Outstanding Orthodontist: No More Artifactual Teeth in Talking Face

Zibo Su, **Ziqi Zhang**, **Kun Wei***, **Xu Yang** and **Cheng Deng**
Xidian University
{tk1076435968, zqzh9116, weikunsk, xuyang.xd, chdeng.xd}@gmail.com

Abstract

Audio-driven talking face synthesis (TFS) enables the creation of realistic speaking videos by combining a single facial image with a speech audio clip. Unlike other facial features that naturally deform during speech, teeth represent unique rigid structures whose shape and size should remain constant throughout the video sequence. However, current methods often produce temporal inconsistencies and artifacts in the teeth region, resulting in a less realistic appearance of the generated videos. To address this, we propose OrthoNet, a plug-and-play framework designed to eliminate unrealistic teeth effects in audio-driven TFS. Our method introduces a Detail-oriented Teeth Aligner module, designed to preserve teeth details and adapt to their shape. It works with a Memory-guided Teeth Stabilizer that integrates a long-term memory bank for global teeth structure and a short-term memory module for local temporal dynamics. Through this framework, OrthoNet acts like an orthodontist for existing Audio2Video methods, ensuring that teeth maintain natural rigidity and temporal consistency even under varying degrees of teeth occlusion. Extensive experiments demonstrate that our method makes the teeth in generated videos appear more natural during speech, significantly enhancing the temporal consistency and structural stability of audio-driven video generation.

1 Introduction

Audio-driven TFS [Prajwal *et al.*, 2020; Xu *et al.*, 2024c; Tian *et al.*, 2025] aims to generate realistic videos of a target person given a single portrait image and an audio speech clip as input. This technology enables automatic generation of synchronized facial animations from audio input, with broad applications in virtual communication, digital entertainment [Xu *et al.*, 2024a], and multimedia content creation. Recent methods such as TalkLip [Wang *et al.*, 2023], SyncTalk [Peng *et al.*, 2024], and TTSF [Jang *et al.*, 2024] have made significant progress in both lip synchronization and the consistency of facial expressions.

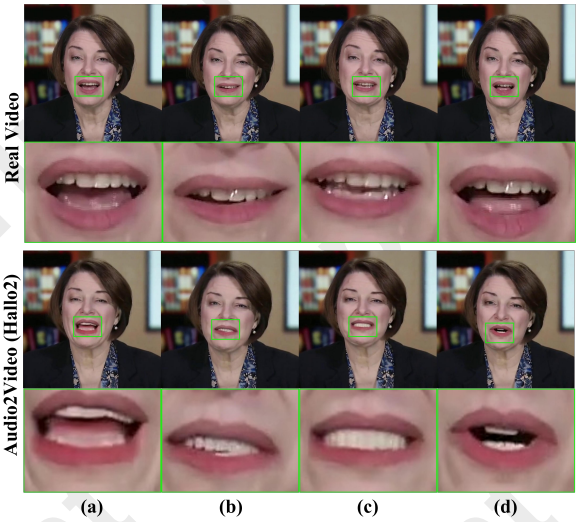


Figure 1: Comparison of real and audio-drive synthesized teeth.

However, a critical yet unexplored challenge remains: maintaining the temporal consistency and realism of teeth appearance under dynamic teeth occlusion during speech. Since the teeth occlusion phenomenon frequently occurs during speech, the generation of the teeth region shifts from an image editing problem to an image inpainting problem, introducing two additional technical challenges. Firstly, teeth inconsistency arises in the generated videos, as shown in Fig. 1 (a), (b), (c), and (d), with unnatural expansion or contraction of the teeth size across frames, leading to instability in the rigid structure. This issue is exacerbated by models that generate frames independently, causing inconsistent regeneration of the teeth region. Secondly, there is the problem of generation hallucination in the teeth region. As shown in Fig.1 (a), significant hallucinations appear in the lower teeth, while Fig.1 (c) shows severe degradation of teeth details and surface textures, with blurred gaps between teeth. Due to insufficient prior knowledge, models may create, remove, or distort teeth structures, resulting in unnatural textures, uneven edges, and unrealistic shapes. This issue is exacerbated by facial movements involving complex interactions between the lips and

*Corresponding author.

teeth. These phenomena make generated videos look obviously fake, negatively impacting viewer experience.

To address these challenges, we propose the **Ortho**dontist **Net**work (**OrthoNet**), which focuses on the teeth region of generated talking face videos to enhance their realism. Inspired by the outstanding orthodontists in aligning and stabilizing teeth, we also propose the Detail-oriented Teeth Aligner and Memory-guided Teeth Stabilizer to eliminate artifacts in the teeth region of TFS. OrthoNet introduces the Detail-oriented Teeth Aligner module to address generation hallucinations. One branch preserves fine details of the target person's teeth region using dense remap convolution (DRM-Conv), while the other adapts the teeth shapes during lip movements through polymorphic kernel convolution (PKConv). Additionally, the Memory-guided Teeth Stabilizer combines long-term memory (LTM) for global teeth features with short-term memory (STM) for fine-grained dynamics during speech, ensuring temporal consistency in the teeth region.

Our key contributions are as follows:

- To the best of our knowledge, our paper is the first to introduce and address the temporal inconsistency and artifact of the teeth region in the TFS task.
- Guided by prior information about the target person's teeth, the Detail-oriented Teeth Aligner module models the structural patterns in the dynamic talking face to eliminate generation hallucinations.
- The Teeth Stabilizer leverages the interaction between LTM and STM to maintain a stable teeth appearance, even with the frequent disturbance caused by the occlusion phenomenon.
- Our method is a plug-and-play framework that can be integrated with other Audio2Video methods to enhance generation realism, as demonstrated by extensive qualitative and quantitative experimental results.

2 Related Work

Audio-driven TFS makes significant strides with deep learning advancements [Tao *et al.*, 2025]. Early methods focus on direct audio-visual mappings, with foundational works like Chen et al. [Chen *et al.*, 2018] and Zhou et al. [Zhou *et al.*, 2019] establishing basic frameworks for lip synchronization and expression generation.

Recent developments bring substantial improvements in synthesis quality and controllability. SyncDiffusion [Zhao *et al.*, 2024] introduces a diffusion-based framework that achieves remarkable temporal coherence and lip-sync accuracy. HeyGen [Li *et al.*, 2024] proposes a attention mechanism specifically designed for handling dynamic facial features during speech. Additionally, SadTalker [Zhang *et al.*, 2023a] develops a robust emotion-aware generation pipeline that significantly enhances the naturalness of facial expressions. Similar approaches for feature enhancement and artifact detection have been explored in other domains [Tao *et al.*, 2021].

The latest research focuses on enhancing fine-grained control and generalization [Xu *et al.*, 2024b]. DiffTalk [Wang *et al.*, 2024] employs a conditional diffusion model that enables precise control over facial attributes while maintaining speech synchronization. VideoReTalking [Xu *et al.*, 2024d] achieves impressive results in cross-identity synthesis through careful disentanglement of speech content and speaker identity. VASA [Xu *et al.*, 2024c] further advances the field by introducing a view-adaptive speaking architecture that handles varying head poses effectively.

Several works specifically target temporal consistency. TalkLip [Wang *et al.*, 2023] introduces a temporal discriminator architecture that significantly improves lip movement stability. SpeechToFace [Zhang *et al.*, 2024] develops an advanced temporal coherence loss that enhances the smoothness of facial transitions. EMO [Tian *et al.*, 2025] proposes an emotional-aware framework that maintains consistency while incorporating expressive variations.

Multi-modal methods also gain prominence. AudioStyle [Chen *et al.*, 2024a] combines audio features with style transfer techniques to achieve more personalized facial animations. StyleTalk [Liu *et al.*, 2024] leverages style-based generation to enhance the visual quality of synthesized faces. These methods demonstrate significant improvements in generating realistic and temporally coherent facial animations.

Despite these advances, existing methods typically treat all facial features uniformly without specific consideration for rigid structures like teeth. While recent works make substantial progress in lip synchronization and overall facial animation quality, they do not address the unique challenges of maintaining teeth temporal consistency throughout the generated video sequence. This oversight often leads to temporal inconsistencies and artifacts in the teeth region, particularly under dynamic teeth occlusion during speech.

3 Method

As shown in Fig. 2, we present OrthoNet, a plug-and-play framework for teeth restoration in talking face videos that operates as a virtual orthodontist. It has two main parts: Detail-oriented Teeth Aligner for extracting precise teeth features and Memory-guided Teeth Stabilizer for maintaining stability over time. Detail-oriented Teeth Aligner uses dual-branch: DRM-Conv preserves fine-grained teeth details, and PKConv captures the overall shape of the teeth, ensuring both small parts and the whole structure are accurately modeled. Memory-guided Teeth Stabilizer has a dual-memory stream inspired by post-orthodontic treatment. LTM maintains the temporal consistency of the teeth, and STM enables teeth adaptation to speech dynamics, similar to how orthodontists use different retainers to retain strength while adapting naturally during speech.

3.1 Detail-oriented Teeth Aligner

The Detail-oriented Teeth Aligner module extracts precise teeth features through a dual-branch method, similar to the method used by orthodontists with different tools to adjust teeth position and shape.

Fine-grained Feature Extraction. Given an input feature map $F_{in} \in \mathbb{R}^{C\times H\times W}$, where C represents the channel dimension and H, W denote spatial dimensions, our

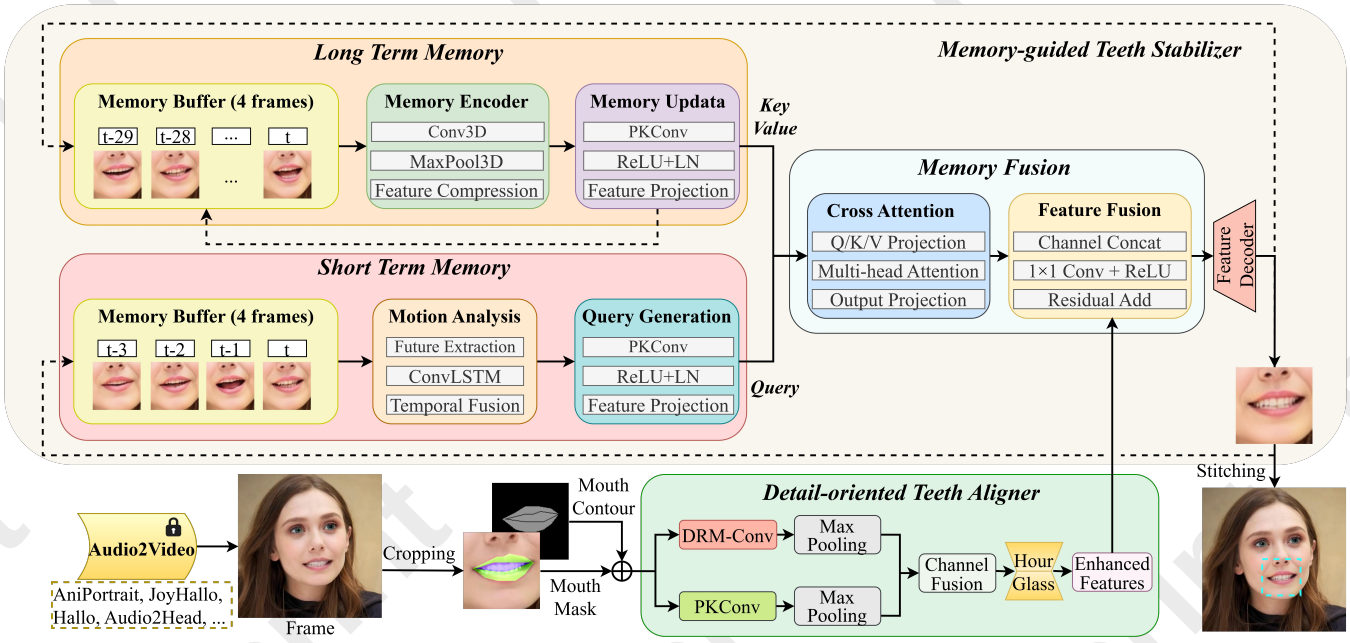


Figure 2: The framework of OrthoNet. Our architecture consists of: Detail-oriented Teeth Aligner - a dual-branch fine-grained feature extraction module with DRM-Conv for detail preservation and PKConv for shape adaptation; and Memory-guided Teeth Stabilizer - combining LTM bank for maintaining global teeth structure consistency and STM module for capturing local temporal dynamics. Cross-attention mechanism fuses information from both memory modules to ensure temporal consistency while preserving teeth details.

dual-branch structure extracts comprehensive teeth features through two complementary pathways: DRM-Conv for preserving fine-grained details and PKConv for dynamic shape adaptation. The DRM-Conv branch, illustrated in Fig. 3a, preserves critical fine-grained details through a space-to-depth transformation followed by specialized convolution operations:

$$F_{drm} = \text{Conv}(\text{Cat}(F_{in}[i:N:scale, j:N:scale]), W_{drm}), \tag{1}$$

where $scale = 2$ is empirically determined to maintain optimal balance between computational efficiency and preservation of critical edge details. The operation $\text{Cat}(\cdot)$ concatenates feature maps along the channel dimension, while W_{drm} represents learnable convolution weights [Sunkara and Luo, 2022] specifically designed for detail enhancement. For handling complex teeth shapes and occlusions, the PKConv branch, depicted in Fig. 3b, employs an adaptive sampling mechanism:

$$F_{pk}(p) = \sum_{k \in P_n} w_k(p) \cdot F_{in}(p + \Delta p_k + f_{offset}(F_{in}, M_{teeth})), \tag{2}$$

where p represents the spatial position, P_n denotes the sampling pattern with n points, $w_k(p)$ are position-specific adaptive weights, and $f_{offset}(F_{in}, M_{teeth})$ computes dynamic offsets [Zhang *et al.*, 2023b] based on both input features and teeth segmentation mask M_{teeth}. The multi-scale features from both branches are integrated through an adaptive fusion mechanism:

$$F_{out} = \sum_{l=1}^{L} \beta_l \cdot \text{Up}(\alpha_l \cdot F^l_{drm} + (1 - \alpha_l) \cdot F^l_{pk}), \tag{3}$$

where $\alpha_l \in [0, 1]$ dynamically balances the contribution of local details and global structure at each scale level l, $\text{Up}(\cdot)$ performs bilinear upsampling to align feature resolutions, and β_l are learnable scale-specific weights that optimize the importance of each resolution level. This multi-scale fusion ensures comprehensive feature extraction across different spatial scales, crucial for accurate teeth alignment planning.

Feature Quality Enhancement. To ensure pixel-level accuracy between the generated teeth region I_g and the ground truth I_{gt}, we employ:

$$\mathcal{L}_{rec} = \alpha_1 \|I_g - I_{gt}\|_1 + \alpha_2 \|I_g - I_{gt}\|_2^2. \tag{4}$$

This reconstruction loss [Li *et al.*, 2023] ensures pixel-level accuracy and perceptual similarity between generated and ground truth images. Additionally, we introduce a teeth perception loss to capture structural characteristics:

$$\mathcal{L}_{teeth} = \beta \cdot L_{struct} + (1 - \beta) \cdot L_{detail}, \tag{5}$$

where:

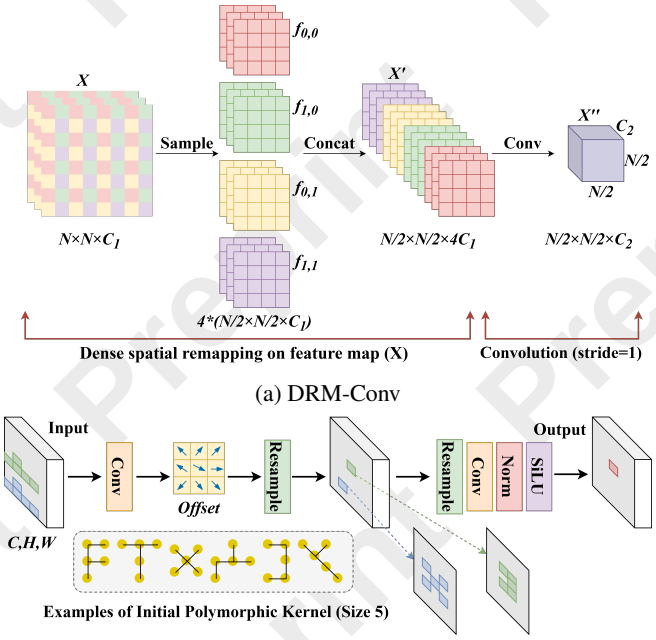
$$\mathcal{L}_{struct} = \frac{1}{N_t} \sum_{p \in M_t} \|\varphi_l(I^p_g) - \varphi_l(I^p_{gt})\|_2^2, \tag{6}$$

$$\mathcal{L}_{detail} = \frac{1}{N_t} \sum_{p \in M_t} \|\varphi_h(I^p_g) - \varphi_h(I^p_{gt})\|_1. \tag{7}$$

Here, M_t denotes the teeth region mask, φ_l and φ_h represent low-level and high-level VGG-19 features respectively, with β set to 0.7.

3.2 Memory-guided Teeth Stabilizer

To maintain temporal consistency while preserving the structural details of teeth, we propose a Memory-guided Teeth Stabilizer that integrates both long-term and short-term temporal dependencies [Xu *et al.*, 2021], similar to how orthodontic retainers stabilize teeth after treatment, as detailed in Algorithm 1.



(a) DRM-Conv

(b) PKConv

Figure 3: Illustrations of DRM-Conv and PKConv.

Long-term Memory Module. The LTM module maintains a feature buffer $B_L \in \mathbb{R}^{30\times C}$ storing teeth structural information from past 30 frames. Quality assessment (Q_{struct}) evaluates each new frame's teeth features before updating the buffer. Only high-quality frames meeting the threshold θ_q are retained through Top-K selection, ensuring the buffer maintains representative exemplars of teeth structure. This adaptive update mechanism prevents degradation from poor-quality frames while preserving stable global features.

Short-term Memory Module. The STM module employs a ConvLSTM architecture to capture frame-to-frame variations in teeth appearance within a 4-frame sliding window (S_{st}). Through specialized PKConv-based queries, it generates hidden states (h_t) that encode local temporal dependencies. This enables focused attention on relevant teeth features and smooth transitions between frames. The STM adapts to dynamic speech patterns while maintaining local consistency.

Memory Integration. The framework integrates LTM and STM features through a cross-attention mechanism. Attention weights (w_{lt}) determine the relevance of long-term features to the current frame. Temporal fusion combines ConvLSTM states with current features, while an adaptive blending factor (α) balances global stability from LTM with local dynamics from STM based on frame quality. This dual-memory method ensures both temporal consistency and natural motion during speech. The temporal consistency is enforced through two key loss components:

First, we utilize an adversarial loss based on WGAN-GP:

$$\mathcal{L}_{adv} = \mathbb{E}_{I_g\sim\mathbb{P}_g}[D(I_g)] - \mathbb{E}_{I_{gt}\sim\mathbb{P}_r}[D(I_{gt})] + \lambda_{gp}\mathcal{L}_{gp}, \quad (8)$$

where $D(\cdot)$ is the discriminator, $\mathbb{P}_g$ and $\mathbb{P}_r$ are distributions of generated and real images, and $\mathcal{L}_{gp}$ is the gradient penalty

Algorithm 1 Memory Bank Update in OrthoNet

Require: Current frame F_t, Long-term buffer B_L, Short-term states S_{st}, Quality threshold θ_q
Ensure: Temporally consistent features F_{out}

$Q_{struct} \leftarrow \text{QualityAssess}(F_t)$
if $Q_{struct} > \theta_q$ **then**
  $f_{teeth} \leftarrow \text{ExtractTeethFeatures}(F_t)$
  $B_{new} \leftarrow \text{Top-K}(B_L \cup \{f_{teeth}\}, K = 30)$
  $B_L \leftarrow B_{new}$
end if
$S_{new} \leftarrow S_{st}[2:] \cup \{F_t\}$
$h_t \leftarrow \text{ConvLSTM}(F_t, S_{new})$
$S_{st} \leftarrow S_{new}$
$w_{lt} \leftarrow \text{SoftmaxAttn}(F_t, B_L)$
$f_{lt} \leftarrow \sum_{i=1}^{30} w_{lt}^i \cdot B_L^i$
$f_{st} \leftarrow \text{TemporalFusion}(h_t, S_{st})$
$\alpha \leftarrow \sigma(Q_{struct})$
$F_{out} \leftarrow \alpha \cdot f_{lt} + (1-\alpha) \cdot f_{st}$
return F_{out}

term. Second, we introduce a temporal consistency loss:

$$\mathcal{L}_{temp} = \underbrace{\|I_g^t - w(I_g^{t-1})\|_1}_{\text{local consistency}} + \gamma \underbrace{\|h_t - f_{lstm}(h_{t-1}, I_g^t)\|_1}_{\text{memory consistency}}, \quad (9)$$

where $w(\cdot)$ represents optical flow warping, h_t is the ConvLSTM hidden state, and $f_{lstm}(\cdot)$ is the state transition function.

The overall loss function is formulated as:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{adv} + \lambda_3\mathcal{L}_{temp} + \lambda_4\mathcal{L}_{teeth}, \quad (10)$$

where $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$ are empirically set to balance different loss terms.

4 Experiments

In this section, we present experimental results evaluating our OrthoNet framework. We introduce evaluation metrics for teeth quality and temporal consistency, then compare against state-of-the-art (SOTA) methods and conduct ablation studies to validate our design choices.

4.1 Experimental Settings

Dataset and Data Processing. We train the proposed method on High-Definition Talking Face (HDTF) [Zhang *et al.*, 2021] dataset and our self-built high-resolution news anchor dataset. HDTF contains videos of 362 different identities with a total duration of 15.8 hours. The reason for choosing HDTF is that the teeth regions in this dataset are relatively clear and well-exposed during speech. Additionally, to enrich the diversity of teeth appearances under different speaking conditions and enhance the model's ability to learn stable and accurate teeth representations, we constructed a supplementary dataset collected from various news channels. This self-built dataset features news anchors with appropriate speaking speeds and high probability of teeth exposure, making it ideal for capturing teeth features. The videos in this dataset have resolutions of 720P or 1080P with a total duration of 24.7 hours. We will publicly release our dataset to facilitate future research. The details of data processing are shown in Fig 4.

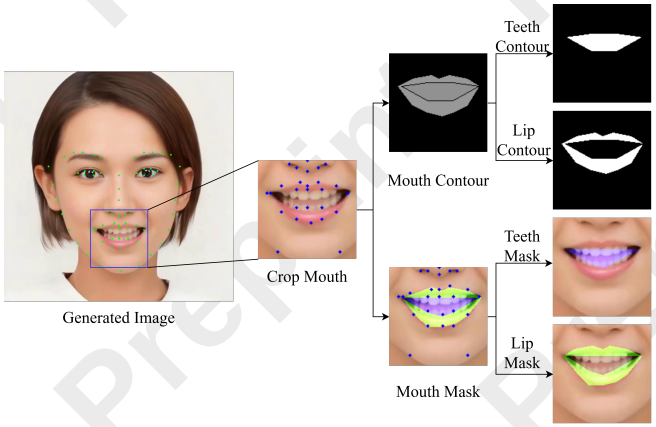


Figure 4: Details regarding data processing. For each video frame that is generated, face landmarks are detected in order to precisely locate the mouth region. This is achieved by identifying crucial points like the left and right corners of the mouth, the tip of the nose, and the jaw. Subsequently, the mouth region is cut out. Moreover, masks for the teeth and lips, along with their contours, are created based on these identified key points. Eventually, the cropped regions and the generated masks are resized to a resolution of 96×96 pixels, which helps to guarantee consistent resolution for all the input data.

Evaluation Indicators. To evaluate teeth generation and temporal consistency in talking face videos, we propose a framework combining traditional and specialized metrics:

- Fréchet Video Distance (FVD):

$$\text{FVD} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}), \quad (11)$$

where μ_r, μ_g are mean feature embeddings, and Σ_r, Σ_g are covariance matrices of real and generated video sequences. Extracted using the I3D network, FVD evaluates temporal coherence of the face region.

- Structural Similarity Index (SSIM):

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}, \quad (12)$$

where μ_x, μ_y are local means, σ_x^2, σ_y^2 are variances, and σ_{xy} is covariance. Constants c_1, c_2 stabilize division.

- Teeth Stability Metric (TSM):

$$\text{TSM} = 1 - \frac{1}{T-1} \sum_{t=1}^{T-1} \frac{1}{3} \left(\Delta r_{pa}(t) + \Delta r_{ar}(t) + \Delta a_n(t)\right). \quad (13)$$

TSM quantifies temporal consistency of teeth shapes across frames, using perimeter-area ratio (r_{pa}), aspect ratio (r_{ar}), and normalized area (a_n).

- Edge Clarity Index (ECI):

$$\text{ECI} = \frac{1}{T} \sum_{t=1}^{T} \frac{1}{N_e} \sum_{(x,y) \in E_t} \frac{\sqrt{G_x^2(x,y) + G_y^2(x,y)}}{G_{\max}}. \quad (14)$$

ECI evaluates sharpness and definition of teeth edges, where E_t are edge pixels (Canny operator), G_x, G_y are Sobel gradients, and $G_{\max}$ is the maximum gradient magnitude.

Method	TSM↑	ECI↑	SSIM↑	FVD↓
AniPortrait	0.753	4.068	0.847	238.5
AniPortrait+OrthoNet	**0.847**	**4.321**	**0.884**	**233.8**
Audio2Head	0.662	3.968	0.899	242.4
Audio2Head+OrthoNet	**0.785**	**4.213**	**0.907**	**237.9**
EchoMimic	0.728	4.013	0.856	347.9
EchoMimic+OrthoNet	**0.843**	**4.276**	**0.889**	**343.5**
JoyHallo	0.713	4.003	0.869	170.8
JoyHallo+OrthoNet	**0.826**	**4.239**	**0.895**	**166.1**
Hallo2	0.686	3.953	0.879	158.2
Hallo2+OrthoNet	**0.813**	**4.249**	**0.906**	**153.6**
Real video	0.882	4.386	0.939	-

Table 1: Comprehensive comparison of different methods with and without OrthoNet enhancement. ↑ indicates higher is better, ↓ indicates lower is better.

Implementation Details. We implement our framework using PyTorch and train it on four A6000 GPUs. The input frames are processed to 96×96 resolution with detected facial landmarks for teeth and lip mask generation. For network architecture, we set scale $= 2$ in DRM-Conv. The memory modules maintain a 30-frame long-term buffer and 4-frame short-term window based on ablation studies. During training, we employ the Adam optimizer with a learning rate of 1×10^{-5} and train for 30,000 epochs with batch size 4. The model is trained on the combined HDTF dataset and our self-built dataset.

4.2 Experimental Results

Quantitative Analysis. We evaluate OrthoNet against five SOTA TFS methods: AniPortrait [Wei *et al.*, 2024], Audio2Head [Wang *et al.*, 2021], EchoMimic [Chen *et al.*, 2024b], JoyHallo [Shi *et al.*, 2024], and Hallo2 [Cui *et al.*, 2024]. For each baseline, we compare its original version with a OrthoNet-enhanced version (denoted as "+OrthoNet"). Table 1 presents quantitative results across all evaluation metrics.

Our framework demonstrates significant improvements in shape stability, with TSM scores increasing by 9.8% to 15.7% across different baselines. Notably, EchoMimic+OrthoNet achieves the largest improvement from 0.728 to 0.843, while Hallo2+OrthoNet reaches the highest absolute score of 0.813 among enhanced methods. In terms of structural quality, the integration of OrthoNet consistently improves teeth edge definition and overall structural similarity, with ECI improvements of 0.2-0.3 points observed across all methods. AniPortrait+OrthoNet achieves particularly strong gains with a 6.2% increase in ECI. The temporal coherence, as measured by FVD scores, also shows consistent enhancement, with JoyHallo+OrthoNet achieving the most substantial improvement through a 2.8% reduction in FVD. As shown in Table 1, our method achieves significant improvements in teeth-specific metrics (TSM, ECI, SSIM) for the mouth region evaluation, while also demonstrating modest gains in global video quality as measured by FVD.

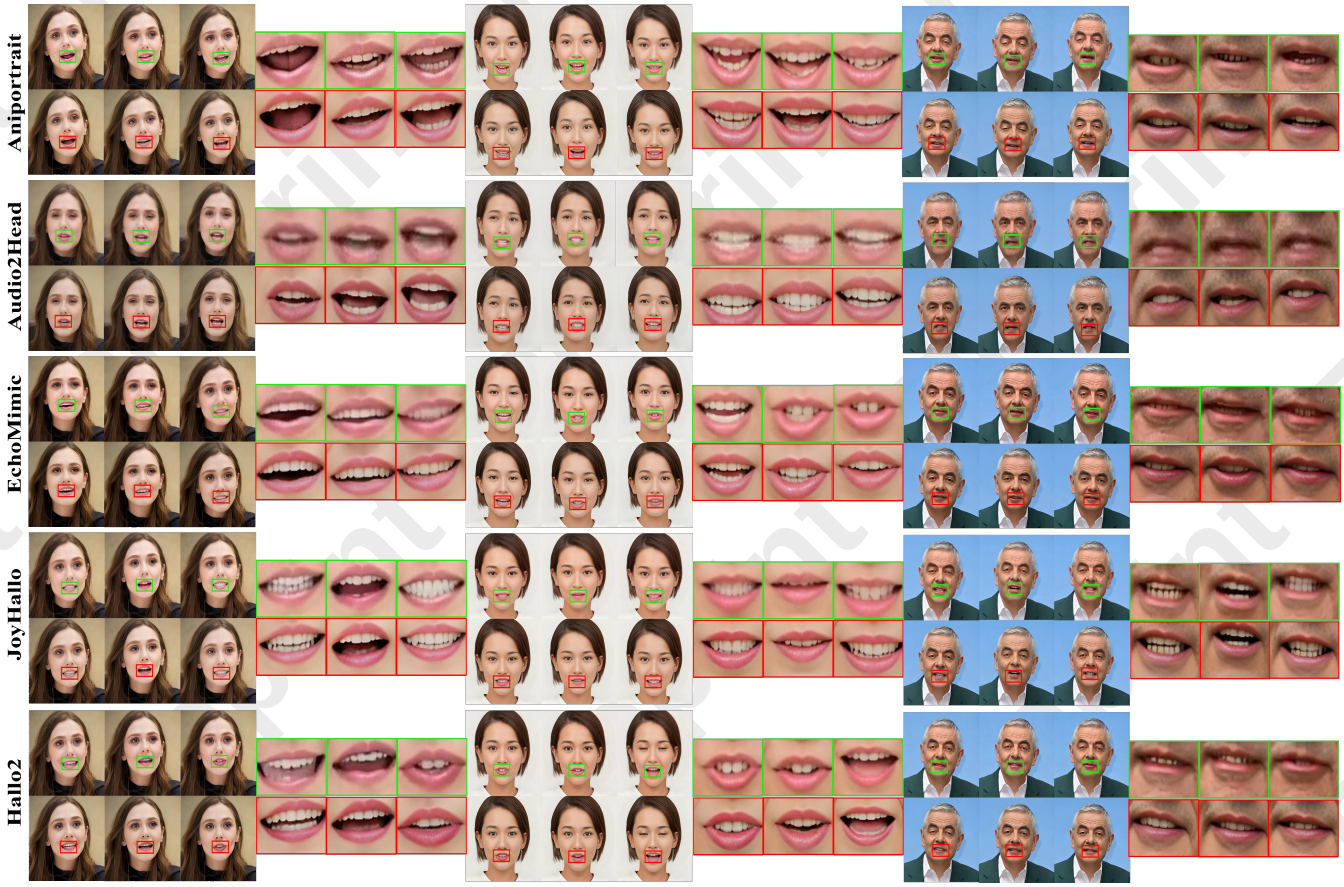


Figure 5: Qualitative comparison. In the images, green boxes highlight results generated by the original baseline methods, while red boxes show results after incorporating our proposed approach. Our OrthoNet framework consistently improves temporal stability and temporal consistency of teeth across different baseline methods, particularly under varying degrees of teeth occlusion.

Qualitative Analysis. Fig. 5 presents visual comparisons highlighting the qualitative improvements brought by OrthoNet. The enhanced geometric consistency is particularly evident in the preservation of teeth shapes and proportions across frames during rapid mouth movements. Our method maintains consistent teeth width and height ratios under varying degrees of teeth occlusion, whereas baseline methods often exhibit noticeable shape fluctuations.

The preservation of fine-grained teeth features represents another significant improvement in our results. OrthoNet successfully maintains consistent teeth edges and surface textures, effectively preventing the common issue of detail loss during mouth motion. Furthermore, our framework exhibits robust performance in handling dynamic teeth occlusion, where OrthoNet successfully reconstructs partially visible teeth while maintaining structural coherence, addressing a major limitation of existing methods.

The memory-guided method also enables more natural transitions in teeth visibility. As shown in Fig. 5, our method eliminates abrupt changes and artifacts during mouth opening/closing sequences, resulting in more realistic mouth animations. This improvement is particularly noticeable in sequences with rapid speech patterns, where baseline methods often struggle to maintain consistency.

Cross-Method Performance Analysis. OrthoNet demonstrates its effectiveness as a general enhancement module through consistent performance improvements across diverse baseline architectures, with enhancement ratios varying by less than 3%. This highlights its adaptability and generalization capability. The framework complements baseline methods, preserving their strengths while improving teeth-related quality. For example, JoyHallo retains its strong lip synchronization performance with a 15.7% gain in teeth stability. Moreover, methods with lower initial teeth generation performance (e.g., EchoMimic) show larger absolute improvements, while high-performing baselines (e.g., Hallo2) achieve notable gains. These results validate OrthoNet's ability to enhance teeth consistency in TFS while maintaining compatibility with diverse baselines, proving the effectiveness of our memory-guided method.

4.3 Ablation Studies

We conduct ablation studies on OrthoNet's three key components—memory modules, feature extraction, and loss functions—to validate their effectiveness (Table 2).

Method	TSM↑	ECI↑	SSIM↑	FVD↓
Memory-guided Teeth Stabilizer Ablation				
w/o Memory	0.695	3.982	0.883	157.8
w/o LTM	0.741	4.087	0.891	156.2
w/o STM	0.775	4.163	0.897	154.8
Full Model	**0.813**	**4.249**	**0.906**	**153.6**
Detail-oriented Teeth Aligner Ablation				
Standard Conv	0.702	3.993	0.884	157.5
w/o DRM-Conv	0.748	4.096	0.892	155.9
w/o PKConv	0.777	4.174	0.897	154.7
Full Model	**0.813**	**4.249**	**0.906**	**153.6**
Loss Function Ablation				
Basic Loss	0.708	4.012	0.885	157.2
w/o $\mathcal{L}_{\text{temp}}$	0.753	4.112	0.893	155.6
w/o $\mathcal{L}_{\text{teeth}}$	0.779	4.169	0.897	154.5
Full Model	**0.813**	**4.249**	**0.906**	**153.6**

Table 2: Ablation studies on different components of OrthoNet using Hallo2 as the baseline method.

LTM	STM	TSM↑	ECI↑	SSIM↑	FVD↓
10	4	0.767	4.128	0.893	155.9
20	4	0.789	4.186	0.898	154.8
30	4	**0.813**	**4.249**	**0.906**	**153.6**
40	4	0.811	4.245	0.905	153.8
30	2	0.785	4.173	0.896	154.9
30	3	0.798	4.208	0.901	154.2
30	**4**	**0.813**	**4.249**	**0.906**	**153.6**
30	5	0.812	4.247	0.905	153.7

Table 3: Analysis of memory size configurations in OrthoNet.

Memory Module Analysis. We evaluate different memory configurations to validate our dual-memory design. Removing both memory modules causes significant degradation (TSM: -14.5% to 0.695), with LTM showing greater impact than STM when ablated individually (TSM: -8.9% vs -4.7%). For optimization of memory capacity, as presented in Table 3, experiments demonstrate that a 30-frame LTM buffer and a 4-frame STM window achieve optimal performance (TSM: 0.813). Smaller LTM sizes (10/20 frames, TSM: 0.767/0.789) or STM windows (2/3 frames, TSM: 0.785/0.798) lead to reduced stability, while larger sizes yield minimal improvements. These results validate both the effectiveness of our dual-memory architecture and its optimal configuration settings. As shown in Fig. 6, the model produces irregular teeth variations and temporal instability without memory modules, particularly evident in the sudden changes of teeth shape and size between consecutive frames, as well as the degradation of fine-grained details during mouth movements.

Feature Extraction Analysis. Our dual-branch design significantly outperforms standard convolution (TSM: -13.7% to 0.702). Between the two specialized branches, DRM-Conv shows greater importance in preserving teeth details, as its removal causes more substantial degradation (TSM: -8.0% to 0.748) compared to PKConv ablation (TSM: -4.4% to 0.777).

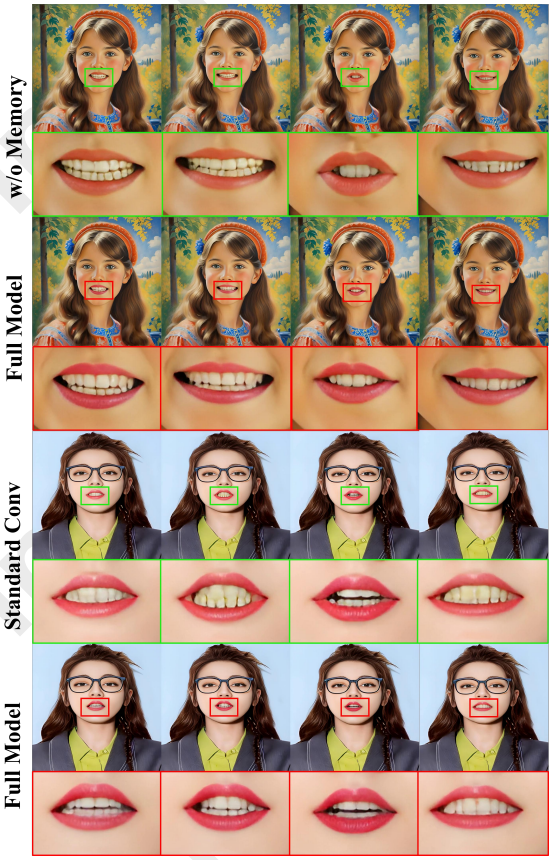


Figure 6: Visualization of partial ablation study results.

This indicates the crucial role of spatially-aware deformable convolution in capturing teeth morphological features. As shown in Fig. 6, standard convolution causes detail loss and unrealistic artifacts, manifesting as blurred teeth edges and inconsistent gaps between teeth that fail to capture the natural structural patterns present in real teeth during speech.

Loss Function Analysis. Using only basic reconstruction and adversarial losses limits the model's performance (TSM: 0.708). The $\mathcal{L}_{\text{teeth}}$ proves important, as its removal leads to notable degradation (TSM: -4.2% to 0.779) compared to temporal loss ablation (TSM: -7.4% to 0.753). The complete loss function achieves the best results across all metrics (TSM: 0.813, FVD: 153.6), validating our loss design.

5 Conclusion

We introduce OrthoNet, a plug-and-play framework for improving teeth realism in audio-driven TFS. It consists of two components: the Teeth Aligner, which preserves detail during mouth movement, and the Teeth Stabilizer, which ensures consistency through a memory system combining structural patterns with motion features. OrthoNet integrates easily into existing synthesis systems and addresses teeth hallucination and inconsistency issues during complex movements. Experimental results demonstrate significant improvements in temporal consistency and structural stability, enhancing visual realism in speech synthesis.

Ethical Statement

This research on talking face synthesis technology focuses solely on enhancing the realism of teeth representation for legitimate applications such as digital entertainment, education, and accessibility. We acknowledge the potential dual-use concerns of synthetic media technologies and emphasize that our work aims to improve visual quality rather than enable deception. All datasets used in this study were ethically sourced from public domain materials or with appropriate permissions. We support responsible development practices, including transparency about synthetic content and the advancement of detection technologies alongside generation capabilities.

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China (No. 2023YFC3305600), the Joint Fund of Ministry of Education of China (8091B02072404), the National Natural Science Foundation of China (62132016, 62171343, and 62406238), the Natural Science Basic Research Program of Shaanxi (2020JC-23), the Fundamental Research Funds for the Central Universities (ZYTS25149), and the National Key Laboratory Foundation of China (Grant No. HTKJ2024KL504011).

References

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

[Chen *et al.*, 2024a] Xingyu Chen, Jun Wu, Youjin Wang, and Hong Zhang. Audiostyle: Audio-driven talking face generation with style transfer. *arXiv preprint arXiv:2401.08742*, 2024.

[Chen *et al.*, 2024b] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. *arXiv preprint arXiv:2407.08136*, 2024.

[Cui *et al.*, 2024] 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. *arXiv preprint arXiv:2410.07718*, 2024.

[Jang *et al.*, 2024] Youngjoon Jang, Ji-Hoon Kim, Junseok Ahn, Doyeop Kwak, Hong-Sun Yang, Yoon-Cheol Ju, Il-Hwan Kim, Byeong-Yeol Kim, and Joon Son Chung. Faces that speak: Jointly synthesising talking face and speech from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8818–8828, 2024.

[Li *et al.*, 2023] Yongyuan Li, Xiuyuan Qin, Chao Liang, and Mingqiang Wei. Hdtr-net: A real-time high-definition teeth restoration network for arbitrary talking face generation methods. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 89–103. Springer, 2023.

[Li *et al.*, 2024] Xinjie Li, Yichen Song, Kaihao Zhang, and Wenhan Yang. Heygen: Speaking style-preserving audio-driven talking face generation. *arXiv preprint arXiv:2401.09920*, 2024.

[Liu *et al.*, 2024] Zhimeng Liu, Xiaodong Xie, Wenqiang Wang, and Yuming Guo. Styletalk: One-shot talking head generation with high-fidelity identity. *arXiv preprint arXiv:2401.09447*, 2024.

[Peng *et al.*, 2024] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024.

[Prajwal *et al.*, 2020] 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*, pages 484–492, 2020.

[Shi *et al.*, 2024] Sheng Shi, Xuyang Cao, Jun Zhao, and Guoxin Wang. Joyhallo: Digital human model for mandarin. *arXiv preprint arXiv:2409.13268*, 2024.

[Sunkara and Luo, 2022] Raja Sunkara and Tie Luo. No more strided convolutions or pooling: A new cnn building block for low-resolution images and small objects. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 443–459. Springer, 2022.

[Tao *et al.*, 2021] Renshuai Tao, Yanlu Wei, Xiangjian Jiang, Hainan Li, Haotong Qin, Jiakai Wang, Yuqing Ma, Libo Zhang, and Xianglong Liu. Towards real-world x-ray security inspection: A high-quality benchmark and lateral inhibition module for prohibited items detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10923–10932, 2021.

[Tao *et al.*, 2025] Renshuai Tao, Manyi Le, Chuangchuang Tan, Huan Liu, Haotong Qin, and Yao Zhao. Oddn: Addressing unpaired data challenges in open-world deepfake detection on online social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 799–807, 2025.

[Tian *et al.*, 2025] 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*, pages 244–260. Springer, 2025.

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

[Wang *et al.*, 2023] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition, pages 14653–14662, 2023.

[Wang *et al.*, 2024] Zhiyuan Wang, Fei Zhao, Tianyi Li, and Jiashuo Zhang. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. *arXiv preprint arXiv:2401.09145*, 2024.

[Wei *et al.*, 2024] Huawei Wei, Zejun Yang, and Zhisheng Wang. Aniportrait: Audio-driven synthesis of photorealistic portrait animation. *arXiv preprint arXiv:2403.17694*, 2024.

[Xu *et al.*, 2021] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021.

[Xu *et al.*, 2024a] Chenghao Xu, Guangtao Lyu, Jiexi Yan, Muli Yang, and Cheng Deng. Llm knows body language, too: Translating speech voices into human gestures. In *ACL*, pages 5004–5013, 2024.

[Xu *et al.*, 2024b] Chenghao Xu, Jiexi Yan, Muli Yang, and Cheng Deng. Rethinking noise sampling in class-imbalanced diffusion models. *IEEE Transactions on Image Processing*, 2024.

[Xu *et al.*, 2024c] 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. *arXiv preprint arXiv:2404.10667*, 2024.

[Xu *et al.*, 2024d] Yuqi Xu, Yue Liu, Qianyu Dong, and Chenxin Xu. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. *arXiv preprint arXiv:2401.07874*, 2024.

[Zhang *et al.*, 2021] 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*, pages 3661–3670, 2021.

[Zhang *et al.*, 2023a] 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*, 2023.

[Zhang *et al.*, 2023b] Xin Zhang, Yingze Song, Tingting Song, Degang Yang, Yichen Ye, Jie Zhou, and Liming Zhang. Akconv: Convolutional kernel with arbitrary sampled shapes and arbitrary number of parameters. *arXiv preprint arXiv:2311.11587*, 2023.

[Zhang *et al.*, 2024] Kai Zhang, Wei Li, Yu Liu, and Feng Jiang. Speechtoface: High-fidelity facial animation synthesis from speech. *arXiv preprint arXiv:2401.11167*, 2024.

[Zhao *et al.*, 2024] Jianhong Zhao, Wei Zhang, Xiaoyu Zhou, and Yuxuan Wang. Syncdiffusion: More realistic talking face generation with neural diffusion models. *arXiv preprint arXiv:2401.12251*, 2024.

[Zhou *et al.*, 2019] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *AAAI Conference on Artificial Intelligence*, 2019.