
Supplementary Material: OmniSync: Towards Universal Lip Synchronization via Diffusion Transformers

Anonymous Author(s)

Affiliation

Address

email

1 Implementation Details

2 The OmniSync framework is built upon a Diffusion Transformer (DiT) architecture designed specifically for universal lip synchronization across diverse visual scenarios.

4 We implement a transformer-based latent diffusion architecture as our core text-to-video (T2V) generation framework. The system begins by encoding video content into latent representations through a 3D Variational Autoencoder (3D-VAE), creating a compressed foundation for our diffusion process. Our design diverges from conventional approaches that typically employ either UNet architectures or transformers with separate 1D temporal attention mechanisms. Instead, we address the inherent limitations of decoupled spatiotemporal processing by integrating full 3D self-attention throughout the model, enabling unified perception of spatiotemporal relationships. This integrated approach enhances physical coherence and visual fidelity in the generated content. In our implementation, we incorporate timestep conditioning by projecting diffusion timesteps to scaling factors and applying RMSNorm to the token representations before each attention module and feed-forward network (FFN), creating a robust framework for high-quality video synthesis. For audio conditioning, we utilize a pretrained Whisper encoder that extracts 1024-dimensional features from speech signals. Text conditioning is handled through a T5-large encoder providing 768-dimensional embeddings.

17 Training was conducted on a combined dataset comprising the MEAD dataset and a 400-hour collection of YouTube videos. We employed the AdamW optimizer with a learning rate of $1e-5$, weight decay of 0.01, and betas of (0.9, 0.999). The model was trained for 80,000 steps using a batch size of 64 across 64 NVIDIA A100 GPUs, with gradient accumulation of 2 steps to effectively double the batch size. The entire training process required approximately 80 hours to complete.

22 When our Dynamic Spatiotemporal Classifier-Free Guidance (DS-CFG) mechanism cannot detect facial landmarks—particularly common with stylized characters or non-human entities—we implement a straightforward fallback strategy. The system performs a center-biased crop of the frame, applying Gaussian-weighted guidance based on the assumption that speaking subjects typically have their mouth positioned near the center. This approximation works effectively since our approach requires only rough localization—the mouth need only be generally positioned in the central region rather than precisely mapped.

29 2 Training Procedure Details

30 A critical aspect of our training methodology is the **timestep-dependent sampling strategy**. The core idea behind this strategy is to provide the model with different types of training data according to the different denoising stages of the diffusion model. This optimizes learning efficiency and final generation quality, especially for the specific task of lip synchronization.

34 Considerations for Early Denoising Stages and Link to Progressive Noise Initialization

35 The primary requirement of the lip synchronization task is to modify only the lip region to match
36 new audio while preserving the head pose, identity features, and background environment of the
37 input video. The early stages of a traditional diffusion process starting from pure random noise
38 (i.e., when the noise level is very high, and t approaches the total number of steps T) are mainly
39 responsible for generating the macroscopic structure of the image, including pose and general identity
40 outlines. However, for lip synchronization, the input video itself already provides this macroscopic
41 structural information. **Since our goal is to *not* change the pose and background, learning this**
42 **already-given information from scratch (starting from random noise) is not only inefficient**
43 **but can also introduce unnecessary variability, potentially leading to discrepancies in pose or**
44 **identity between the final generated video and the original.**

45 Therefore, during inference, we employ the **Progressive Noise Initialization** strategy. This involves
46 directly adding a controlled level of noise (corresponding to a high starting timestep t_{start} , e.g.,
47 $\tau = 0.92$) to the original video frames (which we refer to as "base frames") and then proceeding with
48 the subsequent denoising steps from this t_{start} . This is equivalent to "skipping" the early denoising
49 process from pure noise to t_{start} . **The rationale is that the base video already contains all the**
50 **macroscopic structural information (pose, identity, background) that we wish to preserve. By**
51 **directly noising the base video, we provide the model with an initial state that already possesses**
52 **the correct pose and identity, allowing it to focus the subsequent denoising process more on the**
53 **precise editing of the lip region.**

54 Timestep-Dependent Sampling Strategy in Training

55 To enable the model during the training phase to adapt to this inference mode of "skipping early
56 structure generation" and to effectively learn the specific knowledge required for lip synchronization,
57 we designed the following timestep-dependent sampling strategy:

58 • **For higher noise levels (early diffusion stages, e.g., $t > 850$):** At this stage, even
59 when starting from random noise, the model is primarily learning to construct basic facial
60 structures and poses. To provide the model with the most stable and relevant learning signals,
61 we sample from **pseudo-paired data**. We specifically selected the **MEAD dataset** as our
62 source of pseudo-paired data. The MEAD dataset, recorded under controlled laboratory
63 conditions, offers several crucial advantages:

- 64 – *Fixed Recording Conditions:* Multiple emotional expressions from the same subjects
65 were captured with fixed camera positions and consistent lighting. This results in video
66 sequences where the facial structure and identity remain constant, with variations only
67 in lip shapes and expressions.
- 68 – *Natural Pseudo-Pairs:* Within the same identity, frames from different utterances form
69 natural pseudo-pairs—they maintain nearly identical head poses and environmental
70 conditions but differ in lip configurations.
- 71 – *Multi-View Capture:* The multi-view setup in MEAD further enriches our training
72 data by providing consistent identity representations across different angles. This
73 enables the model to learn pose-invariant facial structures more robustly during the
74 early diffusion timesteps.

75 By using this carefully curated pseudo-paired data, the model, in the early (high-noise)
76 stages, learns how to begin constructing facial features under given (or nearly given) pose
77 and identity conditions, laying a solid foundation for subsequent lip generation. **At this**
78 **point, the model needs to learn how to recover a structure from the noise that is highly**
79 **consistent in pose and identity with the input condition (V_{cd} , the noised version of**
80 **the base video), while preparing for the generation of the target lip shape (the noised**
81 **version of V_{ab}). Therefore, a strong correspondence in pose between the input (x_t from**
82 **noised V_{cd}) and the target (noised V_{ab}) is critical.**

83 • **For lower noise levels (middle and late diffusion stages, e.g., $t \leq 850$):** As the diffusion
84 process enters the middle and late stages, the noise level gradually decreases. At this point, **by**
85 **visualizing x_t at different stages (similar to the analysis in works like VideoJAM [1]), we**
86 **can observe that the main contour information, facial structure, and identity features**

have become relatively clear. In other words, the model has already "seen" the character's pose and general identity.

- *Shift in Model's Learning Focus:* The primary task of the model at this stage is no longer to construct macroscopic structures but to finely sculpt the lip shape and texture details according to the audio signal, ensuring a natural blend with the rest of the face.
- *Rationale for Using Unpaired Data:* Since macroscopic information such as pose and identity is largely established (guaranteed by early-stage learning and/or progressive initialization at inference), the model no longer heavily relies on strict pose pairing between input and target. Therefore, we can transition to sampling from our broader and more diverse collection of **arbitrary/unpaired videos (from the YouTube dataset)**. While this data may not match any specific input video V_{cd} in terms of pose, identity, or scene, it provides a vast number of lip movement samples under different speaking contents and styles. **The model at this stage learns a more generalized "audio-to-lip-shape" mapping and how to apply this mapping to an x_t that already possesses basic contours, to generate lip shapes synchronized with the target audio A_{ab} , and to refine local details and realism.** This strategy significantly expands the diversity of training data, enhancing the model's generalization capabilities to handle various real-world lip synchronization scenarios.

In summary, our timestep-dependent sampling strategy, combined with progressive noise initialization at inference, allows OmniSync to efficiently focus on the core challenges of lip synchronization. By using pseudo-paired data in the early stages to stabilize structural learning and leveraging large-scale unpaired data in the middle and late stages to learn diverse lip expressions, the model ensures both identity and pose consistency in the generated videos and possesses strong lip generation and generalization capabilities.

3 Inference Details

During inference, we implement our flow-matching-based progressive noise initialization strategy with $\tau = 0.92$, followed by 50 denoising steps. This τ value was determined through experimentation that revealed an optimal balance point in the noise-quality tradeoff curve. At $\tau = 0.92$, the model preserves sufficient structural information from the original frames to maintain identity consistency and pose alignment, while still introducing adequate noise to enable flexible modification of lip regions according to audio input. Lower τ values (e.g., 0.85) resulted in excessive preservation of original lip shapes, limiting the model's ability to generate accurate synchronized movements, while higher values (e.g., 0.98) introduced too much noise, compromising identity preservation and creating boundary artifacts. This approach significantly reduces computational requirements compared to standard diffusion processes while maintaining high-quality outputs. The DS-CFG mechanism employs a peak strength (ω_{peak}) of 9.0, base strength (ω_{base}) of 1.0, and temporal decay rate (γ) of 1.5. These parameters balance the influence of audio conditioning across both spatial dimensions and diffusion timesteps.

4 Comparison with Portrait Animation Methods

Our OmniSync framework presents distinct advantages when compared to state-of-the-art Portrait Animation methodologies, such as EMO [2], EchoMimic [3], Hallo [4], and Sonic [5]. While these models can achieve good lip synchronization and generalize across diverse visual styles, they often struggle to fully preserve the unique identity, texture, and dynamic characteristics of source *video* material. This can lead to animations that, despite accurate lip sync, may appear somewhat generic or lose fine-grained subject resemblance.

In contrast, OmniSync's training paradigm, formulated as $(V_{cd}, A_{ab}) \mapsto V_{ab}$, utilizes multi-frame video input (V_{cd}) instead of a single static image. This video input inherently encodes richer information, including subtle facial dynamics and individual speaking styles over a temporal window, beyond static texture details. Conceptually, while the input V_{cd} in our framework serves a conditioning role analogous to the single image in portrait animation approaches, the crucial difference lies in the temporal dimension of V_{cd} , which allows for the preservation of more stylistic and identity-specific information.



Figure 1: **Comparison with portrait animation methods.** Visual comparison between our OmniSync framework and other approaches (EchoMimic, Hallo3, and Sonic), demonstrating our method’s superior ability to preserve identity and natural speaking style while maintaining accurate lip synchronization.

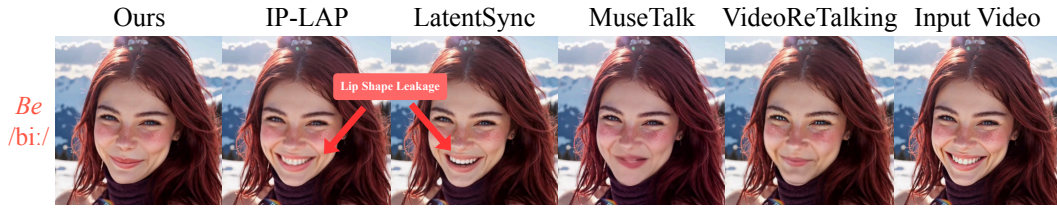


Figure 2: **Comparison of lip synchronization methods** showing how our approach effectively prevents lip shape leakage from the original video, while competing methods struggle with this issue (highlighted in red for IP-LAP and LatentSync).

139 This architectural choice enables OmniSync to maintain strong generalization capabilities while
 140 achieving high fidelity to the source. As shown in Fig. 1, our qualitative comparison indicates that
 141 methods such as EchoMimic [3], Hallo3 [6], and Sonic [5] face challenges in maintaining identity,
 142 which may result in less realistic outcomes, while OmniSync continuously generates synchronized
 143 speech, better preserving the distinctive texture quality and inherent speaking style of the input
 144 video. This holds true even when applied to out-of-distribution subjects, such as non-human or
 145 highly stylized characters. Our model effectively learns to transfer primarily the lip movements
 146 dictated by the target audio, while retaining other visual characteristics of the source video segment.
 147 Consequently, OmniSync’s outputs appear as more natural integrations with the original content.

148 5 Lip Shape Leakage in Lip Synchronization

149 Lip shape leakage occurs when original lip movements from source videos persist in the synchronized
 150 output despite attempts to replace them with new audio-driven movements. This phenomenon
 151 commonly affects traditional lip synchronization methods that rely on masked-frame inpainting.
 152 When using explicit masks to isolate mouth regions, the boundaries between what should be modified
 153 and preserved often become ambiguous, allowing original lip shapes to "leak" into the generated
 154 content. Additionally, the relatively weak conditioning signal provided by audio compared to visual
 155 information exacerbates this problem, as models may default to preserving visual aspects of the
 156 original video rather than fully replacing them.

157 The impact of lip shape leakage is significant for synchronization quality. Viewers perceive temporal
 158 inconsistencies where mouth movements partially match both the original video and target audio,
 159 creating unnatural motion patterns. Critical phonemes requiring distinct mouth shapes (such as
 160 bilabial or labiodental sounds) may not form correctly, reducing visual intelligibility and breaking the
 161 illusion of natural speech. As demonstrated in Fig. 2, methods like IP-LAP and LatentSync frequently
 162 show this limitation, particularly for challenging phonemes that require distinct mouth formations.

163 OmniSync addresses lip shape leakage through its mask-free training paradigm and dynamic spa-
 164 tiotemporal guidance mechanism. Rather than using explicit masks with hard boundaries, our



Figure 3: **Comparison of lip synchronization without (left) and with (right) Video Description conditioning.** Text guidance produces more pronounced lip movements with clearer dental visibility across diverse subjects, as shown in the magnified mouth regions.

165 diffusion-based approach directly modifies frames based on audio conditioning, allowing the model
 166 to determine appropriate modification regions dynamically. Our DS-CFG further mitigates leakage
 167 by applying stronger guidance around mouth regions, ensuring audio conditioning overcomes any ten-
 168 dency for original lip shapes to persist in the output. These innovations effectively eliminate lip shape
 169 leakage, enabling precise and consistent lip synchronization across diverse visual representations,
 170 including stylized characters where traditional face alignment techniques often fail.

171 6 Impact of Video Description on Lip Clarity and Movement Amplitude

172 We conducted an ablation study to evaluate the effect of video description conditioning on lip
 173 synchronization quality. As shown in Fig. 3, we compare results generated without video description
 174 (left column) to those with video description conditioning (right column). The qualitative comparison
 175 clearly demonstrates the impact of textual guidance on lip movement characteristics.

176 During training, we labeled videos with descriptive prompts such as "A person speaking loudly with
 177 clear facial and tooth movements" to establish associations between textual descriptions and specific
 178 lip characteristics. This approach allows the model to learn correlations between descriptive language
 179 and visual speech attributes. At inference time, these text descriptions serve as an interpretable control
 180 mechanism, enabling adjustment of lip clarity and movement amplitude through prompt engineering
 181 without requiring model retraining. The examples demonstrate how this text-guided approach results

182 in more expressive and visually distinct lip synchronization, enhancing overall perceptual quality for
183 viewers.

184 From the visual results, two key improvements are immediately apparent with video description
185 conditioning. First, the lip movements exhibit greater amplitude and expressiveness, particularly
186 evident in the middle and bottom rows where the subjects display more pronounced mouth openings.
187 Second, there is notably improved clarity of dental structures, with teeth being more visible and
188 defined in the right column examples. This enhancement in visual detail contributes significantly to
189 the realism and comprehensibility of the generated speech.

190 7 User Study

191 To comprehensively evaluate the perceptual quality of our proposed OmniSync framework, we
192 conducted an extensive user study involving 39 participants from diverse backgrounds. This study
193 was designed to assess multiple dimensions of lip synchronization quality and compare OmniSync
194 against seven state-of-the-art methods: Wav2Lip, VideoReTalking, TalkLip, IP-LAP, Diff2Lip,
195 MuseTalk, and LatentSync.

196 The study presented participants with 32 video sets, with each set containing eight versions of
197 the same content. These video sets were chosen to represent a wide range of scenarios from our
198 benchmark, including challenging cases with varied head poses, lighting conditions, facial occlusions,
199 and stylistic representations. To prevent order bias, the presentation sequence was randomized for
200 each participant and each video set.

201 Participants evaluated the videos across five key criteria that collectively capture the most important
202 aspects of lip synchronization quality. Lip Sync Accuracy measured how well the mouth movements
203 aligned with the speech audio, with particular attention to phonetic correspondence. Character
204 Identity Preservation assessed how well each method maintained the original character’s appearance
205 and distinctive features throughout the video. Timing Stability evaluated the temporal consistency
206 and naturalness of the generated motion, including the absence of jitter or unnatural transitions.
207 Image Quality focused on the presence or absence of visual artifacts, blurriness, or distortions in the
208 generated content. Video Realism provided an overall assessment of how natural and believable the
209 final result appeared. Each criterion was rated on a 5-point Likert scale, where 1 represented "Very
210 Poor" quality, 2 indicated "Poor" quality, 3 meant "Acceptable" quality, 4 signified "Good" quality,
211 and 5 represented "Excellent" quality.

212 To ensure the reliability and validity of our study, we conducted thorough statistical analyses of the
213 collected data. The internal consistency of ratings was excellent, with a Cronbach’s α coefficient of
214 0.98, indicating high reliability across the evaluation criteria.

215 The qualitative feedback provided by participants offered valuable insights into the perceived strengths
216 and limitations of different approaches. Participants frequently highlighted OmniSync’s natural
217 lip movements for difficult phonemes such as ‘p’ and ‘b’, which often pose challenges for lip
218 synchronization systems. Many noted the consistent identity preservation even during extreme
219 expressions, contrasting this with other methods that showed noticeable identity shifts. The absence
220 of artifacts around the mouth area was another commonly mentioned advantage, with participants
221 comparing this favorably to the boundary issues observed with mask-based approaches.

222 8 Benchmark Construction Details

223 To comprehensively evaluate the performance of lip synchronization methods within the current
224 AI-Generated Content (AIGC) environment, we have meticulously constructed the AIGC-LipSync
225 Benchmark. This benchmark comprises a total of 615 video clips, all generated by leading text-to-
226 video (T2V) or image-to-video (I2V) models. These videos primarily originate from advanced models
227 including Kling, Dreamina, Wan [7], and Hunyuan [8], with all raw video materials downloaded from
228 publicly accessible communities such as Civitai, ensuring a diverse and representative collection.

229 This benchmark has been specifically curated to include a variety of AI-generated content that
230 poses significant challenges for lip synchronization. Among all data, there are approximately 450
231 videos featuring realistic human subjects, around 125 videos of stylized characters with distinct
232 artistic styles, and a smaller set of approximately 40 videos depicting more challenging, atypical

humanoid characters. This composition is designed to span a wide spectrum of visual representations, from photorealistic to highly abstract, thereby enabling a more rigorous test of model generalization capabilities and robustness. These video clips have an average duration of approximately 6 seconds, an average resolution of 976x1409 pixels, and an average frame rate maintained at 30.00 FPS, providing sufficient data for detailed synchronization analysis. As illustrated in Fig. 4 and 5, representative examples from our benchmark showcase its content diversity and the inherent challenges it presents.

9 Ethical Considerations

Our OmniSync framework, while advancing the field of lip synchronization, raises important ethical considerations regarding potential misuse. As with any technology capable of manipulating facial content, there exists risk for creating misleading or deceptive media that could contribute to misinformation or deepfakes. We acknowledge this responsibility and have intentionally focused our research on improving existing video content rather than enabling impersonation or fabrication of speech.

To mitigate potential harms, we recommend implementing watermarking or content provenance solutions when deploying this technology commercially. Additionally, the creation of the AIGC-LipSync Benchmark emphasizes evaluating performance on stylized characters and AI-generated content, steering applications toward creative and entertainment purposes rather than realistic human impersonation.

We are committed to transparency regarding the capabilities and limitations of our approach. The technical advancements presented in OmniSync are published to advance scientific understanding and enable beneficial applications in areas such as film production, accessibility services, and educational content. We encourage the research community to continue developing detection methods for synthetically modified content alongside improvements in generation quality.

10 Limitations

Despite OmniSync’s demonstrated effectiveness across diverse visual scenarios, several limitations remain that present opportunities for future research and development.

First, the system occasionally struggles with very rapid speech patterns, resulting in simplified mouth movements that may not fully capture the nuanced articulation of fast-paced phonemes. This limitation is particularly noticeable for languages with naturally faster speech rates or speakers with rapid delivery styles. The model also shows reduced performance for rare phonemes in non-English languages, reflecting limitations in the linguistic diversity of the training data.

Second, handling multiple talking faces simultaneously represents another significant challenge. The current implementation focuses on individual speakers and lacks mechanisms for differentiating between multiple talking subjects in the same frame. This limitation becomes apparent in conversational scenes or group settings where multiple characters speak, potentially leading to inappropriate synchronization of non-speaking faces or conflicts in attention allocation. Developing multi-subject attention mechanisms and speaker identification components would be necessary to address this limitation effectively.

Furthermore, the current approach does not explicitly model emotional expression, limiting its ability to synchronize subtle emotional cues between speech and facial movements.

By addressing these limitations and pursuing these research directions, future work can build upon the foundation established by OmniSync to develop even more capable, efficient, and universally applicable lip synchronization systems.

References

- [1] H. Chefer, U. Singer, A. Zohar, Y. Kirstain, A. Polyak, Y. Taigman, L. Wolf, and S. Sheynin, “Videogram: Joint appearance-motion representations for enhanced motion generation in video models,” *arXiv preprint arXiv:2502.02492*, 2025.

- 279 [2] L. Tian, Q. Wang, B. Zhang, and L. Bo, “Emo: Emote portrait alive generating expressive portrait
280 videos with audio2video diffusion model under weak conditions,” in *European Conference on*
281 *Computer Vision*. Springer, 2024, pp. 244–260.
- 282 [3] Z. Chen, J. Cao, Z. Chen, Y. Li, and C. Ma, “Echomimic: Lifelike audio-driven portrait animations
283 through editable landmark conditions,” in *Proceedings of the AAAI Conference on Artificial*
284 *Intelligence*, vol. 39, no. 3, 2025, pp. 2403–2410.
- 285 [4] M. Xu, H. Li, Q. Su, H. Shang, L. Zhang, C. Liu, J. Wang, Y. Yao, and S. Zhu, “Hallo: Hierarchi-
286 cal audio-driven visual synthesis for portrait image animation,” *arXiv preprint arXiv:2406.08801*,
287 2024.
- 288 [5] X. Ji, X. Hu, Z. Xu, J. Zhu, C. Lin, Q. He, J. Zhang, D. Luo, Y. Chen, Q. Lin *et al.*, “Sonic:
289 Shifting focus to global audio perception in portrait animation,” *arXiv preprint arXiv:2411.16331*,
290 2024.
- 291 [6] J. Cui, H. Li, Y. Zhan, H. Shang, K. Cheng, Y. Ma, S. Mu, H. Zhou, J. Wang, and S. Zhu, “Hallo3:
292 Highly dynamic and realistic portrait image animation with video diffusion transformer,” *arXiv*
293 *preprint arXiv:2412.00733*, 2024.
- 294 [7] A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng *et al.*,
295 “Wan: Open and advanced large-scale video generative models,” *arXiv preprint arXiv:2503.20314*,
296 2025.
- 297 [8] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang *et al.*,
298 “Hunyuanvideo: A systematic framework for large video generative models,” *arXiv preprint*
299 *arXiv:2412.03603*, 2024.

Non-Human Characters



Landscape Video Footage



Figure 4: **AIGC-LipSync Benchmark Examples (I):** Showcasing non-human characters and landscape-oriented video materials included in the benchmark. This highlights its coverage of diverse subject types and video formats, designed to test the model’s lip synchronization capabilities on non-traditional visual inputs.

Female Character Footage



Male Character Footage



Figure 5: **AIGC-LipSync Benchmark Examples (II)**: Presenting diverse male and female character materials from the benchmark. These are used to evaluate the model’s lip synchronization effectiveness and identity preservation across various human figures, particularly in the rendition of facial features and nuanced expressions.