TALK LIKE YOU: IMITATING HOW YOU SPEAK IN REAL-TIME TALKING HEAD GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In daily life, each person exhibits unique speaking habits, leading to subtle yet consistent lip-shape variations even when pronouncing the same word. Although recent talking head generation methods have achieved impressive visual fidelity and lip synchronization, they largely overlook user-specific customization, especially the motion patterns that characterize individual speaking habits. These habits are difficult to model and capture, as their motion patterns are highly fine-grained and often similar across individuals. As a result, many approaches produce overly uniform facial motions and fail to capture diverse, person-specific articulation patterns. To address this, we propose TalkLikeYou, an efficient framework that imitates how a target person speaks in talking head generation. Our method models habit in motion-space and achieves real-time performance through Flow Matching with only one sampling step during inference. We further adopt a two-stage imitation learning strategy to capture subtle distinctions between habits, allowing users to specify a target habit through either a preset style from the dataset or a reference video. In addition, we introduce a new metric PLAD that projects mouth motions onto representative articulation axes to evaluate imitation accuracy and generation diversity. Extensive experiments demonstrate that TalkLikeYou generates high-quality talking heads in real-time and significantly improves speaking habit imitation compared with prior methods.

1 INTRODUCTION

Audio-driven talking head generation has attracted increasing attention due to its wide range of applications, including digital humans, voice assistants, and film-making. With rapid progress in this field Wang et al. (2024); Zhao et al. (2025), customization has become an important direction for meeting user-specific needs, including imitating how a target person speaks. Such imitation is essential for building personalized virtual avatars that convey a recognizable identity and leave a vivid impression on viewers, as each person's speaking habits are reflected in subtle yet consistent articulatory behaviors in how the lips open, close, stretch, and transition during speech. As shown in Fig. 1(a), even the same word can exhibit slight variations in articulation across individuals, such as the degree of mouth opening when saying "b**u**s", the lip width when pronouncing "l**ea**ve", or the pursing of the lips when saying "f**oo**d". These diverse habitual motion patterns are important parts of how we perceive identity and personality in daily communication Dobs et al. (2016).

Yet imitating a person's speaking habits remains challenging, as they involve fine-grained lip motion patterns that are difficult to **model**. Previous methods Ji et al. (2025); Wang et al. (2025b); Cui et al. (2025); Ki et al. (2025) have made significant progress in generation quality, but they primarily focus on synchronizing lip movements with audio while overlooking subtle motion differences among individuals. They rely on image-space diffusion and typically learn average motion patterns from the dataset without explicitly distinguishing between speakers, resulting in limited diversity and personalization in lip motion. They also suffer from limited efficiency, as they require multiple sampling steps in high-dimensional spaces. Moreover, the differences between habitual motion patterns are often subtle, making them difficult for models to **capture**. Although some methods Ma et al. (2023a); Ye et al. (2024) explore person-specific motion generation, they face limitations in imitation quality and motion coherence or rely on separate training for each individual. In addition, existing metrics are not designed to **measure** speaking-habit similarity and diversity. As a result, habit-aware talking-head generation remains underexplored.

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

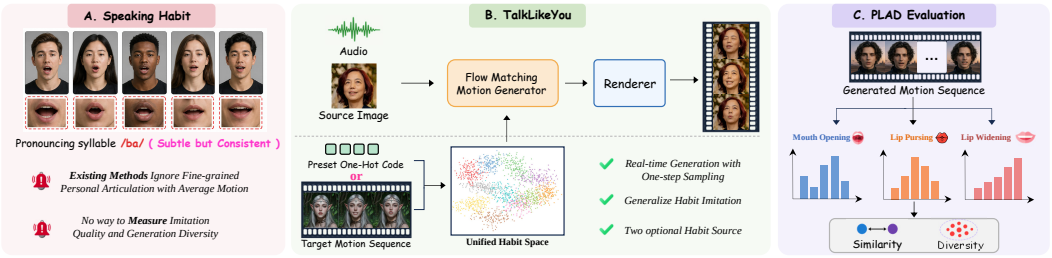


Figure 1: **Overview of TalkLikeYou.** Previous methods often generate uniform motions with limited control over speaking habits. We propose TalkLikeYou for efficient habit imitation in talking head generation, together with a new metric PLAD to evaluate habit similarity and diversity.

To address these issues, we propose TalkLikeYou, an efficient talking head generation framework designed to imitate and control individual speaking habits, as shown in Fig. 1(b). To **model** speaking habits, our framework performs generation in motion space, providing better controllability, temporal coherence, and efficiency than image-space diffusion approaches Tian et al. (2024); Ji et al. (2025). Specifically, we first generate expressive head motions conditioned on target habitual motion and then render them into the final talking-head video. By incorporating Flow Matching Lipman et al. (2022) for motion synthesis, our method generates motions with a single sampling step during inference, enabling real-time performance.

To **capture** fine-grained speaking-habit features for faithful imitation, we adopt a two-stage imitation learning strategy and establish a unified habit space using reference motion sequences and preset habit codes from the dataset. In the first stage, we initially train with only one-hot habit codes, as their explicit distinctions make habit differences easier for the model to learn. After this warm-up, we align reference-based habit features with their corresponding one-hot embeddings, then train on a large-scale video dataset in the second stage to improve generalization and diversity. This also allows our model to handle two scenarios: users can specify the target habit through a reference video or select a preset habit when no reference is available. It enables diverse and customizable generation while maintaining faithful speaking-habit imitation and accurate lip synchronization.

To **measure** the quality of speaking habit imitation and generation diversity, we introduce a new metric Projected Lip-Articulation Distribution (PLAD) as shown in Fig. 1(c). Since speaking habits are reflected in subtle variations in lip articulation, we project motion sequences onto three representative articulation axes and characterize the resulting distributions using histograms. As these distributions become stable for sufficiently long sequences, we compute similarity and diversity scores from histograms to quantitatively evaluate speaking-habit imitation and generation diversity.

In summary, our contributions are as follows:

- We propose TalkLikeYou, a real-time talking-head generation framework capable of imitating specific speaking habits. By applying Flow Matching in motion space, our approach enables one-step motion generation with high coherence and efficiency.
- To improve imitation quality, we design a unified habit space and adopt a two-stage imitation learning strategy to improve imitation generalization and diversity.
- We develop a new metric PLAD to evaluate imitation quality by projecting motion sequences onto lip-articulation axes and measuring similarity and diversity across the histograms.
- Extensive experiments demonstrate that our method achieves state-of-the-art performance in talking-head generation, with superior lip synchronization and speaking-habit imitation.

2 Related Work

2.1 Audio-Driven Talking Head Generation

Talking head generation aims to synthesize realistic facial animations with lip movements synchronized to input audio. Early methods Wang et al. (2025a); Zhang et al. (2024) mainly focus on lower-face synthesis for video dubbing, such as Wav2Lip Prajwal et al. (2020), which introduces a

lip-sync expert network to improve synchronization. Recent approaches utilize modern generative models for better quality. For example, Sonic Ji et al. (2025) and FLOAT Ki et al. (2025) adopt DiT Peebles & Xie (2023) as their backbone, while MultiTalk Kong et al. (2025) is built upon the video generation model WAN Wan et al. (2025). However, these approaches often rely on heavy architectures, leading to high computational cost and long inference time. Moreover, they provide limited controllability for fine-grained motion patterns such as speaking habits.

In parallel, two-stage pipelines have gained attention. They first predict intermediate facial motion representations from audio, such as 3DMM parameters Li et al. (2017) or facial landmarks, and then render lip-synced videos conditioned on them. Among these representations, implicit keypoints show strong expressiveness: face-vid2vid Wang et al. (2021) disentangles them into expression and pose, while LivePortrait Guo et al. (2024) improves controllability with landmark-bounded keypoints. Several methods follow this paradigm: EAT Gan et al. (2023) suffers from limited visual quality, while VASA Xu et al. (2024) and Ditto Li et al. (2025) apply diffusion in motion space but require multiple sampling steps and tend to generate uniform motions without modeling speaking habits. PC-Talk Wang et al. (2025b) provides stronger control and lip synchronization, but its backbone limits generalization. Overall, two-stage methods remain promising, yet still leave room for improvement in efficiency and controllable habit-aware generation.

2.2 Customized Talking Head Generation

Customization has attracted increasing attention in talking-head generation. Several methods focus on emotion control Gan et al. (2023); Tan et al. (2024); Wang et al. (2025b), generating emotional talking-face videos where the emotion source is derived from labels or other inputs such as text and audio. Other works incorporate image editing into this task, including hair transfer Feng et al. (2025); Li et al. (2026) and face swapping Guan et al. (2024) for appearance modification. Some methods Zhu et al. (2025); Wang et al. (2026) further focus on multi-person conversation to improve interaction. However, these works mainly address high-level customization, such as emotion, appearance, or interaction, rather than fine-grained motion control.

Meanwhile, several methods explore more detailed customization, such as imitating person-specific motion patterns, but they still face limitations. StyleTalk Ma et al. (2023a) and DreamTalk Ma et al. (2023b) perform conditioned synthesis with specific motion patterns, yet their control mainly targets emotion-related dynamics rather than habit-related variations in lip shape and articulation. MimicTalk Ye et al. (2024) uses an autoregressive paradigm to reproduce person-specific habits, but it is person-dependent and requires training or adaptation for each individual, limiting generalization to unseen identities. PC-Talk Wang et al. (2025b) can generate style-conditioned results, but may suffer from subtle temporal jitter and unstable motion. Although some 3D facial animation methods Fan et al. (2022); Thambiraja et al. (2023); Zhou et al. (2025) also investigate this direction, their generated 3D motions cannot be directly applied to 2D talking-head generation. Overall, imitating how a target person speaks still leaves substantial room for improvement.

3 Methods

3.1 Overall Framework

To imitate how a target person speaks, our framework first generates habit-aware talking motions and then renders them into the final video, as illustrated in Fig. 2(a). Specifically, we use a motion extractor to obtain implicit keypoint-based motion representations from the source image and the reference video. Following the rendering pipeline Guo et al. (2024), the extracted motion is decomposed into canonical keypoints K_c, expression deformation δ, and pose parameters, including rotation R, translation t, and scale s. The original keypoints K_{ori} are formulated as:

$$K_{ori} = s \cdot (K_c \cdot R + \delta) + t. \tag{1}$$

After obtaining the motion representations, we feed them into a Flow Matching-based motion generator to produce habit-aware motion. The generator is conditioned on habit feature f_h, the audio feature f_{aud}, and the source motion M_{src} used for initialization. It predicts the habitual motion M_{habit}, which is decomposed into the habit-aware rotation R_h and expression deformation δ_h:

$$M_{\text{habit}} = G_\theta(f_h, f_{\text{aud}}, M_{\text{src}}), \quad M_{\text{habit}} \Rightarrow \{R_h, \delta_h\}. \tag{2}$$

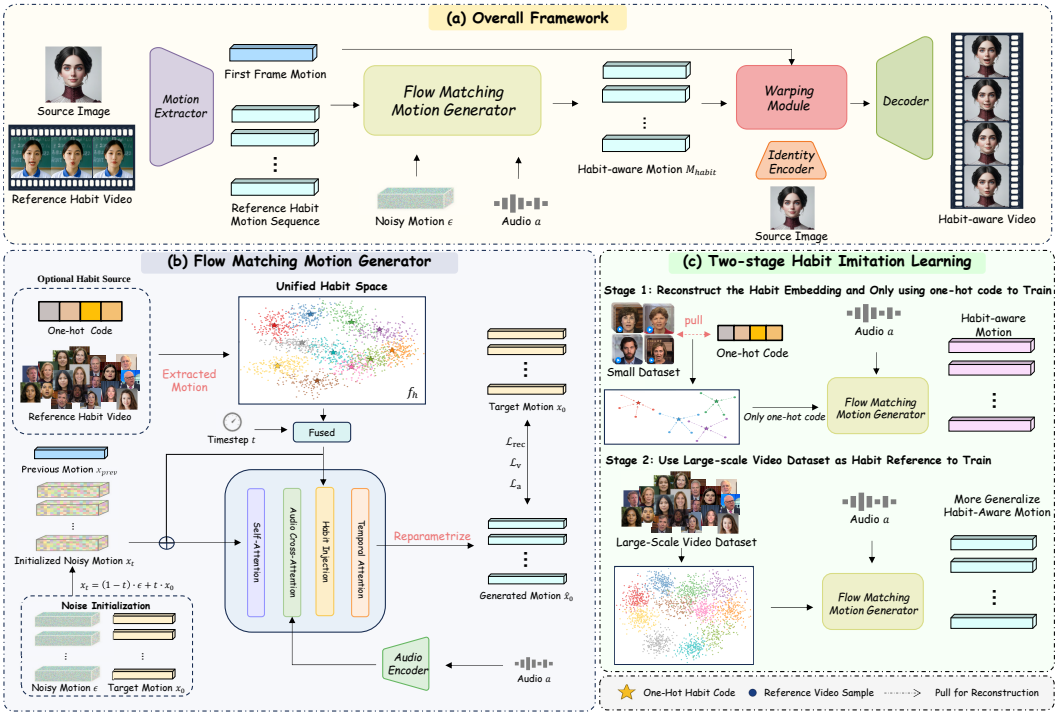


Figure 2: **Framework of TalkLikeYou.** Our method first generates habit-aware motion using the Flow Matching Motion Generator with optional habit sources through one-step sampling, and then renders them into a video. We further adopt a two-stage habit imitation learning strategy for more generalized and stable habit imitation. In the unified habit space, $\star$ denotes the one-hot habit code and $\bullet$ denotes reference-video samples. During the first training stage, we use an alignment loss between reference video features and their corresponding one-hot habit codes.

We then apply the generated motion components to transform the original keypoints into habit-aware driven keypoints K_h:

$$K_h = s \cdot (K_c \cdot R_h + \delta_h) + t. \tag{3}$$

Finally, we render the talking head video from the source image and the generated habit-aware motion. We first encode the source image into an appearance embedding f_a using an identity encoder. A warping decoder then fuses f_a with the original keypoints K_{ori} and the habit-aware driven keypoints K_h to estimate the dense warping field and synthesize the final rendered frame I_{res}:

$$I_{\text{res}} = \text{Decoder}(\text{Warp}(f_a, K_{ori}, K_h)) \,. \tag{4}$$

Since implicit keypoints are semantically associated with facial components, we can selectively edit the lip-related keypoints from the input video to achieve video dubbing. During the training of the Flow Matching Motion Generator, all other parts of the rendering pipeline remain fixed, as they only serve as intermediate representations for motion generation and final video rendering.

3.2 Flow Matching Motion Generator

Model Architecture. Our motion generator aims to produce talking motions that are synchronized with the input audio while reflecting the target speaking habit in the lip region. To this end, we employ a DiT Peebles & Xie (2023) composed of four progressive layers as shown in Fig. 2(b). We first use self-attention to extract intrinsic motion representations, then apply audio cross-attention to incorporate audio features for synchronization. Next, we project either the input habit sequence or the preset one-hot habit code from the training dataset into a unified habit space, and fuse it through a habit-injection layer to enable target-habit imitation. Finally, a temporal attention layer refines sequence-level dynamics and improves temporal consistency. This design enables coherent talking motion generation while preserving the target speaking habit.

Flow Matching Training. For efficient and high-quality generation, we adopt Flow Matching as our training objective. Conventional Flow Matching Lipman et al. (2022) learns a velocity field v_t to transport noisy samples toward the data distribution, enabling faster convergence than denoising-based samplers such as DDIM Song et al. (2020) due to its simple linear trajectory. Although this formulation has shown strong performance in generative models such as FLUX Batifol et al. (2025), we observe that directly predicting the velocity field can introduce noticeable jitter in generated motions. To address this, we reparameterize the model to predict the clean motion x_0 at each timestep, which improves motion coherence and enables one-step sampling during inference. Specifically, at each training step t, we sample noise motion $\epsilon \sim \mathcal{N}(0, I)$ and construct the noisy motion as:

$$x_t = (1-t)\epsilon + tx_0, \quad t \sim \mathcal{U}(0,1). \tag{5}$$

The previous motion x_{prev} is obtained from source M_{src}. Given the noisy motion x_t, audio feature f_{aud}, habit feature f_h, and timestep t, our model predicts the clean motion:

$$M_{\text{src}} \Rightarrow x_{\text{prev}}, \quad \hat{x}_0 = G_\theta(x_t, f_{\text{aud}}, f_h, x_{\text{prev}}, t), \quad \hat{x}_0 \Rightarrow M_{\text{habit}}. \tag{6}$$

During training, we first apply an L_2 reconstruction loss $\mathcal{L}_{\text{rec}}$ between the predicted motion $\hat{x}_0$ and the ground-truth motion x_0. Additionally, we impose a first-frame loss $\mathcal{L}_f$ that aligns the first generated frame with the last frame of the previous motion to ensure a seamless transition between consecutive context windows. The losses $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_f$ are calculated as:

$$\mathcal{L}_{\text{rec}} = \|\hat{x}_0 - x_0\|_2^2, \quad \mathcal{L}_f = \|\hat{x}_0^0 - x_{\text{prev}}\|_2^2. \tag{7}$$

To further enhance temporal consistency, we introduce first-order and second-order temporal derivative losses to constrain frame-to-frame motion dynamics and regularize temporal smoothness:

$$\begin{aligned} \mathcal{L}_v &= \sum_{i=2} \left\| \Delta\hat{x}_0^i - \Delta x_0^i \right\|_2^2, \quad \Delta x_0^i = x_0^i - x_0^{i-1}, \\ \mathcal{L}_a &= \sum_{i=3} \left\| \Delta^2\hat{x}_0^i - \Delta^2 x_0^i \right\|_2^2, \quad \Delta^2 x_0^i = x_0^i - 2x_0^{i-1} + x_0^{i-2}. \end{aligned} \tag{8}$$

where $\hat{x}_0^0$ denotes the first frame of the predicted motion. This constraint helps extend generation to arbitrary-length sequences by maintaining continuity across context windows. The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_v \mathcal{L}_v + \lambda_a \mathcal{L}_a + \mathcal{L}_f. \tag{9}$$

With this motion-spaced Flow Matching paradigm, our model can efficiently generate temporally coherent talking motions with accurate lip synchronization and faithful speaking habit imitation.

3.3 Two-stage Habit Imitation Learning Strategy

Although our framework can generate motions with a target habit from a reference motion sequence, directly training the model from scratch with reference motion sequences leads to unsatisfactory habit imitation. This is because speaking habits are fine-grained lip region patterns that often appear as subtle pronunciation-dependent variations, making them difficult to extract from motion alone. In contrast, one-hot habit codes provide explicit and separable supervision, making habit distinctions easier to learn and naturally supporting preset habit control when no reference video is available. Based on this, we propose a two-stage habit imitation strategy, as shown in Fig. 2(c), to improve habit imitation quality while preserving preset one-hot control.

In the first stage, we train the motion generator on a limited dataset using only one-hot habit codes defined by training dataset, where each identity is assigned a unique code to represent its distinct talking habit. This explicit categorization helps the model learn audio-lip synchronization while distinguishing habit differences, forming a foundational habit space with separable representations. Meanwhile, we train the habit extractor with reference videos from the same dataset by reconstructing the corresponding one-hot habit feature:

$$\begin{aligned} f_h^{\text{oh}} &= E_{\text{oh}}(c), \quad \hat{x}_0 = G_\theta(x_t, f_{\text{aud}}, f_h^{\text{oh}}, x_{\text{prev}}, t), \\ \hat{f}_h^{\text{ref}} &= E_{\text{ref}}(M_{\text{ref}}), \quad \mathcal{L}_{\text{align}} = \left\| \hat{f}_h^{\text{ref}} - f_h^{\text{oh}} \right\|_2^2, \end{aligned} \tag{10}$$

ALGORITHM 1: Projected Lip-Articulation Distribution for Speaking-Habit Evaluation

Input: Generated motion sequences $\{M_i\}_{i=1}^{N}$; lip-articulation projection axes $\mathcal{Q}$.
Output: PLAD histograms $\{\mathcal{H}_i\}_{i=1}^{N}$, pairwise habit similarity scores, and diversity score D_{div}.
for *each motion sequence* M_i **do**
  Extract expression deformation sequence $\delta_i = \{\delta_{i,t}\}_{t=1}^{T}$;
  for *each articulation axis* $q \in \mathcal{Q}$ **do**
    for *each frame* t **do**
      Compute projection scalar $p_{i,t}^{q} = \frac{\langle \delta_{i,t}, q \rangle}{\langle q, q \rangle}$;
    end
    Construct normalized histogram $H_i^q = \mathrm{Hist}(\{p_{i,t}^{q}\}_{t=1}^{T})$;
  end
  Form PLAD descriptor $\mathcal{H}_i = \{H_i^q \mid q \in \mathcal{Q}\}$;
end
for *each pair* (M_i, M_j) **do**
  Compute histogram IoU $S_{\mathrm{IoU}}(M_i, M_j)$ and JS distance $D_{\mathrm{JS}}(M_i, M_j)$ using Eq. 14;
end
Compute group diversity D_{div} using Eq. 15;

where c denotes the one-hot identity code. This encourages reference-based habit features to be projected into a unified habit space, preparing the model for the next stage.

In the second stage, we train the model with reference videos from a large-scale dataset. For each sample, we select two sequences from the same person, where one sequence serves as the target motion and the other serves as the reference motion:

$$\hat{x}_0 = G_\theta(x_t, f_{\text{aud}}, \hat{f}_h^{\text{ref}}, x_{\text{prev}}, t). \tag{11}$$

With the structured habit space learned in the first stage, the model can better imitate reference-derived habit features and learn diverse motion patterns from large-scale data. To preserve compatibility between two conditioning options, we periodically reuse the one-hot codes from the first stage during training, once every ten steps. This strategy improves habit imitation quality while supporting both reference video and preset one-hot habit control.

4 Projected Lip-Articulation Distribution

Existing talking head metrics mainly measure visual quality or lip synchronization, but are not designed to evaluate whether generated motion preserves a specific speaking habit. To evaluate speaking-habit imitation and generation diversity, we introduce Projected Lip-Articulation Distribution (PLAD). Since speaking habits are reflected in subtle and consistent lip articulations, PLAD projects motion sequences onto representative lip-articulation axes and characterizes the resulting distributions using histograms. We assume that a speaking habit is not determined by isolated frames, but by stable statistical tendencies over repeated articulatory movements. Although each frame is influenced by phonetic content, a person tends to show consistent distributional patterns over sufficiently long sequences along these representative axes such as mouth opening, lip stretching, or lip pursing. Thus, motions with the same habit should have similar PLAD histograms, while different habits should yield distinguishable distributions.

Specifically, as shown in Alg. 1, for a generated motion sequence, we extract the lip-region expression deformation sequence $\delta = \{\delta_i\}_{i=1}^{T}$, where each δ_i denotes the implicit expression parameters at frame i. We define a set of lip-articulation projection axes $\mathcal{Q} = \{q_m\}_{m=1}^{M}$, where each axis corresponds to a representative articulation pattern. For each frame, we project the expression parameters onto each articulation axis and obtain a scalar projection:

$$p_i^q = \frac{\langle \delta_i, q \rangle}{\langle q, q \rangle}, \quad q \in \mathcal{Q}. \tag{12}$$

In practice, we construct three representative axes for *mouth opening*, *lip widening*, and *lip pursing*, corresponding to articulations in words such as "b**u**s", "l**ea**ve", and "f**oo**d", respectively. The pro-

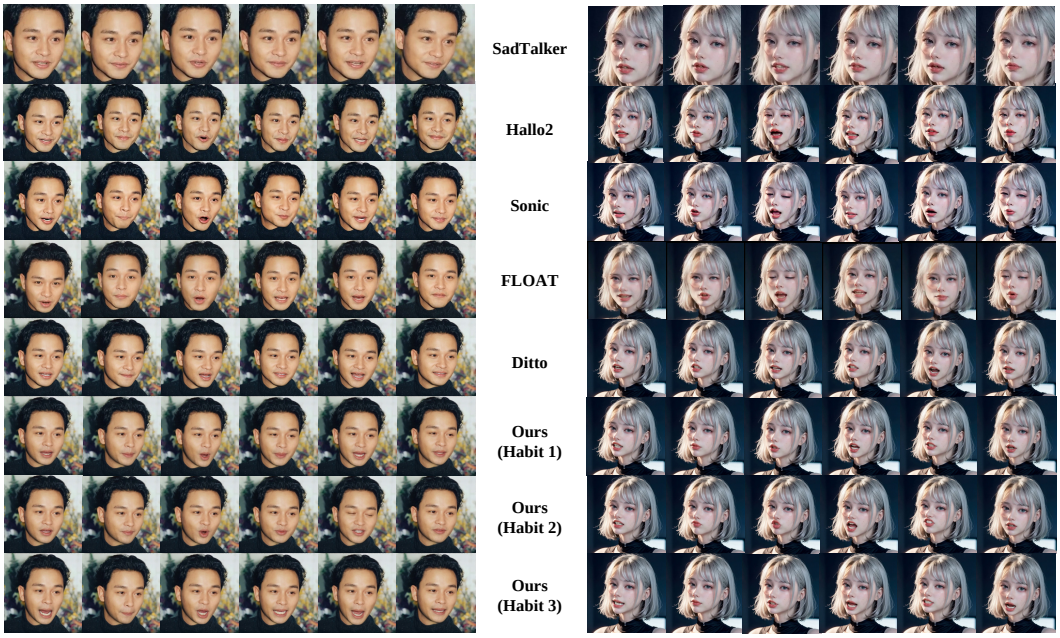


Figure 3: **Comparison Results.** We compare our method with advanced baselines under three selected speaking habits, showing better visual quality and lip synchronization. Please zoom in for details. Video results are provided in the supplementary material.

jection sequence $P^q = \{p_i^q\}_{i=1}^T$ describes how strongly the motion follows the articulation axis q over time. We then construct a normalized histogram from each projection sequence and represent the final PLAD descriptor as:

$$H^q = \mathrm{Hist}(P^q), \quad \mathcal{H} = \{H^q \mid q \in \mathcal{Q}\}. \tag{13}$$

To compare two generated motions, we directly measure the similarity between their histograms. Given two motion sequences A and B, we compute the histogram IoU and JS distance as:

$$S_{\mathrm{IoU}}(A,B) = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \frac{\sum_b \min(H_A^q(b), H_B^q(b))}{\sum_b \max(H_A^q(b), H_B^q(b))}, \quad D_{\mathrm{JS}}(A,B) = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathrm{JS}(H_A^q \,\|\, H_B^q). \tag{14}$$

where b indexes histogram bins and $\mathrm{JS}(\cdot\|\cdot)$ denotes the Jensen-Shannon divergence. A higher histogram IoU and a lower JS distance indicate stronger habit consistency, while lower IoU and higher JS distance indicate better habit separability. Furthermore, to evaluate the diversity of speaking habits generated by a model, we compute the average pairwise JS distance over a group of generated samples. Given N generated motion sequences, the diversity score is defined as:

$$D_{\mathrm{div}} = \frac{2}{N(N-1)} \sum_{1\le i<j\le N} D_{\mathrm{JS}}(M_i, M_j). \tag{15}$$

The diversity score D_{div} reflects whether a model can generate distinguishable speaking habits instead of collapsing to similar distributions.

5 Experiments

5.1 Experimental Settings

Implementation Details. We convert all videos to 25 fps and extract audio and motion representations before training. The Flow Matching Motion Generator is optimized with Adam Kingma & Ba (2014) at a learning rate of 1×10^{-4}, while other modules are fixed as they are initialized from LivePortrait Guo et al. (2024) for its balance between efficiency and quality. The first and second stages

Table 1: Quantitative comparisons with state-of-the-art methods.

Method	HDTF						CelebV-HQ					
	Sync-C↑	Sync-D↓	FID↓	NIQE ↓	FVD ↓	CSIM ↑	Sync-C↑	Sync-D↓	FID↓	NIQE ↓	FVD ↓	CSIM ↑
SadTalk Zhang et al. (2023)	7.66	7.32	87.60	45.72	534.14	0.85	6.36	8.36	109.11	40.82	829.09	0.80
Hallo2 Cui et al. (2025)	7.33	7.97	25.41	13.02	272.57	0.83	6.32	8.73	63.95	13.13	**688.03**	0.75
Sonic Ji et al. (2025)	8.62	**6.70**	23.42	12.91	195.21	0.84	7.68	7.35	66.37	12.88	704.12	0.76
Float Ki et al. (2025)	7.04	8.16	84.70	12.39	382.93	0.82	6.44	8.64	137.23	11.89	854.89	0.70
Ditto Li et al. (2025)	5.23	9.94	21.61	12.62	319.03	0.87	5.80	9.19	62.67	12.13	743.16	0.82
TalkLikeYou (Ours)	**8.73**	6.82	**7.71**	**12.24**	**112.94**	**0.94**	**7.97**	**7.25**	**61.51**	**11.79**	694.40	**0.85**

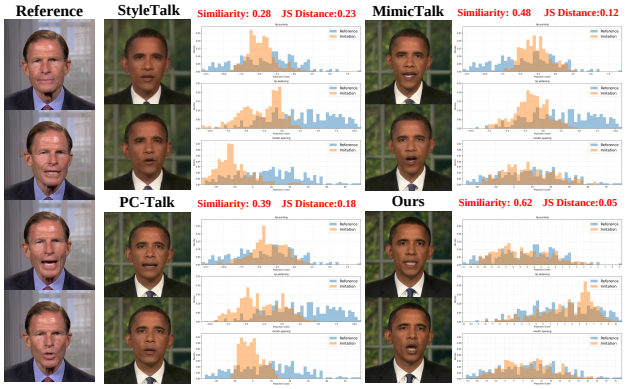


Figure 4: Visualization of habit imitation comparison with related methods.

Table 2: Habit Diversity and Efficiency Comparison in each method.

Method	Diversity	FPS
SadTalker	41.33	10.76
Hallo2	45.56	0.69
Sonic	25.47	0.38
Float	27.19	0.47
Ditto	33.99	29.85
Ours	**71.14**	**32.43**

are trained for 48 and 72 hours on an RTX 3090, respectively. We use HDTF Zhang et al. (2021) in the first training stage, which contains high-quality videos of 380 identities, each represented by a one-hot code as a preset habit option. The second stage is trained on CelebV-HQ Zhu et al. (2022), a large-scale in-the-wild talking head dataset. We adopt the Wav2Lip Prajwal et al. (2020) audio encoder for its audio-lip aligned features. More details are provided in the supplementary.

Baselines. We compare our method with five advanced talking head generation baselines, including SadTalker Zhang et al. (2023), Hallo2 Cui et al. (2025), Sonic Ji et al. (2025), FLOAT Ki et al. (2025), and Ditto Li et al. (2025). For habit imitation, we choose StyleTalk Ma et al. (2023a), MimicTalk Ye et al. (2024) and PC-Talk Wang et al. (2025b), as they are the only relevant open-source methods. Note that MimicTalk requires person-specific training before generating results.

Metrics. We evaluate our method from several aspects. Lip synchronization is measured by Sync-C and Sync-D, which are computed using a pretrained SyncNet Chung & Zisserman (2016). Image quality is evaluated with the reference-based FID Seitzer (2020) metric and the non-reference-based NIQE Mittal et al. (2012) metric. We use FVD Unterthiner et al. (2019) to measure temporal consistency and CSIM Deng et al. (2019a) for identity preservation. For habit imitation, we use the similarity and diversity metrics proposed in our PLAD.

5.2 Evaluation on Lip-Synchronization

Quantitative Results. As shown in Tab. 1, TalkLikeYou achieves strong overall performance on both HDTF and CelebV-HQ. It obtains the best Sync-C scores on both datasets, with Sync-D only slightly behind Sonic on HDTF, indicating accurate lip synchronization. Our method also achieves the lowest FID and NIQE, demonstrating better visual quality and naturalness. In addition, TalkLikeYou achieves the best FVD on HDTF and the highest CSIM scores on both datasets, showing strong temporal consistency and identity preservation. As shown in Tab. 2, our method also has an efficiency advantage over Ditto, as it requires only one sampling step during inference. Overall, these results demonstrate that TalkLikeYou achieves balanced performance across lip synchronization, visual quality, motion stability, identity consistency, and efficiency.

Qualitative Results. As shown in Fig. 3, we compare our method with other baselines under three speaking habits. Although these habits involve subtle lip-shape differences for the same pronunciation, our method consistently achieves accurate lip synchronization. In contrast, existing methods often produce unrealistic results: SadTalker produces rigid motions, while Hallo2 and Sonic show

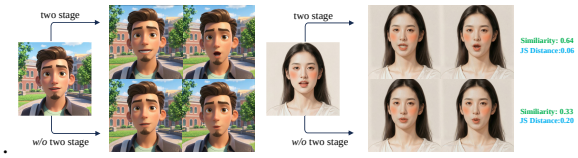


Figure 5: Ablation study on the two-stage habit imitation learning strategy.

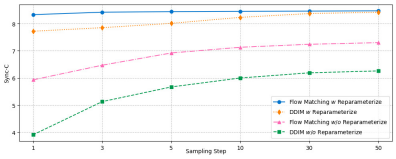


Figure 6: Ablation study of Flow Matching and DDIM with reparameterization.

blurry teeth regions and unstable lip movements. FLOAT and Ditto yield clearer results, but their motions are less natural, with FLOAT producing limited mouth movements and Ditto often over-opening the mouth. Overall, our method generates high-quality results with fluid motion across different speaking habits. Please refer to the supplementary for user study and more visual results.

5.3 Evaluation on Habit Imitation

Similarity. We conduct habit imitation experiments against related methods. As shown in Fig. 4, our method achieves the highest similarity and the lowest JS distance, indicating better imitation of the target speaking habit. The visual clips also show that our results are closest to the reference video, such as in mouth-widening patterns and lower-teeth exposure. More samples are provided supplementary. We also present the histogram of each syllable projection, where our method shows much higher overlap with the reference than other methods, indicating that it more effectively captures the distributional characteristics of the target speaking habit.

Diversity. As shown in Tab. 2, we evaluate the habit diversity of each method. We randomly sample habit features from the habit space and observe that our method achieves a clear advantage over others, indicating that TalkLikeYou can generate more diverse speaking habits rather than collapsing to uniform motions. Furthermore, we observe that each method exhibits characteristic motion patterns, so we can even use motions generated by other methods as references, bridging the motion-distribution gap between each methods. More diversity analysis is provided in the supplementary.

5.4 Ablation Study

Two-stage Habit Imitation Learning Strategy. As shown in Fig. 5, we compare habit imitation results with and without the two-stage learning strategy. The results show that incorporating this strategy improves the generalization of habit motion. In contrast, direct training tends to produce less mouth opening and more uniform motions. The corresponding quantitative metrics further demonstrate the effectiveness of the two-stage learning strategy.

Flow Matching and DDIM with reparameterize. We conduct an ablation study on Flow Matching and reparameterization, as shown in Fig. 6. Flow Matching maintains strong lip-sync performance across sampling steps, while the DDIM-based formulation is less stable, supporting the effectiveness of Flow Matching for efficient motion-space sampling and one-step generation. Removing reparameterization degrades both Flow Matching and DDIM, indicating that conventional velocity or noise prediction leads to unstable generation in our framework. In contrast, reparameterization improves sampling stability and yields clear performance gains. These results validate the importance of both components for accurate and efficient lip-synchronized motion generation.

6 Conclusion

In this paper, we present TalkLikeYou, an efficient talking head generation framework for habit imitation. Our method focuses on generating talking head videos with specific speaking habits, providing flexible customization for user requirements. We use Flow Matching to generate motions conditioned on multiple habit sources, enabling efficient one-step sampling during inference. We further adopt a two-stage habit imitation learning strategy to improve habit generalization. To evaluate habit imitation quality, we propose a new metrics PLAD that measures motion similarity and diversity using histogram features from several representative motion axes. Extensive experiments show that our method achieves strong performance in talking head generation and habit imitation, validating the effectiveness of both our framework and the proposed metrics.

AI USE STATEMENT

In this work, we used generative AI tools solely to assist with writing and polish the manuscript, including correcting grammar, refining wording, and improving readability. We did not use generative AI tools for tasks requiring disclosure under the ICLR policy. We have reviewed all AI-assisted text to ensure its accuracy, preserve the intended meaning, and verify its consistency with our methods and experimental findings. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our method generates high-fidelity talking-face videos that imitate a target person's speaking habits. Since these habits are recognizable aspects of individual behavior, we acknowledge the potential ethical risks and societal concerns associated with misuse, particularly on public platforms. To help mitigate these risks, we are committed to responsible research dissemination, supporting advances in deepfake detection, and promoting the safe and ethical use of generative media technologies.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of our method, training procedures, implementation settings, and evaluation protocols in the main paper and appendix. All pretrained components used in our pipeline are obtained from publicly available implementations, and all datasets used for training and evaluation are publicly accessible. We will release our source code and trained model checkpoints on GitHub to support independent reproduction and further research.

REFERENCES

Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pp. arXiv–2506, 2025.

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

Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. In *International Conference on Learning Representations*, volume 2025, pp. 91659–91671, 2025.

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

Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 0–0, 2019b.

Katharina Dobs, Isabelle Bülthoff, and Johannes Schultz. Identity information content depends on the type of facial movement. *Scientific reports*, 6(1):34301, 2016.

Yingruo Fan, Zhaojiang Lin, Jun Saito, Wenping Wang, and Taku Komura. Faceformer: Speech-driven 3d facial animation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18770–18780, 2022.

Guanwen Feng, Zhiyuan Ma, Yunan Li, Jiahao Yang, Junwei Jing, and Qiguang Miao. Faceedittalker: Controllable talking head generation with facial attribute editing. *arXiv preprint arXiv:2505.22141*, 2025.

Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22634–22645, 2023.

Jiazhi Guan, Zhiliang Xu, Hang Zhou, Kaisiyuan Wang, Shengyi He, Zhanwang Zhang, Borong Liang, Haocheng Feng, Errui Ding, Jingtuo Liu, et al. Resyncer: Rewiring style-based generator for unified audio-visually synced facial performer. In *European Conference on Computer Vision*, pp. 348–367. Springer, 2024.

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

Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 193–203, 2025.

Taekyung Ki, Dongchan Min, and Gyeongsu Chae. Float: Generative motion latent flow matching for audio-driven talking portrait. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14699–14710, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *arXiv preprint arXiv:2505.22647*, 2025.

Jijie Li, Jiankuo Zhao, Xiangyu Zhu, and Zhen Lei. The devil is in the leakage: A disentangled dual-purification framework for high-fidelity hairstyle transfer. *arXiv preprint arXiv:2607.11281*, 2026.

Tianqi Li, Ruobing Zheng, Minghui Yang, Jingdong Chen, and Ming Yang. Ditto: Motion-space diffusion for controllable realtime talking head synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9704–9713, 2025.

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. URL https://doi.org/10.1145/3130800.3130813.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 1896–1904, 2023a.

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

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 484–492, 2020.

Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid, August 2020. Version 0.3.0.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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

Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Christian Theobalt, and Justus Thies. Imitator: Personalized speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 20621–20631, 2023.

Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pp. 244–260. Springer, 2024.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Baiqin Wang, Xiangyu Zhu, Fan Shen, Hao Xu, Shukai Chen, and Zhen Lei. Et-talk: Effective training strategy to enhance synchrony and fidelity for talking face generation. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025a.

Baiqin Wang, Xiangyu Zhu, Fan Shen, Hao Xu, and Zhen Lei. Pc-talk: Precise facial animation control for audio-driven talking face generation. *arXiv preprint arXiv:2503.14295*, 2025b.

Baiqin Wang, Sen Chen, Jiankuo Zhao, Xiangyu Liu, Zhen Lei, and Xiangyu Zhu. Towards flexible, natural, efficient interaction for conversational talking face generation. *arXiv preprint arXiv:2606.31088*, 2026.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10039–10049, 2021.

Zidu Wang, Xiangyu Zhu, Tianshuo Zhang, Baiqin Wang, and Zhen Lei. 3d face reconstruction with the geometric guidance of facial part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1672–1682, 2024.

Sicheng Xu, Guojun Chen, Yu-Xiao Guo, Jiaolong Yang, Chong Li, Zhenyu Zang, Yizhong Zhang, Xin Tong, and Baining Guo. Vasa-1: Lifelike audio-driven talking faces generated in real time. *Advances in Neural Information Processing Systems*, 37:660–684, 2024.

Zhenhui Ye, Tianyun Zhong, Yi Ren, Ziyue Jiang, Jiawei Huang, Rongjie Huang, Jinzheng He, Chen Zhang, Zehan Wang, Xize Chen, et al. Mimictalk: Mimicking a personalized and expressive 3d talking face in minutes. *Advances in neural information processing systems*, 37:1829–1853, 2024.

Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8652–8661, 2023.

Yue Zhang, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. Musetalk: Real-time high quality lip synchronization with latent space inpainting. *arXiv preprint arXiv:2410.10122*, 2024.

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

Jiankuo Zhao, Xiangyu Zhu, Zidu Wang, and Zhen Lei. Stavatar: Soft binding and temporal density control for monocular 3d head avatars reconstruction. *arXiv preprint arXiv:2511.19854*, 2025.

Xukun Zhou, Fengxin Li, Ziqiao Peng, Xinyu Wang, Hongyan Liu, Zhaoxin Fan, and Jun He. Meta-learning empowered meta-face: Personalized speaking style adaptation for audio-driven 3d talking face animation. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025.

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

Yongming Zhu, Longhao Zhang, Zhengkun Rong, Tianshu Hu, Shuang Liang, and Zhipeng Ge. Infp: Audio-driven interactive head generation in dyadic conversations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10667–10677. IEEE, 2025.

A APPENDIX

In the supplementary material, we cover the following aspects:

1. **Further Explanations:** We clarify the speaking-habit imitation task, introduce the Flow Matching formulation, and explain the rationale behind PLAD evaluation.
2. **Implementation Details:** We provide further details on the unified habit space, the PLAD algorithm, and the implementation settings.
3. **Additional Experimental Results:** We present a user study, an analysis of generation diversity, a visualization of the habit space, and additional qualitative results.
4. **Discussion:** We discuss potential directions for future work and ethical considerations for the responsible use of our method.

We strongly encourage readers to watch the supplementary videos to get a clearer sense of our method's motion dynamics and imitation quality. These videos include an overview of our framework, comparisons with existing methods, and demonstrations of speaking-habit imitation. Together, they highlight our method's accurate lip synchronization, faithful habit imitation, and real-time generation capabilities.

B FURTHER EXPLANATIONS

B.1 SPEAKING HABIT IMITATION

Speaking-habit imitation aims to generate talking-head videos that reflect how a target person speaks through characteristic facial behaviors, particularly lip movements. Current video generative models can produce high-fidelity videos but often struggle to capture these fine-grained, person-specific articulation patterns, which require focused modeling of the lip region. Therefore, we adopt a motion-based generation method and use implicit keypoints as the representation for their high expressiveness and landmark-based semantic constraints. Specifically, we introduce a landmark-guided loss during implicit keypoint learning to minimize the distance between projected implicit keypoints and 2D facial landmarks. These properties enable us to focus the motion generator on lip-region dynamics, improving speaking-habit imitation.

Several methods have explored speaker-dependent styles related to speaking habits. However, speaking styles and speaking habits describe different levels of behavior. Speaking style broadly characterizes the overall manner of speaking, including rhythm, intonation, emotion, and expressiveness, and can be shared across individuals. In contrast, speaking habits refer to subtle, person-specific articulation patterns that remain relatively consistent across utterances. In this work, we focus on these habitual lip-motion patterns, which capture finer-grained aspects of individual speaking behavior than overall speaking style.

B.2 FLOW MATCHING

Flow Matching is a generative modeling approach that learns a continuous velocity field to transform a noise distribution into a target data distribution. In our framework, it learns to transport noise to clean motion along a linear interpolation path. As shown in Fig. 7(a), the RMSE relative to clean motion changes linearly with the timestep, whereas DDPM exhibits a nonlinear relationship under its noise schedule. This linear interpolation provides a more direct transport path, facilitating approximation of the learned dynamics with fewer sampling steps. In contrast, conventional diffusion sampling typically relies on repeated denoising updates to progressively recover motion from noise. By combining Flow Matching with direct clean-motion prediction, our generator further simplifies inference and produces the complete motion sequence in a single sampling step, supporting efficient, real-time generation.

B.3 PLAD EVALUATION

Our proposed PLAD uses three representative lip-articulation axes to characterize motion sequences through projection histograms: mouth opening, lip pursing, and lip widening. We consider these

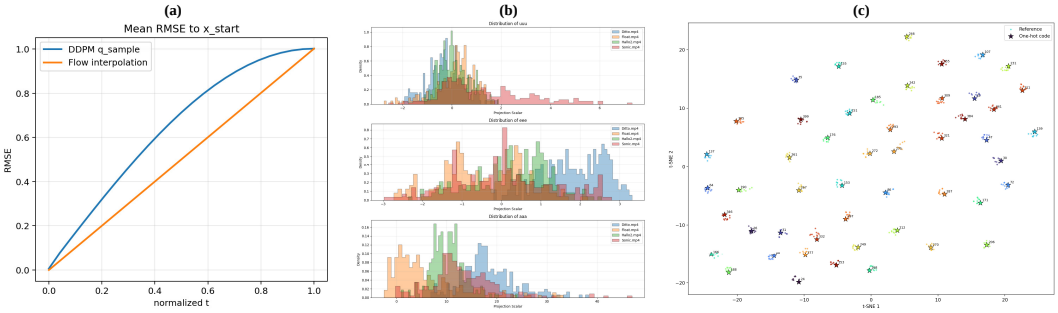


Figure 7: (a) Comparison of generation tendency under DDPM and Flow Matching. (b) PLAD histograms of different methods. (c) t-SNE visualization of the unified habit space, where stars denote preset one-hot habit features and dots denote reference-based habit features.

axes sufficient to capture key characteristics of speaking habits, as they represent common articulation patterns in daily communication, and many lip configurations can be approximated by their linear combinations. These axes thus provide a compact coordinate system for describing lip motion, analogous to the three spatial coordinates. The histogram along each axis reflects a person's habitual articulation tendencies in the corresponding dimension, collectively characterizing their speaking habit.

Existing motion evaluation protocols rarely assess speaking habits explicitly. In 3D facial animation, metrics such as MSE and RPAC primarily measure reconstruction accuracy and overall motion quality, rather than fine-grained, habit-specific articulation patterns. E-FID Deng et al. (2019b) measures distributional differences in expression features across subjects but lacks explicit semantic alignment at the motion level and offers limited insight into specific articulation patterns. In contrast, our PLAD metric projects motion onto three semantically meaningful lip-articulation axes, enabling targeted evaluation of subtle and consistent lip-motion variations that characterize individual speaking habits.

C Implementation Details

C.1 Unified Habit Space

One of our goals for habit-aware generation is to support reference-based habit imitation while also providing preset habit control when a reference video is unavailable. To this end, we project the two types of habit representations into a unified habit space. Specifically, we convert the reference video into a motion sequence and use the habit extractor E_{ref} to obtain its habit representation, while the one-hot habit code is processed by E_{oh} to produce the preset habit feature:

$$f_h = \begin{cases} E_{\text{ref}}(M_{\text{ref}}), & \text{reference video is available,} \\ E_{\text{oh}}(c), & \text{one-hot preset is used.} \end{cases}$$

E_{ref} is based on a Transformer Vaswani et al. (2017), while E_{oh} projects the one-hot code using an MLP. The habit feature f_h is then fed into the motion generator through both the input layer and the habit-injection layer, together with the diffusion timestep t, enabling the generated motion to follow the target speaking habit.

C.2 PLAD Algorithm

We evaluate speaking-habit similarity and generation diversity separately. For similarity, we first extract lip motions from the generated and reference videos using the motion extractor, then construct three projection histograms for each sequence following the procedure described in the main paper. Each sequence contains at least 500 frames, corresponding to 20 seconds at 25 fps, to obtain stable distribution estimates. We then compute histogram IoU and JS distance between the corresponding generated and reference histograms. We represent each histogram as an array of normalized bin values, allowing direct computation without additional sampling. For diversity, we evaluate more than 50 generated videos and compute the average pairwise JS distance between their PLAD descriptors.

Table 3: User Study. The rating is on scale of 1-5 where higher rank demonstrate better results.

Methods	Lip-sync Accuracy	Image Quality	Temporal Consistency	Habit Imitation	Overall Realistic
Ditto Li et al. (2025)	2.2	2.4	2.7	—	2.5
Hallo2 Cui et al. (2025)	2.2	2.8	3.3	—	3.0
Sonic Ji et al. (2025)	4.0	2.4	3.5	—	3.6
SadTalker Zhang et al. (2023)	1.2	1.9	2.4	—	1.9
StyleTalk Ma et al. (2023a)	1.3	1.9	1.6	1.5	1.7
PC-Talk Wang et al. (2025b)	3.1	2.4	3.6	3.3	3.4
TalkLikeYou(Ours)	**4.8**	**4.7**	**4.6**	**4.6**	**4.8**

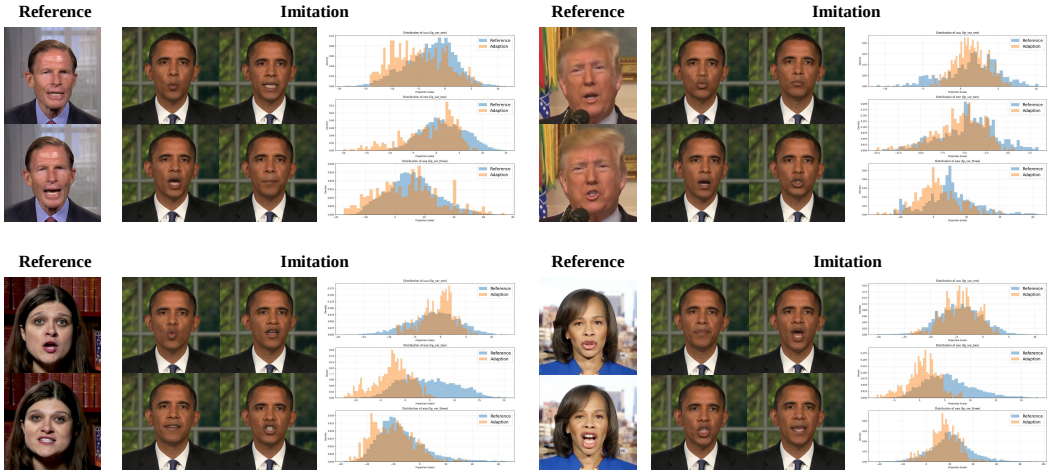


Figure 8: More results on habit imitation.

This score measures the distributional differences among generated motions in the lip-articulation space, with higher values indicating more diverse speaking habits.

C.3 Implementation Settings

We preprocess all videos at 25 fps and extract their audio and motion representations before training. The Flow Matching Motion Generator operates on 18-dimensional lip-motion features and takes a 100-frame reference motion sequence for habit conditioning. The generated motions are then rendered into 512×512 videos using rendering modules initialized from LivePortrait Guo et al. (2024), which offers a favorable balance between efficiency and visual quality. These rendering modules remain fixed while we optimize the motion generator. The motion extractor is also initialized with LivePortrait. We use 1000 flow-matching timesteps and train the generator with Adam Kingma & Ba (2014) at a learning rate of 1×10^{-4}. The two training stages take 48 and 72 hours, respectively, on a single RTX 3090 GPU. The first stage uses HDTF Zhang et al. (2021), which contains high-quality videos of 380 identities, each assigned a one-hot code as a preset habit option. The second stage uses CelebV-HQ Zhu et al. (2022), a large-scale, in-the-wild talking-face dataset, to improve generalization. Throughout training and inference, we use the Wav2Lip Prajwal et al. (2020) audio encoder to extract features aligned with lip movements.

At inference time, our framework achieves 32 FPS with one-step Flow Matching sampling. The rendering pipeline animates a cropped face region and stitches it back into the original frame, supporting both image animation and video dubbing. Before animation, we apply a neutralization procedure to establish a consistent initial lip configuration and mouth-opening distance across identities, reducing the effects of imperfect disentanglement in implicit keypoints. The neutral face retains a slightly open mouth to provide a visible reference for the teeth, helping our warp-based renderer preserve tooth structure and maintain temporal consistency throughout the video.



Figure 9: More comparison on Lip Synchronization with other baselines.

D Additional Experimental Results

D.1 Analysis on Generation Diversity

We use videos generated by other methods as references to investigate whether different methods exhibit distinguishable motion patterns. As shown in Fig. 7(b), the four methods exhibit distinct distributions in the PLAD space. Sonic tends to produce stronger projections along the lip-pursing axis, which may partly explain its high lip-sync scores if the synchronization metric favors such articulation patterns. Hallo2 tends to generate wider mouth movements, whereas FLOAT and Ditto produce smaller lip movements with more concentrated distributions. These differences suggest that each method implicitly learns characteristic lip-motion patterns, analogous to the distinct speaking habits of individuals. More importantly, our method can use these generated videos as references and adapt to their corresponding motion distributions. This shows that TalkLikeYou is not limited to real-world human reference videos, but can also capture method-specific motion patterns from AI-generated videos. In this sense, our framework provides a flexible platform to bridge the motion-distribution gap across different methods and generate talking motions with controllable, method-like speaking habits.

D.2 Visualization of the Habit Space

As shown in Fig. 7(c), we visualize the unified habit space using t-SNE, where stars represent preset one-hot habit features and dots represent features extracted from reference videos. Reference features associated with the same identity cluster around their corresponding preset feature, while different identities form distinct clusters. This suggests that our two-stage learning strategy aligns the two types of habit representations while preserving distinctions among speaking habits, supporting both reference-based imitation and preset habit control within a shared space. Furthermore, by interpolating between two points in the habit space, we can generate a new speaking habit that combines characteristics of both, enabling users to customize speaking habits by blending multiple sources.

D.3 User Study

We conduct a user study to evaluate five key dimensions: lip synchronization, image quality, temporal consistency, habit imitation, and overall realism. For each method, we use 10 generated video



Figure 10: More visualization results on stylize faces.

clips, each lasting more than 5 seconds. The questionnaire follows the Mean Opinion Score protocol with a five-point scale (1–5), where higher scores indicate better performance, and involves more than 20 participants. As shown in Tab. 3, TalkLikeYou achieves the highest scores across all evaluated dimensions, demonstrating its effectiveness in talking-head generation. In particular, our method performs well in habit imitation, faithfully reproducing target speaking habits while maintaining accurate lip synchronization, high visual quality, and temporal consistency. These results further validate the effectiveness of TalkLikeYou in generating realistic, habit-aware talking-head videos.

D.4 More Visible Results

We provide additional qualitative results for a more comprehensive evaluation. As shown in Fig. 10, our method generalizes to diverse portrait styles, including real photographs, AI-generated images, oil paintings, cartoons, and sketches, demonstrating the robustness of our renderer. We also provide further comparisons with baseline methods in Fig. 9, where our method maintains accurate lip synchronization while supporting diverse speaking habits. Moreover, we present additional habit imitation examples in Fig. 8. Our method faithfully imitates the reference speaking habits, with substantial overlap between the generated and reference PLAD histograms, further demonstrating its imitation capability.

E Discussion

Limitations and Future Work. While our method can generate talking faces with specific speaking habits, it may struggle to maintain motion stability for exaggerated articulation patterns, such as excessively wide mouth openings. Our framework also focuses primarily on the mouth region, where speaking habits are readily observable, rather than modeling person-specific behaviors across the entire face. In the future, we will explore incorporating eye movements and head pose dynamics to capture a more comprehensive range of individual speaking behaviors.

Ethical Considerations. Our method generates high-fidelity talking-face videos that imitate a target person's speaking habits. Since these habits are recognizable aspects of individual behavior, we acknowledge the potential ethical risks and societal concerns associated with misuse, particularly on public platforms. To help mitigate these risks, we are committed to responsible research dissemination, supporting advances in deepfake detection, and promoting the safe and ethical use of generative media technologies.