7) or too low (1), confirming the validity of our balanced configuration.

To further investigate the masking effect within  $\mathcal{P}_{\text{robust}}$ , we evaluate different masking ratios  $\mathcal{M}_a$ . As shown in Fig. 3, performance peaks at  $\mathcal{M}_a = 0.2$  and declines as the value deviates. We therefore employ a stochastic masking strategy, sampling  $\mathcal{M}_a$  uniformly from 0.1 to 0.3 during training, which balances regularization with representation capacity.

Finally, we assess the quality of audio-predicted explicit features via t-SNE visualization under two suboptimal settings: (i) a path ratio of 1:7:2, and (ii) a high masking rate  $\mathcal{M}_a = 0.9$ . In both cases (Fig. 4), the audio-predicted features show weaker alignment with image-derived features, whereas our default configuration yields tighter overlap, demonstrating more effective cross-modal feature fusion and disentanglement.

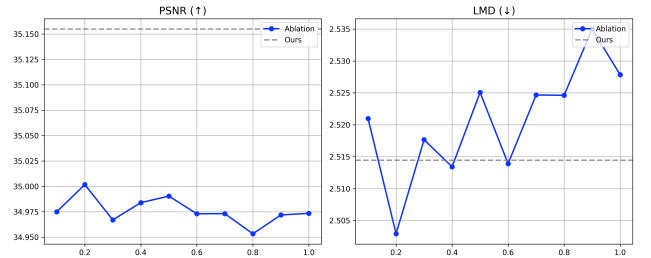


Figure 3. Ablation study of  $\mathcal{M}_a$  in Hybrid Motion Modeling Module.

### 3.2. Application to Few-Shot Personalization

To further verify the versatility and generalizability of our hybrid motion modeling paradigm, we apply it to the state-of-the-art few-shot personalization framework, InStAG [14]. This experiment demonstrates that our approach can function as a universal generalizable prior, enhancing motion fidelity and stability in few-shot talking head synthesis.

**Framework Integration.** We replace the implicit motion

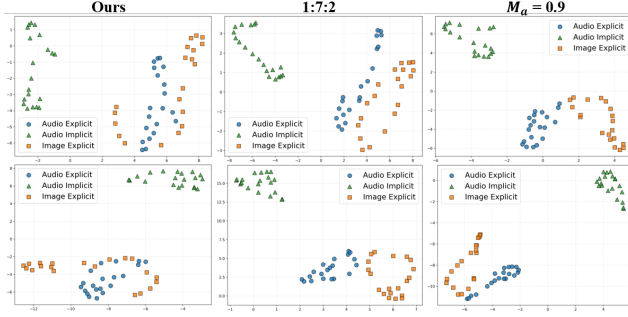


Figure 4. t-SNE visualization of motion features over three 20-frame clips. Each row corresponds to one clip; each column represents a different training setting.

prior in InsTaG with our hybrid motion model. Specifically, the proposed **Generalizable Prior Network** adopts the HM-Talker architecture, incorporating both the Audio-to-Visual Mapper (A2VM) and the Stochastic Feature Pairing (SFP) training strategy. This network learns a universal, anatomically grounded mapping from audio to facial motion. The lightweight **Personalized Adaptation Field** in InsTaG is retained to capture identity-specific nuances through few-shot fine-tuning. The final motion field is derived by combining the outputs of the generalizable prior and the personalized adaptation module.

**Training of the Generalizable Prior.** Following the InsTaG training protocol, we pre-train the Generalizable Prior Network on a multi-identity dataset including “Shaheen”, “May”, “Jae-in”, “Obama”, and “Macron”. The SFP strategy is applied during this pre-training stage to ensure robust cross-speaker generalization.

**Few-Shot Adaptation and Evaluation.** For personalization, we select four unseen subjects from the HDTF dataset [26]. Each model is fine-tuned using a short 10-second reference video, during which only the Personalized Adaptation Field is updated while the Generalizable Prior remains frozen. This setup evaluates how effectively our generalizable prior transfers to novel subjects with minimal supervision.

**Baseline for Comparison.** To ensure fair comparison, we re-implement the InsTaG framework under our unified preprocessing pipeline, denoted as **InsTaG\***. Our hybrid-integrated variant, **InsTaG-HM**, is then compared against InsTaG\* to quantify the improvements contributed by hybrid motion modeling.

**Results and Analysis.** As shown in Tab. 1, our integrated framework **InsTaG-HM** consistently outperforms both the original InsTaG and our re-implementation (InsTaG\*) across four unseen subjects. It achieves higher motion quality and lip synchronization, with Sync-C improving from 4.777 (InsTaG) and 5.332 (InsTaG\*) to **5.458**, and Landmark Distance (LMD) reduced to **3.208**. Although PSNR gains are marginal, this aligns with our focus on en-

| Setting     | PSNR $\uparrow$ | AUE-(L/U) $\downarrow$ | Sync-C $\uparrow$ | LMD $\downarrow$ |
|-------------|-----------------|------------------------|-------------------|------------------|
| InsTaG-HM   | <b>29.799</b>   | 0.95/ <b>0.64</b>      | <b>5.458</b>      | <b>3.208</b>     |
| InsTaG*     | 29.631          | <b>1.01</b> /0.70      | 5.332             | 3.221            |
| InsTaG [14] | 25.594          | <b>1.08</b> /0.82      | 4.777             | 3.270            |

Table 1. Comparison on the few-shot personalization task. We integrate our hybrid motion model into InsTaG (denoted as InsTaG-HM) and compare it against the original method and our re-implementation (InsTaG\*). Results are averaged across four unseen subjects from the HDTF dataset, demonstrating that our model serves as a superior generalizable prior.

hancing motion fidelity rather than static appearance. These results verify that introducing our hybrid motion model as a generalizable prior effectively stabilizes and refines few-shot adaptation, allowing the lightweight personalization module to capture identity-specific details while preserving articulatory precision.

### 3.3. Qualitative Results on Diverse Identities

To further assess the robustness of our framework beyond the core benchmark dataset, we evaluated HM-Talker on five additional subjects from the diverse HDTF dataset [26]. Following our standard per-subject training protocol, we performed self-reconstruction to examine the learned personalized motion fields. As shown in Fig. 5, our method maintains visually coherent reconstructions across a wide range of individuals. It effectively captures distinct articulatory behaviors and accommodates diverse facial traits, such as facial hair and expression dynamics. These results suggest that our approach generalizes well to subjects beyond the primary benchmark, demonstrating the framework’s ability to model personalized motion fields for diverse individuals.

## 4. Discussions

**Why Audio-to-Visual Mapping Works?** While SadTalker predicts audio-to-3DMM motion, its explicit representations exhibit limited identity specificity, often underperforming implicit-only baselines. This limitation arises from its general-purpose objective of synthesizing plausible full-face motion across arbitrary identities, which dilutes identity-dependent accuracy. In contrast, our Audio-to-Visual Mapper (A2VM) establishes a personalized mapping from audio to lower-face visual representations that are strongly correlated with articulatory dynamics. This identity-specific supervision constrains the feature distribution within a compact manifold, enabling a lightweight MLP to effectively capture motion variations. The resulting visual features align seamlessly with our static Gaussian representation, supporting precise and stable motion synthesis. Hence, A2VM favors articulation-accurate, identity-consistent motion generation over generic motion generalization.



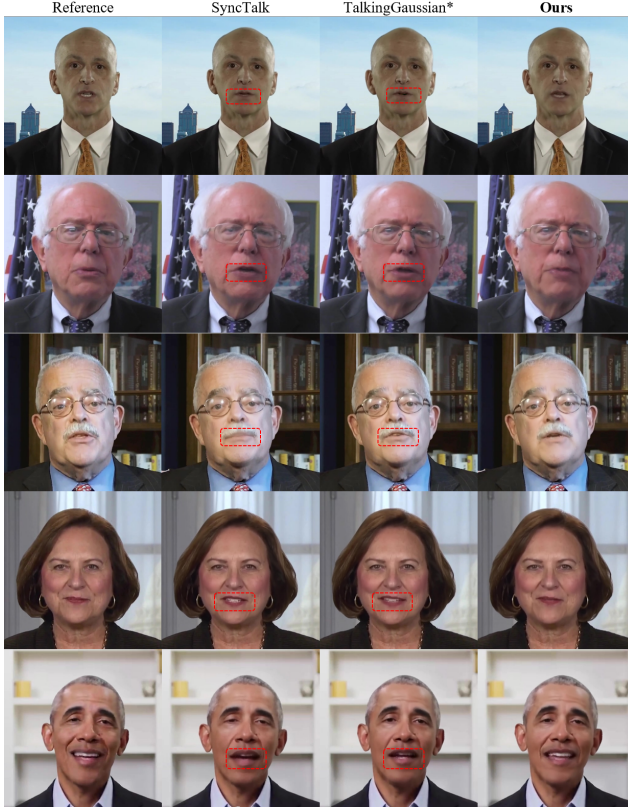


Figure 5. Qualitative results on diverse identities from the HDTF dataset. HM-Talker produces photorealistic and identity-faithful reconstructions, capturing distinct articulatory styles and facial characteristics across subjects.

**Explicit Motion Priors in Application.** Selecting an explicit motion prior naturally involves balancing articulatory detail and geometric stability. In our experiments, Facial Action Units (AUs) capture subtle and interpretable muscle activations that align well with the fine-grained dynamics of speech. However, because they rely primarily on 2D appearance cues, AUs can become less reliable under large head rotations or challenging lighting. In contrast, 3D Morphable Models (3DMMs) impose strong geometric constraints that ensure stable and consistent structure, but their parameters often entangle multiple factors and lack clear semantic meaning. In practice, the choice depends on the application: AUs are preferable when articulatory precision is the priority, whereas 3DMMs are better suited when geometric stability is more critical.

## 5. Limitations and Future Work.

While our method demonstrates strong performance across identity generalization, cross-lingual animation, and robustness settings, several limitations remain. First, the model can still produce minor artifacts in underrepresented cases, especially for subjects with long hair, beards, fine wrinkles, and large mouth openings, which are likely caused

by sparse data coverage in these appearance and motion patterns. Second, artifacts may occur around challenging fusion regions such as the neck and head-torso boundary, where blending different motion and geometry cues is inherently difficult. Third, although our approach improves temporal stability, cross-lip jitter and temporal misalignment can still appear in extreme cross-identity driving scenarios, particularly when there are large anatomical discrepancies or highly diverse speaking styles between source and driving signals. Finally, the current framework still relies on an identity video to construct the initial 3D model, which limits its applicability in ultra-low-input settings. In future work, we plan to expand training data toward more diverse identities and extreme facial conditions, develop stronger boundary-aware fusion and geometry-consistency constraints, and improve temporal modeling for cross-identity audio-driven animation under severe mismatch. We also aim to explore feed-forward 3D reconstruction to support one-image inference and further simplify the deployment pipeline.

## References

- [1] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 157–164. Association for Computing Machinery, 2023. 1
- [2] Hejia Chen, Haoxian Zhang, Shoulong Zhang, Xiaoqiang Liu, Sisi Zhuang, Yuan Zhang, Pengfei Wan, Di Zhang, and Shuai Li. Cafe-talk: Generating 3d talking face animation with multimodal coarse- and fine-grained control, 2025. 1
- [3] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar, 2025. 1
- [4] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 1
- [5] Guanwen Feng, Zhihao Qian, Yunan Li, Siyu Jin, Qiguang Miao, and Chi-Man Pun. Les-talker: Fine-grained emotion editing for talking head generation in linear emotion space, 2025. 1
- [6] Shengjie Gong, Haojie Li, Jiapeng Tang, Dongming Hu, Shuangping Huang, Hao Chen, Tianshui Chen, and Zhuoman Liu. Monocular and generalizable gaussian talking head animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 1
- [7] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 1
- [8] Ricong Huang, Weizhi Zhong, and Guanbin Li. Audio-driven talking head generation with transformer and 3d morphable model. In *ACM International Conference on Multimedia*, page 7035–7039. Association for Computing Machinery, 2022. 1
- [9] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time

- radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1
- [10] Gihoon Kim, Kwanggyoon Seo, Sihun Cha, and Junyong Noh. Nerffacespeech: One-shot audio-driven 3d talking head synthesis via generative prior, 2024. 1
- [11] Dongze Li, Kang Zhao, Wei Wang, Bo Peng, Yingya Zhang, Jing Dong, and Tieniu Tan. Ae-nerf: Audio enhanced neural radiance field for few shot talking head synthesis, 2023. 1
- [12] Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In *IEEE/CVF International Conference on Computer Vision*, pages 7568–7578, 2023. 1
- [13] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. In *European Conference on Computer Vision*, pages 127–145, 2025. 1
- [14] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Jun Zhou, and Lin Gu. Instag: Learning personalized 3d talking head from few-second video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3
- [15] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles, 2023. 1
- [16] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1
- [17] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024. 1
- [18] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *ACM International Conference on Multimedia*, pages 484–492, 2020. 1
- [19] Yasheng Sun, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Zhibin Hong, Jingtuo Liu, Errui Ding, Jingdong Wang, Ziwei Liu, and Koike Hideki. Masked lip-sync prediction by audio-visual contextual exploitation in transformers. In *SIG-GRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 1
- [20] Zhiyao Sun, Tian Lv, Sheng Ye, Matthieu Lin, Jenny Sheng, Yu-Hui Wen, Minjing Yu, and Yong-jin Liu. Diffposetalk: Speech-driven stylistic 3d facial animation and head pose generation via diffusion models. *ACM Transactions on Graphics*, 2024. 1
- [21] Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. *arXiv preprint arXiv:2211.12368*, 2022. 1
- [22] Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *ACM International Conference on Multimedia*, page 1478–1486. ACM, 2021. 1
- [23] Yifan Xie, Tao Feng, Xin Zhang, Xiangyang Luo, Zixuan Guo, Weijiang Yu, Heng Chang, Fei Ma, and Fei Richard Yu. Pointtalk: Audio-driven dynamic lip point cloud for 3d gaussian-based talking head synthesis. In *AAAI Conference on Artificial Intelligence*, 2025. 1
- [24] Chao Xu, Shaoting Zhu, Junwei Zhu, Tianxin Huang, Jiangning Zhang, Ying Tai, and Yong Liu. Multimodal-driven talking face generation via a unified diffusion-based generator, 2023. 1
- [25] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1
- [26] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670, 2021. 3
- [27] Zhimeng Zhang, Zhipeng Hu, Wenjin Deng, Changjie Fan, Tangjie Lv, and Yu Ding. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In *AAAI Conference on Artificial Intelligence*, pages 3543–3551, 2023. 1
- [28] Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023. 1
- [29] Ziqi Zhou, Weize Quan, Hailin Shi, Wei Li, Lili Wang, and Dong-Ming Yan. Gohd: Gaze-oriented and highly disentangled portrait animation with rhythmic poses and realistic expression, 2024. 1