FROM INPAINTING TO EDITING: A SELF-BOOTSTRAPPING PARADIGM FOR CONTEXT-RICH VISUAL DUBBING

Anonymous authors
Paper under double-blind review

Figure 1: Moving beyond mask-inpainting, **X-Dub** redefines visual dubbing as context-rich, full-reference video-to-video editing, which yields precise lip sync and faithful identity preservation, even in challenging scenarios with occlusions and dynamic lighting.

ABSTRACT

Audio-driven visual dubbing aims to synchronize a video's lip movements with new speech, but is fundamentally challenged by the lack of real-world paired training data. Existing methods circumvent this with a mask-based inpainting paradigm, where incomplete context forces models to simultaneously hallucinate missing content (e.g., occlusions) and sync lips, leading to visual artifacts, identity drift, and poor synchronization. In this work, we propose a novel self-bootstrapping paradigm that reframes visual dubbing from an under-specified inpainting task into a well-conditioned video-to-video editing problem. Our approach utilizes a Diffusion Transformer to first generate its own ideal training data: a lip-altered companion video for each sample, forming a context-rich pair with the original. An editor is then trained on these pairs, leveraging the complete and aligned video context to focus solely on precise, audio-driven lip modifications. This context-rich conditioning allows our method to achieve state-of-the-art performance, yielding highly accurate lip sync, faithful identity preservation, and exceptional robustness against challenging in-the-wild scenarios like occlusions and dynamic lighting. We further introduce a timestep-adaptive multi-phase learning strategy that aligns diffusion stages with visual hierarchies, significantly enhancing contextual learning and dubbing quality. Additionally, we propose ContextDubBench, a comprehensive benchmark dataset for robust evaluation in diverse and challenging practical application scenarios. Our visualizations are available at the anonymous page x-dub-lab.github.io, and code will be released to benefit the community.

1 INTRODUCTION

Audio-driven visual dubbing edits pre-recorded talking-head videos to synchronize lip movements with new speech (KR et al., 2019). Unlike audio-driven animation (Tian et al., 2024; Cui et al., 2024a), which generates entire videos from scratch, dubbing is fundamentally an *editing* task: it modifies only the speech-relevant regions while preserving identity and other visual cues from the original video. This uniqueness underpins dubbing's broad applications, from personalized avatars (Thies et al., 2020) to multilingual film translation (Prajwal et al., 2020). Meanwhile, the recent rise of Diffusion Transformers (DiTs) (Peebles & Xie, 2023) has demonstrated their remarkable contextual generation capability in text-to-video (T2V) and image-to-video (I2V) tasks, making them natural candidates for high-fidelity visual dubbing. Yet dubbing presents a distinctive challenge. As a video-to-video task, it requires paired training data where lip movements differ while identity, pose, and environment remain unchanged, which is virtually unattainable in the real world.

To bypass the lack of paired data, existing approaches adopt a self-reconstruction paradigm: they mask the mouth region in a video and train a model to inpaint it conditioned on the corresponding audio and sparse reference frames (Prajwal et al., 2020). However, this yields an under-specified and mismatched context. The model must not only precisely modify lip movements for synchronization, but also hallucinate missing visual information (e.g., facial occlusions) while extracting identity features from reference frames that often exhibit misaligned poses or inconsistent scenes. This divided attention increases learning difficulty and thus commonly leads to lip-sync degradation, which is also exacerbated by mask-boundary lip-shape leakage and weak audio conditioning (Wang et al., 2023; Bigata et al., 2025; Chen et al., 2025). Meanwhile, this dual burden typically induces identity drift and visual artifacts (Zhong et al., 2023; Peng et al., 2025). Ultimately, these failures reveal a fundamental disconnect: the training paradigm forces the model to work with incomplete context, preventing it from leveraging the complete video context yet available during inference. In contrast, DiT excels at I2V animation precisely because training and inference share identical contextual settings with naturally paired data. Thus, we argue that the bottleneck in dubbing lies not in the model architecture, but in the training paradigm's failure to provide suitable contextual data.

We therefore introduce a new perspective: Instead of waiting for ideal data pairs, we let the model generate them for itself. To this end, we present **X-Dub**, a framework built upon a novel **self-bootstrapping paradigm for context-rich dubbing**. Here, a powerful DiT backbone both produces and benefits from its own generated, context-rich video pairs: a DiT-based *generator* first creates a lip-altered video for each training sample. This synthetic-original pair provides ideal training data with consistent pose and identity, varying solely in lip movements. A subsequent DiT-based *editor* then learns dubbing directly from these paired videos, leveraging the rich context they provide. In this way, we transform dubbing from an under-specified inpainting task into a well-conditioned editing problem guided by complete and aligned contextual input.

Concretely, the *generator* is trained with a reconstruction objective on large-scale audiovisual data, adopting a conventional masking setup. It functions not as a direct dubbing solver but as a contextual condition synthesizer, whose outputs serve as conditional inputs rather than supervision targets to prevent artifact learning. This design allows us to optimize for identity preservation and spatiotemporal robustness by sacrificing secondary factors like lip-sync accuracy and generalizability, creating a more reliable contextual condition provider. These paired conditions, though imperfect, provide substantially richer contextual information than raw single reference frames. Consequently, the *editor* faces a significantly simplified task: it can dedicate its full capacity to precise speech-driven lip editing while seamlessly inheriting identity and visual details from the contextual input, ultimately achieving superior lip-sync accuracy, identity preservation, and robustness to visual variations.

This paradigm reframes visual dubbing as targeted editing, i.e., precisely modifying lip movements while preserving global layouts and fine-grained textures for visual consistency. This naturally benefits from progressive learning with specialized supervision for different information types, rather than a monolithic approach. We therefore introduce a **timestep-adaptive multi-phase learning** strategy with LoRA experts, leveraging the inherent tendency of diffusion models to capture distinct information levels at different timesteps (Zhang et al., 2025; Wang et al., 2025). By aligning early, mid, and late diffusion stages with global structure, lip shape, and texture refinement respectively, the model learns complementary objectives at their most effective phases, thereby strengthening lip sync while better maintaining visual consistency.

Finally, we note that existing visual dubbing benchmarks (Afouras et al., 2018; Zhang et al., 2021) are confined to controlled settings with limited motion and diversity, insufficient for evaluating robustness in realistic scenarios. We therefore introduce **ContextDubBench**, a benchmark built from both real-world footage and advanced generative content, encompassing varied motions, environments, styles, and subjects to enable comprehensive evaluation under complex dubbing conditions.

In summary, our contributions are: 1) We propose a **self-bootstrapping dubbing paradigm** that leverages DiT both as a generator of *context-rich* paired input and as an editor trained on them, transforming dubbing from an under-specified inpainting task into a well-conditioned video-to-video editing problem. 2) We propose a **timestep-adaptive multi-phase learning** strategy that disentangles visual information learning across diffusion timesteps, facilitating more effective contextual learning and yielding enhanced lip-sync quality and visual coherence. 3) We construct and release **ContextDubBench**, a benchmark for evaluating dubbing models in complex real-world and generative scenarios. 4) Extensive experiments demonstrate that our method achieves remarkable improvements across all metrics, significantly outperforming existing approaches with more accurate lip sync, superior identity preservation, and exceptional robustness to spatiotemporal variations.

2 Related Work

Visual dubbing. Early visual dubbing methods leverage GANs (Goodfellow et al., 2014) for mask-based inpainting. LipGAN (KR et al., 2019) pioneers this direction with reference-guided synthesis, while Wav2Lip (Prajwal et al., 2020) improves lip sync through SyncNet (Chung & Zisserman, 2016). Subsequent works extend this paradigm: VideoReTalking (Cheng et al., 2022) introduces canonical references to mitigate expression bias, DINet (Zhang et al., 2023) enables high-resolution synthesis via deformation inpainting, and TalkLip (Wang et al., 2023) enhances lip intelligibility using AV-HuBERT (Shi et al., 2022). IP-LAP (Zhong et al., 2023) and StyleSync (Guan et al., 2023) further strengthen identity preservation through landmark- and style-aware optimization.

Recent diffusion-based approaches also exhibit advanced performance. DiffTalk (Shen et al., 2023) and Diff2Lip (Mukhopadhyay et al., 2024) demonstrate the feasibility of diffusion, while MuseTalk (Zhang et al., 2024) achieves real-time synthesis by combining latent diffusion with adversarial training. LatentSync (Li et al., 2024) adapts pre-trained diffusion models with temporal supervision to improve stability. Nevertheless, these methods largely follow a self-reconstruction paradigm based on masked frames and sparse references, which limits contextual richness. By contrast, our approach introduces a *contextual conditioning paradigm*, where paired videos provide informative context, allowing the model to focus on accurate lip editing with stronger stability.

Audio-driven portrait animation. Another related line of work is audio-driven portrait animation, which generates talking videos from still images or text prompts. Recent DiT-based models achieve expressive talking-head (Tian et al., 2024; Cui et al., 2024a), half-body (Cui et al., 2024b; Meng et al., 2025), and full-body results (Wang et al., 2025; Lin et al., 2025). These works demonstrate the power of DiTs for human-centric generation in I2V or T2V paradigms. Visual dubbing instead is a stricter video-to-video editing task: it requires precise speech-driven modifications while preserving other visual cues, enabling seamless integration into recorded videos.

3 Our Approach

As illustrated in Fig. 2, we establish a self-bootstrapping dubbing framework where a DiT model both generates contextual video pairs and learns dubbing from them, thereby reframing dubbing from an under-specified inpainting problem into a well-conditioned video-to-video editing task. We first present the DiT-based *generator*, trained with a mask-based self-reconstruction objective to synthesize lip-varied companion videos that serve purely as contextual inputs (Sec. 3.1). To ensure these pairs provide stable and reliable conditions, we introduce principled construction strategies that prioritize identity preservation and robustness over secondary lip accuracy and generalization (Sec. 3.1.2). On top of such curated contextual pairs, the DiT-based *editor* learns mask-free dubbing as context-driven editing, achieving accurate lip sync, faithful identity retention, and resilience to pose and occlusion variations (Sec. 3.2). Finally, we propose a timestep-adaptive multi-phase learning scheme (Sec. 3.3) that aligns diffusion stages with complementary objectives—structure, lips, and textures—thereby amplifying contextual learning and further enhancing dubbing quality.

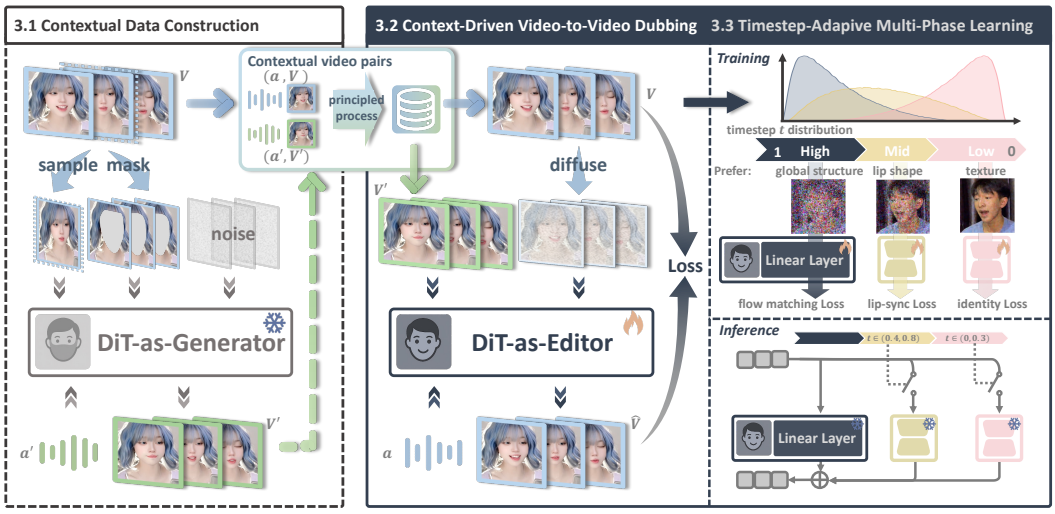

Figure 2: **Overview of X-Dub, our self-bootstrapping dubbing framework.** At its core, our paradigm employs a DiT *generator* to create a lip-altered counterpart for each video, forming a context-rich pair with the original (left). A DiT *editor* then learns mask-free, video-to-video dubbing directly from these ideal pairs, leveraging the complete visual context to ensure accurate lip sync and identity preservation (middle). This contextual learning is further refined by our timestep-adaptive multi-phase learning (right), which aligns different diffusion stages with learning distinct information: global structure, lip movements, and texture details, respectively.

DiT backbone. Our DiT backbone follows the latent diffusion paradigm with a 3D VAE for video compression and a DiT for token sequence modeling (Peebles & Xie, 2023). Each DiT block combines 2D spatial and 3D spatio-temporal self-attention with cross-attention for external conditions.

3.1 Generator: Contextual Condition Constructor

3.1.1 Naïve Mask Dubbing

We implement the DiT-based *generator* under a mask-based self-reconstruction scheme following prior dubbing methods. Given a target video $\boldsymbol{V}_{\text{tgt}}$ with audio $\boldsymbol{a}_{\text{tgt}}$, we apply a facial mask $\boldsymbol{M}$ and reconstruct masked regions $\hat{\boldsymbol{V}}_{\text{tgt}}$ conditioned on $\boldsymbol{a}_{\text{tgt}}$ and a reference frame I_{ref}.

Although this setup yields imperfect dubbing outputs, the *generator* is not designed to solve dubbing directly, but solely to synthesize companion videos as contextual inputs for the *editor* in our paradigm. By embedding lip variations into otherwise consistent frames, the *generator* transforms sparse inpainting contexts into aligned video pairs far stronger than static reference frames.

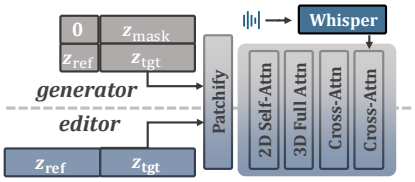

Figure 3: **Conditioning mechanisms for our DiT backbone**. Reference conditions (contextual video for *editor*; single frame for *generator*) and the target video are concatenated into a unified sequence for 3D self-attention. Audio features are injected via cross-attention.

Conditioning mechanisms. As shown in Fig. 3, masked and target frames are encoded by a VAE into $\boldsymbol{z}_{\text{mask}}, \boldsymbol{z}_{\text{tgt}} \in \mathbb{R}^{b\times f\times c\times h\times w}$, and the reference frame into $\boldsymbol{z}_{\text{ref}} \in \mathbb{R}^{b\times c\times h\times w}$. We concatenate $\boldsymbol{z}_{\text{mask}}$ with noised $\boldsymbol{z}_{\text{tgt}}$ channel-wise, and zero-pad $\boldsymbol{z}_{\text{ref}}$ for channel alignment. Concatenating across frames yields the unified DiT input $\boldsymbol{z}_{\text{in}} = \big[\,[\boldsymbol{z}_{\text{mask}}, \boldsymbol{z}_{\text{tgt}}]_{\text{ch}},\ [\boldsymbol{0}, \boldsymbol{z}_{\text{ref}}]_{\text{ch}}\,\big]_{\text{ft}}$, which enables interaction between video and reference tokens via 3D self-attention. Whisper (Radford et al., 2021b) features are injected through cross-attention as audio condition. To extend generation to long videos, we use motion frames (Tian et al., 2024): each segment is conditioned on the last frames of the previous one. During training, the first $m = 2$ frames of $\boldsymbol{z}_{\text{tgt}}$ remain unnoised as motion guidance. Conditional dropout (50%) handles the absence of prior frames in initial segments.

Training objective. We adopt a flow-matching loss $\mathcal{L}_{\text{FM}}$ (details in Sec. B.1), weighted by face and lip masks $\boldsymbol{M}, \boldsymbol{M}_{\text{lip}}$ from DWPose (Yang et al., 2023) via element-wise multiplication ($\odot$):

$$\mathcal{L}_{\text{wFM}} = (1 + \lambda \boldsymbol{M} + \lambda_{\text{lip}} \boldsymbol{M}_{\text{lip}}) \odot \mathcal{L}_{\text{FM}}. \tag{1}$$

Trained in this manner, the *generator* produces a synthetic companion video $\boldsymbol{V}'$ for each real clip $\boldsymbol{V}$ by replacing its original audio $\boldsymbol{a}$ with an alternative $\boldsymbol{a}'$, yielding contextual pairs $(\boldsymbol{V}', \boldsymbol{V})$. Here, $\boldsymbol{V}'$ serves solely as the conditional input for the *editor*.

3.1.2 Principled Pair Construction Strategies

Plain mask-based dubbing inevitably yields imperfect results. We therefore design explicit trade-off strategies to ensure that synthetic companions, while not flawless, provide reliable contextual inputs, turning the *generator* transform from a Naïve dubbber to a reliable contextual condition provider.

To this end, we establish three guiding principles: **1) In-domain quality over generalization.** The *generator* should focus on fidelity within the training distribution rather than broad generalization. **2) Visual consistency under variation.** Companion videos must maintain identity and remain robust to pose, occlusion, and illumination changes. **3) Lip variation over accuracy.** Lip shapes in $\boldsymbol{V}'$ should differ from $\boldsymbol{V}$ to avoid leakage, while tolerating moderate lip-sync inaccuracies.

Accordingly, we implement several strategies. First, we leverage **short-term visual stationarity** by processing videos in brief segments where pose and scene remain relatively stable. Intra-segment reference frames at inference then provide roughly aligned visual context, with motion frames connecting segments into complete videos. While lip accuracy may degrade across segment boundaries, this trade-off favors visual consistency as intended. We also sample alternative audio $\boldsymbol{a}'$ from the same speaker as $\boldsymbol{V}$ to reduce cross-identity conflicts and apply **extended training** beyond nominal convergence to improve identity preservation and lip sync.

To enhance robustness to diverse variations, we incorporate complementary techniques. We handle **occlusions** by annotating and excluding facial occluders from inpainting areas, enabling the generator to be more robust to occlusion scenarios. For **lighting augmentation**, we apply identical relighting to both $\boldsymbol{V}$ and $\boldsymbol{V}'$ in uniformly-lit videos to construct pairs with consistent lighting dynamics. We also perform **quality filtering** using landmark distance and identity similarity metrics to ensure sufficient lip divergence while preserving identity, and supplement with **3D-rendered data** for perfectly aligned pairs. Implementation details are in Sec. B.3.

Together, these principles ensure the *generator* produces contextual pairs that, though not perfect, consistently provide strong and reliable conditions for the *editor*.

3.2 Editor: Context-Driven Video-to-Video Dubber

Given curated pairs $(\boldsymbol{V}', \boldsymbol{V})$, we train a DiT-based *editor* for mask-free dubbing. Unlike the *generator*, the *editor* tackles dubbing directly: given audio $\boldsymbol{a}$ and the companion video $\boldsymbol{V}'$, it learns to produce $\boldsymbol{V}$ as the target, thereby transforming dubbing from a sparse inpainting problem into context-driven editing. In practice, the *editor* surpasses the *generator* across lip accuracy, identity preservation, and robustness, benefiting from the rich contextual input provided by the paired videos.

Contextual conditioning mechanisms. As shown in Fig. 3, the paired reference and target videos are encoded as latents $\boldsymbol{z}_{\text{ref}}, \boldsymbol{z}_{\text{tgt}} \in \mathbb{R}^{b \times f \times c \times h \times w}$. The diffused $\boldsymbol{z}_{\text{tgt}}$ is then concatenated with clean $\boldsymbol{z}_{\text{ref}}$ across frames, forming $\boldsymbol{z}_{\text{in}} \in \mathbb{R}^{b \times 2f \times c \times h \times w}$. Patchifying this sequence enables contextual interaction via 3D self-attention, minimally altering the DiT backbone while fully exploiting its contextual modeling capacity. Audio features and motion frames are integrated identically to Sec. 3.1.

3.3 Timestep-Adaptive Multi-Phase Learning with LoRA Experts

While contextual pairs significantly simplify dubbing, training the *editor* must still balance global structure, precise lip sync, and fine-grained identity. Diffusion models exhibit stage-wise specialization across timesteps (Zhang et al., 2025; Wang et al., 2025), motivating us to introduce a timestep-adaptive multi-phase scheme, where different noise regions target complementary objectives.

Phase partitioning. Following Esser et al. (2024), we shift the timestep sampling distribution to concentrate on different noise levels for each training phase:

$$t_{\text{shift}} = \frac{\alpha t_{\text{orig}}}{1 + (\alpha - 1)t_{\text{orig}}}, \tag{2}$$

where t_{orig} is logit-normal and α sets the shift strength. This yields: 1) high-noise steps for global structure and motion, including background layout, head pose, and coarse identity; 2) mid-noise steps for lip movements; 3) low-noise steps for texture refinement concerning identity details.

High-noise full training. We first train the *editor* under the high-noise distribution with full-parameter optimization. This setting not only facilitates convergence and improves generation quality (Esser et al., 2024), but also encourages the model to learn global structures effectively, thus seamlessly transferring background, head pose, overall identity, and other spatiotemporal dynamics from reference contexts while achieving preliminary lip sync. The objective is the same mask-weighted flow-matching loss $\mathcal{L}_{\text{wFM}}$ as in Eq. 1.

Mid- and low-noise tuning with LoRA experts. We then attach lightweight LoRA modules for mid- and low-noise phases. Since pixel-level constraints are needed, we design a single-step denoising strategy to avoid computational overhead during training:

$$\hat{\boldsymbol{x}}_0 = \mathcal{D}(\boldsymbol{z}_0 + (\boldsymbol{v} - \hat{\boldsymbol{v}}) \cdot \min\{t, t_{\text{thres}}\}), \tag{3}$$

where t_{thres} ensures stable denoising at high noise levels (see Sec. B.5 for detailed derivation).

The *lip expert* operates at mid-noise, supervised by an additional lip-sync loss $\mathcal{L}_{\text{sync}}$ using SyncNet (Chung & Zisserman, 2016) for audio-visual alignment. The *texture expert* works at low-noise with identity loss $\mathcal{L}_{\text{id}}$ computed against references using ArcFace (Deng et al., 2019) and CLIP (Radford et al., 2021a) features. To avoid hurting sync, we randomly disable audio cross-attention (probability 0.5) during texture tuning, computing texture supervision only under silent conditions.

During inference, we manually activate each LoRA within its optimal timestep range: texture expert for $t \in (0, 0.3)$ and lip expert for $t \in (0.4, 0.6)$, ensuring each contributes where most effective.

4 Experiments

Benchmark. To evaluate visual dubbing in practical settings, we construct ContextDubBench, a challenging benchmark of 440 video-audio pairs combining real-world and AI-generated content. Videos feature challenging scenarios like profile views, pose shifts, occlusions, and stylized appearances, while audio includes speech and singing across six languages. Unlike existing controlled-environment datasets, it enables evaluation under complex, realistic conditions, as detailed in Sec. D.

Evaluation metrics. We evaluate generation quality using PSNR, SSIM, Fréchet Inception Distance (FID) for spatial quality, and Fréchet Video Distance (FVD) for temporal consistency. Lip-sync quality is measured by landmark distance (LMD) and SyncNet confidence (Sync-C). Identity preservation is assessed through cosine similarity of ArcFace embeddings (CSIM), CLIP score (CLIPS) for semantic features, and LPIPS for perceptual similarity.

For the more challenging ContextDubBench, we additionally report no-reference perceptual quality metrics, including Natural Image Quality Evaluator (NIQE), Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), and HyperIQA (Su et al., 2020). We also report the overall success rate across all 440 video samples, with failed or entirely unsynchronized generations manually excluded. This metric is crucial for practical dubbing scenarios, as traditional methods often fail completely under visual challenges like stylized characters, occlusions, or extreme poses.

4.1 Quantitative Evaluation

We evaluate our *editor* on both HDTF (Zhang et al., 2021) and ContextDubBench, comparing against state-of-the-art methods including Wav2Lip (Prajwal et al., 2020), VideoReTalking (Cheng et al., 2022), TalkLip (Wang et al., 2023), IP-LAP (Zhong et al., 2023), Diff2Lip (Mukhopadhyay et al., 2024), MuseTalk (Zhang et al., 2024), and LatentSync (Li et al., 2024). We also re-implement a generalizable variant of our *generator*, denoted as *generator**, by removing data-creation-specific

Table 1: **Quantitative results on HDTF.** Top three are highlighted as first , second , and third .

HDTF Dataset									
Method	Visual Quality				Lip Sync		Identity		
	PSNR ↑	SSIM ↑	FID ↓	FVD ↓	Sync-C ↑	LMD ↓	LPIPS ↓	CSIM ↑	CLIPS ↑
Wav2Lip	27.412	0.851	15.475	530.905	7.663	0.896	0.078	0.807	0.842
VideoReTalking	25.189	0.844	11.303	327.886	7.482	1.170	0.056	0.745	0.808
TalkLip	27.024	0.850	17.315	564.307	5.887	0.858	0.060	0.804	0.855
IP-LAP	28.571	0.860	9.026	352.403	5.199	0.934	0.041	0.840	0.899
Diff2Lip	28.716	0.860	12.251	348.290	7.897	0.911	0.036	0.790	0.876
MuseTalk	29.542	0.866	8.123	258.236	6.409	0.741	0.029	0.824	0.884
LatentSync	31.325	0.903	8.042	235.524	8.163	0.821	0.024	0.847	0.902
Ours-*generator**	34.253	0.914	7.873	172.520	8.045	0.670	0.018	0.855	0.917
Ours-*editor*	34.425	0.934	7.031	176.630	8.562	0.630	0.014	0.883	0.923

Table 2: **Quantitative results on ContextDubBench.** "Ref." is short for reference.

FreeSyncBench									
Method	Visual Quality (Ref.)		Visual Quality (No Ref.)			Lip Sync	Identity		Generation
	FID ↓	FVD ↓	NIQE ↓	BRISQUE ↓	HyperIQA ↑	Sync-C ↑	CSIM ↑	CLIPS ↑	Sucess Rate ↑
Wav2Lip	19.330	631.589	6.908	48.397	35.667	5.087	0.738	0.805	62.95%
VideoReTalking	17.535	341.951	6.392	43.112	44.826	5.126	0.684	0.793	59.09%
TalkLip	21.262	550.658	6.284	38.990	34.311	3.213	0.739	0.724	70.45%
IP-LAP	14.891	328.728	6.576	44.879	38.059	2.292	0.797	0.809	57.73%
Diff2Lip	17.126	378.527	6.554	44.059	36.872	4.702	0.705	0.799	71.82%
MuseTalk	17.519	294.312	6.552	43.778	42.335	2.205	0.672	0.753	60.00%
LatentSync	13.602	265.057	6.113	39.154	41.654	6.282	0.801	0.812	59.77%
Ours-*generator**	10.824	224.893	5.920	36.840	48.120	6.514	0.814	0.818	66.05%
Ours-*editor*	9.351	214.298	5.782	29.870	51.960	7.282	0.850	0.839	96.36%

constraints and aligning its setup with the *editor*. This allows a fair comparison between traditional inpainting and our context-rich editing dubbing, isolating paradigm gains from backbone capacity.

Quantitative results in Tab. 1 and 2 show that our *editor* sets a new state of the art. On HDTF, it achieves superior visual quality (FID –12.6%, FVD –25.0%), stronger lip sync (Sync-C +4.9%), and improved identity retention (CSIM +4.3%) over the best prior method. On the more challenging ContextDubBench, the advantages are even more pronounced: our model delivers better visual quality (NIQE 5.78 vs 6.11, BRISQUE 29.9 vs 39.2), higher lip–audio consistency (Sync-C +16.0%), and stronger identity preservation (CSIM +6.1%). Remarkably, it attains a success rate of 96.4%, exceeding the strongest baseline by over 24 points, while most prior methods remain around only 60–70%. This large margin underscores the robustness and practical reliability of our approach in diverse and unconstrained scenarios, as the paired contextual inputs supply complete identity and spatiotemporal cues that allow the model to generalize beyond controlled settings.

Interestingly, our *generator** already surpasses prior methods on HDTF, with clear gains in identity preservation (CLIPS +1.7%) and visual quality (FVD –26.8%). This highlights the strong generative capacity of the DiT backbone and its potential as a contextual synthesizer under tailored principles. More importantly, when trained on synthetic contextual pairs, our *editor* achieves further improvements (CSIM +3.3%, Sync-C +6.4%, and LPIPS –22.2%) while maintaining comparable FVD. These results demonstrate the effectiveness of our self-bootstrapping paradigm: the backbone not only generates paired data but also benefits from it, enabling stronger mask-free dubbing.

To further examine the self-bootstrapping effect, we evaluate the *generator* tailored for data construction. We sampled 20 synthetic pairs unseen during *editor* training, compared with the editor's outputs from the same inputs. As shown in Tab. 4, the *editor* consistently outperforms the constructed pairs in lip sync and visual quality. Notably, it even achieves stronger identity consistency than the training pairs. We attribute this to the fact that slight mismatches in synthetic companions, especially in fine-grained identity details, behave as speech-irrelevant noise that is suppressed during training. Meanwhile, the editor benefits far more from the rich, frame-aligned contextual signals than it is harmed by such noise, resulting in higher identity fidelity and stronger robustness.

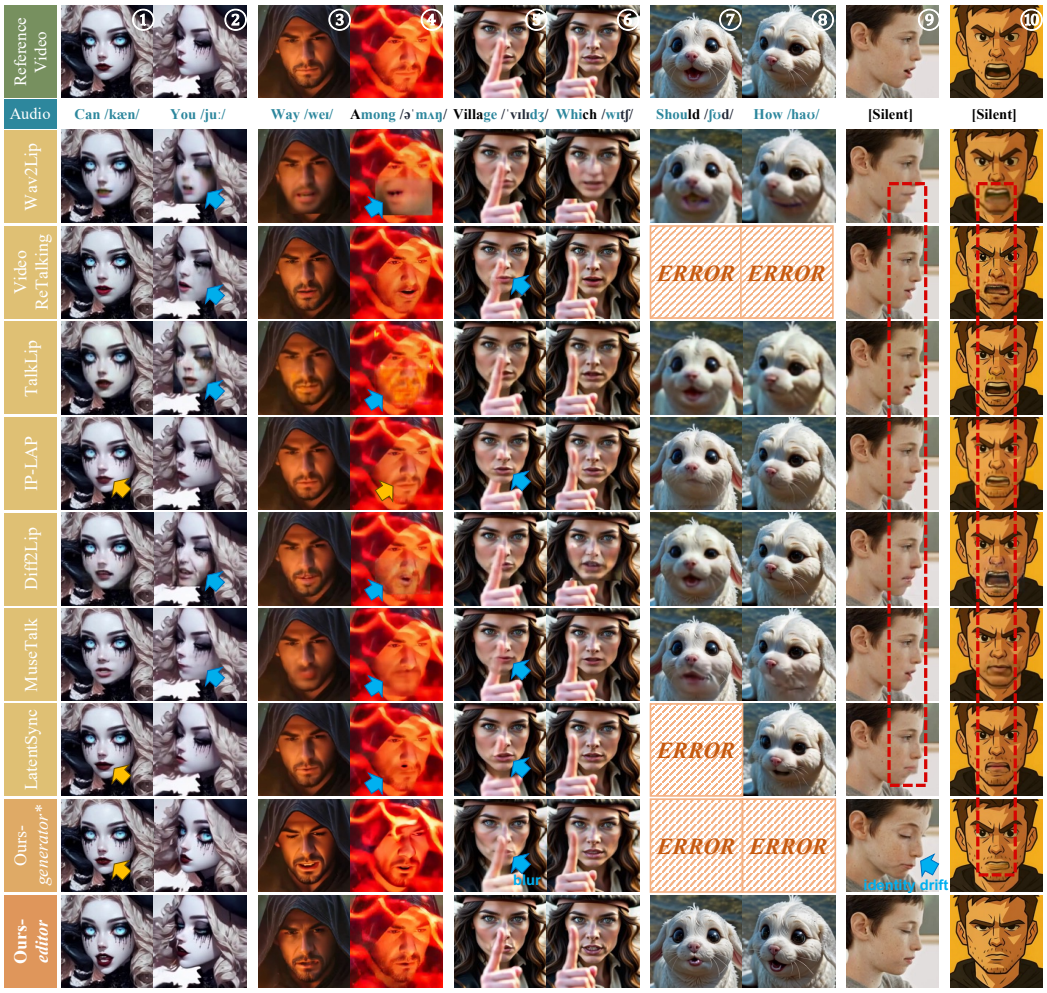

Figure 4: **Qualitative comparisons** across diverse scenarios. Lip-sync errors are marked with yellow, visual artifacts with blue, and lip leakage during silence with red. "ERROR" indicates runtime failure from missing 3DMM or landmarks despite best efforts. Our method exhibits robust performance with superior lip accuracy and identity consistency. Please **zoom in** for details.

Table 3: **User study results** of MOS with 95% confidence intervals.

Method	Realism ↑	Lip Sync ↑	Identity ↑	Overall ↑
Wav2Lip	2.56±0.11	2.80±0.13	3.07±0.14	2.35±0.10
VideoReTalking	3.00±0.09	3.09±0.11	3.58±0.09	3.22±0.11
TalkLip	2.59±0.13	2.08±0.11	3.06±0.11	2.73±0.11
IP-LAP	2.74±0.09	2.49±0.11	3.62±0.11	3.09±0.11
Diff2Lip	2.63±0.11	2.91±0.13	3.22±0.13	2.62±0.12
MuseTalk	2.45±0.10	2.35±0.11	2.98±0.14	2.49±0.11
LatentSync	2.91±0.11	2.81±0.12	3.62±0.11	3.16±0.13
Ours-*generator**	4.28±0.07	3.87±0.09	4.02±0.12	4.48±0.08
Ours-*editor*	4.40±0.06	4.50±0.06	4.40±0.07	4.66±0.05

Table 4: Our *editor* vs. constructed data.

Method	FID ↓	Sync-C ↑	CSIM ↑
Ours-*generator* (constructed data)	7.00	7.88	0.905
Ours-*editor*	6.98	8.97	0.912

Table 5: **Ablation results** on HDTF dataset.

Method	FID ↓	Sync-C ↑	LPIPS ↓	CSIM ↑
Ours-*editor* (full)	7.03	8.56	0.014	0.883
w/ channel concat	6.89	7.49	0.014	0.873
w/ uniform t	18.52	3.85	0.125	0.592
w/o lip tuning	7.00	7.68	0.013	0.875
w/o texture tuning	8.26	8.56	0.018	0.847

4.2 Qualitative Evaluation

Fig. 4 shows qualitative comparisons, where our method consistently produces realistic, lip-synced results across challenging scenarios. Traditional baselines often yield inaccurate lip shapes (Col. 1), visual artifacts (Col. 2), weak robustness to occlusion (Col. 5), and side-view distortions with identity drift (Col. 2&9). Even our *generator**, though using segmentation to handle occlusions, shows

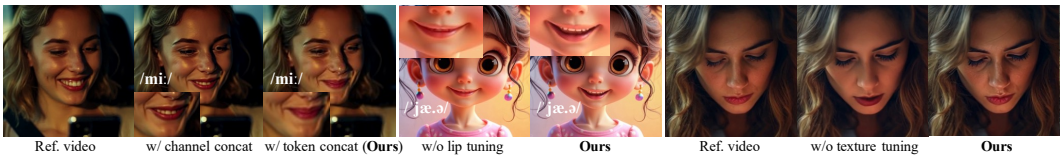

Figure 5: **Ablation results** on video injection and multi-phase learning. Please **Qzoom in** for details.

blur along mask boundaries and remains highly sensitive to mask accuracy. The rightmost column further reveals severe lip-shape leakage in all mask-based methods, where silent frames are corrupted by open-mouth artifacts. In contrast, our editor precisely edits lip movements while preserving identity, remaining robust to spatiotemporal dynamics and even generalizing to non-human characters. Moreover, unlike mask-based pipelines that depend on landmarks or 3DMMs and often fail (marked as "ERROR") on stylized cases, our approach leverages contextual cues to autonomously locate speech-relevant regions, ensuring stable performance across character types and occlusions.

User study. We further conduct a user study with 30 participants on 24 dubbing videos generated by different methods, collecting Mean Opinion Scores (MOS). Each video is rated on a 5-point Likert scale for video realism, lip sync, identity preservation, and overall quality. As shown in Tab. 3, our method achieves clear margins over existing baselines across all aspects. Moreover, our *editor* surpasses *generator*$*$, particularly in identity consistency (4.40 vs. 4.02) and lip sync (4.50 vs. 3.87), validating the self-bootstrapping paradigm and showing that the *editor* delivers perceptually convincing, high-quality dubbing.

4.3 Ablation Study

We conduct ablations on two key components: 1) reference video injection mechanism, and 2) timestep-adaptive multi-phase learning strategy, with results in Tab. 5 and visualization in Fig. 5.

For reference conditioning, replacing our frame-level token concatenation with channel concatenation causes a clear drop in lip sync (Sync-C –12.5%), which is also visible as lip-shape error in Fig. 5. Channel concatenation enforces rigid spatial fusion that conflicts with lip editing, while our token-based design uses self-attention to transfer identity without disturbing lips.

For training, replacing progressive multi-phase sampling with uniform timestep sampling, i.e., learning all noise levels at once, causes severe degradation and even divergence. Stage-wise comparisons further show that removing the lip phase reduces lip sync (–10.3%), with negligible gains in FID and LPIPS, while removing the texture phase weakens fidelity and identity (CSIM –4.1%). These results confirm that the three phases are complementary: high-noise pretraining secures global structure, mid-noise sharpens articulation, and low-noise restores textures and identity. Moreover, the progressive design eases contextual learning by allowing the model to address different information sequentially, rather than struggling with all aspects at once.

5 Conclusion

In this paper, we introduce a novel self-bootstrapping paradigm to address the core challenge in visual dubbing: the absence of paired real-world training data. We argue that instead of relying on masked inpainting, visual dubbing should be reframed as a well-conditioned video-to-video editing task. Built upon this paradigm, we present **X-Dub**, a context-rich dubbing framework, where a DiT model that first acts as a generator to create its own ideal training pairs with complete visual context, and then as an editor that learns from this curated data. This process is further refined by a timestep-adaptive multi-phase learning strategy that disentangles the learning of structure, lips, and texture, enhancing final output quality. Extensive experiments on standard datasets and our new challenging benchmark, ContextDubBench, demonstrate that our method achieves state-of-the-art results. X-Dub shows exceptional robustness in complex, in-the-wild scenarios, significantly outperforming prior works. We believe this work not only sets a new standard for visual dubbing but also offers a valuable insight for other conditional video editing tasks where paired data is scarce.

Ethics Statement

This work presents a self-bootstrapping paradigm for visual dubbing, enabling more accurate and identity-preserving lip synchronization. While such technology can benefit applications in accessibility, education, and multilingual content production, it also raises ethical concerns. In particular, the ability to realistically alter speech and lip movements may facilitate misuse, including the generation of non-consensual content, impersonation, or misinformation. To mitigate these risks, we stress the importance of informed consent, respect for individual privacy, and transparent disclosure of synthetic media. Responsible deployment and adherence to ethical standards are crucial to ensure that advances in visual dubbing contribute positively to society.

References

Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 44(12):8717–8727, 2018.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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.

Hejia Chen, Haoxian Zhang, Shoulong Zhang, Xiaoqiang Liu, Sisi Zhuang, Pengfei Wan, Di ZHANG, Shuai Li, et al. Cafe-Talk: Generating 3d talking face animation with multimodal coarse-and fine-grained control. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Asian conference on computer vision*, pp. 251–263. Springer, 2016.

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*, 2024a.

Jiahao Cui, Hui Li, Yun Zhan, Hanlin Shang, Kaihui Cheng, Yuqi Ma, Shan Mu, Hang Zhou, Jingdong Wang, and Siyu Zhu. Hallo3: Highly dynamic and realistic portrait image animation with diffusion transformer networks. *arXiv e-prints*, pp. arXiv–2412, 2024b.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699, 2019.

Zhiyan Duan, Haotian Fang, Bo Li, Khe Chai Sim, and Ye Wang. The nus sung and spoken lyrics corpus: A quantitative comparison of singing and speech. In *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–9. IEEE, 2013.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Prajwal KR, Rudrabha Mukhopadhyay, Jerin Philip, Abhishek Jha, Vinay Namboodiri, and CV Jawahar. Towards automatic face-to-face translation. In *Proceedings of the 27th ACM international conference on multimedia*, pp. 1428–1436, 2019.

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

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.

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

Rang Meng, Yan Wang, Weipeng Wu, Ruobing Zheng, Yuming Li, and Chenguang Ma. Echomimicv3: 1.3 b parameters are all you need for unified multi-modal and multi-task human animation. *arXiv preprint arXiv:2507.03905*, 2025.

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*, pp. 5292–5302, 2024.

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

Ziqiao Peng, Jiwen Liu, Haoxian Zhang, Xiaoqiang Liu, Songlin Tang, Pengfei Wan, Di Zhang, Hongyan Liu, and Jun He. Omnisync: Towards universal lip synchronization via diffusion transformers. *arXiv preprint arXiv:2505.21448*, 2025.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021a.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021b.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, and Jiwen Lu. Difftalk: Crafting diffusion models for generalized audio-driven portraits animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1982–1991, 2023.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022.

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*, pp. 3667–3676, 2020.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

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

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*, pp. 244–260. Springer, 2024.

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*, pp. 14653–14662, 2023.

MengChao Wang, Qiang Wang, Fan Jiang, and Mu Xu. Fantasytalking2: Timestep-layer adaptive preference optimization for audio-driven portrait animation. *arXiv preprint arXiv:2508.11255*, 2025.

Yu Wang, Xinsheng Wang, Pengcheng Zhu, Jie Wu, Hanzhao Li, Heyang Xue, Yongmao Zhang, Lei Xie, and Mengxiao Bi. Opencpop: A high-quality open source chinese popular song corpus for singing voice synthesis. *arXiv preprint arXiv:2201.07429*, 2022.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4210–4220, 2023.

Shiyi Zhang, Junhao Zhuang, Zhaoyang Zhang, Ying Shan, and Yansong Tang. Flexiact: Towards flexible action control in heterogeneous scenarios. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025.

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 e-prints*, pp. arXiv–2410, 2024.

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

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

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*, pp. 9729–9738, 2023.

A LLM USAGE STATEMENT.

We used large language models (LLMs) solely for linguistic assistance, such as grammar correction and style refinement. No part of the technical content, experimental design, analysis, or conclusions is generated by LLMs. The authors take full responsibility for the content of this paper.

B METHOD AND EXPERIMENT DETAILS

B.1 PRELIMINARY OF FLOW-MATCHING-BASED DIT MODELS.

We adopt a pre-trained T2V DiT model as the backbone for both stages. It follows a latent diffusion paradigm with a 3D causal Variational Auto-Encoder (VAE) (Kingma & Welling, 2013) for video compression and a DiT (Peebles & Xie, 2023) for sequence modeling. Each DiT block interleaves 2D (spatial) self-attention, 3D (spatio-temporal) self-attention, text cross-attention, and feed-forward networks (FFN). Training follows standard flow matching (Esser et al., 2024; Lipman et al., 2022) with the forward process:

$$\boldsymbol{z}_t = (1-t)\,\boldsymbol{z}_0 + t\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I}), \tag{4}$$

and a v-prediction objective to predict $\boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{z}_0$ conditioned on $\boldsymbol{c}$:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{\epsilon}, t}\left[\left\|\boldsymbol{v}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}) - \boldsymbol{v}\right\|_2^2\right]. \tag{5}$$

B.2 DETAILS OF OUR MASK-BASED GENERATOR

Mask setting. Previous mask-based dubbing methods typically employ either smoothly-varying bounding box half-face rectangular masks Prajwal et al. (2020); Cheng et al. (2022); Wang et al. (2023); Zhang et al. (2023) or fixed irregular-shaped masks on affine-transformed facial crops Guan et al. (2023); Li et al. (2024). However, the former's size variations often lead to lip motion information leakage, causing models to learn lip movements from visual occlusion changes rather than the conditional speech, resulting in shortcut learning. The latter constrains jaw position, disrupting pronounced mouth shapes such as wide-open expressions.

Instead, we utilize frame-wise estimated 3D Morphable Model (3DMM) (Retsinas et al., 2024) to obtain full-face masks. Specifically, we maintain each frame's pose, shape, and expression coefficients unchanged except for the jaw opening parameter, which is fixed at a maximum opening value of 0.4. We then project the facial mesh to generate masks. This approach minimizes mask size leakage caused by inter-frame lip variations while providing sufficient editable regions for unrestricted lip control. This strategy facilitates creating synthetic data with lip shapes distinct from the original video, aligning with our data construction principles.

Audio conditioning. Audio features are extracted using the Whisper (Radford et al., 2021b) encoder and then injected via an audio cross-attention layer placed after text cross-attention. Since visual tokens and audio features have different temporal resolutions, for each video frame, we select the corresponding audio feature frames according to the timestamp, together with neighboring frames, forming a temporal window of size $n = 16$. This yields audio tokens $h_a \in \mathbb{R}^{(b\times f)\times n\times c}$, while video tokens are reshaped into $h_V \in \mathbb{R}^{(b\times f)\times(h'\times w')\times c}$, where $h' \times w'$ denotes the visual spatial size after patchfication. Frame-wise cross-attention is then performed between the two modalities.

Reference conditioning. The reference frame I_{ref} is sampled from a different segment of the same video during training to prevent lip-shape leakage, while at inference from the target segment to provide visual cues under a similar head pose.

B.3 DETAILS OF DATA CONSTRUCTION

Short-segment inference. During *generator* inference with a single reference frame, we observe that denoising a long clip of 77 frames (matching the setting used by the backbone and the *editor*) in one pass causes noticeable texture and color drift in the tail frames relative to the first; the drift resets at the first frame of the next clip (see Fig. 6). **Therefore, under a single-reference regime, we conclude that single-pass denoising over long clips is detrimental to identity preservation.**

We hypothesize two contributing factors: 1) the reference frame is anchored at the first position, so later frames become distant in the RoPE index space, amplifying identity drift; and 2) long clips naturally accumulate larger head motion and spatiotemporal changes, which a single reference frame cannot fully constrain.

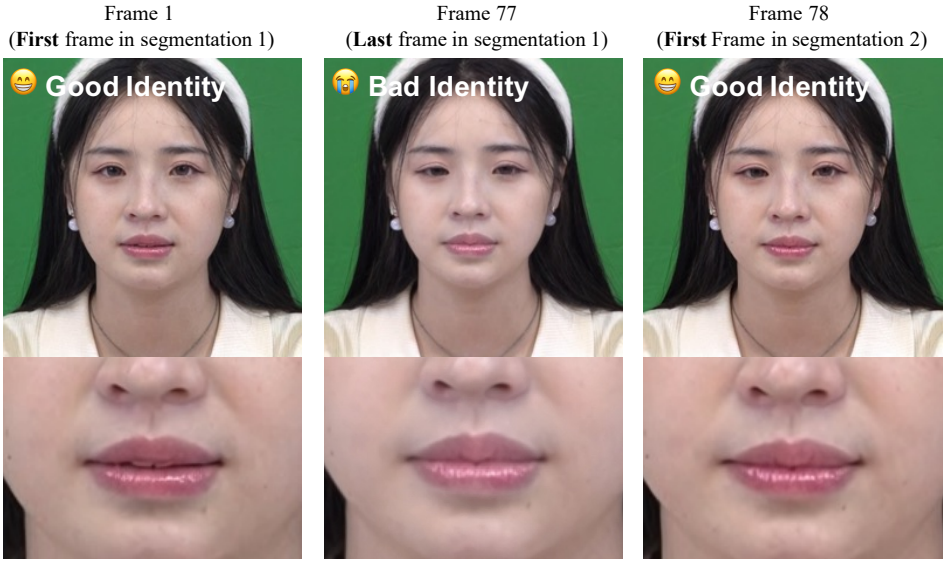

Figure 6: Intra-segment identity drift.

To mitigate this, when constructing contextual pairs with the generator, we adopt short-segment training and inference: we generate clips of 25 frames and bridge adjacent clips with 5 motion frames, then concatenate them to form videos longer than 77 frames for supervising the *editor*, which is trained on 77-frame windows. This short-segment strategy enhances ID preservation, while any slight sacrifice in lip sync accuracy remains within our design guidelines, as shown in Tab. 6.

Table 6: Quantitative results between long-term and short-term processing.

Method	Sync-C (Lip Sync) ↑	CSIM (Identity Preservation) ↑
Long-clip (77 frames)	7.983	0.842
Short-segment (25 frames, +5 overlap)	7.841	0.867

Mask processing with occlusion handling. To enhance the robustness of our generator against occlusions, namely, to maintain consistency with the original video's occlusion patterns and thereby facilitate the editor's ability to naturally inherit them, we introduce an occlusion-handling pipeline First, a vision–language model (VLM) (Bai et al., 2025) is prompted per video with: *"Does any object occlude the person's face? If yes, output **only** a concise description of the object(s). If no, output nothing."* The returned object phrase(s) are then passed to SAM 2 (Ravi et al., 2024) to segment candidate occluders, yielding an occlusion mask M_{occ}. We apply a light manual screening step to remove severely erroneous segmentations.

Finally, we compose the occlusion-aware mask with the original inpainting mask. Let M_{face} be the face mask (foreground 1, background 0), and M_{occ} the occluder mask (1 on occluding objects). The visible-face mask is

$$M_{\text{vis}} = M_{\text{face}} \wedge \neg M_{\text{occ}},$$

and the inpainting mask (where *0* indicates regions to inpaint in our implementation) is

$$M_{\text{inp}} = \neg M_{\text{vis}} \;=\; \neg M_{\text{face}} \;\vee\; M_{\text{occ}},$$

where $\wedge$, $\vee$, and $\neg$ denote logical AND, OR, and NOT, respectively.

While our occlusion annotations can be incomplete or noisy, and using occlusion masks may introduce slight blur near mask boundaries and occasional lip degradation, as shown in the main text, the pipeline still supplies *paired and coherent* references that preserve the scene's occlusion patterns. This supervision encourages the editor to model occlusion–face interactions in context, enabling robust handling of occlusions without labor-intensive manual intervention.

Post-processing. Similar to Zhong et al. (2023), we use a Gaussian-smoothed face mask in post-processing to composite the generator's facial region back onto the original frames, mitigating minor background and boundary artifacts. Concretely, we blur the binary face mask M_{face} with a Gaussian kernel to obtain $\tilde{M}_{\text{face}} \in [0, 1]$, and perform per-frame alpha blending:

$$\boldsymbol{V}_{\text{post}} \;=\; \tilde{M}_{\text{face}} \odot \boldsymbol{V}_{\text{gen}} \;+\; \big(1 - \tilde{M}_{\text{face}}\big) \odot \boldsymbol{V}_{\text{orig}}\,,$$

where $\odot$ denotes element-wise multiplication. This feathered composition keeps backgrounds consistent while preserving sharp facial edits, yielding training pairs with background-aligned context and helping the editor learn background-consistent editing behavior.

Quality filtering. To maintain identity consistency while enforcing distinct lip shapes, we apply two complementary filters to each synthetic-original pair: **1) Identity similarity filter.** We use ArcFace (Deng et al., 2019) to compute cosine similarity between the synthetic and original videos. As a reference, the mean within-speaker similarity across different real segments is 0.812, which is conservative given their differing head motions. Since our paired videos share identical head motion, we adopt a stricter threshold of 0.85 and discard pairs below this value to prevent identity drift. **2) Lip-shape distinction filter.** After aligning faces to a canonical template using the Umeyama algorithm following Deng et al. (2019), we measure the landmark distance over the mouth region between the original and synthetic videos. To ensure sufficient lip-shape variation, we reject pairs with a mouth-region landmark distance below 1.0.

3D talking head rendering data. We leverage Unreal Engine to generate high-quality dubbing pairs. Initially, we acquire the 3D motion representation, which comprises ARKit-based facial expressions and 3D degree-of-freedom (3DOF) head poses. For each dataset entry containing speech audio and 3D motion representation (A, M), we randomly select another entry (A', M'), and replace the speech-correlated coefficients in M with those from M' to form M_{dub}. Both the original dataset entry and its corresponding dubbed version are rendered as follows:

$$\begin{aligned} V &= \mathbf{R}(A, M, I) \\ V_{dub} &= \mathbf{R}(A', M_{dub}, I)' \end{aligned} \tag{6}$$

where $\mathbf{R}$ denotes the Unreal Engine rendering pipeline (following Chen et al. (2025)) and I represents the Unreal Engine Metahuman avatar. To ensure data diversity, we create multiple avatars; however, it is important to note that the same avatar is used for each individual dubbing pair. Ultimately, we collect approximately 10 hours of dubbing pairs, as illustrated in Fig. 7.

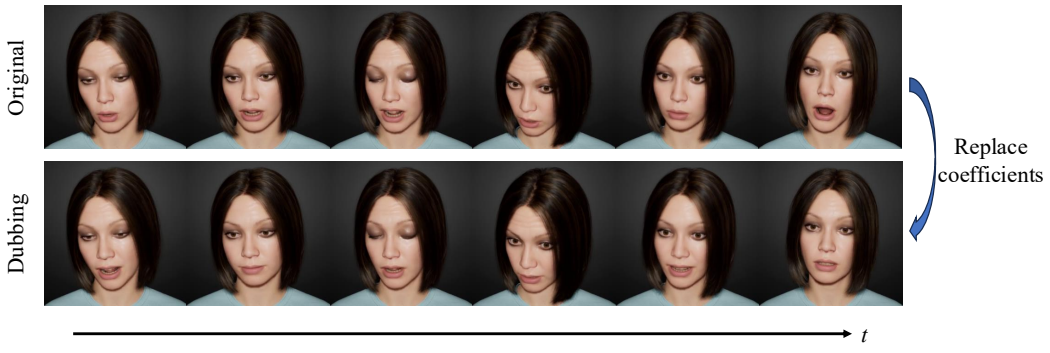

Figure 7: Example of aligned rendered video pairs.

B.4 Details of Our Context-Driven Editor

3D Rotary Position Embedding (RoPE). 3D RoPE is adopted in 3D self-attention of the DiT backbone to distinguish spatial-temporal positions, which we keep unchanged for target tokens. For

reference tokens, inspired by Tan et al. (2024), we adapt RoPE to be temporally-aligned but spatially-shifted. Specifically, a reference token located at (i, j, k), where i, j, and k denote the height, width, and temporal indices, is mapped to $(i + h', j + w', k)$, with (h', w', f') the spatial–temporal sizes after patchification. This design provides two benefits: **(1)** Temporal alignment enables frame-wise consistency preservation of dynamic attributes such as background and head poses; **(2)** Spatial shifting avoids direct overlap that could distort lip movements, and instead encourages the model to capture spatially misaligned yet correlated features like identity information.

B.5 Details of Timestep-Adaptive Multi-Phase Training

Derivation of Eq. (3): Timestep-Constrained Single-Step Denoising. Given the forward diffusion process as in equation 4 and the v-prediction objective $\boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{z}_0$, we derive the single-step denoising formula.

From Eq. 4, we can rearrange to obtain:

$$\boldsymbol{z}_0 = \frac{\boldsymbol{z}_t - t\,\boldsymbol{\epsilon}}{1 - t}. \tag{7}$$

Since $\boldsymbol{v} = \boldsymbol{\epsilon} - \boldsymbol{z}_0$, we have $\boldsymbol{\epsilon} = \boldsymbol{v} + \boldsymbol{z}_0$. Substituting and solving for $\boldsymbol{z}_0$:

$$\begin{aligned} \boldsymbol{z}_0 &= \frac{\boldsymbol{z}_t - t(\boldsymbol{v} + \boldsymbol{z}_0)}{1 - t} = \frac{\boldsymbol{z}_t - t\boldsymbol{v} - t\boldsymbol{z}_0}{1 - t} \\ (1 - t)\boldsymbol{z}_0 &= \boldsymbol{z}_t - t\boldsymbol{v} - t\boldsymbol{z}_0 \\ \boldsymbol{z}_0 &= \boldsymbol{z}_t - t\boldsymbol{v}. \end{aligned} \tag{8}$$

During inference, we use the predicted velocity $\hat{\boldsymbol{v}}$ instead of the true $\boldsymbol{v}$, yielding:

$$\hat{\boldsymbol{z}}_0 = \boldsymbol{z}_t - t\hat{\boldsymbol{v}}. \tag{9}$$

Alternatively, we can express this as:

$$\hat{\boldsymbol{z}}_0 = \boldsymbol{z}_0 + t(\boldsymbol{v} - \hat{\boldsymbol{v}}), \tag{10}$$

which shows the reconstruction error depends on the velocity prediction error scaled by t.

However, when t approaches 1, the velocity prediction error $(\boldsymbol{v} - \hat{\boldsymbol{v}})$ can be amplified, leading to poor reconstructions that result in inaccurate lip sync loss and identity loss computations. To address this, we introduce a timestep constraint:

$$\hat{\boldsymbol{x}}_0 = \mathcal{D}(\boldsymbol{z}_0 + (\boldsymbol{v} - \hat{\boldsymbol{v}}) \cdot \min\{t, t_{\text{thres}}\}), \tag{11}$$

where $\mathcal{D}$ denotes the VAE decoder. Importantly, this clipping is applied only in the denoising computation, not to the actual timestep t used in the model's forward pass. The model still operates with the original t value, enabling it to learn important structural and lip movement information in high- and mid-noise regions. We set $t_{\text{thres}} = 0.6$ in our experiments.

SyncNet supervision. For lip-sync tuning, we adopt a SyncNet (Chung & Zisserman, 2016) comprising a visual encoder S_V and an audio encoder S_a to discriminate temporal alignment between video and audio clips. The lip-sync loss is defined as:

$$\mathcal{L}_{\text{sync}} = \text{CosSim}\big(S_V(\hat{\boldsymbol{x}}_0^{[f:f+8]}),\ S_a(\boldsymbol{a}^{[f:f+8]})\big). \tag{12}$$

This loss is combined with $\mathcal{L}_{\text{mFM}}$ defined in Eq. 1 in a weighted sum to train the lip-sync LoRA:

$$\mathcal{L}_{\text{total}} = (1 + w \cdot \boldsymbol{M} + w_{\text{lip}} \cdot \boldsymbol{M}_{\text{lip}}) \odot \mathcal{L}_{\text{FM}} + w_{\text{sync}} \cdot \mathcal{L}_{\text{sync}}. \tag{13}$$

B.6 Other Implementation Details

We conduct experiments using an internal 1B-parameter T2V model on 32 A100 GPUs, with face-centered videos at 512×512 resolution and 25 fps. For the *generator*, we conduct extended training for 10 epochs on 600 hours of internet audio-video data, sampling 25 frames with lr=1e-5 and batch size 256. After inference and curation, we obtain 400 hours of video pairs, totaling 800 hours/

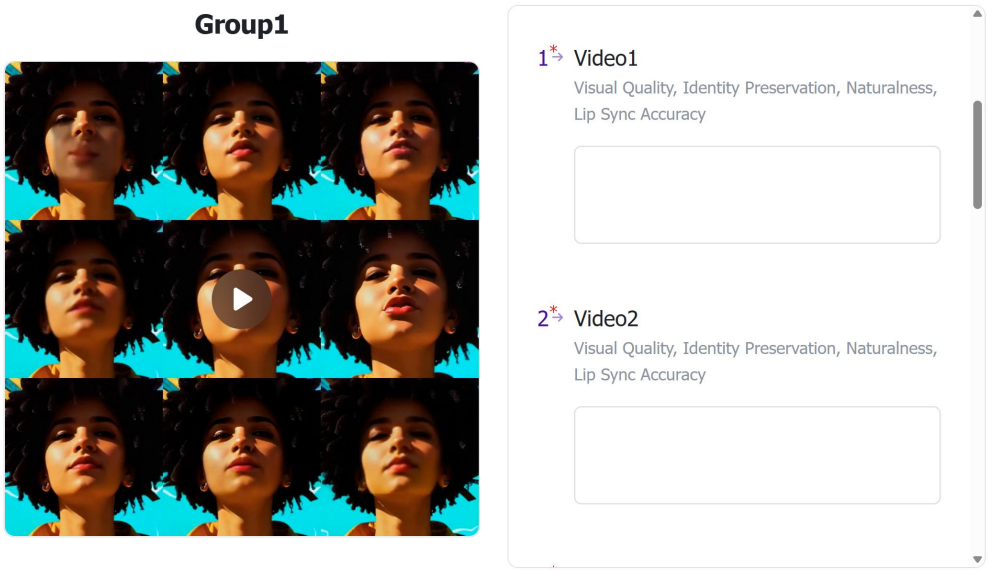

Figure 8: Screenshots of the rating interface of the user study.

For the *editor*, we begin with full-parameter training for 2 epochs on 77-frame samples with lr=1e-5, batch size 256, and timestep shift $\alpha = 4$, followed by LoRA expert training for 0.5 epochs with lr=5e-6 and batch size 64. To reduce computational cost, we decode 4 tokens into 13-frame segments for pixel-level loss computation. Timestep shifts are set to $\alpha = 1.5$ for the lip expert and $\alpha = 0.25$ for the texture expert. Loss weights are set as $w = w_{\text{lip}} = 0.3$ for masks, and 0.05 for SyncNet, CLIP, and ArcFace loss.

C Details of User Study

The user study involved 30 participants. Each participant received compensation of approximately 15 USD for completing a session that lasted 40–50 minutes, which aligns with the average hourly wage. For reference, Fig. 8 provides screenshots of the rating interface used in the study.

D ContextDubBench

To thoroughly evaluate our framework, we construct ContextDubBench benchmark, a challenging benchmark comprising 440 video-audio pairs. The dataset is carefully designed with the following composition:

Audio data. The audio component includes both speech and singing. For speech, we randomly sampled 350 clips from Common Voice (Ardila et al., 2019), spanning six languages and dialects: 170 in English, 60 in Mandarin, 30 in Cantonese, 30 in Japanese, 30 in Russian, and 30 in French. For singing, we incorporated 60 English clips from NUS-48E (Duan et al., 2013) and 30 Mandarin clips from Opencpop (Wang et al., 2022). Each segment lasts between 7 and 14 seconds and captures a wide range of speaking rates, pitch levels, accents, and vocal styles, ensuring rich phonetic and linguistic diversity.

Video data. The video set combines real-world recordings and AI-generated content from publicly available sources with proper copyright clearance (e.g., Civitai, Mixkit, Pexels). It contains 291 clips

Figure 9: ContextDubBench benchmark Examples (I): Showcasing non-human characters with diverse morphological variations.

of natural human subjects, 108 clips of stylized characters with distinct artistic features, and 41 clips of non-human or humanoid entities with durations ranging from 2 to 9 seconds. Representative samples are shown in Fig. 9, Fig. 10, and Fig. 11. Unlike conventional datasets, which are typically captured under controlled conditions, ContextDubBench is explicitly designed to reflect real-world challenges. The dataset incorporates dynamic lighting, partial occlusions, identity-preserving transformations, and substantial variations in pose and motion. By embedding these factors, ContextDubBench more faithfully captures the diversity and unpredictability of real-world scenarios, providing a rigorous testbed for evaluating lip-synchronization models. Illustrative examples are shown in Fig. 12.

Figure 10: ContextDubBench benchmark Examples (II): Showcasing stylized characters with distinctive visual designs.

Figure 11: ContextDubBench benchmark Examples (III): Showcasing real-world human appearances in practical conditions.

Figure 12: Samples from ContextDubBench benchmark showing lighting variations, identity-preserving changes, and occlusions, highlighting complex real-world scenarios.