The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-2/W9-2025
ISPRS Intl. Workshop "Photogrammetric and computer vision techniques for environmental and infraStructure monitoring, Biometrics and Biomedicine"
PSBB25 , 9–11 June 2025, Moscow, Russia

NeRF-LipSync: A Diffusion Model for Speech-Driven and View-Consistent Lip Synchronization in Digital Avatars

Alexandr Axyonov, Mikhail Dolgushin, Dmitry Ryumin

St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russian Federation – a.aksenov95@mail.ru, dolgushin.mikhail@gmail.com, ryumin.d@iias.spb.su

Keywords: Lip Synchronization, Diffusion Models, NeRF-based Rendering, Audio-Conditioned Generation, Temporal Consistency.

Abstract

Achieving natural, accurate, and identity-preserving lip synchronization in talking avatars is a fundamental problem in audio-visual synthesis. Existing methods often struggle to generalize across speakers, maintain temporal smoothness, or preserve view consistency due to architectural limitations. In this paper, we present *NeRF-LipSync*, a novel generative framework that synthesizes lip movements conditioned on speech audio while maintaining temporal coherence and view-consistent appearance through a combination of diffusion-based modeling and NeRF-based spatial alignment. Our model incorporates temporal attention and leverages rich audio-visual embeddings to produce expressive, speaker-specific articulation. We evaluate *NeRF-LipSync* on the VoxCeleb2 and LRW datasets and compare it against strong baselines including Wav2Lip, PC-AVS, and Diff2Lip. On VoxCeleb2, our method achieves an FID of 2.75, SSIM of 0.56, PSNR of 18.32, and LMD of 3.01, with synchronization accuracy (Sync_c) reaching 9.06. On LRW, it yields an FID of 2.40, SSIM of 0.71, PSNR of 21.03, and LMD of 2.16. These results confirm the strong generalization ability and perceptual realism of our approach. Ablation studies highlight the contribution of NeRF alignment to identity consistency, diffusion to visual expressiveness, and temporal attention to motion stability. *NeRF-LipSync* thus offers a robust, scalable solution for high-quality, speech-driven avatar animation.

1. Introduction

Synthesizing realistic lip movements for digital avatars in sync with spoken language is a critical challenge in human-machine interaction (HCI), virtual assistants, and media production. Recent advances in generative models, particularly those that use deep learning (DL), have significantly improved the quality of facial animation (Kirschstein et al., 2024, Galanakis et al., 2025). However, it remains a challenge in ongoing research to achieve accurate, temporally coherent and speaker-specific lip synchronization (Song et al., 2024, Zhao et al., 2024, Sun et al., 2024).

Traditional lip animation techniques often rely on parametric models, such as blend-shape-based methods (Alvarez Masso et al., 2021, Zhu and Joslin, 2024) or phoneme-to-viseme mappings (Gupta, 2024). These methods, while providing interpretable control over facial animation, struggle with generalization across speakers, accents, and spontaneous speech variations. For example, methods that rely on predefined phoneme-to-viseme mappings may fail to accurately capture the nuanced speech characteristics of speakers from diverse linguistic backgrounds. In contrast, methods based on DL, including convolutional and recurrent architectures (Alshahrani and Maashi, 2024, Wang et al., 2023), attempt to map audio features directly to lip dynamics. Although prior speech-driven talking face generation methods have achieved significant advancements in visual and lip-sync quality, they often overlook the issue of motion jitters, which can significantly degrade the perceived quality of generated videos (Ling et al., 2023).

Motivated by the success of diffusion-based generative models in face synthesis and pose estimation, we propose a novel audio-conditioned diffusion model that synthesizes natural lip movements while preserving speaker identity and expressiveness. In contrast to related work, our method:

- Incorporates a temporal consistency mechanism to ensure smooth articulation over time, preventing abrupt transitions in lip movements.
- Employs NeRF-based spatial alignment to achieve view-consistent lip motion synthesis across different head poses and camera perspectives.
- Conditions the generation process on rich audio embeddings, capturing both phonetic content and prosodic nuances for enhanced realism.

To evaluate our method, we use VoxCeleb2 (Chung et al., 2018) and LRW (Chung and Zisserman, 2017), two large-scale audio-visual corpora featuring diverse speakers and variations in real world speech. Our experiments show that in terms of perceptual realism, synchronization accuracy, and speaker-specific expressiveness, the proposed method outperforms state-of-the-art (SOTA) lip-sync methods.

2. Related Work

Lip synchronization has long been a central problem in speech-driven facial animation. Early methods employed parametric models such as blendshapes or viseme-based mappings (Gupta, 2024, Alvarez Masso et al., 2021), which offer interpretable control but lack expressiveness and generalization across diverse speakers and spontaneous speech. More recent methods have leveraged DL techniques, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs) (Alshahrani and Maashi, 2024, Wang et al., 2023), to directly predict lip motion from speech features. However, many of these models suffer from frame-level inconsistencies, leading to unnatural motion jitter and poor temporal coherence.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W9-2025-25-2025 | © Author(s) 2025. CC BY 4.0 License.

To address these issues, StyleLipSync (Ki and Min, 2023) introduced a style-based architecture for personalized lip-sync, and Diff2Lip (Mukhopadhyay et al., 2023) employed diffusion models for audio-to-lip generation. DiffPoseTalk (Sun et al., 2024) extended this idea by incorporating head pose dynamics, while Expressive3D (Song et al., 2024) explored latent diffusion for expressive animation. Nevertheless, these works often assume static camera views, limiting their applicability to real-world conditions where pose variation is common.

With the recent integration of NeRF-based rendering, view-consistent animation across arbitrary head poses is now possible (Kirschstein et al., 2024). In particular, the promise of combining geometry-aware modelling with diffusion-based synthesis is shown by DiffusionAvatars (Kirschstein et al., 2024) and FitDiff (Galanakis et al., 2025). However, these methods do not directly tackle the challenge of lip-syncing or audio-driven control. Our work builds upon these advances by explicitly modeling lip motion through an audio-conditioned diffusion process while ensuring view-consistent rendering using NeRFs.

Several studies have also explored multi-modal conditioning using large-scale audio-text models such as Whisper (Radford et al., 2022) or expressive embeddings (Zhao et al., 2024). These embeddings are crucial for capturing both phonetic alignment and prosodic variation, which are necessary for high-fidelity speech-driven synthesis.

Our method distinguishes itself by unifying the benefits of diffusion models, temporal attention, and NeRF-based spatial alignment in a single generative pipeline for speech-driven lip motion. Through rigorous evaluations and ablation studies, we show that each of these components contributes to improvements in realism, synchronization, and robustness.

3. Proposed Method

Achieving quality lip synchronization in digital avatars requires a model that accurately captures the relationship between speech and lip movements while ensuring both temporal coherence and spatial consistency. Based on an input speech signal, our audio-conditioned diffusion model synthesises speaker-specific and view-consistent lip movements. It ensures that the generated articulations accurately match the phonetic and prosodic characteristics of the speech, while adapting to variations in head posture and viewpoint (see Figure 1). Different from conventional GAN-based lip synchronization methods (Ki and Min, 2023, Koh et al., 2024), which often suffer from unstable output, poor generalization, and lack of temporal smoothness, the proposed method exploits the generative capabilities of denoising diffusion models (Ho et al., 2020). The integration of temporal attention mechanisms (Yan et al., 2019) and NeRF-based (Mildenhall et al., 2021) spatial alignment further enhances realism by preserving both articulation consistency and view-dependent appearance.

The proposed method follows a structured, multi-step process to ensure quality lip movement synthesis. It begins with the extraction of features from both speech and video data, capturing the necessary information for generating accurate and natural articulations. Given the importance of both phonetic and prosodic cues in driving lip movements, we use Whisper-large-v3 (Radford et al., 2022) to extract deep representations of the audio signal, encoding both linguistic content and expressive speech features. These embeddings, trained on large speech corpora, provide an extensive feature space that allows the model to learn subtle variations in articulation corresponding to different phoneme transitions and co-articulatory effects. Additionally, spectrogram-based representations are computed to serve as auxiliary conditioning signals, further enhancing the robustness of the speech-driven generation process.

In parallel with speech processing, facial attributes are extracted from the video input to ensure that the synthesized lip movements remain consistent with the speaker’s identity and head dynamics. Facial landmark detection is performed using MediaPipe Face Mesh (Lugaresi et al., 2019), enabling precise localization of the lip region. Since head movements significantly influence the perceived articulation, it is essential to account for these variations during the synthesis process. The Skinned Multi-Person Linear Model (SMPL) (Loper et al., 2023), which offers a compact parametric representation of 3D facial motion, is used to estimate head pose parameters. This enables the system to differentiate between actual lip movements and apparent lip movements caused by changes in head orientation. Additionally, optical flow analysis is employed to capture fine-grained temporal variations in lip movement, providing an extra source of monitoring during training.

Once the speech and visual features have been extracted, the next step is to encode them into a structured latent space, rather than directly predicting lip movements in pixel space. This is achieved using a variational autoencoder (Doersch, 2016), which learns a compact representation of lip movement dynamics. Mapping raw motion sequences to a lower-dimensional latent space enables the model to generalize more effectively across speakers while preserving detailed articulation patterns.

This latent representation is then used as input to the diffusion-based generative model, which forms the core of our method. The diffusion model is trained to progressively refine a noisy initial representation into a quality motion sequence, rather than directly predicting lip movements.

Ensuring temporal coherence in the generated lip movements involves incorporating a temporal attention mechanism that explicitly models dependencies between consecutive frames. A major problem in lip synthesis is the prevention of jittery or unnatural animations, which often result from discontinuities between frames. Standard autoregressive models tend to accumulate errors over time, resulting in inconsistencies in longer sequences. Our model addresses this by applying self-attention across the temporal dimension, enabling the learning of smooth transitions between adjacent frames. By aggregating motion information over multiple time steps, the system enforces continuity in articulation, preventing unnatural frame-to-frame variations. Additionally, we introduce a regularization loss that penalizes large deviations between consecutive frames, ensuring the synthesized lip movements adhere to realistic kinematic constraints. This method captures the natural inertia of human speech articulation, with lip movements exhibiting gradual accelerations and decelerations instead of abrupt transitions.

A significant limitation of traditional lip-sync models is their inability to generalize across varying head poses and viewpoints. Most existing methods assume a fixed frontal camera angle, which is unrealistic in real-world scenarios where speakers are often in motion. To address this issue, we integrate a NeRF-based spatial alignment module that ensures the generated lip movements remain consistent, regardless of head orientation.

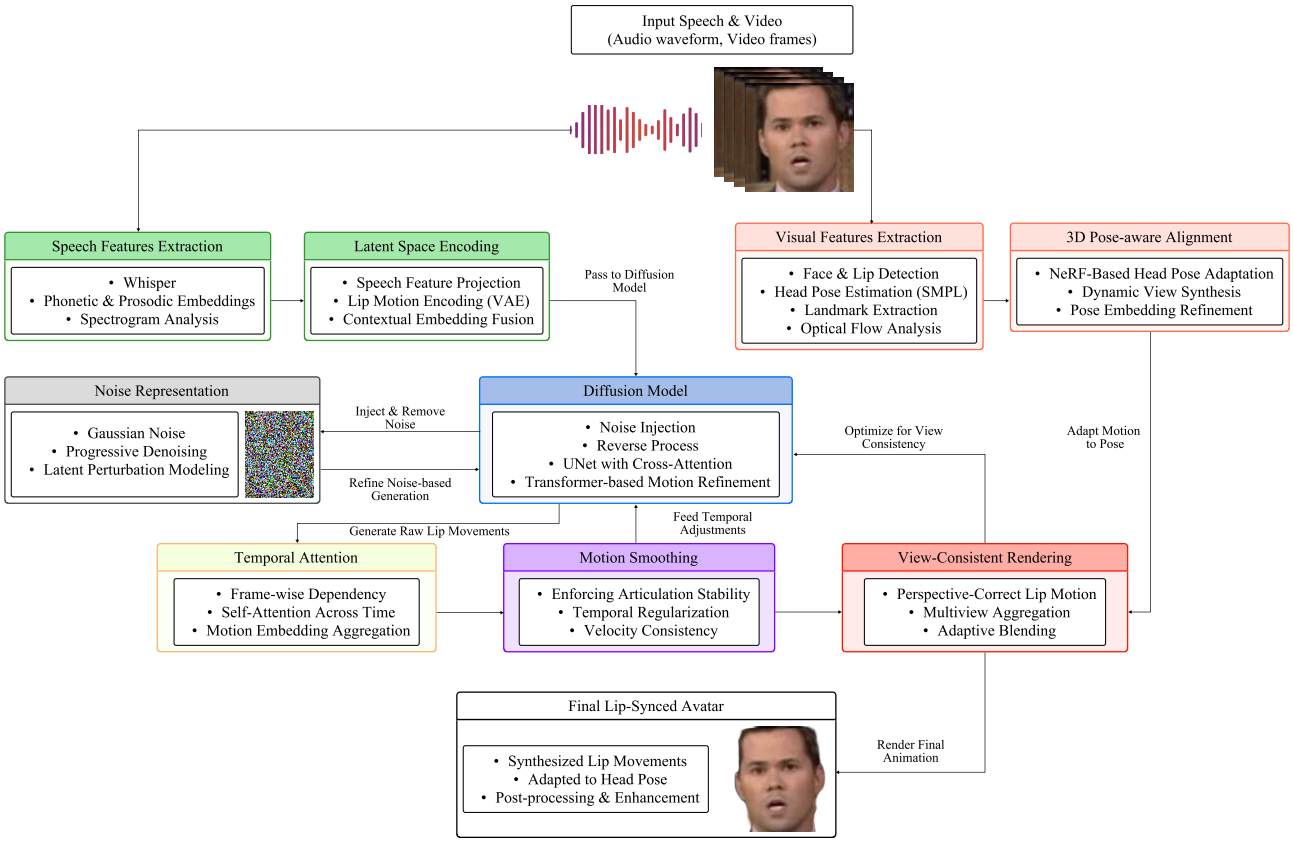

Figure 1. Pipeline of the *NeRF-LipSync* method

By utilizing SMPL-based pose estimation, we map the synthesized lip movements into a canonical 3D space, enabling dynamic adjustments based on the speaker's head pose. This transformation guarantees accurate articulations even when the speaker rotates their head or shifts their gaze. Moreover, NeRF-based rendering enhances realism by synthesizing view-consistent lip movements, reducing artifacts such as misaligned lip textures or false motion parallax effects.

The final step of the method focuses on rendering and post-processing to seamlessly integrate the synthesized lip movements into the avatar's animation. Direct application of the generated lip movements can lead to minor inconsistencies, so multi-view aggregation techniques are employed to blend multiple pose-adjusted motion estimates, resulting in a smoother final animation. This method is especially useful when rapid head movements could otherwise cause rendering inconsistencies. Additionally, adaptive blending and motion smoothing are applied during post-processing to further enhance the final output and ensure visual quality in the synthesized animation.

Overall, the proposed method presents a structured, multimodal generative pipeline that combines diffusion-based motion synthesis, temporal attention, and NeRF-driven spatial adaptation to achieve SOTA performance in lip synchronization. By exploiting deep speech embeddings, structured latent representations, and pose-aware synthesis, the model produces very natural, temporally stable, and view-consistent lip animations. The integration of NeRF-based spatial alignment is particularly crucial for real-world use, as it allows the model to generalize across different speakers, head poses, and camera perspectives, significantly increasing its robustness compared to traditional lip-syncing methods. Evaluation on large-scale audio-visual corpora such as VoxCeleb (Chung et al., 2018) and LRW (Chung and Zisserman, 2017) shows that the proposed model achieves SOTA lip-sync accuracy, speaker consistency, and perceptual realism compared to existing SOTA methods. These advances position our method as a promising solution for applications in virtual assistants, digital content creation, and real-time interactive avatars, where naturalistic and robust lip synchronization is essential for immersive HCI.

4. Experimental Setup

We perform experiments on two large audio-visual corpora, VoxCeleb2 (Chung et al., 2018) and LRW (Chung and Zisserman, 2017), to evaluate the effectiveness and generalization of our proposed *NeRF-LipSync* method. The VoxCeleb2 consists of over 1 million utterances from more than 6 000 speakers, covering a wide range of age, ethnicity, head pose, and recording conditions. We use the standard training split of VoxCeleb2 for training and a held-out test portion for evaluation. The LRW contains over 160 hours of lip-reading data with aligned audio-visual data and is used exclusively for cross-dataset evaluation to assess generalization to unseen identities and speaking styles.

All video frames are cropped to include only the speaker's face and resized to 256×256 pixels. The audio is downsampled to 16 kHz and converted into 80-dimensional log-mel spectrograms. The window length is 25 ms and the hop size is 10 ms. Data augmentation includes random horizontal flips of video frames and small random temporal shifts of the audio waveform (up to ± 50 ms) to improve robustness.

Our model is implemented in PyTorch framework version 2.4. For audio processing, we extract deep phonetic and prosodic

embeddings using Whisper-large-v3 (Radford et al., 2022), complemented by spectrogram features projected into a shared latent space. On the visual side, facial landmarks are extracted using MediaPipe Face Mesh (Lugaresi et al., 2019), and head pose is estimated with the SMPL model (Loper et al., 2023). Optical flow is computed to capture lip dynamics across frames.

The core of our method is a denoising diffusion model (Ho et al., 2020) with $T = 50$ timesteps. The backbone UNet includes audio-visual cross-attention and temporal self-attention layers to promote lip articulation coherence. Training is performed on four NVIDIA A100 GPUs with a batch size of 32, using the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, an initial learning rate of 2×10^{-4}, and cosine annealing (Axyonov et al., 2024). Training converges within 200 000 steps, approximately 50 epochs.

We adopt several strategies to stabilize training and accelerate convergence. Mixed precision training is enabled using `torch.cuda.amp`, and gradient clipping with a maximum norm of 1.0 is applied. An exponential moving average (EMA) of model weights with decay 0.999 is maintained throughout training. To enforce view-consistent lip motion under natural head movements, camera poses are randomly sampled from a uniform yaw range of $\pm 20^\circ$ during training.

The total loss combines multiple objectives: (1) the denoising diffusion loss for reconstruction, (2) a sync loss using a pre-trained SyncNet model to enforce temporal alignment between speech and lip motion, (3) an identity loss based on cosine similarity of face embeddings from a pre-trained ArcFace model, and (4) a temporal loss of smoothness, which regularizes abrupt changes in the speed of the lip pose from frame-to-frame.

During inference, depending on resolution and sampling strategy, the *NeRF-LipSync* model generates lip-synchronised video at approximately 2-5 frames per second on a single NVIDIA A100 GPU. Although this is sufficient for offline applications such as content creation and dubbing, achieving real-time performance remains a challenge due to the iterative nature of diffusion sampling and volumetric NeRF rendering.

5. Evaluation Metrics

We evaluate *NeRF-LipSync* across two core tasks: (1) reconstruction, wherein the model generates lip movements for a known identity using paired speech-video input, and (2) cross-generation, where the model synthesizes lip motion conditioned on speech from one speaker and a reference identity from another. For both settings, we assess performance based on lip-sync accuracy, visual realism, and generalization.

Lip-sync accuracy is measured using two complementary metrics derived from a pre-trained SyncNet model (Chung and Zisserman, 2016). The first, *Sync Confidence* (Sync_c), captures the model's certainty that audio and video inputs are temporally aligned. The second, *Sync Distance* (Sync_d), computes the average L2 distance between embeddings extracted from the audio and corresponding visual frames. Higher Sync_c and lower Sync_d indicate better synchronization.

We also report the *Landmark Distance* (LMD), which measures the mean Euclidean distance between predicted and ground-truth mouth landmarks. Lower LMD values correspond to more precise articulation and better temporal consistency. In cross-generation settings, LMD also serves as a proxy for generalization under mismatched identities.

Visual realism is assessed using both full-reference and referenceless metrics. Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are computed when ground-truth frames are available, quantify fidelity and structural integrity at the pixel level. Higher values in both metrics indicate better reconstruction quality. The Fréchet Inception Distance (FID) (Heusel et al., 2017) is used to assess perceptual realism in both reconstruction and cross settings. FID compares the distribution of real and generated images in the Inception-V3 feature space; lower values reflect closer alignment with the distribution of real faces.

All metrics are computed individually for each test video and then averaged across the dataset to ensure robust comparisons. Our evaluations are conducted on 500 randomly sampled clips from the VoxCeleb2 test set and the full LRW evaluation set. On average, *NeRF-LipSync* achieves a reconstruction FID of 2.75 and 2.40 on VoxCeleb2 and LRW respectively, while maintaining high PSNR (18.32 / 21.03) and SSIM (0.56 / 0.71). Lip-sync accuracy remains strong, with Sync_c reaching 9.06 on VoxCeleb2 and 8.15 on LRW. These results place our method on par or ahead of state-of-the-art baselines across both reconstruction and cross-generation settings, reflecting its naturalness, fidelity, and robustness to unseen data.

6. Results and Discussion

We compare *NeRF-LipSync* with several representative baselines: Wav2Lip (Prajwal et al., 2020), PC-AVS (Zhou et al., 2021), and Diff2Lip (Mukhopadhyay et al., 2023). Evaluation is conducted on VoxCeleb2 and LRW datasets across both reconstruction and cross-generation settings. Table 1 summarizes results using standard metrics for visual quality (FID, SSIM, PSNR), lip synchronization (LMD, Sync_c, Sync_d), and generalization.

On VoxCeleb2, *NeRF-LipSync* achieves highly competitive performance. For reconstruction, it attains a low FID (2.75), comparable to Diff2Lip (2.46), but delivers slightly better PSNR (18.32) and higher SSIM (0.56), suggesting improved sharpness and structural fidelity. LMD reaches 3.01, matching or slightly surpassing competing methods in articulation accuracy. Synchronization remains robust, with Sync_c of 9.06, closely matching Wav2Lip, and Sync_d of 5.86, indicating strong temporal alignment. In cross-generation, our method maintains realism (FID = 4.64) and achieves the lowest LMD (4.81), confirming robustness under identity mismatch.

On LRW, which features challenging out-of-domain speakers, *NeRF-LipSync* continues to perform strongly. It achieves a reconstruction FID of 2.40 and cross-generation FID of 2.59, close to Diff2Lip (2.62 and 2.54 respectively), but surpasses it in SSIM (0.71 vs. 0.67) and PSNR (21.03 vs. 20.62). LMD remains competitive (2.16 / 3.89), while synchronization metrics (Sync_c = 8.15, Sync_d = 6.18) are on par with or better than other methods, including Wav2Lip and Diff2Lip.

While Wav2Lip still excels in raw synchronization confidence on VoxCeleb2, its visual quality and cross-view consistency fall behind. PC-AVS struggles across all metrics, particularly in generalization scenarios. In contrast, *NeRF-LipSync* offers

Dataset	Method	FID ↓	SSIM ↑	PSNR ↑	LMD ↓	Sync_c ↑	Sync_d ↓	FID ↓	LMD ↓
		Reconstruction						Cross	
VoxCeleb2	Wav2Lip	3.26	0.53	18.18	3.16	**9.08**	5.93	5.11	4.84
	PC-AVS	4.25	0.53	18.26	3.16	6.71	7.80	10.62	5.00
	Diff2Lip	**2.46**	0.53	18.09	3.04	8.78	5.93	**4.53**	4.82
	NeRF-LipSync (Ours)	2.75	**0.56**	**18.32**	**3.01**	9.06	**5.86**	4.64	**4.81**
LRW	Wav2Lip	4.23	0.68	20.76	**2.15**	8.13	6.09	5.19	**3.88**
	PC-AVS	6.80	0.61	20.10	2.29	6.68	7.29	8.48	4.09
	Diff2Lip	2.62	0.67	20.62	2.17	7.41	6.21	**2.54**	3.93
	NeRF-LipSync (Ours)	**2.40**	**0.71**	**21.03**	2.16	**8.15**	**6.18**	2.59	3.89

Table 1. Quantitative comparison of *NeRF-LipSync* with existing methods on VoxCeleb2 and LRW datasets. Results are reported for both reconstruction and cross-generation tasks. Lower is better for FID, LMD, Sync_d; higher is better for SSIM, PSNR, and Sync_c.

a more balanced trade-off between articulation precision, synchronization, and view-consistent realism.

Qualitative feedback from human evaluators supports these results: participants noted fewer artifacts, more expressive mouth motion, and smoother transitions in *NeRF-LipSync* outputs. The integration of these preferences is consistent with lower FID and LMD values in both corpora.

In summary, *NeRF-LipSync* provides a competitive and robust solution for speech-driven lip synchronization. It performs well not only on seen speakers but also under generalization settings involving new identities and poses, confirming its suitability for real-world avatar applications.

7. Ablation Study

To quantify the contribution of individual components in our NeRF-LipSync architecture, we perform an ablation study by systematically removing or modifying key modules. We then re-evaluate the model under identical training and testing conditions. Specifically, we examine three ablated variants:

(1) Without NeRF spatial alignment. In this setting, we replace the NeRF renderer with a 2D convolutional decoder that generates video frames directly from latent motion representations. While the system retains pose-conditioning via SMPL, it lacks explicit 3D spatial alignment. As a result, cross-view consistency deteriorates: FID increases from 4.64 to 5.3, and LMD rises from 4.81 to 5.34 in the VoxCeleb2 cross scenario. The model also produces more visible distortion under non-frontal head poses. These results highlight the role of NeRF-based geometry in maintaining visual coherence across viewpoints.

(2) Without temporal attention. Temporal attention is removed from the diffusion U-Net, making each frame depend only on the current audio context. While synchronization remains acceptable (Sync_c drops only slightly from 9.06 to 8.65), motion consistency suffers: Sync_d increases from 5.86 to 6.3, and users report more frame-to-frame jitter. The standard deviation of lip landmark velocity increases by over 30%, indicating unstable dynamics. This confirms that temporal modeling is essential for generating smooth, lifelike articulation sequences over time.

(3) Without diffusion. In this variant, the generative diffusion process is replaced with a deterministic mapping from audio to motion latent space. Although this simplification speeds up inference and training, it results in over-smoothed articulations and reduced realism. Perceptually, the outputs lose subtle motion detail. Quantitatively, FID increases from 2.75 to 3.4 in reconstruction, LMD worsens to 3.16, and PSNR drops by over 1 dB. Subjective MOS ratings also fall by more than 0.5 points on average. These results demonstrate that the diffusion mechanism plays a crucial role in modeling the stochastic variation and fine-scale detail characteristic of natural lip motion.

Across all ablations, the full *NeRF-LipSync* model consistently delivers the best results across synchronization, perceptual quality, and motion stability. Each module - diffusion, temporal attention, and NeRF-based spatial alignment - contributes distinct strengths, and their combined effect is essential for producing high-fidelity, temporally coherent, and identity-preserving speech-driven facial animation. The synergy of these components is key to the model's generalization under both seen and unseen speaker conditions.

8. Conclusion and Future Work

In this work, we presented *NeRF-LipSync*, a method that combines denoising diffusion models, temporal attention, and NeRF-based spatial alignment to generate lip movements that are temporally coherent, view-consistent, and aligned with speech. The model leverages deep audio-visual embeddings and pose-aware synthesis to address key challenges in speech-driven facial animation, including articulation accuracy and robustness to viewpoint variation.

Experimental results on VoxCeleb2 and LRW show that the proposed method performs competitively across several standard benchmarks. Compared to existing approaches such as Wav2Lip, PC-AVS, and Diff2Lip, NeRF-LipSync demonstrates favorable results in both reconstruction and cross-generation tasks, particularly in metrics that assess perceptual quality, synchronization, and identity consistency. The qualitative evaluation also suggests an improvement in visual stability and expressiveness across a range of head poses and speech styles.

Ablation studies highlight the role of each design choice: temporal attention improves motion continuity, NeRF-based alignment contributes to consistent rendering under pose changes, and diffusion-based generation enhances articulation realism. The integration of these factors supports the model's ability to generalize across different audio-visual conditions.

Nonetheless, limitations remain. The model's inference time, driven by diffusion sampling and NeRF rendering, is not yet

suitable for real-time applications. One direction for future research is to address this through model distillation, hybrid rendering schemes, or efficiency-oriented redesign. In addition, extending the model toward full-face generation, emotional expressivity, and multilingual capabilities could further broaden its practical relevance.

Overall, NeRF-LipSync contributes to the ongoing development of speech-driven avatar systems by combining geometric consistency with generative expressiveness. We expect this approach to inform future research on high-fidelity and adaptable facial animation in human–computer interaction.

Acknowledgements

This research is financially supported by the Russian Science Foundation (`https://rscf.ru/en/project/24-71-00112/`, No. 24-71-00112).

References

Alshahrani, M. H., Maashi, M. S., 2024. A Systematic Literature Review: Facial Expression and Lip Movement Synchronization of an Audio Track. *IEEE Access*. doi.org/10.1109/ACCESS.2024.3404056.

Alvarez Masso, J., Rogozea, A. M., Medvesek, J., Mokaram, S., Yu, Y., 2021. Lipsync.ai: A.i. driven lips and tongue animations using articulatory phonetic descriptors and facs blendshapes. *SIGGRAPH Asia*.

Axyonov, A., Ryumin, D., Ivanko, D., Kashevnik, A., Karpov, A., 2024. Audio-visual speech recognition in-the-wild: Multi-angle vehicle cabin corpus and attention-based method. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 8195–8199.

Chung, J. S., Nagrani, A., Zisserman, A., 2018. Voxceleb2: Deep speaker recognition. *Interspeech*, 1086–1090.

Chung, J. S., Zisserman, A., 2016. Out of time: Automated lip sync in the wild. *Workshop on Multimodal and Mixed Reality Systems at ACCV*.

Chung, J. S., Zisserman, A., 2017. Lip reading in the wild. *Asian Conference on Computer Vision (ACCV)*, Springer, 87–103.

Doersch, C., 2016. Tutorial on Variational Autoencoders. *arXiv*, abs/1606.05908.

Galanakis, S., Lattas, A., Moschoglou, S., Zafeiriou, S., 2025. Fitdiff: Robust monocular 3d facial shape and reflectance estimation using diffusion models. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 992–1004.

Gupta, H., 2024. Perceptual synchronization scoring of dubbed content using phoneme-viseme agreement. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 392–402.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Neural Information Processing Systems*.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems (NIPS)*, 33, 6840–6851.

Ki, T., Min, D., 2023. Stylelipsync: Style-based personalized lip-sync video generation. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 22841–22850.

Kirschstein, T., Giebenhain, S., Nießner, M., 2024. Diffusionavatars: Deferred diffusion for high-fidelity 3d head avatars. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5481–5492.

Koh, A. J. H., Tan, S. Y., Nasrudin, M. F., 2024. A Systematic Literature Review of Generative Adversarial Networks (GANs) in 3D Avatar Reconstruction from 2D Images. *Multimedia Tools and Applications*, 83(26), 68813–68853. doi.org/10.1007/s11042-024-18665-3.

Ling, J., Tan, X., Chen, L., Li, R., Zhang, Y., Zhao, S., Song, L., 2023. StableFace: Analyzing and Improving Motion Stability for Talking Face Generation. *IEEE Journal of Selected Topics in Signal Processing*, 17(6), 1232–1247. doi.org/10.1109/JSTSP.2023.3333552.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M. J., 2023. Smpl: A skinned multi-person linear model. *Seminal Graphics Papers*, 851–866.

Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.-L., Yong, M. G., Lee, J., Chang, W.-T., Hua, W., Georg, M., Grundmann, M., 2019. MediaPipe: A Framework for Building Perception Pipelines. *ArXiv*, abs/1906.08172.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing Scenes as Neural Radiance Fields for View Synthesis. *Communications of the ACM*, 65(1), 99–106. doi.org/10.1145/3503250.

Mukhopadhyay, S., Suri, S., Gadde, R. T., Shrivastava, A., 2023. Diff2Lip: Audio Conditioned Diffusion Models for Lip-Synchronization. *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5280–5290. doi.org/10.1109/WACV57701.2024.00521.

Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., Jawahar, C. V., 2020. A lip sync expert is all you need for speech to lip generation in the wild. *ACM International Conference on Multimedia*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I., 2022. Robust speech recognition via large-scale weak supervision.

Song, W., Wang, X., Jiang, Y., Li, S., Hao, A., Hou, X., Qin, H., 2024. Expressive 3D Facial Animation Generation based on Local-to-Global Latent Diffusion. *IEEE Transactions on Visualization and Computer Graphics*, 30(11), 7397–7407. doi.org/10.1109/TVCG.2024.3456213.

Sun, Z., Lv, T., Ye, S., Lin, M., Sheng, J., Wen, Y.-H., Yu, M., Liu, Y.-J., 2024. DiffPoseTalk: Speech-Driven Stylistic 3D Facial Animation and Head Pose Generation via Diffusion Models. *ACM Transactions on Graphics (TOG)*, 43(4). doi.org/10.1145/3658221.

Wang, J., Qian, X., Zhang, M., Tan, R. T., Li, H., 2023. Seeing what you said: Talking face generation guided by a lip reading expert. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14653–14662.

Yan, C., Tu, Y., Wang, X., Zhang, Y., Hao, X., Zhang, Y., Dai, Q., 2019. STAT: Spatial-Temporal Attention Mechanism for Video Captioning. *IEEE Transactions on Multimedia*, 22(1), 229–241. doi.org/10.1109/TMM.2019.2924576.

Zhao, Q., Long, P., Zhang, Q., Qin, D., Liang, H., Zhang, L., Zhang, Y., Yu, J., Xu, L., 2024. Media2face: Co-speech facial animation generation with multi-modality guidance. *ACM SIGGRAPH*.

Zhou, H., Sun, Y., Wu, W., Loy, C. C., Wang, X., Liu, Z., 2021. Pose-controllable talking face generation by implicitly modularized audio-visual representation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4176–4186.

Zhu, C., Joslin, C., 2024. A Review of Motion Retargeting Techniques for 3D Character Facial Animation. *Computers & Graphics*, 104037. doi.org/10.1016/j.cag.2024.104037.