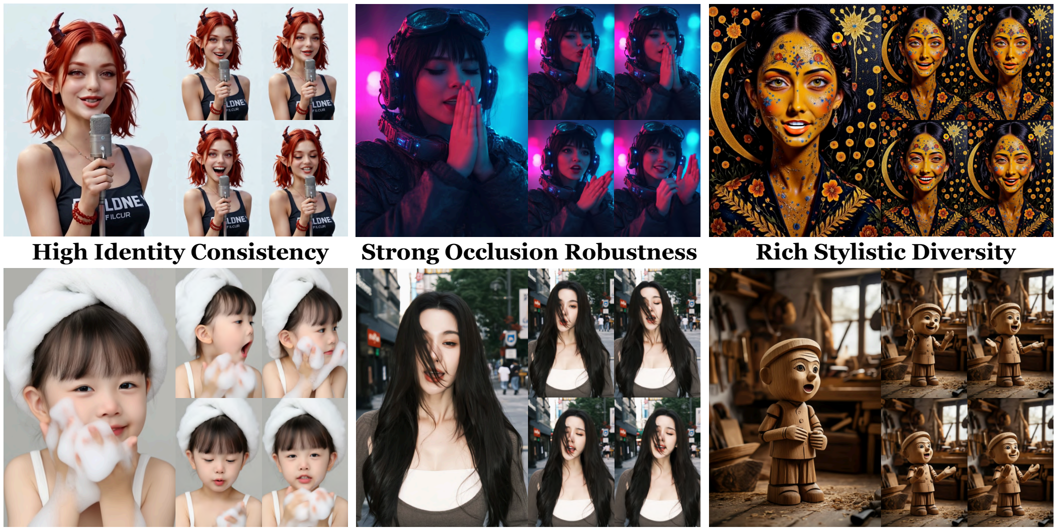
# OmniSync: Towards Universal Lip Synchronization via Diffusion Transformers

**Ziqiao Peng**[1*] **Jiwen Liu**[2†] **Haoxian Zhang**[2] **Xiaoqiang Liu**[2] **Songlin Tang**[2]
**Pengfei Wan**[2] **Di Zhang**[2] **Hongyan Liu**[3✉] **Jun He**[1✉]
[1]Renmin University of China [2]Kling Team, Kuaishou Technology [3]Tsinghua University
[†]Project Leader [✉]Corresponding Author

https://ziqiaopeng.github.io/OmniSync/



**High Identity Consistency** **Strong Occlusion Robustness** **Rich Stylistic Diversity**

Figure 1: **OmniSync** demonstrates universal lip synchronization capabilities, effectively handling facial occlusion, while maintaining visual consistency and generating accurate lip movements.

## Abstract

Lip synchronization is the task of aligning a speaker's lip movements in video with corresponding speech audio, and it is essential for creating realistic, expressive video content. However, existing methods often rely on reference frames and masked-frame inpainting, which limit their robustness to identity consistency, pose variations, facial occlusions, and stylized content. In addition, since audio signals provide weaker conditioning than visual cues, lip shape leakage from the original video will affect lip sync quality. In this paper, we present OmniSync, a universal lip synchronization framework for diverse visual scenarios. Our approach introduces a mask-free training paradigm using Diffusion Transformer models for direct frame editing without explicit masks, enabling unlimited-duration inference while maintaining natural facial dynamics and preserving character identity. During inference, we propose a flow-matching-based progressive noise initialization to ensure pose and identity consistency, while allowing precise mouth-region editing. To address the weak conditioning signal of audio, we develop a Dynamic Spatiotemporal Classifier-Free Guidance (DS-CFG) mechanism that adaptively adjusts guidance strength over time and space. We also establish the AIGC-LipSync Benchmark, the first evaluation suite for lip synchronization in diverse AI-generated videos.

[*] Work done during an internship at Kling Team, Kuaishou Technology.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Extensive experiments demonstrate that OmniSync significantly outperforms prior methods in both visual quality and lip sync accuracy, achieving superior results in both real-world and AI-generated videos.

## 1 Introduction

Lip synchronization, matching mouth movements with speech audio, is essential for creating compelling visual content across film dubbing [51], digital avatars [32, 31, 55, 5], and teleconferencing [24, 28, 53]. With the rise of AI-generated content, this technology has evolved from a specialized technique to a fundamental aspect of the video generation landscape [47, 34, 20]. Despite significant advances in text-to-video (T2V) models [4, 2, 48, 16, 39] creating increasingly realistic footage, achieving precise and natural lip synchronization remains an unsolved challenge.

Traditional lip synchronization approaches rely heavily on reference frames combined with masked-frame inpainting [33, 51, 11, 10]. This methodology extracts appearance information from reference frames to inpaint masked regions in target frames—a process that introduces several critical limitations. These methods struggle with head pose variations, identity preservation, and artifact elimination, especially when target poses differ significantly from references [30, 29].

Furthermore, the dependence on explicit masks cannot fully prevent unwanted lip shape leakage, compromising synchronization quality and restricting applicability across diverse visual representations [1]. The challenges intensify in the context of audio-driven generation. Unlike strong visual cues, audio signals provide relatively weak conditioning, making precise lip synchronization difficult [40]. Additionally, existing methods rely on face detection and alignment [3] techniques that break down when applied to stylized characters and non-human entities, precisely the diverse content that modern text-to-video models excel at generating.

This technical gap is compounded by the absence of standardized evaluation frameworks for lip sync in stylized videos. Current benchmarks [52, 42] focus almost exclusively on photorealistic human faces in controlled settings, failing to capture the visual diversity inherent in AI-generated videos.

To address these challenges, we introduce OmniSync, a universal lip synchronization framework designed for diverse videos. Our approach eliminates reliance on reference frames and explicit masks through a diffusion-based direct video editing paradigm. In addition, we establish AIGC-LipSync Benchmark, the first comprehensive evaluation framework for lip synchronization across diverse AIGC contexts. OmniSync's technical approach is built upon three key innovations:

First, we implement a mask-free training paradigm using Diffusion Transformers (DiT) [26] for direct cross-frame editing. Our model learns a mapping function $(V_{cd}, A_{ab}) \mapsto V_{ab}$, where $V$ represents video frames and $A$ represents audio. The indices $(a : b, c : d)$ represent different segments sampled from the same video. The model modifies only speech-relevant regions according to target audio without requiring explicit masks or references. This approach enables unlimited-duration inference while maintaining natural facial dynamics and preserving character identity.

Second, we introduce a flow-matching-based progressive noise initialization strategy during inference. Rather than beginning with random noise [38], we inject controlled noise into original frames using Flow Matching [19], then execute only the final denoising steps. This approach maintains spatial consistency between source and generated frames while allowing sufficient flexibility for precise mouth region modifications, effectively mitigating pose inconsistency and identity drift.

Third, we develop a dynamic spatiotemporal Classifier-Free Guidance (CFG) framework [13] that provides fine-grained control over the generation process. By adaptively adjusting guidance strength across both temporal and spatial dimensions: temporally reducing guidance strength as denoising progresses, and spatially applying Gaussian-weighted control centered on mouth-relevant regions. This balanced approach ensures precise lip synchronization without disturbing unrelated areas.

Our contributions can be summarized as follows:

- A universal lip synchronization framework that eliminates reliance on reference frames and explicit masks, enabling accurate speech synchronization across diverse visual representations.
- A flow-matching-based progressive noise initialization strategy during inference, effectively stabilizing the early denoising process and mitigating pose inconsistency and identity drift.

- A dynamic spatiotemporal CFG framework that provides fine-grained control over audio influence, addressing the weak signal problem in audio-driven generation.
- A comprehensive AIGC-LipSync Benchmark for evaluating lip synchronization in AI-generated content, including stylized characters and non-human entities.

## 2 Related Work

### 2.1 Audio-driven Lip Synchronization

**GAN-based Lip Synchronization.** Traditional GAN-based [9] methods [33, 41, 7, 23, 12] have established important foundations in lip synchronization. Wav2Lip [33] pioneered the use of pretrained SyncNet to supervise generator training, setting a benchmark for subsequent research. DINet [51] enhanced synchronization quality by performing spatial deformation on reference image feature maps, better preserving high-frequency details. IP-LAP [54] introduced a two-stage approach that first infers landmarks from audio before rendering them into facial images. ReSyncer [10] incorporated 3D mesh priors for facial motion, effectively reducing artifacts.

**Diffusion-based Lip Synchronization.** Recent advances in diffusion models [25, 50, 17, 22] have enabled significant progress in audio-driven lip synchronization. LatentSync [17] represents an end-to-end framework based on audio-conditioned latent diffusion models without intermediate motion representation. SayAnything [22] employs a denoising UNet architecture that processes video latents with audio conditioning. MuseTalk [50] proposes a novel sampling strategy that selects reference images with head poses closely matching the target.

However, these methods still rely on reference frames combined with masked-frame inpainting, leading to head pose limitations, identity preservation issues, and blurry edge generation. Our OmniSync framework addresses these limitations through a mask-free training paradigm that enables application across diverse visual representations.

### 2.2 Audio-driven Portrait Animation

Audio-driven portrait animation [38, 46, 15, 6, 14, 27, 37, 49] differs fundamentally from lip sync. Portrait animation [45, 8] follows an image-to-video framework without constraints on head poses or facial expressions, eliminating the need to integrate generated content back into original video. This approach is unsuitable for post-generation lip synchronization in video generation pipelines. In contrast, lip synchronization [33, 17] operates within a video-to-video framework, modifying only lip movements while maintaining compatibility with existing footage. This represents a more constrained task, requiring precise modification of lip regions while preserving surrounding facial features.

Recent models like OmniHuman-1 [18] and Mocha [44] use audio directly as a conditioning signal for image-to-video or text-to-video frameworks. However, due to limitations in talking head datasets, their generative capabilities don't match the versatility of advanced video generation models. This gap highlights why specialized lip synchronization for AI-generated videos remains critical.

## 3 Method

### 3.1 Overview

In this section, we present OmniSync, a universal lip synchronization framework designed for diverse visual content (Fig. 2). Our approach comprises three key components: 1) a mask-free training paradigm that eliminates dependency on reference frames and explicit masks, 2) a flow-matching-based progressive noise initialization strategy for enhanced inference stability, and 3) dynamic spatiotemporal Classifier-Free Guidance (CFG) that optimizes lip sync while preserving facial details. The following subsections provide comprehensive explanations of each component.

### 3.2 Mask-Free Training Paradigm

Traditional lip synchronization methods [33, 50] rely on masked-frame inpainting, isolating the mouth region before generating content based on audio input. Despite their prevalence, these approaches

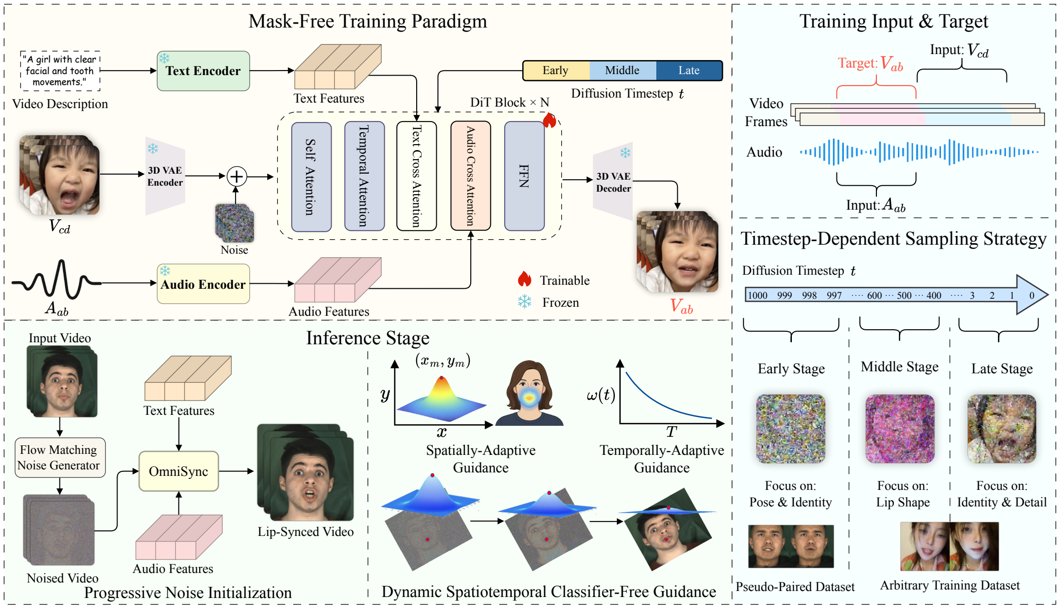


Figure 2: **Overview of OmniSync.** A mask-free training paradigm employs timestep-dependent sampling to predict the lip-synchronized targets $V_{ab}$. During inference, progressive noise initialization and dynamic spatiotemporal CFG ensure consistent head pose and precise lip synchronization.

produce boundary artifacts and struggle with identity preservation. Crucially, they require explicit face detection and alignment—techniques that fail with stylized characters and non-human entities.

An alternative approach is direct frame editing, which aims to transform frames according to target audio without relying on masks or references. However, this approach requires perfectly paired training data with identical head poses and identity—differing only in lip movements. Such paired data is extremely rare and would severely restrict the model's generalizability to diverse visual results.

To address these limitations, we leverage the progressive denoising characteristic of diffusion models, introducing a novel training strategy that varies data sampling based on diffusion timesteps. This allows for stable learning without requiring perfectly paired examples. Our goal is to learn a conditional generation process mapping $(V_{cd}, A_{ab}) \mapsto V_{ab}$ through iterative denoising, where $V$ represents video frames and $A$ represents audio.

We employ Flow Matching [19] as our training objective. Given an input video segment $V_{cd}$ from frames $c$ to $d$, and a target audio segment $A_{ab}$ from frames $a$ to $b$, our model generates the corresponding video segment $V_{ab}$ via the diffusion process:

$$x_{t-1} = \text{DiT}(x_t, V_{cd}, A_{ab}, t), \tag{1}$$

where $x_t$ represents the noised version of target video $V_{ab}$ at timestep $t$, and DiT denotes our diffusion transformer, which predicts the denoised state at timestep $t-1$.

The CFM loss used for training is defined as:

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t,x_t,V_{cd},A_{ab},V_{ab}} \left[ \| v_\theta(x_t, V_{cd}, A_{ab}, t) - u_t(x_t|V_{ab}) \|_2^2 \right], \tag{2}$$

where $v_\theta(x_t, V_{cd}, A_{ab}, t)$ is the learned velocity field predicted by DiT with conditioning on input video $V_{cd}$ and target audio $A_{ab}$, $u_t(x_t|V_{ab})$ is the conditional velocity field typically defined as $u_t(x_t|V_{ab}) = (V_{ab} - x_t)/(1-t)$ for the linear interpolation path $x_t = (1-t)x_0 + tV_{ab}$.

**Timestep-Dependent Sampling Strategy.** A critical insight in our approach is recognizing that the diffusion process can be decomposed into distinct phases, each with different learning requirements. Specifically, early timesteps focus on generating fundamental facial structure, including pose and identity information; middle timesteps primarily generate lip movements driven by audio; while late timesteps refine identity details and textures. To capitalize on this natural progression, we utilize different datasets for distinct timesteps.

For early timesteps (approximately $t \approx T$), responsible for generating overall facial structure, we employ pseudo-paired data from controlled laboratory settings. These samples maintain nearly identical pose information, with variations only in lip movements, providing stable learning signals for structural features and ensuring pose alignment between input and output.

For middle and late timesteps, we transition to more diverse data, sampling from arbitrary videos. During middle timesteps, the model learns to generate lip shapes guided by audio input, whereas in late timesteps (approximately $t \approx 0$), it focuses on refining identity details and ensuring texture consistency. This timestep-dependent training strategy can be formalized as:

$$p(V_{cd}, V_{ab}|t) = \begin{cases} p_{\text{pseudo-paired}}(V_{cd}, V_{ab}) & \text{if } t > t_{\text{threshold}}, \\ p_{\text{arbitrary}}(V_{cd}, V_{ab}) & \text{otherwise.} \end{cases} \tag{3}$$

Here, $p_{\text{pseudo-paired}}$ indicates sampling from controlled datasets with minimal pose variations, while $p_{\text{arbitrary}}$ signifies sampling from our diverse collection of videos. The conditional generation process can be expressed mathematically as:

$$p_\theta(V_{ab}|V_{cd}, A_{ab}) = \int p_\theta(V_{ab}|x_0) p_\theta(x_0|V_{cd}, A_{ab}) dx_0, \tag{4}$$

where $p_\theta(V_{ab}|x_0)$ represents the mapping from the fully denoised state to the output video, and $p_\theta(x_0|V_{cd}, A_{ab})$ captures the relationship between input conditions and the denoised state. Here, $x_0$ refers to the completely denoised latent representation (at timestep $t = 0$).

This progressive training approach aligns well with the natural learning progression of diffusion models. By providing appropriate training signals at each stage, we enable stable learning even without perfectly paired data, allowing our model to generalize effectively to diverse real-world scenarios while maintaining identity consistency.

### 3.3 Progressive Noise Initialization

Standard diffusion-based generation [38] typically begins from random noise (timestep $T$) and progressively denoises toward the final output (timestep 0). However, this approach often results in subtle but noticeable pose misalignments between generated content and original video frames, creating undesirable boundary artifacts and compromising identity preservation.

The fundamental issue lies in error accumulation during the early stages of diffusion. Even minor deviations in early timesteps—when basic facial structure is being formed—can lead to significant misalignments in the final output. This problem is relevant for lip synchronization, where the goal is to modify only speech-relevant regions while maintaining perfect spatial consistency elsewhere. To address this challenge, we introduce a flow-matching-based progressive noise initialization strategy that transforms the traditional diffusion process.

**Flow-Matching Noise Initialization.** Rather than starting the diffusion process from random noise at timestep $T$, we initialize from original video frames with a controlled level of noise. This simulates an intermediate state in the diffusion trajectory, corresponding to a normalized parameter $\tau$. The initialization is performed by adding this controlled noise to the original video frame:

$$x_{\text{init}} = \text{FM}_{\text{add}}(V_{\text{source}}, \tau) = (1 - \tau)V_{\text{source}} + \tau\epsilon, \tag{5}$$

where $x_{\text{init}}$ is the initial noised state derived using the parameter $\tau$, $V_{\text{source}}$ is the source video frame, and $\epsilon \sim \mathcal{N}(0, I)$ is random noise. Let $t_{\text{start}}$ be the discrete timestep corresponding to this initialization point ($T$ is the total number of diffusion steps, and $\tau \in [0, 1]$).

This initialization strategy provides two significant advantages. First, it bypasses the early stages of diffusion (from $T$ down to $t_{\text{start}}$) where general facial structure is formed. This ensures that head pose and global structure are directly inherited from the source frame. Second, it reduces computational requirements by performing denoising only for the remaining steps, from $t_{\text{start}}$ down to 0.

The complete progressive denoising process can be expressed as:

$$x_t = \begin{cases} x_{\text{init}} & \text{if } t = t_{\text{start}}, \\ \text{DiT}(x_{t+1}, V_{\text{source}}, A_{\text{target}}, t+1) & \text{if } t_{\text{start}} > t \geq 0, \end{cases} \tag{6}$$

where $A_{\text{target}}$ is the target audio used to guide the denoising process, and $t$ here represents discrete diffusion timesteps.

This approach effectively creates a two-stage process: (1) initialization using the flow-matching-inspired noise addition (Eq. 5) to reach a state equivalent to timestep $t_{\text{start}}$, and (2) guided denoising from $t_{\text{start}}$ to 0 that focuses on modifying mouth regions according to the target audio while preserving the overall facial structure, identity features, and head pose from the source frame. By skipping the early noisy stages where basic structures form, we maintain spatial consistency while allowing sufficient flexibility for precise mouth region modifications.

### 3.4 Dynamic Spatiotemporal Classifier-Free Guidance

Audio-driven lip synchronization faces a fundamental challenge: audio signals provide relatively weak conditioning compared to visual cues [40]. Standard Classifier-Free Guidance (CFG) [13] can enhance audio conditioning, but applying uniform guidance across spatial and temporal dimensions creates an unavoidable dilemma: higher guidance scales produce more accurate lip sync but introduce texture artifacts, while lower scales preserve visual fidelity but yield less precise lip movements.

To resolve this tension, we introduce Dynamic Spatiotemporal Classifier-Free Guidance (DS-CFG), a novel approach that provides fine-grained control over the generation process across both spatial and temporal dimensions. Our method applies varying guidance strengths to different regions of the frame and different stages of the diffusion process, achieving an optimal balance between lip synchronization accuracy and overall visual quality.

**Spatially-Adaptive Guidance.** The key insight for spatial adaptation is that audio information primarily affects the mouth region, while other facial areas should remain largely unchanged. We implement this through a Gaussian-weighted spatial guidance matrix that concentrates guidance strength around speech-relevant regions:

$$\mathbf{G}_{\text{spatial}}(x, y) = \omega_{\text{base}} + (\omega_{\text{peak}} - \omega_{\text{base}}) \cdot \exp\left(-\frac{(x - x_m)^2 + (y - y_m)^2}{2\sigma^2}\right) \tag{7}$$

where $(x_m, y_m)$ represents the mouth center, $\sigma$ controls the spread of the Gaussian distribution, $\omega_{\text{base}}$ is the baseline guidance strength applied to non-mouth regions, and $\omega_{\text{peak}}$ is the peak strength applied at the mouth center. This spatial adaptation ensures that audio conditions strongly influence lip and surrounding regions while minimally affecting other facial features.

**Temporally-Adaptive Guidance.** We observe that audio conditioning plays different roles at different stages of the diffusion process. In early diffusion timesteps, strong guidance helps establish correct lip shapes, while in later stages, excessive guidance can disrupt fine texture details. [2, 43, 35] To address this, we implement a temporally decreasing guidance schedule:

$$\omega(t) = \omega_{\text{peak}} \cdot \left(\frac{t}{T}\right)^{\gamma} \tag{8}$$

where $t$ is the current diffusion timestep, $T$ is the total number of timesteps, $\omega_{\text{peak}}$ is the maximum guidance scale, and $\gamma$ controls the decay rate, with a value of 1.5. This temporal adaptation ensures strong guidance during early and middle diffusion stages when coarse structures are formed, gradually reducing influence during later stages when fine details and textures are refined.

**Unified Dynamic Spatiotemporal CFG.** Combining both spatial and temporal adaptations, our DS-CFG approach modifies the standard CFG formulation to:

$$\hat{\epsilon}_\theta(x_t, c, t) = \epsilon_\theta(x_t, \emptyset, t) + \mathbf{G}_{\text{spatial}} \cdot \omega(t) \cdot (\epsilon_\theta(x_t, c, t) - \epsilon_\theta(x_t, \emptyset, t)) \tag{9}$$

where $\epsilon_\theta(x_t, c, t)$ and $\epsilon_\theta(x_t, \emptyset, t)$ are the noise predictions with and without conditioning, respectively.

Through this DS-CFG, our method achieves precise control over the generation process, effectively addressing the weak audio signal problem in audio-driven generation.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets.** We trained OmniSync using the MEAD dataset [42] and a 400-hour dataset collected from YouTube. MEAD's controlled laboratory recordings with diverse facial expressions but minimal

Table 1: Quantitative comparison with previous methods on HDTF Dataset.

| **HDTF Dataset** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **Full Reference Metrics** | | | **No Reference Metrics** | | | **Lip Sync** | |
| | FID ↓ | FVD ↓ | CSIM ↑ | NIQE ↓ | BRISQUE ↓ | HyperIQA ↑ | LMD ↓ | LSE-C ↑ |
| Wav2Lip [33] | 14.912 | 543.340 | 0.852 | 6.495 | 53.372 | 45.822 | 10.007 | 7.630 |
| VideoReTalking [7] | 11.868 | 379.518 | 0.786 | 6.333 | 50.722 | 48.476 | 8.848 | 7.180 |
| TalkLip [41] | 16.680 | 691.518 | 0.843 | 6.377 | 52.109 | 44.393 | 15.954 | 5.880 |
| IP-LAP [54] | 9.512 | 325.691 | 0.809 | 6.533 | 54.402 | 50.086 | 7.695 | 7.260 |
| Diff2Lip [25] | 12.079 | 461.341 | 0.869 | 6.261 | 49.361 | 48.869 | 18.986 | 7.140 |
| MuseTalk [50] | 8.759 | 231.418 | 0.862 | 5.824 | 46.003 | 55.397 | 8.701 | 6.890 |
| LatentSync [17] | 8.518 | 216.899 | 0.859 | 6.270 | 50.861 | 53.208 | 17.344 | **8.050** |
| **Ours** | **7.855** | **199.627** | **0.875** | **5.481** | **37.917** | **56.356** | **7.097** | 7.309 |

Table 2: Quantitative comparison with previous methods on AIGC-LipSync Benchmark.

| **AIGC-LipSync Benchmark** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **Full Reference Metrics** | | | **No Reference Metrics** | | | **Generation Success Rate** | |
| | FID ↓ | FVD ↓ | CSIM ↑ | NIQE ↓ | BRISQUE ↓ | HyperIQA ↑ | All Videos ↑ | Stylized Characters ↑ |
| Wav2Lip [33] | 22.989 | 562.245 | 0.727 | 5.392 | 42.816 | 50.511 | 71.38% | 26.67% |
| VideoReTalking [7] | 20.439 | 329.460 | 0.669 | 5.947 | 45.047 | 48.645 | 48.78% | 7.78% |
| TalkLip [41] | 31.180 | 619.179 | 0.754 | 5.239 | 41.692 | 50.608 | 52.36% | 34.44% |
| IP-LAP [54] | 14.686 | 247.402 | 0.796 | 5.546 | 45.153 | 53.174 | 45.53% | 6.67% |
| Diff2Lip [25] | 23.542 | 403.149 | 0.692 | 5.440 | 42.442 | 50.335 | 74.63% | 36.67% |
| MuseTalk [50] | 17.668 | 297.621 | 0.667 | 4.935 | 36.017 | 58.334 | 92.20% | 67.78% |
| LatentSync [17] | 15.374 | 263.111 | 0.751 | 5.342 | 41.917 | 54.648 | 74.96% | 35.56% |
| **Ours** | **10.681** | **211.350** | **0.808** | **4.588** | **25.485** | **61.906** | **97.40%** | **87.78%** |

head movement provided ideal data for training early denoising stages, while the YouTube dataset enhanced generalization across varied real-world conditions for middle and late stages.

To address the limitations of existing benchmarks that focus solely on real-world videos with frontal views and stable lighting, we created the AIGC-LipSync Benchmark. This comprehensive evaluation framework comprises 615 human-centric videos generated by state-of-the-art text-to-video models such as Kling, Dreamina, Wan [39], and Hunyuan [16]. The benchmark specifically captures challenging visual scenarios such as large facial movements, profile views, variable lighting, occlusions, and stylized characters—conditions that traditional benchmarks fail to address. Details about benchmark construction can be found in the supplementary materials.

**Implementation Details.** We implement our OmniSync model using the Diffusion Transformer architecture. The model is trained on a combined dataset for 80,000 steps using AdamW optimizer [21] with a learning rate of 1e-5. Training is completed in 80 hours using 64 NVIDIA A100 GPUs with a batch size of 64. Audio features are extracted via a pre-trained Whisper encoder, and text conditioning utilizes a T5 encoder. Training employs the timestep-dependent sampling threshold $t_{\text{threshold}} = 850$. The experimental results indicate that excessive thresholds induce significant misalignment while insufficient values will leak the original lip shape. During inference we adopt our flow-matching-based progressive noise initialization starting at $\tau = 0.92$, followed by 50 denoising steps.

### 4.2 Quantitative Evaluation

We evaluate OmniSync against state-of-the-art methods including Wav2Lip [33], VideoReTalking [7], TalkLip [41], IP-LAP [54], Diff2Lip [25], MuseTalk [50], and LatentSync [17] using a comprehensive suite of metrics. For visual quality assessment, we employ FID (Fréchet Inception Distance) to measure frame-level fidelity, FVD (Fréchet Video Distance) for temporal consistency, and CSIM (Cosine Similarity) to quantify identity preservation. Perceptual quality is assessed using no-reference metrics including NIQE (Natural Image Quality Evaluator), BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator), and HyperIQA [36]. For audio-visual synchronization, we measure LMD (Landmark Distance) between predicted and ground truth facial landmarks in the mouth region, and LSE-C (Lip Sync Error - Confidence) to evaluate lip synchronization quality.

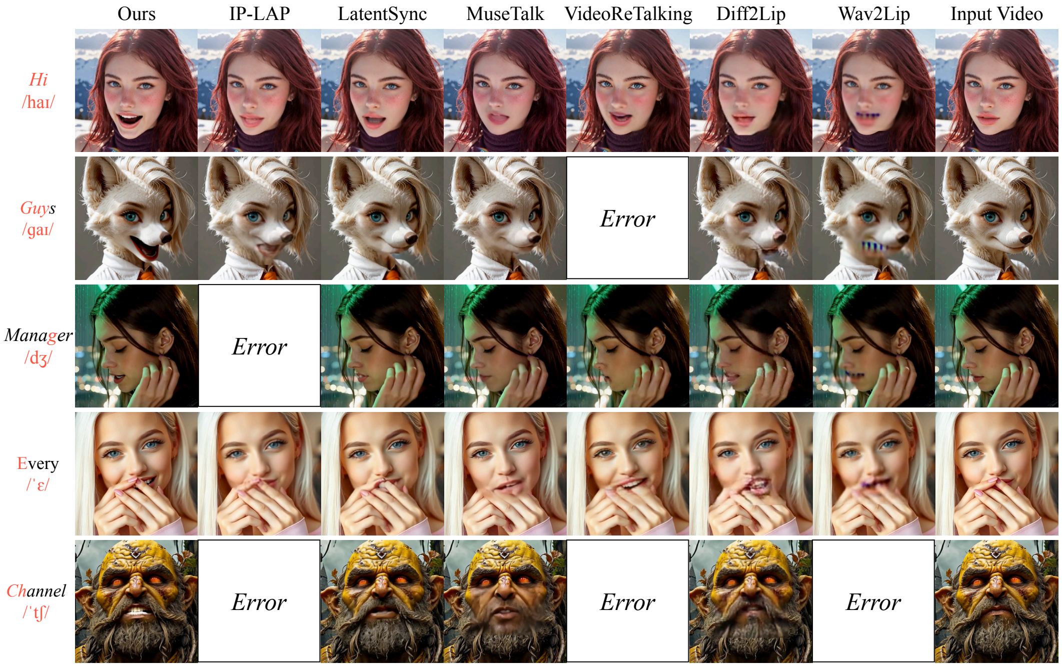


Figure 3: **Qualitative comparison** with previous methods across diverse subjects and phonemes. Our approach produces more accurate lip synchronization and better identity preservation.

Table 3: **User study** results comparing various audio-driven lip-sync methods.

| **Metric** | **Wav2Lip** | **Video ReTalking** | **TalkLip** | **IP-LAP** | **Diff2Lip** | **MuseTalk** | **Latent Sync** | **Ours** |
|---|---|---|---|---|---|---|---|---|
| Lip Sync Accuracy | 2.684 | 2.769 | 1.940 | 2.359 | 3.000 | 2.632 | 3.812 | **3.923** |
| Identity Preservation | 2.410 | 2.786 | 2.222 | 2.991 | 2.889 | 3.197 | 3.658 | **4.128** |
| Timing Stability | 2.556 | 2.376 | 2.162 | 3.034 | 2.812 | 3.145 | 3.581 | **4.043** |
| Image Quality | 2.017 | 2.607 | 1.889 | 3.171 | 2.402 | 3.094 | 3.632 | **4.051** |
| Video Realism | 2.120 | 2.419 | 1.838 | 2.316 | 2.479 | 2.761 | 3.453 | **3.872** |

For the AIGC-LipSync benchmark, we report the Generation Success Rate across all 615 videos and specifically for stylized characters. This metric shows the percentage of videos that are successfully synchronized and pass human verification. This evaluation is essential for universal lip synchronization in AI-generated content, where traditional metrics may not fully capture the challenges of stylized characters, extreme poses, and other atypical visual conditions.

The experimental results in Tab. 1 and Tab. 2 demonstrate that our approach achieves superior performance on multiple metrics. On the HDTF dataset, our method reduced FID by 7.8% and FVD by 8.0% compared to LatentSync, with a remarkable 23.2% improvement in BRISQUE over Diff2Lip. For lip synchronization, we achieved the lowest LMD, outperforming IP-LAP by 7.8%, while LatentSync maintained a slight edge in LSE-C due to its SyncNet-based loss constraint.

On the challenging AIGC-LipSync benchmark, OmniSync demonstrated exceptional capabilities with a 30.5% FID reduction and 19.7% FVD reduction compared to LatentSync, alongside improved identity preservation. Most significantly, our method achieved a 97.40% Generation Success Rate across all videos—substantially higher than MuseTalk (92.20%) and other methods (below 75%). For stylized characters, our success rate of 87.78% outperformed MuseTalk (67.78%), demonstrating OmniSync's capability to handle diverse visual representations including stylized characters.

### 4.3 Qualitative Evaluation

We present qualitative comparisons between OmniSync and existing methods in Fig. 3. Our approach produces more natural facial expressions and superior lip synchronization. Due to lip shape leakage, IP-LAP [54] and LatentSync [17] frequently fail at mouth shape modification, resulting in poor lip synchronization effects. MuseTalk [50] and VideoReTalking [7] modify lip movements but

Table 4: **Ablation study** for our method.

| Methods | FID ↓ | FVD ↓ | CSIM ↑ | NIQE ↓ | BRISQUE ↓ | HyperIQA ↑ | LSE-C ↑ |
|---|---|---|---|---|---|---|---|
| Ours | **15.710** | **287.168** | **0.814** | **5.321** | **29.588** | **57.288** | 7.06 |
| w/o Timestep-Dependent Sampling Strategy | 21.552 | 549.768 | 0.727 | 5.462 | 30.346 | 56.204 | 7.00 |
| w/o Progressive Noise Initialization | 16.731 | 361.282 | 0.805 | 5.349 | 29.789 | 56.511 | 7.03 |
| w/ Low Static CFG | - | - | - | 5.359 | 29.724 | 56.568 | 4.16 |
| w/ High Static CFG | 22.725 | 348.335 | 0.782 | 5.473 | 29.678 | 56.289 | **7.10** |

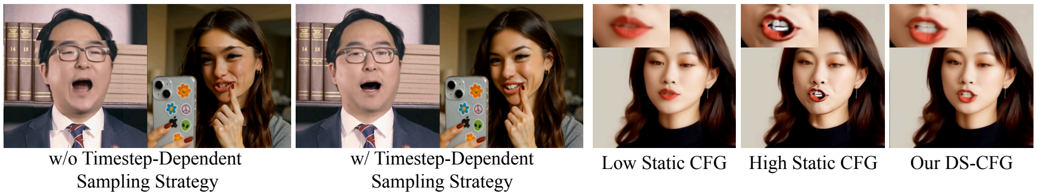

w/o Timestep-Dependent Sampling Strategy
w/ Timestep-Dependent Sampling Strategy
Low Static CFG
High Static CFG
Our DS-CFG

Figure 4: **Ablation study** for timestep-dependent sampling strategy and different CFG settings.

frequently lose identity and visual quality. Diff2Lip [25] and Wav2Lip [33] commonly exhibit lip sync errors, mouth artifacts, and identity drift, particularly in challenging or stylized cases. In contrast, OmniSync consistently maintains identity details and generates realistic, expressive lip movements, demonstrating robust performance. Our approach effectively balances audio and visual cues, addressing the challenge of weak audio conditioning.

**User Study.** To assess perceptual quality, we conducted a user study with 39 participants evaluating 32 video sets generated by OmniSync and seven competing methods, with a standardized Cronbach's $\alpha$ coefficient of 0.98. Participants rated each video on a 5-point Likert scale across five criteria: Lip Sync Accuracy, Character Identity preservation, Timing Stability, Image Quality, and Video Realism. As shown in Tab. 3, OmniSync outperformed all competitors across all metrics, achieving superior scores in Lip Sync Accuracy (3.923 vs. 3.812 for LatentSync), Character Identity (4.128 vs. 3.658), Timing Stability (4.043 vs. 3.581), Image Quality (4.051 vs. 3.632), and Video Realism (3.872 vs. 3.453). These results confirm OmniSync's superior ability to generate high-quality talking videos.

### 4.4 Ablation Study

To clarify the contributions of each core component in our framework, we conduct an ablation study targeting three key modules: the timestep-dependent data sampling strategy, progressive noise initialization, and the Dynamic Spatiotemporal Classifier-Free Guidance (DS-CFG) mechanism. Quantitative results are presented in Tab. 4, and corresponding visual examples are shown in Fig. 4.

Removing the timestep-dependent sampling strategy results in a significant drop in identity preservation and pose consistency, with a 10.7% decrease in CSIM and substantial increases in FID and FVD. As shown in Fig. 4, without this sampling strategy, the generated faces often exhibit clear mismatches with the original image, including noticeable facial misalignment issues. This validates our design choice of aligning pseudo-paired data with early diffusion steps, which proves critical for generating structurally stable outputs. Similarly, removing progressive noise initialization leads to evident temporal inconsistencies and an increase in FVD, confirming the importance of our flow-matching initialization in preserving spatial anchoring and motion coherence.

We also compare our proposed DS-CFG with both low and high static CFG settings. As illustrated in Fig. 4, low CFG provides insufficient audio conditioning, resulting in under-articulated lip movements (LSE-C: 4.16), whereas high CFG improves synchronization (LSE-C: 7.10) but introduces noticeable artifacts and distortions in facial details. In contrast, DS-CFG achieves an optimal balance by applying strong localized guidance in early diffusion stages and gradually reducing it in later steps. These results confirm that dynamic control across temporal and spatial dimensions is essential for producing expressive and visually coherent lip synchronization in generative video content.

## 5 Conclusion

In this paper, we introduce OmniSync, a universal lip synchronization framework for diverse content that addresses critical limitations of traditional approaches. Our three key innovations—a mask-free training paradigm eliminating mask dependencies, a flow-matching-based progressive noise initialization strategy ensuring identity preservation, and dynamic spatiotemporal Classifier-Free Guidance balancing synchronization with visual quality—collectively enable precise lip movements across diverse visual representations. To support systematic evaluation in this field, we establish the AIGC-LipSync Benchmark, the first comprehensive framework for assessing lip synchronization in varied AIGC contexts. Extensive experiments demonstrate OmniSync's superior performance across challenging scenarios, establishing a robust foundation for integrating precise lip synchronization into the broader AI video generation ecosystem.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (NSFC) under Grant Nos. 62436010, 72572090, 62172421 and 62572474. This work was also supported by the Outstanding Innovative Talents Cultivation Funded Programs 2023 of Renmin University of China.

## References

[1] Antoni Bigata, Rodrigo Mira, Stella Bounareli, Michał Stypułkowski, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Keysync: A robust approach for leakage-free lip synchronization in high resolution. *arXiv preprint arXiv:2505.00497*, 2025.

[2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017.

[4] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[5] Hejia Chen, Haoxian Zhang, Shoulong Zhang, Xiaoqiang Liu, Sisi Zhuang, Yuan Zhang, Pengfei Wan, Di Zhang, and Shuai Li. Cafe-talk: Generating 3d talking face animation with multimodal coarse-and fine-grained control. *arXiv preprint arXiv:2503.14517*, 2025.

[6] Zhiyuan Chen, Jiajiong Cao, Zhiquan Chen, Yuming Li, and Chenguang Ma. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2403–2410, 2025.

[7] 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*, pages 1–9, 2022.

[8] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[10] Jiazhi Guan, Zhiliang Xu, Hang Zhou, Kaisiyuan Wang, Shengyi He, Zhanwang Zhang, Borong Liang, Haocheng Feng, Errui Ding, Jingtuo Liu, et al. Resyncer: Rewiring style-based generator for unified audio-visually synced facial performer. In *European Conference on Computer Vision*, pages 348–367. Springer, 2024.

[11] 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*, pages 1505–1515, 2023.

[12] Anchit Gupta, Rudrabha Mukhopadhyay, Sindhu Balachandra, Faizan Farooq Khan, Vinay P Namboodiri, and CV Jawahar. Towards generating ultra-high resolution talking-face videos with lip synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5209–5218, 2023.

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[14] Xiaozhong Ji, Xiaobin Hu, Zhihong Xu, Junwei Zhu, Chuming Lin, Qingdong He, Jiangning Zhang, Donghao Luo, Yi Chen, Qin Lin, et al. Sonic: Shifting focus to global audio perception in portrait animation. *arXiv preprint arXiv:2411.16331*, 2024.

[15] Jianwen Jiang, Chao Liang, Jiaqi Yang, Gaojie Lin, Tianyun Zhong, and Yanbo Zheng. Loopy: Taming audio-driven portrait avatar with long-term motion dependency. *arXiv preprint arXiv:2409.02634*, 2024.

[16] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[17] Chunyu Li, Chao Zhang, Weikai Xu, Jinghui Xie, Weiguo Feng, Bingyue Peng, and Weiwei Xing. Latentsync: Audio conditioned latent diffusion models for lip sync. *arXiv preprint arXiv:2412.09262*, 2024.

[18] Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. Omnihuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. *arXiv preprint arXiv:2502.01061*, 2025.

[19] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[20] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024.

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[22] Junxian Ma, Shiwen Wang, Jian Yang, Junyi Hu, Jian Liang, Guosheng Lin, Kai Li, Yu Meng, et al. Sayanything: Audio-driven lip synchronization with conditional video diffusion. *arXiv preprint arXiv:2502.11515*, 2025.

[23] Yifeng Ma, Suzhen Wang, Zhipeng Hu, Changjie Fan, Tangjie Lv, Yu Ding, Zhidong Deng, and Xin Yu. Styletalk: One-shot talking head generation with controllable speaking styles. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 1896–1904, 2023.

[24] Ming Meng, Yufei Zhao, Bo Zhang, Yonggui Zhu, Weimin Shi, Maxwell Wen, and Zhaoxin Fan. A comprehensive taxonomy and analysis of talking head synthesis: Techniques for portrait generation, driving mechanisms, and editing. *arXiv preprint arXiv:2406.10553*, 2024.

[25] Soumik Mukhopadhyay, Saksham Suri, Ravi Teja Gadde, and Abhinav Shrivastava. Diff2lip: Audio conditioned diffusion models for lip-synchronization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5292–5302, 2024.

[26] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[27] Ziqiao Peng, Yi Chen, Yifeng Ma, Guozhen Zhang, Zhiyao Sun, Zixiang Zhou, Youliang Zhang, Zhengguang Zhou, Zhaoxin Fan, Hongyan Liu, et al. Actavatar: Temporally-aware precise action control for talking avatars. *arXiv preprint arXiv:2512.19546*, 2025.

[28] Ziqiao Peng, Yanbo Fan, Haoyu Wu, Xuan Wang, Hongyan Liu, Jun He, and Zhaoxin Fan. Dualtalk: Dual-speaker interaction for 3d talking head conversations. *arXiv preprint arXiv:2505.18096*, 2025.

[29] Ziqiao Peng, Wentao Hu, Junyuan Ma, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Hui Tian, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk++: High-fidelity and efficient synchronized talking heads synthesis using gaussian splatting. *arXiv preprint arXiv:2506.14742*, 2025.

[30] Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 666–676, 2024.

[31] Ziqiao Peng, Yihao Luo, Yue Shi, Hao Xu, Xiangyu Zhu, Hongyan Liu, Jun He, and Zhaoxin Fan. Selftalk: A self-supervised commutative training diagram to comprehend 3d talking faces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5292–5301, 2023.

[32] Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, and Zhaoxin Fan. Emotalk: Speech-driven emotional disentanglement for 3d face animation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 20687–20697, 2023.

[33] KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492, 2020.

[34] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[35] Michał Stypułkowski, Konstantinos Vougioukas, Sen He, Maciej Zięba, Stavros Petridis, and Maja Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5091–5100, 2024.

[36] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3667–3676, 2020.

[37] Zhiyao Sun, Ziqiao Peng, Yifeng Ma, Yi Chen, Zhengguang Zhou, Zixiang Zhou, Guozhen Zhang, Youliang Zhang, Yuan Zhou, Qinglin Lu, et al. Streamavatar: Streaming diffusion models for real-time interactive human avatars. *arXiv preprint arXiv:2512.22065*, 2025.

[38] Linrui Tian, Qi Wang, Bang Zhang, and Liefeng Bo. Emo: Emote portrait alive generating expressive portrait videos with audio2video diffusion model under weak conditions. In *European Conference on Computer Vision*, pages 244–260. Springer, 2024.

[39] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[40] Cong Wang, Kuan Tian, Jun Zhang, Yonghang Guan, Feng Luo, Fei Shen, Zhiwei Jiang, Qing Gu, Xiao Han, and Wei Yang. V-express: Conditional dropout for progressive training of portrait video generation. *arXiv preprint arXiv:2406.02511*, 2024.

[41] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14653–14662, 2023.

[42] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *European conference on computer vision*, pages 700–717. Springer, 2020.

[43] Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernández Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *arXiv preprint arXiv:2404.13040*, 2024.

[44] Cong Wei, Bo Sun, Haoyu Ma, Ji Hou, Felix Juefei-Xu, Zecheng He, Xiaoliang Dai, Luxin Zhang, Kunpeng Li, Tingbo Hou, et al. Mocha: Towards movie-grade talking character synthesis. *arXiv preprint arXiv:2503.23307*, 2025.

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

[46] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. *arXiv preprint arXiv:2406.08801*, 2024.

[47] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[48] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[49] Guozhen Zhang, Zixiang Zhou, Teng Hu, Ziqiao Peng, Youliang Zhang, Yi Chen, Yuan Zhou, Qinglin Lu, and Limin Wang. Uniavgen: Unified audio and video generation with asymmetric cross-modal interactions. *arXiv preprint arXiv:2511.03334*, 2025.

[50] Yue Zhang, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. Musetalk: Real-time high quality lip synchronization with latent space inpainting. *arXiv preprint arXiv:2410.10122*, 2024.

[51] 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, pages 3543–3551, 2023.

[52] Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3661–3670, 2021.

[53] Zongzheng Zhang, Jiawen Yang, Ziqiao Peng, Meng Yang, Jianzhu Ma, Lin Cheng, Huazhe Xu, Hang Zhao, and Hao Zhao. Morpheus: A neural-driven animatronic face with hybrid actuation and diverse emotion control. *arXiv preprint arXiv:2507.16645*, 2025.

[54] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2023.

[55] Xukun Zhou, Fengxin Li, Ziqiao Peng, Kejian Wu, Jun He, Biao Qin, Zhaoxin Fan, and Hongyan Liu. Meta-learning empowered meta-face: Personalized speaking style adaptation for audio-driven 3d talking face animation. *arXiv preprint arXiv:2408.09357*, 2024.

# NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction fully and accurately reflect the scope and contributions of the paper, including the development of a reference-free diffusion transformer framework (OmniSync), a flow-matching-based noise initialization strategy, a dynamic spatiotemporal CFG method, and the introduction of the AIGC-LipSync Benchmark. All claims are well-supported by theoretical exposition and extensive experimental results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The supplementary materials discuss the limitations of our method. We highlight, for example, that while our framework shows strong generalization on diverse AIGC scenarios, performance on certain rare, highly stylized, or non-human characters may still pose challenges.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper is focused on algorithmic and applied contributions; no formal theorems or proofs are required.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details necessary to reproduce the main experimental results are included in the Sec. 4.1, such as model architecture, dataset details, evaluation metrics, training schedules, hardware, and hyperparameters.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Due to ongoing anonymization and preparation, the benchmark will be made publicly available upon publication.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all experimental settings in the Sec. 4.1 and 4.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: This paper reports the statistical significance of user study .

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources (GPU type and count, training time, batch size) required for the experiments are detailed in the Sec. 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The broader impacts of our work are discussed in the paper. OmniSync enables more robust and flexible lip synchronization in AI-generated video, benefiting content creation, accessibility, and virtual communication. We acknowledge the potential risk of misuse (e.g., deepfakes) and recommend responsible deployment and further safeguards.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: No high-risk models or datasets are publicly released at submission time.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external datasets and code used in this work are properly cited and credited, and are used in accordance with their respective licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new AIGC-LipSync Benchmark is thoroughly described in the paper. Upon publication, documentation and access instructions will be provided to the community.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We described the details of the user study in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: No IRB review was required.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.