

FlexLTS: A Unified Paradigm for Content-Accurate Lip-to-Speech Synthesis with Optional Lip Reading

Zhechen Liu^{1,2}, Shuang Yang^{1,2}, Shiguang Shan^{1,2}, Xilin Chen^{1,2}

¹ State Key Laboratory of AI Safety, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

² University of Chinese Academy of Sciences, Beijing 100190, China

liuzhechen20a@mailsucas.ac.cn {shuang.yang, sgshan, xlchen}@ict.ac.cn

Abstract

Lip-to-speech (LTS) synthesis aims to generate intelligible speech from silent lip movements, but the ambiguous visual-to-phoneme mapping makes content-accurate synthesis challenging. Existing methods mainly follow two routes: (1) direct synthesis from lip movements and (2) synthesis dependent on an external lip-reading model. The former often suffers from content errors, while the latter improves content accuracy at the cost of computational overhead and error propagation. To address these limitations, we introduce FlexLTS¹, a unified paradigm for content-accurate LTS with optional lip reading. It generates accurate speech without requiring lip reading, while still benefiting from lip-reading text when available. FlexLTS learns phonetic prior embeddings from training transcripts and employs a visual-centric prior fusion mechanism to selectively incorporate relevant phonetic priors into the visual stream, yielding content-enriched semantic representations. Moreover, our proposed Progressive Textual Dropout strategy further promotes prior internalization by reducing reliance on explicit text during training. Experiments show that FlexLTS achieves state-of-the-art content accuracy and speech quality on the popular LRS2-BBC and LRS3-TED benchmarks and also performs well in cross-dataset zero-shot evaluation. Notably, without using any lip-reading model, FlexLTS surpasses existing lip-reading-aided methods in content accuracy.

1 Introduction

Lip-to-speech synthesis aims to generate intelligible speech directly from silent lip movements and has applications in assistive communication, silent video dubbing, and forensic analysis. The core goal of LTS is to generate speech that is both intelligible and content-accurate, so that the synthesized speech faithfully preserves the speaker’s intended

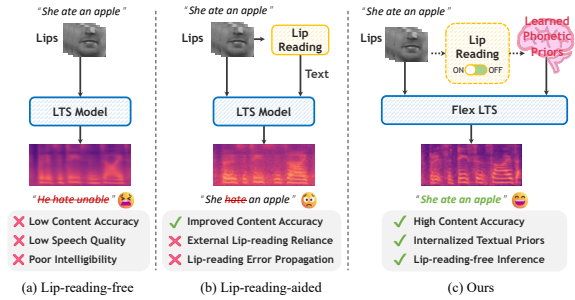


Figure 1: Comparison of lip-to-speech synthesis paradigms. Existing methods either synthesize speech from lip movements alone or use text supplied by an external lip-reading model at inference time. FlexLTS introduces a new paradigm that internalizes phonetic priors into learnable parameters to improve content accuracy while making external lip reading optional rather than mandatory.

message. However, this goal is fundamentally challenged by the ambiguous one-to-many mapping from visual lip movements to phonemes, where a single visual pattern may correspond to multiple distinct sounds, such as /p/, /b/, and /m/. This ambiguity makes content-accurate synthesis from lip movements alone especially difficult.

Driven by recent advances in deep learning, a number of lip-reading-free methods (Akbari et al., 2018; Hegde et al., 2022; Le Cornu and Milner, 2015, 2017; Mira et al., 2022b) have achieved promising results on limited-vocabulary datasets (Harte and Gillen, 2015; Cooke et al., 2006). These methods directly model the mapping from lip sequences to speech without using a lip-reading model or any text-derived semantic cues at inference time. This design keeps the pipeline simple, but also makes content modeling difficult: the model must recover phonetic content solely from an ambiguous visual modality. As a result, these methods still struggle to maintain high content accuracy when extended to open-vocabulary and real-world scenarios.

To improve content recovery from ambiguous lip

¹<https://jasonliu20020101.github.io/FlexLTS/>

movements, one emerging direction is to leverage textual information to refine the generated speech content. Recent methods employ transcripts predicted by a pre-trained lip-reading model to guide synthesis, either through classifier guidance (Yemini et al., 2024) or as an additional input (Pham et al., 2025). We refer to methods that use only lip movements as external inputs during inference as lip-reading-free (LR-free), and those that additionally depend on lip-reading text as lip-reading-aided (LR-aided). These LR-aided methods show that textual cues are effective for resolving visual ambiguity. However, they require an additional lip-reading model to produce text before or during synthesis, introducing computational overhead beyond the synthesis model itself. Moreover, since lip-reading predictions can contain recognition errors, incorrect transcripts may be propagated to the synthesizer and distort the generated content. Therefore, their gains in content accuracy remain dependent on both the availability and reliability of an external lip-reading model.

To overcome this dependence, we aim to make accurate content recovery a capability of the synthesis model itself, rather than an outcome contingent on external lip-reading predictions. Specifically, we use training transcripts to learn reusable phoneme-level cues, termed phonetic priors, within the LTS model. These priors help recover content from lip movements when no lip-reading text is available. When such text is provided, it further adapts the priors to the current utterance. In this way, content-accurate synthesis becomes available without requiring an external lip-reading model, while retaining additional textual guidance when it is available.

In this work, we propose FlexLTS, a framework that operationalizes this new paradigm. FlexLTS learns reusable phonetic prior embeddings from training transcripts. These embeddings complement ambiguous visual cues even when no lip-reading text is provided. To make this possible, we introduce Progressive Textual Dropout, which randomly replaces training transcripts with a global token and teaches the model to activate global phonetic priors without explicit text. When lip-reading text is available, the same framework activates content-specific phonetic priors instead. The visual stream then selectively incorporates the most relevant phonetic priors to form content-enriched semantic representations.

We evaluate FlexLTS under both LR-free and

LR-aided settings on two challenging in-the-wild benchmarks (Afouras et al., 2018a,b). Experimental results show that FlexLTS achieves state-of-the-art content accuracy and speech quality. More importantly, even in the LR-free setting, FlexLTS surpasses existing LR-aided methods in content accuracy. Cross-dataset evaluation further demonstrates that FlexLTS transfers effectively to unseen domains without target-domain adaptation.

Our contributions can be summarized as follows:

- We introduce a unified paradigm for content-accurate lip-to-speech synthesis with optional lip reading. It supports accurate synthesis without an external lip-reading model and can further exploit lip-reading text when available.
- We propose FlexLTS to realize this paradigm. It learns phonetic prior embeddings from training transcripts, promotes their internalization through Progressive Textual Dropout, and selectively fuses visual-relevant phonetic priors to form content-enriched representations.
- Experiments on LRS2-BBC and LRS3-TED demonstrate state-of-the-art content accuracy and speech quality. Notably, FlexLTS in the LR-free setting surpasses existing LR-aided methods in content accuracy.

2 Related Works

2.1 Lip-reading

Lip-reading, or Visual Speech Recognition (VSR), aims to predict speaking content from lip movements. Although recent methods have substantially improved VSR (Ahn et al., 2024; Ma et al., 2023; Zhu et al., 2023; Ma et al., 2021; Shi et al., 2022; Haliassos et al., 2023; Zhang et al., 2024; Chang et al., 2024), reliable recognition remains difficult because visual speech is ambiguous and high-performing models require large-scale annotated data. For instance, on LRS3-TED (Afouras et al., 2018b), methods trained on about 2,000 hours achieve roughly 20% WER (Ma et al., 2023; Ahn et al., 2024), whereas a reported 12.8% WER requires 100,000 hours (Chang et al., 2024). Consequently, LR-aided LTS systems introduce an additional inference-time model and may propagate incorrect lip-reading transcripts into synthesized speech. FlexLTS instead learns phonetic priors from training transcripts, without requiring lip-reading text for LR-free synthesis.

2.2 Lip-to-Speech

Lip-to-Speech (LTS) synthesizes speech from silent lip movements. However, similar visual articulations may correspond to different phonemes, often degrading content accuracy and perceptual quality. Early works mainly studied this task on constrained, small-scale datasets (Cooke et al., 2006; Harte and Gillen, 2015), using CNN-based networks or rule-based systems (Le Cornu and Milner, 2015, 2017; Akbari et al., 2018). Recent approaches based on Conformers (Gulati et al., 2020), GANs (Goodfellow et al., 2020), and diffusion models (Ho et al., 2020) have significantly advanced LTS on large-scale, in-the-wild datasets (Vougioukas et al., 2019; Mira et al., 2022a; Kim et al., 2021; Choi et al., 2023a; Wang and Zhao, 2022; Hegde et al., 2022). Most of these methods follow the LR-free setting: they directly model the mapping from lip sequences to speech without using an external lip-reading model at inference time. While this design keeps the inference pipeline simple, it provides no explicit semantic cues to resolve visual ambiguity, making content-accurate synthesis difficult in open-vocabulary scenarios.

To improve content recovery, a recent line of work incorporates textual information from a pre-trained lip-reading model. These LR-aided methods use lip-reading transcripts to guide a diffusion-based generator (Yemini et al., 2024) or directly fuse them with visual features (Pham et al., 2025). Although they improve content accuracy, their synthesis pipeline mandatorily depends on an additional lip-reading model, which introduces extra computational overhead and exposes the generator to upstream recognition errors. Different from both routes, FlexLTS is designed as a unified framework to support both LR-free and LR-aided inference.

3 Method

As shown in Figure 2, FlexLTS is organized around three requirements of optional lip reading: the model must store useful textual cues without requiring text at inference time, adapt these cues to the current lip movements, and synthesize speaker-specific speech from the resulting content representation. Accordingly, FlexLTS consists of three components: a Phonetic Prior Encoder, a Prior-Fused Visual Encoder, and a Speech Generator. The Phonetic Prior Encoder learns reusable phonetic prior embeddings from training transcripts and activates

them as phonetic priors according to the available textual condition. The Prior-Fused Visual Encoder keeps the visual stream as the backbone and fuses relevant phonetic priors into visual representations, and the Speech Generator converts the resulting semantic features into speech. With this design, the same model supports both lip-reading-free (LR-free) and lip-reading-aided (LR-aided) inference.

3.1 Unified Inference Setting

FlexLTS supports LR-free and LR-aided synthesis with the same model:

$$Y = f(V, C), \quad (1)$$

where V is the lip sequence, Y is the generated speech, and C is the condition used to activate phonetic priors. For LR-free inference, $C = \text{global}$, where global is a special token that activates global phonetic priors. For LR-aided inference, $C = \hat{T}$, where $\hat{T}$ is optional text predicted by an external lip-reading model and activates content-specific phonetic priors. The two modes share the same visual backbone and speech generator, and differ only in the condition sent to the Phonetic Prior Encoder.

3.2 Phonetic Prior Encoder

The Phonetic Prior Encoder is the interface between transcript supervision during training and optional textual conditioning at inference time. Its role is not to predict text, but to convert phoneme-level textual information into a reusable prior representation that can be used with or without lip-reading text.

Prior activation. We adopt a Q-Former-like query-based design (Li et al., 2023). The encoder maintains learnable phonetic prior embeddings $\Lambda \in \mathbb{R}^{N_p \times d}$, whose entries serve as queries. Given an input phoneme sequence, a lightweight Transformer encoder (Vaswani et al., 2017) produces textual features $f_T \in \mathbb{R}^{L_t \times d}$. The phonetic prior embeddings attend to the textual features, where f_T serves as keys and values:

$$\hat{\Lambda} = \text{LN}(\Lambda + \text{CrossAttn}(\Lambda, f_T, f_T)), \quad (2)$$

$$P = \text{LN}(\hat{\Lambda} + \text{FFN}(\hat{\Lambda})). \quad (3)$$

We refer to the output P as phonetic priors. Given lip-reading text, the encoder produces text-conditioned phonetic priors P_{text} . When no text is available, a special global token activates the

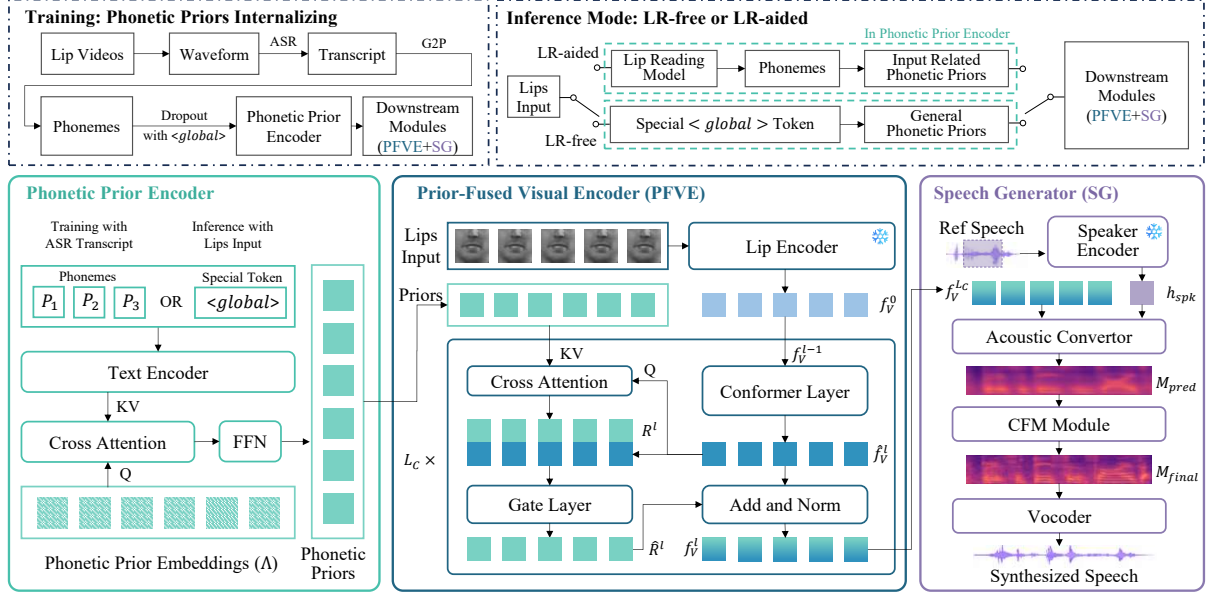


Figure 2: Architecture overview of FlexLTS. The dashed path denotes the optional external lip-reading process; the Phonetic Prior Encoder itself does not take lip inputs.

shared phonetic prior embeddings into global phonetic priors P_{global} . Since the global token contains no utterance-specific content, P_{global} represents the corpus-level phonetic regularities stored in the learnable prior embeddings rather than a transcript of the current utterance.

Prior internalization. To better internalize phoneme-level information into the model parameters, we introduce Progressive Textual Dropout (PTD) during training. For each training sample, we obtain a transcript by applying ASR to the training audio and convert it into a phoneme sequence. The phoneme sequence is used to activate content-specific phonetic priors and supervise the phonetic prior embeddings. However, always providing the full phoneme sequence may cause the model to over-rely on the textual input instead of using the information captured by the phonetic prior embeddings. Therefore, we gradually replace the phoneme sequence with the special global token:

$$p_{\text{drop}}(s) = \min \left(p_{\text{drop}}^{\max}, \frac{s}{\eta S_{\text{total}}} \cdot p_{\text{drop}}^{\max} \right), \quad (4)$$

where s is the current training step, S_{total} is the total number of steps, $p_{\text{drop}}^{\max}$ is the maximum dropout probability, and η controls the schedule. This strategy couples two learning signals. When phoneme sequences are provided, the learnable embeddings acquire phoneme-level regularities through transcript-conditioned activation. When

the global token is used, the model must generate speech from visual input while activating the same embeddings without utterance-level text. As a result, textual supervision is converted into reusable phonetic prior embeddings that support LR-free inference and can still be conditioned on lip-reading text in LR-aided inference.

3.3 Prior-Fused Visual Encoder

The Prior-Fused Visual Encoder keeps the visual stream as the main path and fuses phonetic priors only as complementary cues. This design makes the module visually grounded: phonetic priors help resolve ambiguity, but they do not replace visual evidence. We adopt the visual encoder from a pre-trained, frozen AV-HuBERT model (Shi et al., 2022) to extract initial lip features f_V^0 . The encoder then stacks visual-centric gated fusion layers. In each layer, the visual features f_V^{l-1} first model temporal dependencies with a Conformer block (Gulati et al., 2020), and then retrieve relevant phonetic priors:

$$\hat{f}_V^l = \text{Conformer}(f_V^{l-1}), \quad (5)$$

$$R^l = \text{CrossAttn}(\hat{f}_V^l, P, P), \quad (6)$$

where P is either P_{global} or P_{text} . A gate then determines how much of the retrieved information should be fused into the visual stream:

$$g^l = \sigma(\text{Gate}(\text{Concat}(\hat{f}_V^{l-1}, R^l))), \quad (7)$$

$$\hat{R}^l = g^l \odot R^l, \quad (8)$$

$$f_V^l = \text{LN}(\hat{f}_V^{l-1} + \hat{R}^l). \quad (9)$$

Here, $\hat{f}_V^{l-1}$ is the processed visual feature, R_l is the retrieved phonetic-prior feature, and g_l is the predicted fusion gate; σ , $\odot$, and LN denote sigmoid, element-wise multiplication, and layer normalization, respectively. This layer keeps the visual stream in control: visual features decide which priors are relevant and how strongly they should be incorporated. Therefore, even when all utterances share the same global phonetic priors in LR-free inference, different lip movements can still select different prior information and form content-enriched semantic representations. This visual-centric selection makes the internalized priors adaptive rather than a fixed global bias.

3.4 Speech Generator

The Speech Generator transforms the content-enriched semantic representation into waveform while preserving speaker characteristics. Following V2SFlow (Choi et al., 2025a), we use a lightweight pre-trained speaker encoder (Wan et al., 2018) to extract a speaker embedding from a 0.5-second reference speech segment. The speaker embedding is then fused with the semantic representation before acoustic generation.

3.4.1 Acoustic Converter

Let f_V^{LC} denote the output of the Prior-Fused Visual Encoder and $\mathbf{h}_{\text{spk}}$ denote the speaker embedding. We first concatenate them and project the fused representation into the acoustic space:

$$\mathbf{h}_{\text{acoustic}} = \text{Linear}(\text{Concat}(f_V^{LC}, \mathbf{h}_{\text{spk}})), \quad (10)$$

The Acoustic Converter then maps $\mathbf{h}_{\text{acoustic}}$ to a mel-space acoustic feature $\mathbf{M}_{\text{pred}} \in \mathbb{R}^{T \times D_{\text{mel}}}$ by L_A layers of conformer:

$$\mathbf{M}_{\text{pred}} = \text{Linear}(\text{Conformer}(\mathbf{h}_{\text{acoustic}})). \quad (11)$$

Here, $\mathbf{M}_{\text{pred}}$ denotes a mel-space acoustic feature conditioned on both speech content and speaker identity, rather than the final mel-spectrogram.

3.4.2 CFM Module

Conditioned on $\mathbf{M}_{\text{pred}}$, we employ a Conditional Flow Matching (CFM) module (Lipman et al., 2023) to generate the final mel-spectrogram $\mathbf{M}_{\text{final}}$. We adopt the efficient Optimal-Transport CFM variant (Lipman et al., 2023; Mehta et al., 2024), where the target vector field $\mathbf{u}_t$ simplifies to

$$\mathbf{u}_t = \mathbf{x}_1 - (1 - \sigma_{\min})\mathbf{x}_0, \quad (12)$$

where $\mathbf{x}_0 \sim p_0$ denotes Gaussian noise, $\mathbf{x}_1 \sim q$ represents the target data, and $\sigma_{\min}$ is a small constant that stabilizes training. Following Matcha-TTS (Mehta et al., 2024; Kim et al., 2025), the CFM module is implemented with a U-Net conditioned on $\mathbf{M}_{\text{pred}}$. Finally, the generated mel-spectrogram is converted into waveform using a pre-trained HiFi-GAN vocoder (Kong et al., 2020; Choi et al., 2025a).

3.5 Training and Inference

The entire model is trained end-to-end by jointly optimizing a mel-space supervision loss and a CFM loss. Specifically, the mel-space supervision loss directly measures the difference between the predicted and ground-truth mel-spectrograms:

$$\mathcal{L}_{\text{Mel}} = \|\mathbf{M}_{\text{pred}} - \mathbf{M}_{\text{gt}}\|_1, \quad (13)$$

where $\mathbf{M}_{\text{gt}}$ denotes the ground-truth mel-spectrogram. Meanwhile, the CFM loss supervises the prediction of the target vector field:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, q(\mathbf{x}_1), p_0(\mathbf{x}_0)} \|\mathbf{u}_t - v_t(\mathbf{x}_t | \boldsymbol{\mu}; \theta)\|_2^2, \quad (14)$$

where t is the continuous flow time, $\mathbf{x}_t$ is the intermediate noisy sample at time t , $\boldsymbol{\mu}$ denotes the condition derived from $\mathbf{M}_{\text{pred}}$, $v_t(\cdot | \boldsymbol{\mu}; \theta)$ is the predicted vector field parameterized by θ , and $\mathbf{u}_t$ is the target vector field defined in Eq. 12. The overall training objective is given by:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Mel}} + \lambda_2 \mathcal{L}_{\text{CFM}}. \quad (15)$$

where λ_1 and λ_2 are loss weights.

At inference time, the input is always a lip sequence. Depending on the inference mode, FlexLTS uses the global token for LR-free inference or lip-reading-derived phoneme sequences for LR-aided inference. Given the resulting $\mathbf{M}_{\text{pred}}$ as the condition, the CFM decoder generates the final mel-spectrogram using classifier-free guidance (CFG) (Ho and Salimans, 2021). Starting from $\mathbf{x}_0 \sim \mathcal{N}(0, I)$, we solve the generative ODE using first-order Euler updates:

$$\mathbf{x}_{t+\epsilon} = \mathbf{x}_t + \epsilon \{ \beta v_t^c(\mathbf{x}_t) + (1 - \beta) v_t^u(\mathbf{x}_t) \}, \quad (16)$$

where ϵ is the step size and β is the guidance scale. For brevity, $v_t^c(\mathbf{x}_t)$ denotes $v_t(\mathbf{x}_t | \boldsymbol{\mu}; \theta)$, the conditional vector field, while $v_t^u(\mathbf{x}_t)$ denotes $v_t(\mathbf{x}_t | \emptyset; \theta)$, the unconditional vector field, where $\emptyset$ represents the null condition. The generated mel-spectrogram is converted to waveform by a pre-trained vocoder (Kong et al., 2020; Choi et al., 2025a).

4 Experimental Settings

4.1 Datasets and Metrics

We evaluate our method on two large-scale in-the-wild benchmarks: LRS2-BBC (Afouras et al., 2018a) and LRS3-TED (Afouras et al., 2018b). LRS2-BBC contains about 224 hours of transcribed BBC videos with unconstrained head motion and open-domain vocabulary. LRS3-TED contains over 439 hours of TED/TEDx speech from diverse speakers and topics. For cross-dataset zero-shot evaluation, all models are trained only on LRS3-TED and directly tested on LRS2-BBC, Wild-VSR (Djilali et al., 2024), and CelebV-Dub (Sung-Bin et al., 2025), without target-domain training or fine-tuning.

We use both objective and subjective metrics. For objective evaluation, we measure content accuracy using Word Error Rate (WER) from an ASR model (Ma et al., 2023), following prior LTS works (Pham et al., 2025; Yemini et al., 2024); additional WER results from Whisper (Radford et al., 2023) are provided in the appendix. We evaluate speech quality with UTMOS (Saeki et al., 2022) and DNSMOS (Reddy et al., 2021), intelligibility with ESTOI (Jensen and Taal, 2016), audio-visual synchronization with LSE-D and LSE-C (Chung and Zisserman, 2016), and inference efficiency with Real-Time Factor (RTF) following (Pham et al., 2025). For subjective evaluation, we conduct a MOS study with 20 experienced participants. Each participant rates naturalness, intelligibility, quality, and synchronization on a 1–5 scale for 165 samples, consisting of 15 randomly selected samples from each of 11 methods. We report MOS with 95% confidence intervals.

4.2 Implementation Details

We use a frozen AV-HuBERT (Large) (Shi et al., 2022) as the Lip Encoder (pre-trained general version, not fine-tuned for lip reading). The Prior-Fused Visual Encoder contains $L_C = 4$ visual-centric gated fusion layers of dimension 512. The Phonetic Prior Encoder uses a 2-layer Transformer Text Encoder of dimension 512 and $N_p = 128$ learnable phonetic prior embeddings with $d = 512$. The Speech Generator uses an Acoustic Converter with $L_A = 4$ layers, followed by conditional flow matching. During inference, we set $steps = 5$, $\beta = 1.3$, and $\tau = 0.5$. Details of data pre-processing, training settings, and baseline implementation are provided in the appendix.

5 Experimental Results

5.1 Objective Results

Overall Comparison. Table 1 compares our proposed FlexLTS with existing methods on LRS2-BBC and LRS3-TED. Across both LR-free and LR-aided settings, FlexLTS achieves the best content accuracy and intelligibility in terms of WER and ESTOI. In the LR-free setting, FlexLTS reduces WER to 22.1% on LRS2-BBC and 20.6% on LRS3-TED, outperforming previous LR-free methods by large margins. Notably, even without using any lip-reading model, FlexLTS already surpasses existing LR-aided methods in content accuracy on both datasets.

FlexLTS also achieves strong speech quality and audio-visual synchronization. Its DNSMOS and UTMOS scores are superior or competitive with the best prior systems, and its LSE-D/LSE-C results remain close to the strongest synchronization baselines. These results indicate that improving content accuracy through phonetic priors does not sacrifice speech quality or synchronization.

Comparing the two inference modes further shows the effectiveness of the unified design. Existing LR-free and LR-aided methods exhibit a clear gap because they are built as separate paradigms. In contrast, FlexLTS uses the same backbone and speech generator in both modes; only the condition for activating phonetic priors changes. As a result, the gap between LR-free and LR-aided performance is small, especially on LRS3-TED, where WER differs by only 0.1%.

LR-free Setting. In the LR-free setting, models must infer speech content from lip movements without external lip-reading text. FlexLTS achieves the best WER and ESTOI on both datasets. Compared with the strongest previous LR-free WER results, FlexLTS reduces WER from 35.6% to 22.1% on LRS2-BBC and from 27.9% to 20.6% on LRS3-TED. This improvement supports the central role of phonetic prior embeddings: during training, they internalize phoneme-level textual priors, and during LR-free inference, global phonetic priors provide useful references for resolving visual ambiguity.

LR-aided Setting. In the LR-aided setting, FlexLTS further exploits text predicted by an external lip-reading model. It achieves the best WER, ESTOI, DNSMOS, and UTMOS on both datasets. On LRS2-BBC, FlexLTS reduces WER from 22.6% to 17.8% compared with LipVoicer,

Table 1: Objective evaluation results on the LRS2-BBC and LRS3-TED test sets. **Bold** and underlined values denote the best and second-best results in each setting, respectively. ‡ indicates results from feeding the GT mel-spec to the vocoder. †: Results reported in (Pham et al., 2025).

Method	LRS2-BBC						LRS3-TED						RTF↓
	WER(%)↓	ESTOI↑	DNSMOS↑	UTMOS↑	LSE-D↓	LSE-C↑	WER(%)↓	ESTOI↑	DNSMOS↑	UTMOS↑	LSE-D↓	LSE-C↑	
Ground Truth	4.8	1.000	2.488	3.054	6.223	8.451	1.0	1.000	2.716	3.572	6.822	7.806	–
Vocoder‡	5.4	0.849	2.458	2.727	6.276	8.279	1.3	0.879	2.740	3.205	6.834	7.694	–
<i>Lip-reading-free</i>													
Multitask (Kim et al., 2023)	56.5	0.283	1.127	1.381	7.937	7.472	35.6	0.294	1.111	1.327	8.085	7.004	0.014
Intelligible (Choi et al., 2023b)	36.2	<u>0.403</u>	2.020	2.136	6.473	8.527	28.0	<u>0.397</u>	2.422	2.671	6.803	7.982	0.013
DiffV2S (Choi et al., 2023a)	53.3	0.283	2.048	2.428	6.867	7.690	37.5	0.284	2.609	3.068	7.240	7.433	–
V2SFlow (Choi et al., 2025a)	<u>35.6</u>	0.219	2.406	3.393	6.297	8.614	<u>27.9</u>	0.379	2.637	3.624	<u>6.710</u>	<u>8.165</u>	0.019
Faces2Voices (Kim et al., 2025)	36.9	0.228	<u>2.552</u>	3.881	6.039	8.884	28.1	0.353	2.759	3.993	6.445	8.477	0.021
FlexLTS (ours)	22.1	0.596	2.680	<u>3.496</u>	<u>6.193</u>	<u>8.704</u>	20.6	0.567	<u>2.728</u>	<u>3.678</u>	6.796	8.083	0.015
<i>Lip-reading-aided</i>													
LipVoicer (Yemini et al., 2024)	<u>22.6</u>	0.244	2.193	2.162	7.721	6.896	<u>22.3</u>	0.209	2.458	2.443	8.287	6.321	14.117†
RESOUND (Pham et al., 2025)	25.6	<u>0.424</u>	<u>2.445</u>	<u>2.997</u>	5.929	8.904	25.8	<u>0.401</u>	<u>2.638</u>	<u>3.150</u>	6.521	8.232	<u>0.301</u>
FlexLTS (ours)	17.8	0.595	2.701	3.533	<u>6.175</u>	<u>8.733</u>	20.5	0.568	2.885	3.734	<u>6.787</u>	<u>8.092</u>	0.283

Table 2: Subjective evaluation results on the LRS2-BBC test set. Results are presented with 95% confidence interval. ‡ indicates results from feeding the GT mel-spec to the vocoder.

Method	Naturalness↑	Intelligibility↑	Quality↑	Synchronization↑
Ground Truth	4.19±0.11	3.98±0.13	3.95±0.11	4.46±0.11
Vocoder‡	3.99±0.11	3.82±0.11	3.66±0.12	4.41±0.09
<i>Lip-reading-free</i>				
Multitask (Kim et al., 2023)	1.54±0.10	1.68±0.09	1.47±0.07	2.86±0.13
Intelligible (Choi et al., 2023b)	2.52±0.14	2.52±0.13	1.88±0.09	3.41±0.10
DiffV2S (Choi et al., 2023a)	2.95±0.14	2.78±0.15	3.01±0.12	3.59±0.12
V2SFlow (Choi et al., 2025a)	4.09±0.12	3.90±0.13	4.04±0.11	4.21±0.10
Faces2Voices (Kim et al., 2025)	4.16±0.11	3.97±0.10	4.23±0.10	4.75±0.13
FlexLTS (ours)	4.27±0.09	4.14±0.09	<u>4.09±0.09</u>	<u>4.57±0.11</u>
<i>Lip-reading-aided</i>				
LipVoicer (Yemini et al., 2024)	2.60±0.12	2.80±0.13	2.47±0.11	2.92±0.15
RESOUND (Pham et al., 2025)	3.24±0.14	3.34±0.13	2.77±0.12	4.63±0.09
FlexLTS (ours)	4.23±0.09	4.31±0.08	4.11±0.09	<u>4.61±0.10</u>

and improves ESTOI from 0.424 to 0.595 compared with RESOUND. These results show that FlexLTS does not merely avoid dependence on lip reading; it can also effectively use lip-reading text when available by activating content-specific phonetic priors.

5.2 Subjective Results

Table 2 presents subjective evaluation results on LRS2-BBC. FlexLTS obtains the best Naturalness and Intelligibility scores in both LR-free and LR-aided settings, showing that the gains in objective content metrics are also reflected in human perception. In the LR-free setting, FlexLTS achieves 4.27 in Naturalness and 4.14 in Intelligibility, outperforming all previous systems. In the LR-aided setting, FlexLTS further improves Intelligibility to 4.31. For Quality and Synchronization, FlexLTS remains highly competitive with the strongest baselines. These results suggest that incorporating phonetic priors into the visual stream improves content recovery while preserving naturalness, quality, and audio-visual alignment.

Table 3: Ablation studies of training strategy on the LRS2-BBC test set. Results are presented as “LR-aided/LR-free”. Abbreviations: TD for Text Dropout; PTD for Progressive Text Dropout. LSE-D/-C are omitted for negligible variations.

Method	WER(%)↓	ESTOI↑	DNSMOS↑	UTMOS↑
No Strategy	18.6 / 44.0	0.444 / 0.388	<u>2.585</u> / 2.438	3.625 / 3.220
+ TD ($p_{drop} = 0.4$)	<u>18.0</u> / <u>23.9</u>	<u>0.589</u> / <u>0.586</u>	2.557 / <u>2.545</u>	3.490 / 3.438
+ TD ($p_{drop} = 0.8$)	35.9 / 36.9	0.413 / 0.411	2.540 / 2.534	3.484 / <u>3.483</u>
+ PTD (ours, $p_{drop}^{max} = 0.8$)	17.8 / 22.1	0.595 / 0.596	2.701 / 2.680	<u>3.533</u> / 3.496

Table 4: Ablation studies of FlexLTS on the LRS2-BBC test set. Results are presented as “LR-aided/LR-free”. PPE: Phonetic Prior Embeddings; TS: Textual Stream; GF: Gated Fusion; LF: Low-level Fusion; HF: High-level Fusion; CFM: CFM Module. LSE-D/-C are omitted for negligible variations.

Method	WER(%)↓	ESTOI↑	DNSMOS↑	UTMOS↑
FlexLTS	17.8 / 22.1	0.595 / 0.596	2.701 / 2.680	<u>3.533</u> / 3.496
<i>LTS Paradigm</i>				
–w/o PPE	19.7 / 43.3	0.441 / 0.392	2.558 / 2.461	3.550 / 3.263
–w/o TS	– / 37.0	– / 0.413	– / 2.527	– / 3.432
<i>Model Structure</i>				
–w/o GF	20.1 / 29.4	0.528 / 0.517	2.639 / 2.625	3.428 / 3.362
–w/o LF	18.4 / 25.0	0.561 / 0.556	<u>2.680</u> / <u>2.667</u>	3.493 / 3.451
–w/o HF	19.5 / 26.2	0.554 / 0.549	2.673 / 2.663	<u>3.500</u> / 3.455
–w/o CFM	<u>17.9</u> / 21.1	<u>0.587</u> / <u>0.588</u>	2.157 / 2.150	2.524 / 2.485

5.3 Ablation Studies

We conduct ablation studies on LRS2-BBC to analyze the proposed paradigm, model structure, and training strategy. All ablated variants have comparable parameter sizes; further analyses, including speaker embeddings, text quality, and ASR models, are provided in the appendix.

LTS Paradigm. We first examine whether phonetic prior embeddings are necessary. In Table 4, *w/o PPE* removes phonetic prior embeddings and directly uses the Text Encoder output as phonetic priors, while keeping a comparable parameter bud-

get. This variant degrades WER from 17.8% / 22.1% to 19.7% / 43.3%, with a particularly large drop in the LR-free setting. This shows that direct text features are not sufficient for robust LR-free synthesis; the learnable phonetic prior embeddings are essential for internalizing reusable phoneme-level priors.

The *w/o TS* variant removes the textual branch and uses only visual features, also under a comparable parameter budget. It therefore only applies to LR-free inference. Its WER increases to 37.0%, confirming that the textual-prior pathway provides complementary information beyond the visual stream.

Model Structure. We next study how phonetic priors should be fused into the visual stream. Removing gated fusion (*w/o GF*) replaces selective fusion with a simpler integration strategy and increases WER to 20.1% / 29.4%. This confirms that the visual stream should not receive phonetic priors indiscriminately; it must select relevant information according to the current lip movements.

We further study fusion at different visual levels. The full model shares the same phonetic priors across Low-level Fusion (LF) and High-level Fusion (HF). Retaining only HF (*w/o LF*) or LF (*w/o HF*) degrades performance, indicating that the priors provide complementary lower-level phonetic cues and higher-level contextual information. Finally, removing the CFM module (*w/o CFM*) mainly reduces DNSMOS and UTMOS, confirming its importance for speech quality.

Training Strategy. Table 3 evaluates the effect of textual dropping. Without text dropping, the model over-relies on explicit text and performs poorly in LR-free inference, resulting in a large LR-aided/LR-free WER gap. Fixed Text Dropout with $p_{\text{drop}} = 0.4$ reduces this gap but remains sub-optimal, while $p_{\text{drop}} = 0.8$ weakens text learning and hurts both modes. In contrast, Progressive Textual Dropout allows the model to learn from phoneme sequences before adapting to the global condition, achieving the best overall performance and reducing the WER gap to 4.3%.

5.4 Generalization Analysis

Compatibility with Alternative Visual Encoders. Following most existing LTS methods, we adopt AV-HuBERT Large (Shi et al., 2022) as the default visual encoder. However, the proposed phonetic-prior learning and fusion mechanisms are not tai-

Table 5: Compatibility evaluation on LRS2-BBC. Results are reported as “LR-aided/LR-free”.

Visual Encoder	WER(%)↓	DNSMOS↑	UTMOS↑	ESTOI↑
AV-HuBERT Large (Shi et al., 2022)	17.8 / 22.1	2.701 / 2.680	3.533 / 3.496	0.595 / 0.596
USR2.0 Huge (Haliassos et al., 2026)	16.9 / 19.7	2.710 / 2.698	3.606 / 3.565	0.587 / 0.591

Table 6: Cross-dataset evaluation. All models are trained only on LRS3-TED. Results are reported as “LR-aided/LR-free”.

Model	Test Data	WER(%)↓	UTMOS↑	DNSMOS↑
FlexLTS (ours)	LRS3-TED	20.5 / 20.6	3.734 / 3.678	2.885 / 2.728
	LRS2-BBC	21.4 / 24.7	3.421 / 3.407	2.655 / 2.592
	WildVSR	53.2 / 55.6	3.342 / 3.318	3.123 / 3.124
	CelebV-Dub	40.7 / 43.9	3.188 / 3.200	2.714 / 2.728
FlexLTS (w/o TS)	LRS3-TED	- / 34.3	- / 3.624	- / 2.671
	LRS2-BBC	- / 38.6	- / 3.243	- / 2.594
	WildVSR	- / 71.2	- / 3.179	- / 2.646
	CelebV-Dub	- / 62.6	- / 3.148	- / 2.685
RESOUND (Pham et al., 2025)	LRS3-TED	25.8 / -	3.150 / -	2.638 / -
	LRS2-BBC	27.4 / -	3.090 / -	2.417 / -
	WildVSR	64.5 / -	3.127 / -	2.630 / -
	CelebV-Dub	51.9 / -	2.984 / -	2.533 / -
V2SFlow (Choi et al., 2025a)	LRS3-TED	- / 27.9	- / 3.624	- / 2.637
	LRS2-BBC	- / 38.0	- / 3.380	- / 2.629
	WildVSR	- / 59.7	- / 3.242	- / 2.605
	CelebV-Dub	- / 53.4	- / 3.059	- / 2.593

lored to AV-HuBERT and can be integrated with representations produced by different visual encoders without changing the overall framework. To evaluate this compatibility, we replace AV-HuBERT Large with USR2.0 Huge (Haliassos et al., 2026) while keeping the remaining components unchanged. As shown in Table 5, USR2.0 achieves lower WER and better DNSMOS and UTMOS in both LR-aided and LR-free settings, while maintaining comparable ESTOI. These results show that FlexLTS works effectively with different visual encoders and can benefit from stronger visual representations.

Zero-shot Generalization. Table 6 reports cross-dataset results that evaluate the models’ zero-shot generalization. All models are trained only on the 433-hour LRS3-TED and directly tested on LRS2-BBC, WildVSR and CelebV-Dub without target-domain training or fine-tuning. On LRS2-BBC, FlexLTS achieves 21.4%/24.7% WER in LR-aided/LR-free inference, compared with 27.4% for RESOUND and 38.0% for V2SFlow. It also consistently achieves lower WER than the corresponding baselines on WildVSR and CelebV-Dub.

The FlexLTS (w/o TS) variant is the same *w/o TS* ablation model reported in Table 4. Its consistently higher WER further confirms that the Textual Stream helps learn transferable phonetic priors. Meanwhile, FlexLTS maintains competitive UTMOS and DNSMOS, indicating that the content improvements do not compromise speech quality.

Table 7: Comparison with a visually conditioned TTS method on LRS3-TED. “Reported” uses the evaluation setup in the original paper; the other two rows use the same VSR input and ASR evaluator.

Method	VSR Model	ASR Evaluator	WER(%)
AlignDiT (Choi et al., 2025b) (reported)	Same as LipVoicer	Whisper-large-v3	19.5
AlignDiT (re-evaluated)	Auto-AVSR	Whisper-medium	21.6
FlexLTS (LR-aided)	Auto-AVSR	Whisper-medium	20.5

Table 8: Qualitative Analysis of FlexLTS. GT: ground truth text; VSR: lip-reading transcript; Ours: ASR transcript of our generated audio. Underlined: content of interest; **Red** : wrong content; **Green** : correct content.

Type	Sentence	Phoneme Sequence
<i>Example-A: Phonetic Correction</i>		
GT	BACK TO THE HISTORY BOOKS	... HH IH S T ER IY B UH K S
VSR	BACK TO THE HISTORY BOX	... HH IH S T ER IY B AA K S
Ours	BACK TO THE HISTORY BOOKS	... HH IH S T ER IY B UH K S
<i>Example-B: Contextual Refinement</i>		
GT	... GOOD REASONS AND BAD REASONS	... R IY Z AH N Z <u>AH N D</u> B AE D ...
VSR	... GOOD REASONS EMPIRE REASONS	... R IY Z AH N Z <u>EH M</u> P AY ER ...
Ours	... GOOD REASONS <u>AND BAD</u> REASONS	... R IY Z AH N Z <u>AH N D</u> B AE D ...
<i>Example-C: Misleading</i>		
GT	TIME NOW FOR OUR NEXT GUEST	T AY M <u>N AW</u> F AO R AW ER ...
VSR	TIME <u>NOW</u> FOR OUR NEXT GUEST	T AY M <u>N AW</u> F AO R AW ER ...
Ours	TIME <u>OUT</u> FOR OUR NEXT GUEST	T AY M <u>AW T</u> F AO R AW ER ...

Comparison with Visually Conditioned TTS Methods. LTS is a visually conditioned speech synthesis task, so we further compare FlexLTS with AlignDiT (Choi et al., 2025b), a recent TTS method that uses video, text, and reference audio. Since its reported 19.5% WER is obtained with different VSR and ASR models, we re-evaluate the released model using the same Auto-AVSR input and Whisper-medium evaluator as our LR-aided setting. As shown in Table 7, FlexLTS achieves 20.5% WER, compared with 21.6% for AlignDiT under the controlled protocol. This result further supports the effectiveness of phonetic-prior conditioning.

5.5 Qualitative Analysis

Effect of Phonetic Priors. Table 8 provides representative examples of how FlexLTS uses phonetic priors to improve content generation. In Example-A, the model corrects the lip-reading transcript from “BOX” to “BOOKS”, recovering the correct vowel phoneme. In Example-B, it restores the phrase “AND BAD” from a misleading transcript, showing that phonetic priors can help recover more coherent content. Example-C presents a failure case, where the model changes “TIME NOW” to “TIME OUT”. This suggests that phonetic priors may occasionally favor a more likely phrase when the visual evidence is ambiguous.

Learned Prior-Phoneme Correlations. Figure 3a visualizes the similarity between phonetic

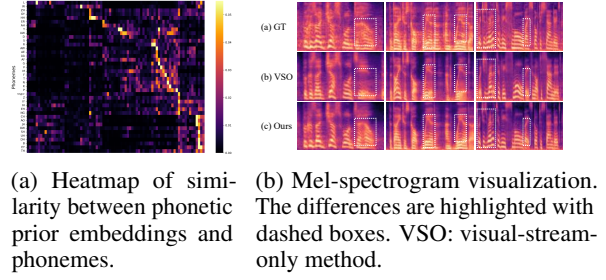


Figure 3: Visualization Results.

prior embeddings and phoneme embeddings. The heatmap shows that different embeddings correlate with different phonemes, suggesting that the learned parameters capture phoneme-related regularities. Visually similar phonemes, such as /AW/, /AE/, and /AA/, share some correlated embeddings, while other embeddings provide discriminative cues. This supports the role of phonetic prior embeddings as a reusable reference space for resolving visual ambiguity.

Mel-spectrogram Visualization. Figure 3b compares mel-spectrograms from ground-truth audio, the visual-stream-only baseline (w/o TS in Table 4), and our LR-free FlexLTS. The visual-stream-only baseline produces blurrier spectrograms and misses fine details, whereas FlexLTS recovers clearer structures that better match the ground truth. This is consistent with the objective and subjective results, where phonetic priors improve intelligibility while maintaining high speech quality.

6 Conclusion

In this paper, we presented FlexLTS, a unified paradigm for content-accurate lip-to-speech synthesis with optional lip reading. FlexLTS learns phonetic prior embeddings from training transcripts, promotes their internalization through Progressive Textual Dropout, and selectively fuses relevant priors into the visual stream. Thus, the same model generates content-accurate speech without external lip reading in LR-free inference, while further exploiting available lip-reading text in LR-aided inference. Experiments on LRS2 and LRS3 demonstrate state-of-the-art content accuracy and speech quality, with FlexLTS in the LR-free setting outperforming existing LR-aided methods in content accuracy. These results show that content-accurate LTS can benefit from textual cues without requiring an external lip-reading model.

Limitations

Although FlexLTS substantially improves content accuracy, the ambiguity of silent lip movements cannot be fully eliminated. In the LR-free setting, when visual evidence is insufficient, learned phonetic priors may favor plausible but incorrect content, as illustrated by our qualitative failure case. In the LR-aided setting, optional lip-reading text can further improve synthesis, but erroneous transcripts may still mislead the generated content. In addition, our model is trained and evaluated only on English datasets; its effectiveness in other languages remains to be studied. Future work may investigate multilingual LTS as well as more reliable prior selection and fusion under uncertain visual or textual conditions.

Ethical Considerations

The authors used generative AI tools solely to polish the language and improve the clarity of the manuscript. All research ideas, technical content, analyses, and conclusions were independently developed and verified by the authors.

Our work uses audio-visual recordings of human speech, which raise privacy, consent, and security concerns. All datasets used in this study are licensed for academic research and were used in accordance with their applicable terms. Nevertheless, lip-to-speech systems could be misused for non-consensual speech reconstruction, privacy-invasive surveillance, voice impersonation, or misleading forensic applications. Accordingly, such systems should be used only with the explicit consent of both the individual depicted in the video and the speaker whose reference audio is used. Because generated speech may contain content errors, it should not be regarded as reliable forensic evidence. Appropriate safeguards include clearly labeling synthesized speech, maintaining auditable records of system use, restricting access to sensitive inputs and models, and complying with applicable laws and dataset licenses.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China (No. 62276247, 62461160331, U24A20332).

References

- Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018a. Deep Audio-Visual Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8717–8727.
- Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018b. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496.
- Young Jin Ahn, Jungwoo Park, Sangha Park, Jonghyun Choi, and Kee-Eung Kim. 2024. SyncVSR: Data-Efficient Visual Speech Recognition with End-to-End Crossmodal Audio Token Synchronization. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 867–871.
- Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. 2018. Lip2AudSpec: Speech reconstruction from silent lip movements video. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2516–2520.
- Oscar Chang, Hank Liao, Dmitriy Serdyuk, Ankit Shah, and Olivier Siohan. 2024. Conformer is All You Need for Visual Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 10136–10140.
- Jeongsoo Choi, Joanna Hong, and Yong Man Ro. 2023a. DiffV2S: Diffusion-based Video-to-Speech Synthesis with Vision-guided Speaker Embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7812–7821.
- Jeongsoo Choi, Ji-Hoon Kim, Jinyu Li, Joon Son Chung, and Shujie Liu. 2025a. V2SFlow: Video-to-Speech Generation with Speech Decomposition and Rectified Flow. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.
- Jeongsoo Choi, Ji-Hoon Kim, Kim Sung-Bin, Tae-Hyun Oh, and Joon Son Chung. 2025b. Aligndit: Multimodal aligned diffusion transformer for synchronized speech generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10758–10767.
- Jeongsoo Choi, Minsu Kim, and Yong Man Ro. 2023b. Intelligible Lip-to-Speech Synthesis with Speech Units. In *Proceedings of the Annual Conference of the International Speech Communication Association*, volume 2023, pages 4349–4353.
- Joon Son Chung and Andrew Zisserman. 2016. Out of time: automated lip sync in the wild. In *Proceedings of the Asian Conference on Computer Vision*, pages 251–263.
- Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. 2006. An audio-visual corpus for speech perception and automatic speech recognition.

- The Journal of the Acoustical Society of America*, 120(5):2421–2424.
- Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Eustache LeBihan, Haithem Boussaid, Ebtesam Al-mazrouei, and Merouane Debbah. 2024. Do VSR Models Generalize Beyond LRS3? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6635–6644.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative Adversarial Networks. *Communications of the ACM*, 63(11):139–144.
- Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 5036–5040.
- Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, and Maja Pantic. 2023. Jointly Learning Visual and Auditory Speech Representations from Raw Data. In *International Conference on Learning Representations*.
- Alexandros Haliassos, Rodrigo Mira, and Stavros Petridis. 2026. Pay Attention to CTC: Fast and Robust Pseudo-Labeling for Unified Speech Recognition. In *International Conference on Learning Representations*.
- Naomi Harte and Eoin Gillen. 2015. TCD-TIMIT: An Audio-Visual Corpus of Continuous Speech. *IEEE Transactions on Multimedia*, 17(5):603–615.
- Sindhu B Hegde, KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2022. Lip-to-Speech Synthesis for Arbitrary Speakers in the Wild. In *Proceedings of the ACM International Conference on Multimedia*, pages 6250–6258.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 33:6840–6851.
- Jonathan Ho and Tim Salimans. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS Workshop*.
- Jesper Jensen and Cees H Taal. 2016. An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022.
- Ji-Hoon Kim, Jeongsoo Choi, Jaehun Kim, Chaeyoung Jung, and Joon Son Chung. 2025. From Faces to Voices: Learning Hierarchical Representations for High-quality Video-to-Speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15874–15884.
- Minsu Kim, Joanna Hong, and Yong Man Ro. 2021. Lip to Speech Synthesis with Visual Context Attentional GAN. *Advances in Neural Information Processing Systems*, 34:2758–2770.
- Minsu Kim, Joanna Hong, and Yong Man Ro. 2023. Lip-to-Speech Synthesis in the Wild with Multi-task Learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.
- Thomas Le Cornu and Ben Milner. 2015. Reconstructing Intelligible Audio Speech from Visual Speech Features. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 3355–3359.
- Thomas Le Cornu and Ben Milner. 2017. Generating Intelligible Audio Speech from Visual Speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(9):1751–1761.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pages 19730–19742.
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations*.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5.
- Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2021. End-to-end Audio-visual Speech Recognition with Conformers. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7613–7617.
- Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. 2024. Matcha-TTS: A Fast TTS Architecture with Conditional Flow Matching. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 11341–11345.
- Rodrigo Mira, Alexandros Haliassos, Stavros Petridis, Björn W Schuller, and Maja Pantic. 2022a. SVTS: Scalable Video-to-Speech Synthesis. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 1836–1840.

- Rodrigo Mira, Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, Björn W Schuller, and Maja Pantic. 2022b. End-to-End Video-to-Speech Synthesis Using Generative Adversarial Networks. *IEEE Transactions on Cybernetics*, 53(6).
- Long-Khanh Pham, Thanh VT Tran, Minh-Tan Pham, and Van Nguyen. 2025. RESOUND: Speech Reconstruction from Silent Videos via Acoustic-Semantic Decomposed Modeling. *Proceedings of the Annual Conference of the International Speech Communication Association*.
- Artem Ploujnikov and Mirco Ravanelli. 2022. SoundChoice: Grapheme-to-Phoneme Models with Semantic Disambiguation. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 486–490.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the International Conference on Machine Learning*, pages 28492–28518.
- Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2021. DNSMOS: A Non-intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing*.
- Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 4521–4525.
- Bowen Shi, Wei-Ning Hsu, Kushal Lakhota, and Abdelrahman Mohamed. 2022. Learning Audio-visual Speech Representation by Masked Multimodal Cluster Prediction. In *International Conference on Learning Representations*.
- Kim Sung-Bin, Jeongsoo Choi, Puyuan Peng, Joon Son Chung, Tae-Hyun Oh, and David Harwath. 2025. VoiceCraft-Dub: Automated Video Dubbing with Neural Codec Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14623–14632.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30:5998–6008.
- Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2019. Video-driven speech reconstruction using generative adversarial networks. In *Proceedings of the Annual Conference of the International Speech Communication Association*, pages 4125–4129.
- Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4879–4883.
- Yongqi Wang and Zhou Zhao. 2022. FastLTS: Non-Autoregressive End-to-End Unconstrained Lip-to-Speech Synthesis. In *Proceedings of the ACM International Conference on Multimedia*, pages 5678–5687.
- Yochai Yemini, Aviv Shamsian, Lior Bracha, Sharon Gannot, and Ethan Fetaya. 2024. LipVoicer: Generating Speech from Silent Videos Guided by Lip Reading. In *International Conference on Learning Representations*.
- Yuanhang Zhang, Shuang Yang, Shiguang Shan, and Xilin Chen. 2024. ES3: Evolving Self-Supervised Learning of Robust Audio-Visual Speech Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27069–27079.
- Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. 2023. VATLM: Visual-Audio-Text Pre-Training with Unified Masked Prediction for Speech Representation Learning. *IEEE Transactions on Multimedia*, 26:1055–1064.

A Implementation Details

A.1 Data Pre-processing

Dataset Description. Both LRS2-BBC and LRS3-TED are English, in-the-wild audio-visual speech datasets containing lip videos, speech audio, and transcriptions. LRS2-BBC contains approximately 224 hours of BBC programming. LRS3-TED contains approximately 439 hours of TED/TEDx talks. We follow the standard dataset splits in all experiments and do not introduce new collected data.

Visual Data Processing. For the visual modality, we adopt the processing pipeline from AV-HuBERT (Shi et al., 2022). Specifically, for each frame in a 25 FPS video, we first extract a 96×96 grayscale image centered on the lip region. This region is then center-cropped to 88×88 pixels. Finally, we normalize each video sequence using its own mean and standard deviation (z-score normalization).

Audio Data Processing. For the audio modality, we follow the methodology of V2SFlow (Choi et al., 2025a) and Intelligible (Choi et al., 2023b). We extract 80-dimensional log mel-spectrograms from the 16 kHz raw audio. The spectrograms are computed using a window length of 640 and a hop length of 160. Furthermore, to facilitate model optimization during training, we apply z-score normalization to the log mel-spectrograms using a mean of -5.5 and a standard deviation of 2.0 computed from all training data.

Textual Data Processing. For the textual modality, we first use the Whisper model (Radford et al., 2023) to perform Automatic Speech Recognition (ASR) on the ground truth audio, generating the corresponding transcripts. These transcripts are then converted into phoneme sequences using the grapheme-to-phoneme (g2p) toolkit from Sound-Choice (Ploujnikov and Ravanelli, 2022). We explicitly preserve the spaces between words in the resulting phoneme sequence (i.e. <sp>).

Subjective Evaluation Details Our subjective study involved 20 experienced experts recruited via email. They were asked to evaluate all 11 methods (including GT) with 15 randomly selected test samples (165 samples in total). The evaluation strictly followed a blind protocol: all samples were anonymized and all participants had no prior exposure to any samples. Aligning with prior works, we

report MOS with 95% confidence intervals computed from the t-distribution, rather than the standard deviation.

A.2 Model Architecture

Phonetic Prior Encoder. The Text Encoder is implemented as a 2-layer Transformer Encoder with positional encodings and a hidden dimension of 512. The feed-forward network used for phonetic prior activation has an intermediate dimension of 2048.

Prior-Fused Visual Encoder. For the visual backbone, we use a pre-trained AV-HuBERT Large model (Shi et al., 2022) with frozen parameters. We use its general pre-trained version rather than the version fine-tuned for lip reading. Visual features are extracted from the 24th layer and projected into 512-dimensional representations through a linear layer followed by a 2-layer Transformer Encoder. The subsequent visual-centric gated fusion stack contains 4 layers, each employing a 512-dimensional Conformer block (Gulati et al., 2020) to model the visual sequence while incorporating relevant phonetic priors. Each fusion gate is implemented as a linear layer followed by a sigmoid activation.

Speech Generator. For the *Speaker Encoder*, we use a pretrained light-weight model following (Choi et al., 2025a). Before fusion, the speaker embedding is repeated along the temporal dimension and concatenated with the semantic features along the channel dimension. For the *Acoustic Converter*, we use a stack of 4 Conformer layers (512-dim, kernel size 31) and a final linear layer to generate an 80-dimensional coarse-grained mel-spectrogram. For the *CFM Module*, we adopt a Transformer-based Conditional Flow Matching (CFM) model based on Matcha-TTS (Mehta et al., 2024). This module consists of 2 Downsample, 2 Mid, and 2 Upsample Blocks. It takes the time step t , the noisy sample $\mathbf{x}_t$, and the condition μ (i.e. coarse mel-spectrogram $\mathbf{M}_{\text{pred}}$) to predict the velocity field $\mathbf{u}_t$ (Eq. 12). The hyperparameter is set to $\sigma = 10^{-4}$, consistent with prior work. For the *Vocoder*, to ensure a fair comparison, we use the official pre-trained HiFi-GAN (Kong et al., 2020) provided by previous studies (Choi et al., 2025a).

A.3 Training Details

Our model was trained on 8 NVIDIA RTX 3090 (24GB) GPUs using mixed-precision (FP16) and

Distributed Data Parallel (DDP). We trained the model for 40k iterations using the AdamW optimizer (Loshchilov and Hutter, 2019). The learning rate was warmed up over the first 30% of iterations to a peak of 5×10^{-5} and then decayed following a cosine schedule. We employed a dynamic batching strategy with a maximum global batch size of 256, ensuring that each GPU processed no more than 320 seconds of audio data per batch. The loss terms were balanced with weights $\lambda_1 = \lambda_2 = 10$. To enable classifier free guidance (Ho and Salimans, 2021), we randomly dropped the condition μ in the CFM module with a probability of $p = 0.2$ during training. Additionally, we implemented the progressive dropout strategy for the input text sequence as described in Sec. 3.2, where the sequence is randomly replaced with a single <global> token using parameters $p_{\text{seq}}^{\text{max}} = 0.8$ and $\eta = 0.6$. It is important to note that, by default, our model was trained using ASR-transcribed text rather than ground truth text; a detailed comparison regarding the training text is provided in our ablation study (Table 12).

A.4 Inference Details

For inference, we employ classifier-free guidance (Ho and Salimans, 2021) with a guidance scale of $\beta = 1.3$ and a temperature of $\tau = 0.5$. We use an Euler ODE solver with 5 sampling steps. Evaluation is performed under two distinct settings: “LR-free” and “LR-aided”. In the “LR-free” setting, the model receives the lip sequence and a <global> token as text input. In the “LR-aided” setting, the text input is a transcription generated by a popular lip-reading model (Ma et al., 2023) which has been widely used in LR-aided LTS methods (Pham et al., 2025; Yemini et al., 2024). For every LR-aided method, the reported RTF includes the complete inference pipeline, including the runtime of the external lip-reading model.

A.5 Comparison methods

We compare our method against state-of-the-art Lip-to-Speech methods, categorized into two groups: (1) LR-free methods, which synthesize speech solely from lip movements (Multitask (Kim et al., 2023), Intelligible (Choi et al., 2023b), DiffV2S (Choi et al., 2023a), V2SFlow (Choi et al., 2025a), Faces2Voices (Kim et al., 2025)); and (2) LR-aided methods, which leverage an additional lip-reading model (Ma et al., 2023) for textual guidance (LipVoicer (Yemini et al., 2024), RE-

SOUND (Pham et al., 2025)). For evaluation, we use results from official implementations (Choi et al., 2023a, 2025a; Yemini et al., 2024; Kim et al., 2025) or from our own reproductions (Kim et al., 2023; Choi et al., 2023b; Pham et al., 2025).

B Additional Ablation Studies

B.1 Sensitivity to the Number of Phonetic Priors

Table 9: Sensitivity analysis of the number of phonetic priors N_p on LRS2.

N_p	LR-aided WER(%)↓	LR-free WER(%)↓
64	18.4	24.7
128	17.8	22.1
256	17.6	22.4

Table 9 shows that $N_p = 64$ provides insufficient capacity, whereas increasing N_p from 128 to 256 yields no consistent improvement across the two inference settings while increasing the parameter and computational costs. We therefore use $N_p = 128$ as a practical trade-off.

B.2 Effect of the Learned Phonetic Prior Embeddings

Table 10: Ablations of learned phonetic priors on LRS2. Randomized PPE starts from the trained full model and randomizes only the PPE parameters before inference.

Method	LR-aided WER(%)↓	LR-free WER(%)↓
Full FlexLTS	17.8	22.1
w/o Textual Stream	–	37.0
Randomized PPE	30.3	30.4

As shown in Table 10, randomizing only the learned PPE parameters increases LR-aided/LR-free WER from 17.8%/22.1% to 30.3%/30.4%, while all other trained components remain unchanged. Together with the visual-only variant, this controlled intervention confirms that PPE stores useful phonetic information and that the visual stream learns to retrieve it during inference.

Fusion-level variants. In the full model, all L_C Prior-Fused Visual Encoder layers receive phonetic priors. In *w/o HF*, only the first layer receives the priors; in *w/o LF*, only the last layer receives them. All remaining layers are standard Conformer layers that process visual features only, and the variants use comparable parameter counts.

B.3 Effect of Speaker Embedding

Table 11: Evaluation of *Speaker Embedding* on the LRS2 dataset. Results are presented as performance under “LR-aided/LR-free”. “Same-Emb” means using the same speaker embedding during inference for all tested samples.

Method	WER(%)↓	ESTOI↑	DNSMOS↑	UTMOS↑
<i>testing the effect of speaker embedding</i>				
Same-Emb	18.1 / 22.9	0.594 / 0.595	2.708 / 2.684	3.550 / 3.481
FlexLTS(ours)	17.8 / 22.1	0.595 / 0.596	2.701 / 2.680	3.533 / 3.496

We examine whether the speaker embedding provides unintended utterance-level content information. In *Same-Emb*, all test samples are synthesized using the same speaker embedding, removing any sample-specific information that could be carried by this input. As shown in Table 11, replacing the original speaker embeddings with a shared embedding causes only minor changes in WER and ESTOI under both LR-aided and LR-free settings, while speech quality remains comparable. This indicates that the content accuracy of FlexLTS is mainly derived from the visual stream and phonetic priors, rather than from content leakage through the speaker embedding. Therefore, the speaker embedding primarily controls speaker characteristics and does not account for the reported content-generation improvements.

B.4 Effect of Training Text Quality

Table 12: Evaluation of *training text quality* on the LRS2 dataset. Results are presented as performance under “LR-aided/LR-free”. “Quality (WER(%))” here denotes the WER of training texts, serving as a metric for training text quality; ASR means automatic speech recognition (our method); GT denotes ground truth; VSR denotes visual speech recognition (i.e. lip-reading).

Method	Quality (WER(%))↓	WER(%)↓	ESTOI↑	DNSMOS↑	UTMOS↑
<i>Training text</i>					
GT	0.0	16.9 / 20.6	0.599 / 0.601	2.699 / 2.686	3.538 / 3.478
VSR	26.7	19.2 / 22.2	0.593 / 0.591	2.714 / 2.706	3.551 / 3.507
ASR(ours)	15.2	17.8 / 22.1	0.595 / 0.596	2.701 / 2.680	3.533 / 3.496

Table 12 investigates the effect of training text quality on our model’s performance. In this ablation study, we train our FlexLTS model using text sources with varying WER, including Ground Truth (GT) and transcripts from ASR (Radford et al., 2023) and VSR (Ma et al., 2023) models. On the one hand, the results show higher-quality text, such as GT, enables the model to learn a more effective textual prior, which in turn helps to better

resolve the one-to-many ambiguity during inference. On the other hand, even when trained on the VSR transcripts which has a WER of as high as 26.7%, FlexLTS still achieves performance remarkably close to that of the model trained on GT text, with only a slight performance degradation of about 2.3%/1.6% (from 16.9%/20.6% to 19.2%/22.2%). This demonstrates that FlexLTS is able to learn a textual prior even from imperfect transcripts, alleviating the need for costly, manually annotated ground-truth text and enabling training on more general audio-visual datasets.

B.5 Effect of Inference Text Quality

Table 13: Evaluation of *inference text quality* on the LRS2 dataset. Note that all results are obtained from the FlexLTS model in the “LR-aided” setting. “Quality (WER(%))” here denotes the average WER of all input texts for inference. Other abbreviations are kept consistent with Table 12, where “[MODEL]” denotes the model from which the input text is obtained.

Method	Quality (WER(%))↓	WER(%)↓	ESTOI↑	DNSMOS↑	UTMOS↑
<i>Inference text</i>					
GT	0.0	11.3	0.605	2.705	3.542
ASR[Whisper (Radford et al., 2023)]	4.8	11.4	0.605	2.698	3.540
VSR[Conformer (Ma et al., 2021)]	31.9	22.0	0.589	2.695	3.530
VSR[Auto-AVSR (Ma et al., 2023)](ours)	16.7	17.8	0.595	2.701	3.533

Table 13 evaluates how the quality of the textual input at inference time affects the output. In the LR-aided setting, we feed our model with text inputs of different quality (measured by WER), including ground truth (GT), outputs from an ASR model (ASR[Whisper (Radford et al., 2023)]), and predictions from two VSR models: a stronger one (VSR[Auto-AVSR (Ma et al., 2023)]) and a moderately performing one (VSR[Conformer (Ma et al., 2021)]). This experiment is designed to assess the impact of input text quality on model performance and to provide an empirical upper bound for the proposed method. The results indicate that improving the quality of the inference text (i.e., lowering its WER) significantly boosts the content accuracy of the generated audio. Notably, even when the provided inference text is of low quality—for instance, where the WER of the provided text by VSR [Conformer (Ma et al., 2021)] reaches 31.9%—the WER of the speech generated by FlexLTS remains at only 22.0%, substantially lower than that of the provided inference text. This demonstrates that our method has effectively learned textual priors during training to assist in the final speech synthesis, rather than relying solely on the input text information to accomplish the lip-to-speech task.

B.6 Evaluation with Different ASR models

Table 14: Evaluation of the generated speech samples with different ASR Models. Here, “A-WER” and “W-WER” are computed using pre-trained ASR models from Auto-AVSR (Ma et al., 2023) and Whisper-medium (Radford et al., 2023), respectively. $\ddagger$ indicates results from feeding the ground truth mel-spectrogram to the vocoder.

Method	A-WER(%) $\downarrow$	W-WER(%) $\downarrow$
Ground Truth	4.8	4.8
Vocoder $\ddagger$	5.4	5.0
<i>LR-free</i>		
Multitask (Kim et al., 2023)	56.5	58.4
Intelligible (Choi et al., 2023b)	36.2	39.1
DiffV2S (Choi et al., 2023a)	53.3	59.9
V2SFlow (Choi et al., 2025a)	35.6	37.2
Faces2Voices (Kim et al., 2025)	36.9	38.2
FlexLTS (ours)	22.1	22.8
<i>LR-aided</i>		
LipVoicer (Yemini et al., 2024)	22.6	24.3
RESOUND (Pham et al., 2025)	25.6	26.7
FlexLTS (ours)	17.8	18.1

Table 14 presents a comparative analysis of content accuracy evaluated by different ASR models. For fairness and consistency with prior works (Yemini et al., 2024; Pham et al., 2025), we use the ASR model from Auto-AVSR (Ma et al., 2023) as our primary evaluation tool throughout the main paper. To validate the reliability of this metric and provide a more comprehensive assessment, we also performed evaluation with another widely-used ASR model, Whisper-medium (Radford et al., 2023). The results demonstrate that our approach not only consistently achieves the optimal WER across both ASR evaluators but also exhibits the highest stability. Specifically, the performance discrepancy between the two ASR models is minimal for our method, with a gap of only 0.7% in the lipreading-free setting and 0.3% in the lipreading-aided setting. This variance is significantly lower than that of competing methods evaluated under different ASR backends. These findings underscore the robustness of our generated speech against variations in ASR evaluators.

B.7 Effect of the Progressive Textual Dropout(PTD) Strategy

Table 15 evaluates the proposed progressive textual dropout strategy (PTD, described in Sec. 3.2) across different settings. To ensure a fair comparison, we re-train the RESOUND model with different dropout strategies, and evaluate it under both LR-aided and LR-free settings. As shown in the

Table 15: Evaluation of *PTD strategy* across different methods on the LRS2 dataset. Results are presented as performance under “LR-aided/ LR-free”. * The results are obtained by our implementations to fulfill different settings.

Method	WER(%) $\downarrow$	ESTOI $\uparrow$	DNSMOS $\uparrow$	UTMOS $\uparrow$
<i>No Textual Dropout</i>				
RESOUND* (Pham et al., 2025)	25.6 / 44.7	0.424 / 0.397	2.445 / 2.376	2.997 / 2.766
FlexLTS	18.6 / 44.0	0.444 / 0.388	2.585 / 2.438	3.625 / 3.220
<i>with Progressive Textual Dropout(PTD)</i>				
RESOUND* (Pham et al., 2025)(+PTD)	25.0 / 35.9	0.435 / 0.417	2.420 / 2.409	3.003 / 2.950
FlexLTS(+PTD, ours)	17.8 / 22.1	0.595 / 0.596	2.701 / 2.680	3.533 / 3.496

table, FlexLTS surpasses RESOUND across nearly all four metrics when using no textual dropout under both the LR-aided and LR-free scenarios, with only one exception of a marginal decrease in ESTOI under the LR-free setting. When our proposed PTD strategy is introduced, the results show that the proposed strategy not only improves the effect under the LR-aided setting where the strategy acts effectively as data augmentation, but more importantly, it consistently helps existing text-aware LTS models to work under the LR-free setting. For example, when applied to RESOUND (Pham et al., 2025), the WER in the LR-free setting decreases significantly from 44.7% to 35.9%, accompanied by improvements in ESTOI and UTMOS. Together with the advantage of the textual knowledge prior-integrated modules in FlexLTS, incorporating PTD further enhanced the ability of FlexLTS to tackle the text-free setting, substantially reducing the WER of FlexLTS from 44.0% to 22.1%.

C Visualization

For the prior-phoneme similarity heatmap in Figure 3a, we obtain each phoneme representation by feeding the corresponding phoneme token into the Text Encoder. We then compute dot products between each phoneme representation and every embedding in the PPE module. Arranging these scores by phoneme and prior indices produces the heatmap and reflects their association strengths.

In Figure 4, we compare the mel-spectrograms of audio generated by FlexLTS against those from the competing models listed in Table 1. The results of “Vocoder” is obtained from feeding the ground truth mel-spectrogram to the vocoder. Visually, our generated mel-spectrogram most closely matches the ground truth, preserving both its overall structure and finer details with high fidelity. Specifically, our generated spectrograms exhibit sharper time-frequency structures with significantly less background noise, indicating superior audio

fidelity. More critically, the spectral distribution of our method aligns more closely with the GT. While some baseline methods produce visually clear spectrograms, their temporal-spectral patterns often deviate from the GT shape, suggesting potential semantic mismatches (i.e., generating words different from the target content). In contrast, our approach leverages learned textual priors to effectively guide the model in maintaining linguistic coherence and content accuracy during synthesis. This mechanism effectively ensures that our generated speech is not only clear but also semantically consistent with the target, resulting in significant advantages over existing methods.

D Demo Videos

We provide video samples generated by our proposed FlexLTS and other LTS methods in the project page: <https://jasonliu20020101.github.io/FlexLTS/>. Specifically, we provide test samples from part of the dataset, as well as samples generated from real-world videos not included in the dataset.

For the test set samples, we randomly select 25 samples from the *test split* of each dataset (LRS2-BBC (Afouras et al., 2018a) and LRS3-TED (Afouras et al., 2018b)). All selected samples are *unseen* speakers, i.e., the speakers do not appear in the training set. For each sample, we include the results of FlexLTS under both the LR-free and LR-aided settings, together with the outputs of seven prior methods and the ground-truth (GT) video, resulting in 500 videos in total. We also provide the corresponding transcription for each sample to facilitate comparison of content accuracy across different generated videos.

For real-world samples, we randomly selected some video clips from online sources. We guarantee that none of these samples appear in any training data, including data used for pre-training modules. We also present the original real videos and the outputs generated by our method.

E Artifact Use and Dataset Content Considerations

Licenses and intended use. Our study uses publicly released research artifacts, including the AV-HuBERT model (Shi et al., 2022), the LRS2-BBC and LRS3-TED datasets (Afouras et al., 2018a,b). AV-HuBERT is used under the License Agreement for non-commercial research purposes. LRS2-BBC

is accessed for non-commercial academic research under the BBC Research & Development Data Sharing Agreement. LRS3-TED are used in accordance with the terms accompanying their public releases. Our use of these artifacts is limited to academic research on lip-to-speech synthesis and is consistent with their intended research use. We do not redistribute the original datasets or release derived data for non-research applications.

Privacy and content considerations. LRS2-BBC and LRS3-TED are existing in-the-wild audio-visual speech datasets collected from BBC and TED/TEDx videos, respectively. Because lip-to-speech synthesis requires facial lip movements and speech audio, these datasets inherently contain potentially identifying visual and vocal information. We do not collect additional personal information, construct identity annotations, or attempt speaker identification. We use only the official benchmark data under their access conditions and do not redistribute the original audio-visual samples. Since naturally occurring broadcast and talk content may potentially contain sensitive or offensive material, publicly displayed qualitative/demo examples should be screened before release to avoid exposing such content or unnecessary personal references.

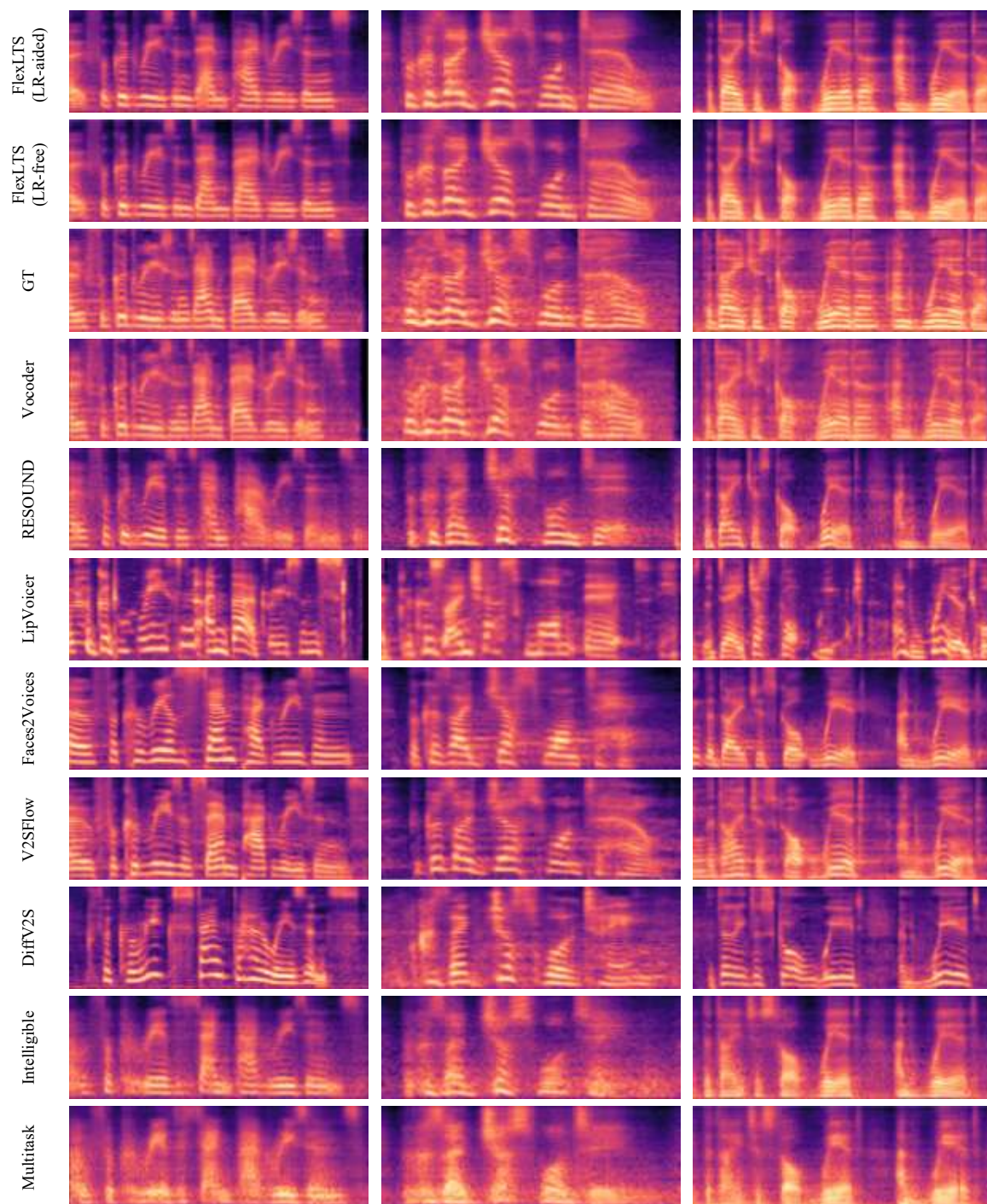


Figure 4: Mel-spectrogram visualization on LRS2-BBC dataset.