OR-TALK: OCCLUSION-ROBUST LIP SYNCHRONIZATION VIA SELECTIVE APPEARANCE TRANSPORT

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-driven lip synchronization modifies mouth motion according to a target speech signal while preserving identity and head motion. Under facial occlusion, existing methods often distort or remove foreground occluders, since preserving occluder appearance conflicts with audio-driven control of mouth shape. We propose OR-Talk, which formulates occluder recovery as an appearance transport problem. Based on this formulation, we introduce Occlusion Adaptive Transport (OAT) to selectively transfer the appearance cues needed for occluder reconstruction from the identity features to the decoder. To avoid leaking mouth information during this process, we construct paired identity conditions with matched head poses but different mouth shapes. This design preserves occluder appearance while keeping the mouth shape controlled by the input audio. Experiments show competitive performance on standard benchmarks and substantially improved robustness to facial occlusion. Under 50–100% mouth occlusion, OR-Talk improves PSNR by 6.65 dB and reduces OCC-E by 51.6% compared with X-Dub.

1 INTRODUCTION

Lip synchronization aims to modify mouth shape in a portrait video according to a target audio signal while preserving the original identity and head motion. It has been widely applied to visual dubbing (Prajwal et al., 2019; Xie et al., 2021), digital avatars (Thies et al., 2020; Zielonka et al., 2023), and video conferencing (Wang et al., 2021; Mukhopadhyay et al., 2024). Existing methods have explored GAN-based generation (Prajwal et al., 2020; Guan et al., 2023; Zhang et al., 2023b; Zhou et al., 2020; Park et al., 2022), diffusion-based synthesis (Bigioi et al., 2024; Du et al., 2023; Zhang et al., 2023a; Li et al., 2024; Yu et al., 2026; 2024), and motion-appearance disentanglement (Zhang et al., 2024; Yu et al., 2024; Tan et al., 2024; Yao et al., 2024), substantially improving visual fidelity, audio-visual synchronization, and controllability. Despite these advances, most existing methods focus on clearly visible faces and rarely consider foreground occlusions. Such occlusions are common in practical scenarios, where failure to preserve foreground objects can introduce noticeable artifacts and degrade generation quality. This makes robustness to facial occlusion an important problem in practical lip synchronization.

Facial occlusion introduces a challenge to audio-driven lip synchronization. Preserving an occluding object requires fine-grained appearance cues from the reference identity, whereas directly transferring dense identity features may also expose the reference mouth shape, providing a visual shortcut. This creates an inherent conflict between occluder appearance preservation and audio-driven mouth shape. Based on this observation, we argue that occluder recovery should be viewed as a selective appearance transport process, where only appearance information relevant to occluder reconstruction is transferred to the generated frame while mouth articulation remains controlled by the target audio.

To this end, we propose OR-Talk, an occlusion-robust lip synchronization framework. At its core, we introduce Occlusion Adaptive Transport (OAT), which selectively retrieves and transfers appearance cues from the identity branch to the generated frame. To prevent the appearance transport from leaking the reference mouth shape, we construct pose-aligned identity conditions with different mouth shapes. To further provide a suitable representation for selective transport, OR-Talk separately models identity appearance, head motion, and audio-driven mouth articulation through identity, motion, and audio branches and introduces Adaptive Audio Condition Injection (AACI) to coordinate visual

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.

and audio features during generation. In addition, we establish an occlusion-specific evaluation protocol and introduce Occlusion Error (OCC-E) to directly quantify reconstruction quality in occluded regions.

In summary, our main contributions are as follows:

- We identify the tension between occluder appearance preservation and audio-driven mouth control as a key challenge in occluder recovery.
- We introduce occluder recovery as selective appearance transport under audio-driven motion generation.
- OAT realizes this formulation through deformable cross-attention over sampled identity features, trained with pose-aligned identity conditions. Experiments show improved occlusion robustness and competitive performance on standard benchmarks.

2 Related Work

2.1 Audio-Driven Lip Synchronization

Early audio-driven lip synchronization methods primarily focused on reconstructing the mouth region while preserving the surrounding face. Wav2Lip (Prajwal et al., 2020) established effective audio-visual synchronization supervision, while DINet (Zhang et al., 2023b) and IP-LAP (Zhong et al., 2023) improved reconstruction through reference feature deformation and landmark guidance. Later work sought greater control over facial motion. VideoReTalking (Cheng et al., 2022) separates expression normalization from lip synchronization, while GC-AVT (Liang et al., 2022) and EDTalk (Tan et al., 2024) disentangle facial motion factors to enable finer control over articulation, pose, and expression.

More recent work improves synthesis quality and efficiency through latent-space generation and stronger separation of appearance and motion. MuseTalk (Zhang et al., 2024) performs efficient latent-space inpainting, while LatentSync (Li et al., 2024) strengthens audio conditioning in diffusion-based synthesis to reduce visual shortcuts. FD2Talk (Yao et al., 2024) and Make Your Actor Talk (Yu et al., 2024) further separate facial factors or speech-driven motion from appearance generation. Despite this progress, foreground occluder preservation remains less explored. X-Dub (He et al., 2025) uses generative bootstrapping to train a mask-free dubbing model that can preserve foreground occluders. OR-Talk instead treats occluder recovery as selective appearance transport, using pose-aligned identity conditions with different mouth shapes to reduce reference mouth leakage during training.

2.2 Occlusion-Robust Generation and Deformable Feature Fusion

Facial occlusion has been extensively studied in face segmentation, recognition, and reconstruction, where synthetic occluders are commonly introduced to improve robustness to foreground objects (Yin & Chen, 2022; Voo et al., 2022). FaceOcc (Yin & Chen, 2022) collects diverse real-world occlusions including hands, glasses, and microphones, while Voo et al. (2022) propose both naturalistic and random synthetic occlusion generation strategies and show that diverse synthetic occlusions can improve robustness even to unseen occluders. In lip synchronization, the problem is more constrained because the model must preserve the foreground occluder while simultaneously replacing the original mouth shape with a new one driven by the target audio.

Deformable convolution and deformable attention provide flexible mechanisms for aggregating spatially displaced features. Deformable Convolutional Networks (Dai et al., 2017) introduce learnable sampling offsets to adapt convolutional receptive fields to object geometry, and DCNv2 (Zhu et al., 2019) further improves this mechanism with modulation of sampling responses. Deformable DETR (Zhu et al., 2020) extends this idea to attention by attending to a small set of learned sampling locations around each query, substantially reducing the cost of dense global attention. Deformable Attention Transformer (Xia et al., 2022) further demonstrates that data-dependent sparse sampling allows attention to focus on relevant regions while suppressing interference from unrelated features. Inspired by these properties, OR-Talk introduces OAT to selectively retrieve occlusion-related appearance features from the identity branch through region-constrained sparse sampling.

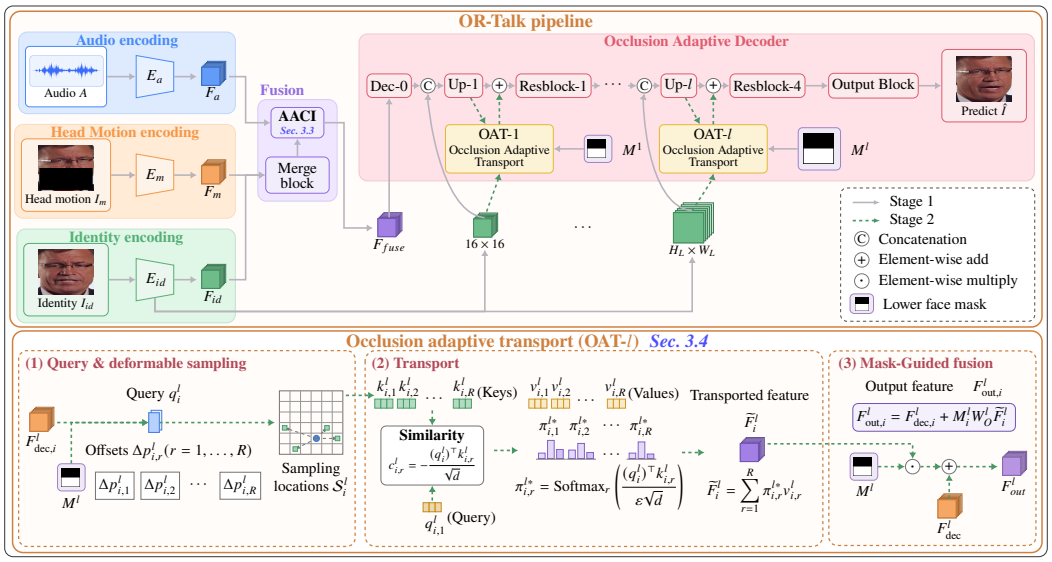


Figure 1: Overview of the proposed method.

3 METHOD

3.1 OVERVIEW

As shown in Figure 1, the proposed framework consists of an audio encoder E_a, a head motion encoder E_m, an identity encoder E_{id}, and an Occlusion Adaptive Decoder (OAD) D. The audio encoder extracts speech representations F_a to drive lip movements. The head motion encoder extracts F_m to provide head motion and background context. The identity encoder extracts a global identity feature F_{id} together with multiscale dense appearance features F_{dense}. F_m and F_{id} are first fused through a merge block to form a unified visual representation. The AACI module then introduces audio information into this representation, and the OAD decodes the resulting features to generate realistic face frames with accurate lip synchronization.

The model is trained in two stages. Stage I learns a base model that separates identity appearance, head motion, and audio-driven lip shape. In Stage II, the Stage I model is frozen, and only the components introduced for occlusion robustness are optimized. To support this adaptation, synthetic occluders are randomly applied to both the identity image and the masked head motion image to simulate different occlusion conditions. To prevent target mouth information from leaking through the identity branch, we also construct pose-aligned identity images with different mouth shapes. Based on these, OAT selectively transfers multiscale identity features to the decoder for better occluder reconstruction. During inference, we use the full model obtained after Stage II training.

Given an identity image I_{id}, a masked head motion image I_m, audio A, and a lower face mask M, the generation process is formulated as:

$$\hat{I}^{(s)} = \begin{cases} G_1(I_{\mathrm{id}}, I_m, A), & s = 1, \\ G_2(I_{\mathrm{id}}, I_m, M, A), & s = 2, \end{cases} \tag{1}$$

where s denotes the training stage. The generated frame is expected to preserve identity, follow the target head motion, and maintain accurate audio-visual synchronization.

3.2 SEPARATE CONDITION REPRESENTATION LEARNING

We use an identity encoder, a head motion encoder, and an audio encoder to explicitly separate identity appearance, head motion, and audio conditions, reducing their entanglement in audio-driven lip synchronization.

Audio encoder. The audio encoder extracts speech representations to drive mouth shapes and produces a latent audio feature F_a. To support spatial alignment and fusion with visual features, we map F_a into a compact 16×16 spatial representation. This representation connects speech cues with local mouth regions and nearby facial context, allowing more effective fusion with identity and motion features during decoding.

Head motion encoder. The head motion encoder extracts head pose and background context from a masked head motion image to guide the target head motion during generation. To prevent visual shortcuts, the mouth region is masked before encoding, removing access to the target mouth shape and avoiding ground truth lip leakage. The encoder therefore captures motion information, including head pose and partial background layout, from the remaining visible regions. This design preserves the target head motion while allowing the mouth shape to be controlled by the audio representation.
Identity encoder. The identity encoder extracts representations that preserve identity from the reference face. It produces a global identity feature and dense identity features. The global feature captures overall facial appearance, while the dense features provide detailed spatial cues for local facial reconstruction.

3.3 Adaptive Audio Condition Injection

Audio and visual features play different roles in lip synchronization. The motion feature F_m provides the target head pose, while the audio feature F_a mainly determines the phoneme dependent mouth shape. Directly concatenating or uniformly injecting these features requires the decoder to implicitly determine their relative contributions. Too much audio information may overwrite visual structure and cause unstable facial appearance, while an overly strong visual condition may weaken the audio-driven mouth shape. Inspired by the adaptive fusion mechanism (Dai et al., 2021), we design an AACI module to adaptively balance audio and visual features based on their joint representation.

After merging F_{id} and F_m into a unified visual representation F_v and spatially aligning F_v with F_a, we compute their joint representation $F_{\mathrm{av}} = F_v + F_a$ and predict an adaptive modulation weight:

$$w = \sigma\left(\Psi_{\mathrm{local}}(F_{\mathrm{av}}) + \Psi_{\mathrm{global}}(F_{\mathrm{av}})\right), \tag{2}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The local branch Ψ_{local} uses two convolutional layers to capture local relations, while the global branch Ψ_{global} applies global average pooling followed by a two-layer MLP to capture global information. The fused feature is then computed as:

$$F_{\mathrm{fused}} = w \odot F_v + (1 - w) \odot F_a, \tag{3}$$

where $\odot$ denotes element wise multiplication. The learned weight w provides an adaptive balance between visual and audio features.

3.4 Occluder Preservation in Lip Synchronization

Building on the separate condition encoding in Stage I, Stage II further focuses on preserving foreground occluders during lip synchronization. The model needs to preserve the appearance of the occluder while keeping the mouth shape controlled by the input audio. To achieve this, we treat occluders as appearance information carried by the identity image and formulate their reconstruction as an appearance transport process. Relevant appearance features are then selectively transferred from the identity branch to the corresponding regions of the generated frame.

To support appearance transport in Stage II, the identity image should be spatially aligned with the masked head motion image I_m. Directly using a target frame with the same pose would also reveal the target mouth shape. We therefore use the factorized control learned in Stage I to construct pose-aligned identity images with different mouth shapes. These images preserve the target pose while changing the original mouth shape, providing suitable identity conditions without exposing the target mouth content. Based on these pairs, we synthesize occluders at the same locations on both the identity image and the masked head motion image using the same occluder and geometric transformation.

3.4.1 Pose-Aligned Paired Data Construction

To construct identity conditions that preserve the target head pose while containing different mouth shapes, we use the Stage I model to synthesize a pose-aligned identity image as:

$$I_{\mathrm{id}}^{\mathrm{fake}} = G_1(I_{\mathrm{id}}, I_m, A'), \tag{4}$$

where G_1 is the Stage I model, I_{id} is the identity image, I_m is the masked head motion image that provides the target pose, and A' is sampled from a different segment of the target audio A. Thus, $I_{\mathrm{id}}^{\mathrm{fake}}$ preserves the target head pose while containing a different mouth shape.

During Stage II training, fake identity images and synthetic occlusion are introduced with probability ρ_{occ}. The Stage II model is trained as:

$$\hat{I}^{(z)} = G_2\left(\tilde{I}_{\mathrm{id}}^{(z)}, \tilde{I}_m^{(z)}, M, A\right), \tag{5}$$

where $z \sim \mathrm{Bernoulli}(\rho_{\mathrm{occ}})$ is a binary indicator that determines whether the pose-aligned fake identity image is used and synthetic occlusion is applied, and ρ_{occ} denotes the occlusion probability. $\hat{I}^{(z)}$ denotes the generated frame under the condition z. G_2 is the Stage II model, which takes the identity input $\tilde{I}_{\mathrm{id}}^{(z)}$, the masked head motion input $\tilde{I}_m^{(z)}$, the lower face mask M, and the audio condition A as input. When $z = 0$, no synthetic occlusion is applied. When $z = 1$, synthetic occluders are first added to the pose-aligned identity image and the unmasked head motion image. The lower-face mask is then applied to the occluded head motion image. The same occlusion augmentation is also applied to the ground truth frame to construct the corresponding supervision target.

3.4.2 Occlusion Adaptive Transport

Although the pose-aligned paired data reduce direct mouth leakage, they do not determine which identity features should be transferred to the generated frame. Direct skip fusion may transfer irrelevant appearance features. We therefore propose OAT to perform appearance transport between the identity branch and the decoder.

At the l-th skip level, OAT transfers appearance features from the dense identity feature F_{dense}^l to the decoder feature F_{dec}^l, where $F_{\mathrm{dense}}^l, F_{\mathrm{dec}}^l \in \mathbb{R}^{B\times C_l\times H_l\times W_l}$. To avoid unrestricted identity injection, the transport is activated only within the resized lower face mask $M^l \in \{0,1\}^{H_l\times W_l}$. For each decoder location p_i^l satisfying $M_i^l = 1$, we first construct a query from the current decoder feature $q_i^l = W_Q^l F_{\mathrm{dec},i}^l$ and use it to predict R deformable sampling offsets, $\Delta p_{i,r}^l = W_{\Delta p}^{l,r} q_i^l, \quad r = 1,\dots R$. Each offset specifies a candidate source location $p_i^l + \Delta p_{i,r}^l$ on the dense identity feature map. We then retrieve the corresponding keys and values from these sampled locations:

$$\begin{aligned} k_{i,r}^l &= W_K^l F_{\mathrm{dense}}^l\left(p_i^l + \Delta p_{i,r}^l\right), \\ v_{i,r}^l &= W_V^l F_{\mathrm{dense}}^l\left(p_i^l + \Delta p_{i,r}^l\right). \end{aligned} \tag{6}$$

We define the local transport cost as:

$$c_{i,r}^l = -\frac{(q_i^l)^\top k_{i,r}^l}{\sqrt{d}}. \tag{7}$$

We optimize an entropy regularized local appearance matching objective:

$$\pi_i^{l*} = \arg\min_{\pi_i^l\in\Delta^R}\left[\sum_{r=1}^R \pi_{i,r}^l c_{i,r}^l + \varepsilon\sum_{r=1}^R \pi_{i,r}^l \log \pi_{i,r}^l\right], \tag{8}$$

where Δ^R denotes the probability simplex and ε is the entropy regularization coefficient. This objective encourages the transport weights to favor sampled identity features with lower matching costs while maintaining a smooth local distribution. As derived in Appendix A, equation 8 admits the following closed-form solution:

$$\pi_{i,r}^{l*} = \mathrm{Softmax}_r\left(\frac{(q_i^l)^\top k_{i,r}^l}{\varepsilon\sqrt{d}}\right). \tag{9}$$

Thus, the local appearance transport can be efficiently implemented as deformable cross-attention over the sampled identity features. The transported appearance feature is computed as:

$$\widetilde{F}_i^l = \sum_{r=1}^R \pi_{i,r}^{l*} v_{i,r}^l. \tag{10}$$

Finally, OAT injects the transported feature through mask-guided residual fusion:

$$F_{\mathrm{out},i}^l = F_{\mathrm{dec},i}^l + M_i^l W_O^l \widetilde{F}_i^l. \tag{11}$$

where W_O^l, $W_{\Delta p}^{l,r}$, W_K^l, W_V^l and W_Q^l are learnable projections. The deformable attention retrieves spatially displaced appearance features, while the mask restricts their injection to the designated decoder region. Together with the pose-aligned paired data construction in Section 3.4.1, OAT transfers appearance cues associated with occlusion without directly exposing the target mouth shape.

3.5 Training Objectives

Following existing works (Prajwal et al., 2020; Zhang et al., 2023b; Chen et al., 2025), we train the proposed framework using a combination of reconstruction, perceptual, adversarial, and synchro-

nization objectives. The overall training objective is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_\mathrm{p} \mathcal{L}_\mathrm{p} + \lambda_\mathrm{m} \mathcal{L}_\mathrm{m} + \lambda_\mathrm{g} \mathcal{L}_\mathrm{g} + \lambda_\mathrm{s} \mathcal{L}_\mathrm{s}, \tag{12}$$

where λ_1, λ_p, λ_m, λ_g and λ_s are the corresponding loss weighting coefficients. Detailed formulations of these objectives are provided in Appendix B.

4 Experiments

4.1 Experimental setup

Datasets. We train our model on HDTF (Zhang et al., 2021), LRS2 (Afouras et al., 2018), and VoxCeleb2 (Chung et al., 2018). We exclude videos with poor audio-visual synchronization, low-resolution faces, or partially visible faces. For evaluation, we randomly select 30 videos from each dataset before training and exclude them from the training set. We additionally use 30 videos from CelebV-HQ (Zhu et al., 2022) for cross-identity evaluation. We also construct an occlusion-specific test set to evaluate robustness to facial occlusion. To avoid ground truth mouth leakage from the original videos, we first reconstruct the mouth region in the identity frames and then add synthetic occluders. More details are provided in Appendix C.

Implementation Details. All video frames are cropped to the face region and resized to 256×256. The audio is resampled to 16 kHz and converted into mel-spectrogram features. Our method is implemented in PyTorch and trained using the Adam optimizer (Kingma & Ba, 2014) on two NVIDIA A800 GPUs. Stage I trains the base lip synchronization model, while Stage II freezes the base model and optimizes only the OAT modules. We train the model at learning rate 0.0001 with batch size 4. The decoder uses four OAT layers. The loss weights λ_1, λ_p, λ_m, λ_g, and λ_s are set to 0.92, 1.0, 20.0, 0.04, and 0.04, respectively. The parameter ρ_{occ} is set to 0.06.

Evaluation Metrics. We evaluate visual quality using PSNR (Huynh-Thu & Ghanbari, 2008), SSIM (Wang et al., 2004), FID (Heusel et al., 2017), and FVD (Unterthiner et al., 2018). Perceptual and identity consistency are measured using LPIPS (Zhang et al., 2018) and CSIM (Deng et al., 2019). Audio-visual synchronization is evaluated using LSE-C, LSE-D (Prajwal et al., 2020), and mouth landmark distance (LMD) (Guan et al., 2023). In addition, we introduce OCC-E to evaluate occluder preservation. OCC-E is defined as:

$$\text{OCC-E} = \frac{\left\| M_\mathrm{occ} \odot \left(\hat{I} - I_\mathrm{gt} \right) \right\|_2}{\sqrt{\|M_\mathrm{occ}\|_1}}, \tag{13}$$

where $\hat{I}, I_\mathrm{gt} \in [0, 255]^{H \times W \times 3}$ and M_occ is broadcast across RGB channels. Lower OCC-E indicates better occluder reconstruction.

4.2 Comparison with State-of-the-Art Methods

We compare OR-Talk with several state-of-the-art lip synchronization methods, including Wav2Lip (Prajwal et al., 2020), DINet (Zhang et al., 2023b), IP-LAP (Zhong et al., 2023), MuseTalk (Zhang et al., 2024), VideoReTalking (Cheng et al., 2022), LatentSync (Li et al., 2024), and X-Dub (He et al., 2025). Quantitative results on HDTF, LRS2 and VoxCeleb2 are presented in Table 1.

In terms of visual quality, OR-Talk achieves the best PSNR, SSIM, FID, and FVD across all three datasets, outperforming the strongest baseline in overall visual quality. OR-Talk also achieves the lowest LPIPS while maintaining competitive CSIM scores, indicating strong perceptual similarity and identity preservation. In terms of lip synchronization, OR-Talk achieves the lowest LMD on all three datasets, the best LSE-C and LSE-D on LRS2, and the lowest LSE-D on VoxCeleb2. Overall, OR-Talk provides a strong balance among visual quality, identity consistency and audio-visual synchronization. Qualitative comparisons are provided in Appendix D.2.

We further evaluate cross-identity generalization on CelebV-HQ with results reported in Table 2. OR-Talk achieves 29.49 PSNR, 0.92 SSIM, 10.53 FID, and 62.92 FVD. It also achieves the lowest LMD (1.46) and the best LSE-C (6.85) and LSE-D (6.75). Its CSIM (0.88) is slightly lower than LatentSync (0.89) but remains higher than the scores achieved by most other methods.

4.3 Evaluation under Different Mouth Occlusion Ratios

We evaluate occlusion robustness using a benchmark with three mouth occlusion levels, covering 0%, 0–50%, and 50–100% of the mouth region. The 0% group contains no mouth occlusion but may

Table 1: Quantitative comparison on HDTF, LRS2, and VoxCeleb2.

Test Set	Method	Visual Quality				Persona		Lip Sync		
		PSNR↑	SSIM↑	FID↓	FVD↓	LPIPS↓	CSIM↑	LSE-C↑	LSE-D↓	LMD↓
HDTF	Wav2Lip	21.70	0.78	16.22	137.60	0.18	0.83	9.00	6.72	2.59
	DINet	25.75	0.87	10.46	120.84	0.11	0.88	8.07	7.03	1.82
	IP-LAP	27.89	0.90	11.25	79.01	0.09	**0.96**	8.68	6.35	1.31
	MuseTalk	25.78	0.87	9.39	158.92	0.12	0.88	7.93	7.20	1.56
	VideoReTalking	24.26	0.83	10.98	107.17	0.14	0.88	8.62	6.47	2.50
	LatentSync	29.41	0.91	5.93	58.95	0.08	0.94	**9.43**	**5.76**	1.36
	X-Dub	24.30	0.83	13.57	144.64	0.14	0.89	8.35	7.34	1.93
	OR-Talk	**29.98**	**0.93**	**4.77**	**35.13**	**0.05**	**0.96**	9.24	5.94	**1.10**
LRS2	Wav2Lip	21.90	0.79	38.75	245.43	0.19	0.74	7.82	5.79	4.48
	DINet	23.16	0.82	44.52	558.84	0.20	0.63	4.52	7.13	5.24
	IP-LAP	26.53	0.87	41.65	295.46	0.15	0.85	5.40	5.59	3.99
	MuseTalk	23.23	0.82	45.52	448.10	0.20	0.66	5.98	7.11	4.55
	VideoReTalking	24.27	0.84	39.66	205.47	0.16	0.80	7.61	5.70	4.73
	LatentSync	28.60	0.91	23.44	158.62	0.12	0.83	7.96	5.72	2.51
	X-Dub	24.88	0.85	32.97	251.39	0.17	0.80	7.93	7.04	3.42
	OR-Talk	**29.29**	**0.92**	**23.26**	**116.72**	**0.11**	**0.88**	**8.18**	**5.54**	**2.34**
VoxCeleb2	Wav2Lip	21.30	0.77	29.31	248.46	0.17	0.71	**9.20**	6.63	4.25
	DINet	23.41	0.83	27.49	340.19	0.15	0.69	6.86	8.06	4.05
	IP-LAP	26.87	0.88	21.92	202.27	0.11	**0.89**	6.96	7.82	2.98
	MuseTalk	22.31	0.80	39.58	368.47	0.18	0.64	7.29	7.83	4.52
	VideoReTalking	24.27	0.85	21.97	193.40	0.12	0.85	8.80	6.57	3.49
	LatentSync	27.67	0.90	17.41	129.84	0.10	0.85	8.36	6.62	2.60
	X-Dub	22.88	0.82	28.31	259.88	0.18	0.73	7.67	8.05	3.96
	OR-Talk	**28.63**	**0.92**	**15.56**	**116.31**	**0.08**	0.88	8.90	**6.01**	**2.35**

Table 2: Quantitative comparison of cross-identity generation on CelebV-HQ.

Method	Visual Quality				Persona		Lip Sync		
	PSNR↑	SSIM↑	FID↓	FVD↓	LPIPS↓	CSIM↑	LSE-C↑	LSE-D↓	LMD↓
Wav2Lip	21.63	0.74	34.56	183.46	0.20	0.78	6.46	7.63	3.14
DINet	23.60	0.81	24.15	255.81	0.14	0.75	4.23	9.29	3.34
IP-LAP	21.45	0.75	38.87	232.00	0.17	0.85	3.53	9.58	4.77
MuseTalk	21.41	0.74	39.13	286.13	0.19	0.69	4.63	8.82	3.45
VideoReTalking	24.17	0.80	22.67	118.77	0.15	0.86	6.84	7.05	3.00
LatentSync	27.46	0.87	13.71	96.05	0.10	**0.89**	5.83	7.53	1.80
X-Dub	24.02	0.82	24.09	148.45	0.15	0.86	5.86	8.27	2.08
OR-Talk	**29.49**	**0.92**	**10.53**	**62.92**	**0.07**	0.88	**6.85**	**6.75**	**1.46**

include occluders elsewhere on the face. Table 3 reports the quantitative results. OR-Talk achieves the best performance across all reported metrics. As mouth occlusion increases, most competing methods show worse FVD, LMD, and OCC-E scores, while OR-Talk maintains better performance across all evaluated occlusion levels. Under the most challenging 50–100% mouth occlusion setting, OR-Talk achieves a PSNR of 31.01, SSIM of 0.95, FID of 17.63, FVD of 72.02, LMD of 2.12, and OCC-E of 4.83. Compared with X-Dub, OR-Talk improves PSNR by 6.65 dB and reduces FID, FVD, LMD, and OCC-E by 4.00, 164.97, 0.63, and 5.15, respectively. The lower OCC-E indicates more accurate reconstruction of the occluder, while the low LMD suggests that OR-Talk can still recover plausible mouth motion even when the mouth region is heavily occluded.

Figure 2 shows qualitative comparisons under facial occlusion. Most competing methods produce incomplete occluders or texture errors around the mouth. X-Dub preserves the occluders more completely, but their colors differ from those in the input. OR-Talk better preserves their structure and colors. We also evaluate OR-Talk on samples with natural occlusions from AIGC-LipSync (Peng et al., 2025) with qualitative results provided in Appendix D.3.

4.4 Ablation Study

4.4.1 Analysis of Identity, Motion, and Audio Control

To examine whether OR-Talk provides independent control over identity, head motion, and audio-driven mouth shape, we vary each condition independently, as shown in Figure 3a. The first row shows the original configuration. In the second row, changing only the identity image leaves the head motion and mouth shape unchanged, suggesting that the identity branch mainly controls appearance. In the third row, swapping the head motion image and the identity image changes the head pose while preserving the identity and audio-driven mouth shape. In the fourth row, changing only the audio

Table 3: Quantitative comparison under different mouth occlusion ratios.

Method	0%						0–50%						50–100%					
	PSNR↑	SSIM↑	FID↓	FVD↓	LMD↓	OCC-E↓	PSNR↑	SSIM↑	FID↓	FVD↓	LMD↓	OCC-E↓	PSNR↑	SSIM↑	FID↓	FVD↓	LMD↓	OCC-E↓
Wav2Lip	17.08	0.67	48.43	387.58	3.34	38.07	14.95	0.57	94.49	783.42	5.85	40.39	14.60	0.54	111.58	997.12	8.81	39.59
DINet	22.54	0.81	31.86	206.13	2.87	19.49	17.33	0.69	81.30	672.89	6.66	34.09	16.31	0.66	97.26	754.45	8.70	36.55
IP-LAP	23.78	0.84	33.29	320.41	2.07	21.57	22.10	0.79	56.37	382.76	3.85	19.41	21.66	0.78	67.24	524.49	5.34	20.12
VideoReTalking	15.08	0.58	58.58	747.57	6.57	30.00	14.42	0.54	92.33	1161.50	8.04	29.21	15.32	0.60	98.29	1101.34	7.82	25.41
MuseTalk	24.29	0.84	24.81	112.40	1.83	11.98	22.01	0.78	56.94	240.99	3.25	13.50	21.95	0.77	69.48	345.79	4.37	13.72
LatentSync	24.66	0.86	28.95	117.35	1.76	15.81	20.37	0.77	63.44	356.29	3.88	19.98	19.71	0.75	79.42	601.01	6.53	20.67
X-Dub	24.04	0.82	19.50	190.13	2.19	8.57	23.86	0.82	23.33	218.01	2.20	9.75	24.36	0.83	21.63	236.99	2.75	9.98
OR-Talk	**27.74**	**0.91**	**14.27**	**84.41**	**1.56**	**6.72**	**28.97**	**0.93**	**18.26**	**104.76**	**2.18**	**5.15**	**31.01**	**0.95**	**17.63**	**72.02**	**2.12**	**4.83**



Figure 2: Qualitative comparison under facial occlusions.

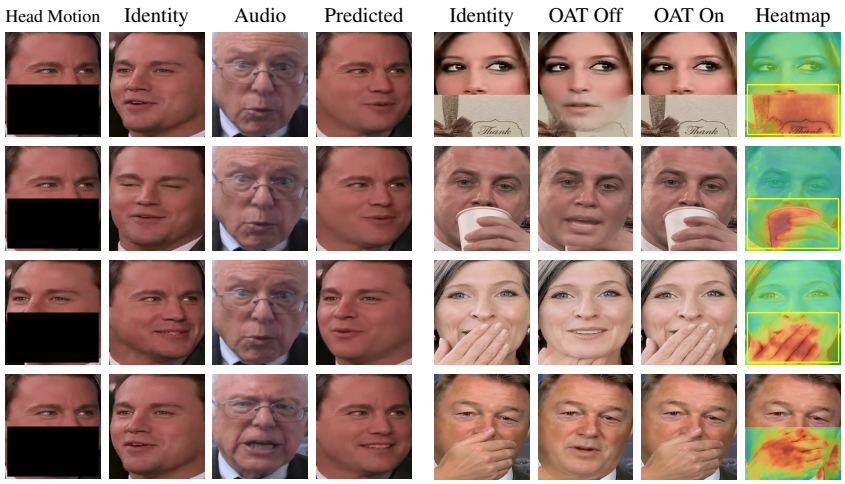
alters the mouth shape, whereas the identity and head motion remain stable. These results indicate that the three branches provide largely independent control over identity appearance, head motion, and mouth shape.

4.4.2 Effect of Occlusion Augmentation and Occlusion Adaptive Transport

We compare the full model with OR-Talk w/o Stage II and OR-Talk w/o OAT. The latter retains Stage II occlusion augmentation but removes the OAT modules. We also train Wav2Lip256 and DINet-OneFrame with the same occlusion augmentation, denoted by Wav2Lip256† and DINet-OneFrame†. Table 4 reports the results under 50–100% mouth occlusion.

For Wav2Lip256, FID decreases from 101.85 to 42.43 and OCC-E decreases from 26.43 to 10.64. For DINet-OneFrame, FID decreases from 93.44 to 31.84 and OCC-E decreases from 27.08 to 10.57. These results show that occlusion augmentation improves occluder preservation. Compared with the full OR-Talk, OR-Talk w/o Stage II shows lower performance across all metrics. In particular, FID increases from 17.63 to 90.26, while OCC-E increases from 4.83 to 27.14.

OR-Talk w/o OAT achieves competitive results, with FID 28.98 and OCC-E 7.20, but performs worse than the full OR-Talk model. Introducing OAT further reduces FID to 17.63 and OCC-E to 4.83, while also improving PSNR, SSIM, FVD, and LMD. These results show that occlusion augmentation improves occluder preservation, while OAT provides further gains in reconstruction quality.



(a) Identity-motion-audio factorized control. (b) OAT heatmap and occlusion recovery.

Figure 3: Qualitative ablation studies. (a) Analysis of identity, motion, and audio control. (b) Visualization of the OAT heatmap and its effectiveness in occlusion recovery.

Table 4: Ablation study under 50%–100% Mouth Occlusion.

Method	PSNR↑	SSIM↑	FID↓	FVD↓	LMD↓	OCC-E↓
Wav2Lip256	19.07	0.67	101.85	845.25	6.39	26.43
DINet-OneFrame	18.72	0.70	93.44	1206.51	8.36	27.08
OR-Talk w/o Stage II	18.81	0.70	90.26	1045.53	6.91	27.14
Wav2Lip256†	23.07	0.81	42.43	<u>154.80</u>	3.59	10.64
DINet-OneFrame†	<u>23.66</u>	<u>0.91</u>	<u>31.84</u>	157.01	<u>3.24</u>	<u>10.57</u>
OR-Talk w/o OAT	26.19	0.87	28.98	139.63	2.79	7.20
OR-Talk	**31.01**	**0.95**	**17.63**	**72.02**	**2.12**	**4.83**

4.4.3 Effect of Occlusion Adaptive Transport

To validate the effectiveness of OAT, we compare the generated results with and without OAT and visualize the corresponding feature heat maps, as shown in Figure 3b. Without OAT, the model largely relies on the audio-driven generation prior to fill the masked lower face region, causing the occluders to disappear or lose their original structure. With OAT enabled, appearance cues from the identity branch are selectively transported into the decoder, leading to substantially better preservation of the occluders.

The heatmaps show that OAT responses are mainly concentrated on the occluder in the lower face region, with substantially weaker transport in visible facial areas. This suggests that OAT does not simply copy dense identity features, but selectively transfers appearance information where it is needed. As a result, the shape and texture of the occluders are better recovered, while the surrounding facial content remains largely preserved.

5 Conclusion

We present OR-Talk for robust audio-driven lip synchronization under facial occlusion. By formulating occluder recovery as selective appearance transport, OAT preserves the appearance of occluders without disrupting audio-driven mouth shape. Pose-aligned identity conditions reduce mouth information leakage from the reference, while separate modeling of identity, head motion, and audio improves control over generation. Experiments show clear improvements in occluder reconstruction under severe occlusion while maintaining competitive visual quality and lip synchronization in nonoccluded settings, highlighting the effectiveness of selective appearance transport in balancing occluder preservation and audio-driven mouth control.

AI USE STATEMENT

We used generative AI tools, including ChatGPT, for language editing and translation. We additionally used these tools to assist with manuscript organization, LaTeX formatting, title suggestions, and auxiliary code editing. All AI-assisted outputs were critically reviewed, revised, and verified by the authors. Experimental results and conclusions were independently validated by the authors, who take full responsibility for the final content of this work.

ETHICS STATEMENT

OR-Talk aims to improve lip synchronization under facial occlusion for applications such as video dubbing and digital avatars. The same technology could be misused to create misleading videos, impersonate individuals, or alter speech without consent. Its improved ability to preserve occluders may also make manipulated content harder to recognize visually. We therefore emphasize obtaining appropriate consent from depicted individuals and clearly disclosing generated or edited content. Any release of data or models should account for privacy, applicable licenses, and potential misuse.

REPRODUCIBILITY STATEMENT

Section 3 describes the model architecture, pose-aligned identity condition construction, Occlusion Adaptive Transport, and training objectives. Section 4.1 provides the datasets, preprocessing procedures, implementation details, and evaluation metrics, including OCC-E. The ablation studies in Section 4.4 examine the effects of the main design choices. Together, these descriptions provide the basis for reproducing our training and evaluation.

REFERENCES

T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. In *arXiv:1809.02108*, 2018.

Dan Bigioi, Shubhajit Basak, Michał Stypułkowski, Maciej Zieba, Hugh Jordan, Rachel McDonnell, and Peter Corcoran. Speech driven video editing via an audio-conditioned diffusion model. *Image and Vision Computing*, 142:104911, 2024.

Bo Chen, Tao Liu, Qi Chen, Xie Chen, and Zilong Zheng. Imtalker: Efficient audio-driven talking face generation with implicit motion transfer. *arXiv preprint arXiv:2511.22167*, 2025.

Kun Cheng, Xiaodong Cun, Yong Zhang, Menghan Xia, Fei Yin, Mingrui Zhu, Xuan Wang, Jue Wang, and Nannan Wang. Videoretalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022.

Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *arXiv preprint arXiv:1806.05622*, 2018.

Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 764–773. Ieee, 2017.

Yimian Dai, Fabian Gieseke, Stefan Oehmcke, Yiquan Wu, and Kobus Barnard. Attentional feature fusion. In *2021 IEEE winter conference on applications of computer vision (WACV)*, pp. 3559–3568. IEEE, 2021.

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, 2019.

Chenpeng Du, Qi Chen, Tianyu He, Xu Tan, Xie Chen, Kai Yu, Sheng Zhao, and Jiang Bian. Dae-talker: High fidelity speech-driven talking face generation with diffusion autoencoder. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 4281–4289, 2023.

Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1505–1515, 2023.

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.

Xu He, Haoxian Zhang, Hejia Chen, Changyuan Zheng, Liyang Chen, Songlin Tang, Jiehui Huang, Xiaoqiang Liu, Pengfei Wan, and Zhiyong Wu. From inpainting to editing: Unlocking robust mask-free visual dubbing via generative bootstrapping. *arXiv preprint arXiv:2512.25066*, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008.

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

Chunyu Li, Chao Zhang, Weikai Xu, Jingyu Lin, Jinghui Xie, Weiguo Feng, Bingyue Peng, Cunjian Chen, and Weiwei Xing. Latentsync: Taming audio-conditioned latent diffusion models for lip sync with syncnet supervision. *arXiv preprint arXiv:2412.09262*, 2024.

Borong Liang, Yan Pan, Zhizhi Guo, Hang Zhou, Zhibin Hong, Xiaoguang Han, Junyu Han, Jingtuo Liu, Errui Ding, and Jingdong Wang. Expressive talking head generation with granular audio-visual control. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3377–3386. IEEE, 2022.

Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Soumik Mukhopadhyay, Saksham Suri, Ravi Teja Gadde, and Abhinav Shrivastava. Diff2lip: Audio conditioned diffusion models for lip-synchronization. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5280–5290. IEEE, 2024.

Se Jin Park, Minsu Kim, Joanna Hong, Jeongsoo Choi, and Yong Man Ro. Synctalkface: Talking face generation with precise lip-syncing via audio-lip memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2062–2070, 2022.

Ziqiao Peng, Jiwen Liu, Haoxian Zhang, Xiaoqiang Liu, Songlin Tang, Pengfei Wan, Di Zhang, Hongyan Liu, and Jun He. Omnisync: Towards universal lip synchronization via diffusion transformers. *Advances in Neural Information Processing Systems*, 38:11405–11424, 2025.

K. R. Prajwal, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and C. V. Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 1428–1436, 2019. doi: 10.1145/3343031.3351066.

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.

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Advances in neural information processing systems*, 32, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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.

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*, pp. 716–731. Springer, 2020.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Kenny TR Voo, Liming Jiang, and Chen Change Loy. Delving into high-quality synthetic face occlusion segmentation datasets. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4711–4720. IEEE, 2022.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10034–10044. IEEE, 2021.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4784–4793. IEEE, 2022.

Tianyi Xie, Liucheng Liao, Cheng Bi, Benlai Tang, Xiang Yin, Jianfei Yang, Mingjie Wang, Jiali Yao, Yang Zhang, and Zejun Ma. Towards realistic visual dubbing with heterogeneous sources. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 1739–1747, 2021.

Ziyu Yao, Xuxin Cheng, and Zhiqi Huang. Fd2talk: Towards generalized talking head generation with facial decoupled diffusion model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3411–3420, 2024.

Xiangnan Yin and Liming Chen. Faceocc: A diverse, high-quality face occlusion dataset for human face extraction. *arXiv preprint arXiv:2201.08425*, 2022.

Runyi Yu, Tianyu He, Ailing Zhang, Yuchi Wang, Junliang Guo, Xu Tan, Chang Liu, Jie Chen, and Jiang Bian. Make your actor talk: Generalizable and high-fidelity lip sync with motion and appearance disentanglement. *arXiv preprint arXiv:2406.08096*, 2024.

Tan Yu, Qian Qiao, Le Shen, Ke Zhou, Jincheng Hu, Dian Sheng, Bo Hu, Haoming Qin, Jun Gao, Changhai Zhou, et al. Soulx-flashhead: Oracle-guided generation of infinite real-time streaming talking heads. *arXiv preprint arXiv:2602.07449*, 2026.

Chenxu Zhang, Chao Wang, Jianfeng Zhang, Hongyi Xu, Guoxian Song, You Xie, Linjie Luo, Yapeng Tian, Xiaohu Guo, and Jiashi Feng. Dream-talk: Diffusion-based realistic emotional audio-driven method for single image talking face generation. *arXiv preprint arXiv:2312.13578*, 2023a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Yue Zhang, Zhizhou Zhong, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. Musetalk: Real-time high-fidelity video dubbing via spatio-temporal sampling. *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.

Zhimeng Zhang, Zhipeng Hu, Wenjin Deng, Changjie Fan, Tangjie Lv, and Yu Ding. Dinet: Deformation inpainting network for realistic face visually dubbing on high resolution video. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 3543–3551, 2023b.

Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, and Guanbin Li. Identity-preserving talking face generation with landmark and appearance priors. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738. IEEE, 2023.

Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makeittalk: speaker-aware talking-head animation. *ACM Transactions On Graphics (TOG)*, 39(6):1–15, 2020.

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.

Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9300–9308. IEEE, 2019.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Instant volumetric head avatars. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4574–4584. IEEE, 2023.

A Derivation of the Cross-Attention Transport

For clarity, we omit the spatial and layer indices. The local entropy regularized transport problem is formulated as:

$$\min_{\pi \in \Delta^R} \sum_{r=1}^{R} \pi_r c_r + \varepsilon \sum_{r=1}^{R} \pi_r \log \pi_r. \tag{14}$$

Introducing a Lagrange multiplier λ for $\sum_r \pi_r = 1$ gives:

$$\mathcal{J}(\pi, \lambda) = \sum_r \pi_r c_r + \varepsilon \sum_r \pi_r \log \pi_r + \lambda \left(\sum_r \pi_r - 1 \right). \tag{15}$$

The stationary condition is $c_r + \varepsilon(\log \pi_r + 1) + \lambda = 0$, which yields:

$$\pi_r = \frac{\exp(-c_r/\varepsilon)}{\sum_{r'=1}^{R} \exp(-c_{r'}/\varepsilon)}. \tag{16}$$

Substituting $c_r = -q^\top k_r / \sqrt{d}$ gives:

$$\pi_r = \mathrm{Softmax}_r \left(\frac{q^\top k_r}{\varepsilon \sqrt{d}} \right), \tag{17}$$

which corresponds to the deformable cross-attention weights used in OAT.

B Detailed Training Objectives

We provide the detailed definitions of the training objectives introduced in the main paper.

Reconstruction Loss. We adopt an L_1 reconstruction loss to enforce pixel-level consistency between the generated frame $\hat{I}$ and the ground truth frame I_{gt}:

$$\mathcal{L}_1 = \left\| \hat{I} - I_{\mathrm{gt}} \right\|_1. \tag{18}$$

Perceptual Loss. To improve perceptual quality and alleviate over-smoothed textures, we employ a pretrained VGG network (Simonyan & Zisserman, 2014), following existing works (Siarohin et al., 2019). Let $\phi_j(\cdot)$ denote the feature map extracted from the j-th VGG layer. The full frame perceptual loss is defined as:

$$\mathcal{L}_{\mathrm{p}} = \sum_j \left\| \phi_j(\hat{I}) - \phi_j(I_{\mathrm{gt}}) \right\|_1. \tag{19}$$

Since the mouth region is particularly important for audio-driven lip synchronization, we further introduce a mouth-region perceptual loss as:

$$\mathcal{L}_{\mathrm{m}} = \sum_j \left\| \phi_j \left(\mathcal{C}_M(\hat{I}) \right) - \phi_j \left(\mathcal{C}_M(I_{\mathrm{gt}}) \right) \right\|_1, \tag{20}$$

where $\mathcal{C}_M(\cdot)$ denotes the operation that extracts the mouth region based on the mouth landmarks.

Adversarial Loss. We apply a perceptual quality discriminator Γ to the generated and real faces to encourage sharper and more realistic details in the mouth region. We adopt the hinge adversarial loss (Lim & Ye, 2017) and apply spectral normalization (Miyato et al., 2018) to stabilize discriminator training. The discriminator objective is formulated as:

$$\mathcal{L}_\Gamma = \mathbb{E}_{I_{\mathrm{gt}}} \left[\max(0, 1 - \Gamma(I_{\mathrm{gt}})) \right] + \mathbb{E}_{\hat{I}} \left[\max(0, 1 + \Gamma(\hat{I})) \right]. \tag{21}$$

The corresponding generator adversarial loss is defined as:

$$\mathcal{L}_{\mathrm{g}} = -\mathbb{E}_{\hat{I}} \left[\Gamma(\hat{I}) \right]. \tag{22}$$

SyncNet Loss. Following Wav2Lip (Prajwal et al., 2020), we employ a pretrained SyncNet to enforce synchronization between the generated mouth motion and the input speech. The synchronization loss is defined as:

$$\mathcal{L}_{\mathrm{s}} = 1 - \frac{S_v(\hat{\mathbf{I}}_{t:t+T-1})^\top S_a(A_{t:t+T-1})}{\left\| S_v(\hat{\mathbf{I}}_{t:t+T-1}) \right\|_2 \left\| S_a(A_{t:t+T-1}) \right\|_2}, \tag{23}$$

Table 5: Quantitative evaluation of the reconstructed reference frames.

Evaluation Region	SSIM ↑	L_1 ↓
Full Face	0.979	0.009
w/o Mouth Region	**0.991**	**0.005**

Table 6: Effect of the occlusion probability ρ_{occ} under 50–100% mouth occlusion.

ρ_{occ}	OCC-E↓	SSIM↑	LMD↓	PSNR↑
0.02	5.54	0.95	2.15	30.77
0.04	5.24	0.95	2.07	31.16
0.06	4.83	0.95	2.12	31.01
0.08	5.42	0.95	2.19	31.07
0.10	5.21	0.95	2.23	31.10
0.30	4.82	0.95	2.16	31.56
0.60	4.61	0.95	2.26	31.89
1.00	4.33	0.96	2.49	31.99

where $S_v(\cdot)$ and $S_a(\cdot)$ denote the visual and audio encoders of SyncNet. Minimizing $\mathcal{L}_{\text{s}}$ encourages the generated lip movements to remain synchronized with the phonetic content of the input audio.

C Occlusion Evaluation Dataset

To evaluate the robustness of our method to facial occlusions while avoiding ground truth lip shape leakage, we construct a facial occlusion test set based on the HDTF test set. We first use LivePortrait (Guo et al., 2024) to reconstruct the mouth motion of each video and paste only the reconstructed mouth region back onto the original frames. In this way, the facial appearance, head pose and background are largely preserved, while the original mouth shape is changed from the reference frames. As shown in Table 5, the reconstructed frames achieve an SSIM of 0.991 and an L_1 distance of 0.005 outside the mouth region. The difference heatmaps in Figure 4 further show that the reconstruction-induced changes are mainly concentrated around the mouth region, indicating that the non-mouth facial content is well preserved. Following the occlusion synthesis strategy (Voo et al., 2022), we overlay hand occluders and randomly sampled objects on both the original and reconstructed frames using identical spatial placements and geometric transformations. Based on the ratio of the mouth region covered by the occluder, we divide the samples into three groups, 0%, 0–50% and 50–100%. The occluded original frames are used as ground truth targets, while the corresponding occluded reconstructed frames are used as model inputs. This paired construction enables a fair evaluation of occlusion robustness without directly exposing the ground truth mouth shape.

D Additional Experiments

D.1 Sensitivity to Occlusion Probability

We analyze the sensitivity of OR-Talk to the occlusion probability ρ_{occ} under severe mouth occlusion, as shown in Table 6. Increasing ρ_{occ} generally improves occluder reconstruction, with lower OCC-E and higher PSNR. However, a larger occlusion probability gradually reduces the accuracy of mouth motion, as reflected by the increase in LMD. At $\rho_{\text{occ}} = 0.06$, the model achieves OCC-E 4.83, LMD 2.12, and PSNR 31.01. Further increasing ρ_{occ} improves OCC-E and PSNR, but leads to worse LMD. We therefore set $\rho_{\text{occ}} = 0.06$ in all experiments to balance occluder reconstruction and mouth shape accuracy.

D.2 Qualitative Comparison with State-of-the-Art Methods

We provide additional qualitative comparisons with state-of-the-art lip synchronization methods on the HDTF, LRS2 and VoxCeleb2 datasets in Figure 5. The first row presents the input audio to-

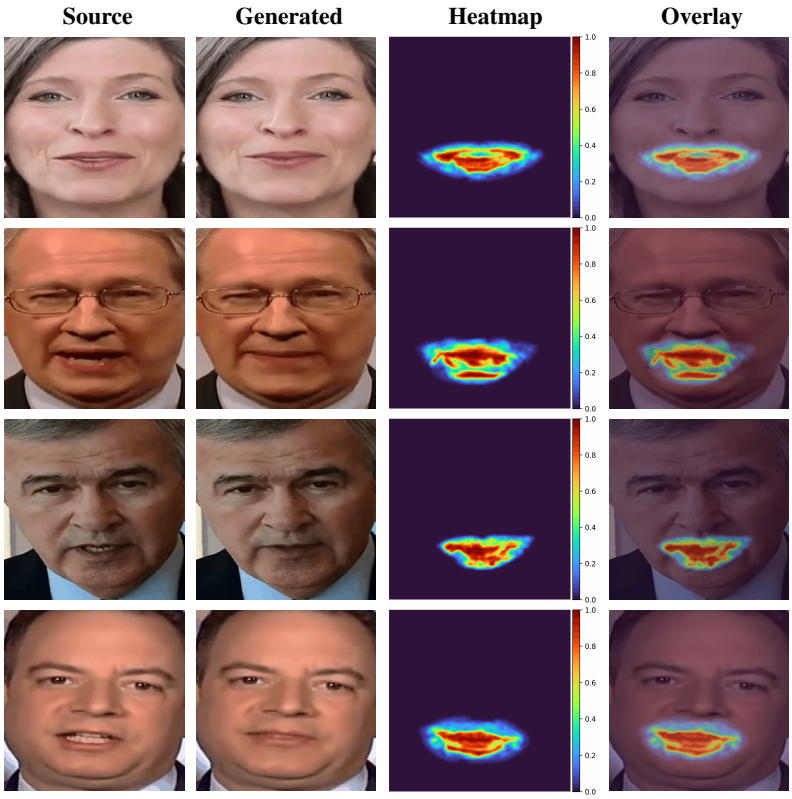


Figure 4: Visualization of the differences between the source and generated frames. From left to right, each row shows the source frame, the corresponding generated frame, the pixel-wise difference heatmap and the heatmap overlaid on the generated frame. Warmer colors indicate larger differences. The differences are primarily concentrated around the mouth region, indicating that the model mainly modifies the lip area while preserving the rest of the facial appearance.

gether with the corresponding video frames, while the second row shows the input video frames. The remaining rows present the generated results of different methods. Compared with existing approaches, OR-Talk produces clearer and more complete mouth structures and generates lip shapes that are more consistent with the corresponding audio conditions. In addition, OR-Talk better preserves facial details outside the mouth region, reducing visual artifacts and maintaining coherent facial appearance across different datasets. These qualitative results further demonstrate the effectiveness and generalization ability of OR-Talk across diverse identities and speech conditions.

D.3 Qualitative Evaluation on AIGC-LipSync

We further evaluate OR-Talk on occluded samples from the AIGC-LipSync benchmark to examine its generalization under more realistic facial occlusions.

As shown in Figure 6, several competing methods tend to distort or overwrite the occluder when synthesizing the mouth region and visible artifacts may also appear around occlusion boundaries. Both X-Dub and OR-Talk are able to preserve and reconstruct the occluder relatively well. However, X-Dub often exhibits insufficient color recovery in the generated results, resulting in noticeable appearance inconsistency. In contrast, OR-Talk better preserves the color and appearance of the occluders while maintaining plausible audio-driven mouth motion. These results demonstrate that OR-Talk achieves more consistent occlusion recovery under challenging real world conditions. Here, “Inference Failed” indicates that some baseline methods are unable to complete inference on certain test videos due to preprocessing failures or errors during the inference process.



Figure 5: Qualitative comparison with state-of-the-art lip synchronization methods on the HDTF, LRS2 and VoxCeleb2 datasets.

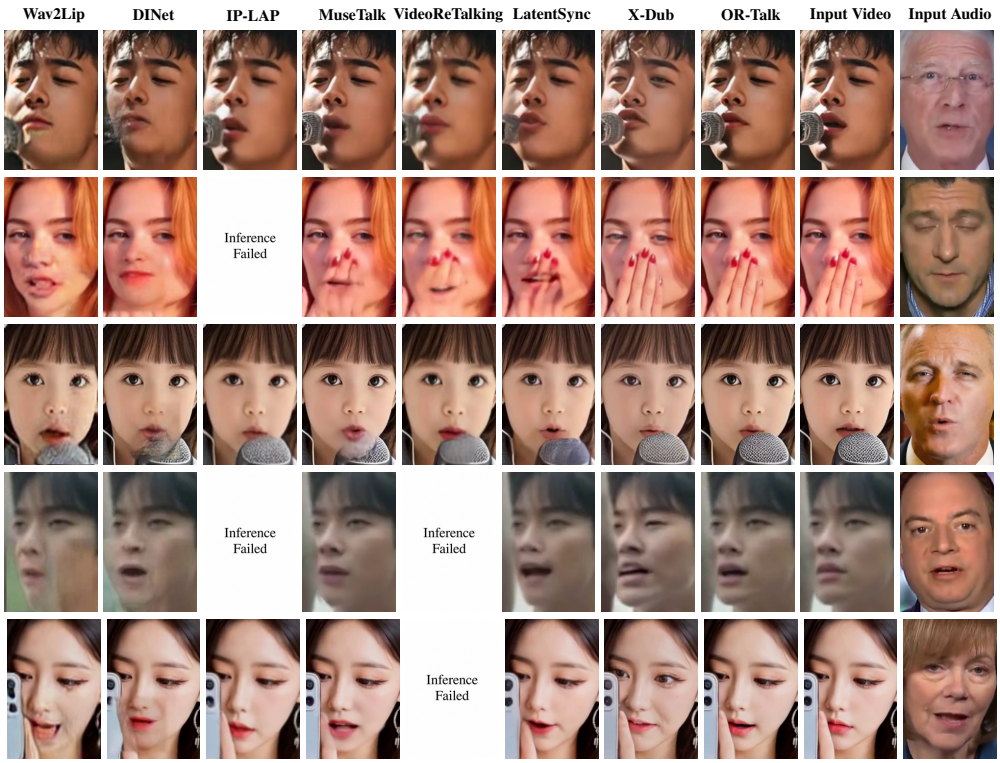


Figure 6: Qualitative Comparison on Occluded Samples from AIGC-LipSync