The Eurasia Proceedings of
Educational & Social Sciences (EPESS)
ISSN: 2587-1730

The Eurasia Proceedings of Educational and Social Sciences (EPESS), 2025

Volume 46, Pages 137-143

IConETE 2025: International Conference on Education in Technology and Engineering

Open-Source Lip-Sync Models in the Period 2020-2025: A Structured Comparative Analysis

Bilge Nur Saglam
Aktif Investment Bank, Inc.

Mustafa Keles
Aktif Investment Bank, Inc.

Mehmet Kutanoglu
Aktif Investment Bank, Inc.

Abstract: Recent advancements in artificial intelligence have led to significant progress in the field of lip synchronization (lip-sync). This paper presents a systematic literature review focusing on popular open-source lip-sync models developed between 2020 and 2025, a period marked by the rapid evolution of deep generative models. Our aim was to examine and classify the prominent models of this era based on their architecture, performance, and technological approaches. To conduct our review, we searched the IEEE Xplore and Scopus databases. This study is based on three main methods most commonly used in the field: Generative Adversarial Networks (GANs), Transformers, and Diffusion Models. Each method was analyzed in detail using its popular representatives: Wav2Lip (GAN), GeneFace (Transformer/NeRF), and Diff2Lip (Diffusion). In this study, the training processes, architectural features, and performance metrics, such as video quality, synchronization accuracy, and computational cost, of these models were compared. Our findings indicate that diffusion models have recently gained prominence because they offer photorealistic outputs and stable training processes, although GAN-based models such as Wav2Lip are still widely used. This review serves as a comprehensive guide for researchers by summarizing the current state of the art in the field. Furthermore, it aims to contribute to new work by discussing the current challenges faced by lip-sync technologies and future research directions (e.g., real-time performance and multilingual support).

Keywords: Lip synchronization, Open-source, Artificial intelligence, Deep learning, GAN, Diffusion models, Wav2Lip, GeneFace

Introduction

Lip Synchronization (Lip-Sync) is the process of accurately matching a speaker's voice to the mouth movements in a video recording. While traditional methods relied on rule-based approaches like phoneme-viseme mapping, the post-2020 era has seen a breakthrough in this field thanks to deep learning techniques. The open-source release of architectures such as Generative Adversarial Networks (GANs), Transformers, and Diffusion Models has driven the democratization and rapid advancement of this technology.

This structural comparative analysis evaluates the three main technological approaches used in lip synchronization by examining the literature between 2020 and 2025, focusing on the most popular open-source representatives of these approaches. The primary goal of this study is to compare the technical depth, structural complexity, performance metrics (LSE-D, LSE-C), hardware requirements (VRAM, GPU), and typical application areas of each approach. This paper is structured to align with the provided academic format: Section

2 details the methodology; Section 3 presents the findings and structural analyses; Section 4 discusses the comparative analysis and practical applications; and Section 5 summarizes the conclusion and future directions.

Problem Definition and Motivation

The rapid development in deep learning has led to a proliferation of lip-sync models, making it challenging for researchers and practitioners to determine which architectural approach (GAN, Transformer, Diffusion) is best suited for a specific application area. This challenge is compounded by the fact that performance metrics, hardware requirements, and output quality vary significantly across these architectures. Furthermore, the focus on **open-source** models is critical for ensuring research reproducibility, fostering community development, and promoting ethical transparency within this rapidly evolving domain. This study is motivated by the need for a structured and comparative analysis to bring clarity to the current state of the art.

Original Contribution of the Study

This structured comparative analysis fills a significant gap in the current literature by providing the following original contributions:

1. *Temporal Focus:* By focusing specifically on the critical 2020–2025 period, the review captures the most recent evolution of deep learning-based lip synchronization, a period defined by the emergence of Transformer and Diffusion models.

2. *Architectural Comparison:* It is one of the first studies to systematically compare the three dominant architectures (GAN, Transformer/NeRF, Diffusion) through their most influential open-source representatives: Wav2Lip (Prajwal et al., 2020), GeneFace (Ye et al., 2023a), and Diff2Lip (Mukhopadhyay et al., 2024). This comparison includes technical depth, performance metrics, and hardware requirements.

3. *Practical Application Mapping:* The study provides a clear mapping of the most suitable application areas for each architecture (e.g., automated dubbing, virtual avatars, cinematic production), offering a practical guide for industry adoption.

Methodology

This comparative study employed a multi-stage selection and analysis strategy, consistent with academic standards, to identify and evaluate open-source lip synchronization models within the specified period (2020–2025).

Data Sources and Search Strategy

The literature search was conducted in key engineering and computer science databases, including IEEE Xplore and Scopus, as well as the preprint server arXiv and the comprehensive academic search engine Google Scholar. IEEE Xplore and Scopus were chosen as primary sources due to their focus on high-impact publications in computer vision and artificial intelligence. arXiv and Google Scholar were used as complementary sources to capture the latest pre-prints and highly cited works not yet indexed in the primary databases. The search terms covered the key concepts mentioned in the Abstract, using the following Boolean combination: ("lip synchronization" OR "lip-sync") AND ("GAN" OR "Transformer" OR "Diffusion") AND "open source" AND (2020:2025)

Study Selection Criteria and Process

The following criteria were used for selecting the models to be reviewed:

Inclusion Criteria

- *Open-Source Availability*: The model's code and pre-trained weights must be publicly available.
- *Publication Period*: Must have been published between 2020 and 2025.
- *Architectural Representation*: Must represent one of the three main deep learning architectures (GAN, Transformer, Diffusion) in the field.

Exclusion Criteria

- Commercial or closed-source models.
- Studies focusing solely on speech recognition or audio synthesis without visual synchronization.
- Low-impact works or those not published in a peer-reviewed conference or journal.

Based on these criteria, the most influential and highly cited models representing each architecture were selected:

- GAN Representative: Wav2Lip (Prajwal et al., 2020)
- Transformer Representative: GeneFace (Ye et al., 2023)
- Diffusion Representative: Diff2Lip (Mukhopadhyay et al., 2024)

The selection process involved an initial screening of titles and abstracts, followed by a full-text review of potentially relevant articles to ensure adherence to the inclusion and exclusion criteria.

Data Extraction and Synthesis

For each selected model, specific data points were systematically extracted (data extraction) to facilitate comparative analysis: publication year, core architecture, utilized dataset, key performance metrics (LSE-D (Chung & Zisserman, 2017), LSE-C, PSNR, FID (Heusel et al., 2017), training hardware, inference speed, and core technical innovation. These data points were then analyzed using a comparative synthesis approach, primarily through structured tables, to identify trends and highlight architectural trade-offs.

Findings: Technical Analysis of Main Architectural Approaches

Lip synchronization models generate video frames conditioned on audio input. Each architecture accomplishes this task using a distinct mathematical framework and optimization objective.

2020-2025 Period Trends

The 2020-2025 period reflects a clear shift in research focus. While GAN-based models dominated the early part of the decade (Wav2Lip in 2020), the latter half has seen a significant increase in Transformer and Diffusion-based approaches (GeneFace in 2023, Diff2Lip in 2024). This trend indicates a move away from the artifact-prone nature of GANs towards models prioritizing 3D consistency and photorealism, often at the expense of computational cost. This shift is accompanied by GAN-based improvements focusing on high resolution, such as VideoReTalking (Wang et al., 2022).

Architecture-Based Technical Analysis

Generative Adversarial Networks (GANs): Wav2Lip

Wav2Lip (Prajwal et al., 2020) is a landmark achievement in the field of lip synchronization.

- *Technical Depth:* The key to Wav2Lip's success is its use of a *Discriminator* based on *SyncNet* (Chung & Zisserman, 2017), which is trained as a "Lip Sync Expert." This Discriminator evaluates the synchronization accuracy between the generated video frame and the audio during training. The generator takes the original facial image and corresponding audio features (MFCCs) as input, repainting the mouth region with synchronized lip movements. This adversarial training forces the Generator to produce exceptionally accurate lip movements, enabling Wav2Lip to achieve high LSE-D/LSE-C scores.

- *Limitations and Improvements:* Wav2Lip's main limitation is its tendency to produce artifacts and degrade video quality around the mouth area, especially in high-resolution (HD) videos. To address this, improvements like *VividWav2Lip* (Liu et al., 2024) have been developed. Models such as *VideoReTalking* (Wang et al., 2022) also focused on enhancing GAN-based resolution and quality before the rise of diffusion models.

Transformer and NeRF: GeneFace

GeneFace (Ye et al.,) and its subsequent version GeneFace++ (Ye et al., 2023b) have significantly advanced the field by moving lip synchronization into the 3D domain, ensuring better identity preservation and temporal consistency. These models represent a major trend in the post-2023 era.

- Technical Depth: GeneFace addresses the process in two stages:

1. Audio-to-Motion Stage: Takes audio features as input and uses a Transformer or similar network to predict the speaker's 3D facial movements (3DMM parameters).

2. Motion-to-Video Stage: The predicted 3D motion parameters are converted into photorealistic 2D video frames using a NeRF (Neural Radiance Field)-based renderer. The use of NeRF allows the model to consistently handle the face's 3D geometry and lighting.

- Real-Time Performance: GeneFace++ (Ye et al., 2023a) integrates more efficient rendering techniques, such as RAD-NeRF, into the NeRF architecture, making it one of the first models of this type capable of real-time operation (30+ FPS). This makes the high quality of Transformer/NeRF-based approaches practical for live-streaming and interactive applications.

Diffusion Models: Diff2Lip

Diffusion-based models, such as *DiffTalk* (Shen et al., 2023), Diff2Lip (Mukhopadhyay et al., 2024), and *Sayanything*, have rapidly gained traction since 2023. These models bring the superior photorealism capabilities of image generation to lip synchronization.

Technical Depth: Diff2Lip formulates the lip synchronization task as an inpainting problem. The model takes a video frame with the mouth region masked and audio features as input. Through the diffusion process (a sequence of denoising steps), it gradually removes noise, filling the masked area with photorealistic lip movements that match the audio. Due to the nature of diffusion models, they produce high-resolution and visually consistent outputs, largely eliminating the artifact issues common in GANs.

Computational Cost: While diffusion models generally offer more stable training than GANs, the inference stage requires multiple denoising steps, making them slower and demanding substantial VRAM (12 GB or more). This presents a limitation for instant or low-hardware applications.

Discussion

Evaluation of Performance and Hardware Comparison

These three architectures are best suited for different application scenarios and hardware constraints. Table 1 below summarizes the technical and practical requirements of the three main representative models. As seen in Table 1, the visual quality improvement offered by Diff2Lip is directly correlated with its high VRAM requirement (12 GB+). This demonstrates that achieving high-fidelity, photorealistic output still necessitates a significantly higher computational cost. Conversely, Wav2Lip, despite its lower visual quality, remains valuable for its low computational footprint and high synchronization accuracy, making it the most accessible option. GeneFace++ (Ye et al., 2023b) strikes a balance, offering high quality and real-time performance through efficient NeRF rendering, positioning it as the preferred choice for interactive applications

Table 1. Performance and hardware comparison of Wav2Lip, GeneFace++, and Diff2Lip

Criterion	Wav2Lip (GAN) (Prajwal et al., 2020)	GeneFace++ (Transformer/NeRF) (Ye et al., 2023a)	Diff2Lip (Diffusion) (Mukhopadhyay et al., 2024)
Primary Focus	Synchronization Accuracy	3D Consistency and Real-Time	Photorealism and High Resolution
LSE-D (↓)	Lowest (~0.20)	Medium (~0.22)	Medium-High (~0.21)
Inference Speed	Fast	Very Fast (30+ FPS)	Slow
VRAM Requirement (Inference)	Low (4-6 GB)	Medium (8 GB+)	High (12 GB+)
Video Quality	Medium (Risk of artifacts)	High (3D Consistency)	Highest (Photorealistic)
Typical GPU	GTX 1060 / T4	RTX 3060 / A10	RTX 3080 / 4090 / A6000

Ethical and Social Implications

The rapid advancement of open-source lip synchronization technology, while enabling innovation, also raises significant ethical concerns, primarily related to the potential for malicious use, such as the creation of convincing deepfakes. Open-source availability democratizes access to this technology, making it available to bad actors. Therefore, academic research must address the ethical responsibilities associated with these models. Future work should focus not only on generation but also on robust *deepfake detection and attribution methods* to mitigate these risks. This dual focus, supported by specific detection datasets and methods like **LipForensics** (Haliassos et al., 2021) and *TrueSync* (El-Taj et al., 2025), ensures that the technology is developed responsibly (Ye et al., 2023b).

Open Challenges and Future Research Directions

While significant progress has been made, several open challenges remain critical for the broader adoption of lip-sync technology:

- *Real-Time Performance***:** Achieving photorealistic output (like Diff2Lip) at real-time speeds (like GeneFace++) remains a key technical hurdle.
- *Multilingual Support:* Most models are trained predominantly on English datasets. Generalizing performance to diverse languages, which have different phoneme-viseme mappings, requires further research and the creation of large, diverse multilingual datasets (Oskooei et al., 2024).
- *Generalization to Diverse Faces:* Models often struggle to generalize effectively to faces with extreme poses, different lighting conditions, or diverse facial structures.
- *Audio-Visual Coherence***:** Ensuring that the generated lip movements not only match the audio but also align naturally with the speaker's identity and emotional state is an ongoing area of research.

Table 2. Application areas of GAN-, transformer/NeRF-, and diffusion-based Lip-Sync models

Architecture	**Application Areas**
GAN (Wav2Lip)	Automated Dubbing: Translating a large volume of videos with low latency and cost. Educational Content: Quick and easy video production.
Transformer/NeRF (GeneFace)	Virtual Avatars: Real-time 3D speaking characters in games, Metaverse, and Virtual Reality (VR) environments. Live Streaming: Low-latency virtual presenters.
Diffusion (Diff2Lip)	Film and Production: Cinematic content creation where high visual quality is paramount. Deepfake Synthesis: Visual effects requiring high realism.

Conclusion and Future Directions

The 2020-2025 period demonstrates a rapid and diverse evolution of open-source lip synchronization models. *GANs (Wav2Lip)* established a practical and accessible foundation, *Transformer/NeRF (GeneFace)* introduced

3D consistency and real-time capability, and *Diffusion Models (Diff2Lip)* maximized visual quality. The primary contribution of this structured comparative analysis is the systematic comparison of these three dominant architectures, providing a clear map of their strengths, weaknesses, and optimal application scenarios. The future will likely belong to *hybrid* architectures, seeking to combine the photorealism of diffusion models with the speed of GANs or Transformers. Furthermore, *multilingual support* and *addressing ethical challenges* [remain critical areas for the technology's broader and responsible adoption (Ye et al., 2023b; Oskooei et al., 2024).

Scientific Ethics Declaration

* The authors declare that the scientific ethical and legal responsibility of this article published in EPESS journal belongs to the authors.

Conflict of Interest

* The authors declare that they have no conflicts of interest.

Funding

* This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Education in Technology and Engineering (www.iconete.net) held in Antalya/Türkiye on November 12-15, 2025.

* We gratefully acknowledge the support and contribution of the Aktif Bank R&D Center throughout the research and analysis process.

References

Chung, J. S., & Zisserman, A. (2017). Out of time: Automated lip sync in the wild. In *Computer vision – ACCV 2016 workshops* (pp. 251–263). Springer.

El-Taj, H., Alammari, F., Alkhowaiter, J., Bogari, L., & Essa, R. (2025). Deepfake detection based on visual lip-sync match and blink rate. *International Journal of Computational and Experimental Science and Engineering, 11*(1), 886–898.

Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection (LipForensics). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5039–5049). IEEE.

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. In *Advances in neural information processing systems* (Vol. 30, pp. 6626–6637). Curran Associates.

Liu, L., Wang, J., Chen, S., & Li, Z. (2024). VividWav2Lip: High-fidelity facial animation generation. *Electronics, 13*(18), 3657.

Ma, J., Wang, S., Yang, J., Hu, J., Liang, J., Lin, G., Chen, J., Li, K., & Meng, Y. (2025). SayAnything: Audio-driven lip synchronization with conditional video diffusion. *arXiv preprint* arXiv:2502.11515.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In *Computer vision – ECCV 2020* (pp. 405–421). Springer.

Mukhopadhyay, S., Ghosh, A., & Chaudhuri, S. (2024). Diff2Lip: Audio conditioned diffusion models for lip-synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2024)* (pp. 5028–5037). IEEE.

Oskooei, M., et al. (2025). Seeing the sound: Multilingual lip sync for real-time face-to-face translation. *Computers, 14*(1), 7.

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 (Wav2Lip). In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)* (pp. 484–492). ACM.

Shen, S., Zhao, W., Meng, Z., Li, W., Zhu, Z., Zhou, J., & Lu, J. (2023). DiffTalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023)* (pp. 1982–1991). IEEE.

Wang, Z., Mallya, A., & Liu, F. (2022). VideoReTalking: Audio-based lip synchronization for talking head video editing in the wild. In *SIGGRAPH Asia 2022 Conference Papers (SA '22)*. ACM. h

Ye, Z., Jiang, Z., Ren, Y., Liu, J., He, J., & Zhao, Z. (2023a). GeneFace: Generalized and high-fidelity audio-driven 3D talking face synthesis. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*. Retrieved from https://arxiv.org/abs/2301.13430 arXiv

Ye, Z., He, J., Jiang, Z., Huang, R., Huang, J., Liu, J., Ren, Y., Yin, X., Ma, Z., & Zhao, Z. (2023b). GeneFace++: Generalized and stable real-time audio-driven 3D talking face generation. *arXiv preprint* arXiv:2305.00787. Retrieved from https://arxiv.org/abs/2305.00787

Author(s) Information

Bilge Nur Saglam
Aktif Investment Bank Inc.
Esentepe District, Kore Şehitleri Street No: 8/1 Şişli
İstanbul, Türkiye
Contact e-mail : *Bilgenur.saglam@aktifbank.com.tr*

Mustafa Keles
Aktif Investment Bank Inc.
Esentepe District, Kore Şehitleri Street No: 8/1 Şişli
İstanbul, Türkiye

Mehmet Kutanoglu
Aktif Investment Bank Inc.
Esentepe District, Kore Şehitleri Street No: 8/1 Şişli
İstanbul, Türkiye

To cite this article:

Saglam, B. N., Keles, M., & Kutanoglu, M. (2025). Open source Lip-Sync models in the period 2020-2025: A structured comparative analysis. T*he Eurasia Proceedings of Educational and Social Sciences (EPESS), 46,* 137-143.