Listening to the Slides: Multimodal LLM Speech Inpainting with Acoustic-Aware Loss

Yanir Edri and Bracha Laufer-Goldshtein
School of Electrical and Computer Engineering, Tel-Aviv University, Tel-Aviv, Israel
{yaniredri@mail, blaufer@tauex}.tau.ac.il

Abstract

Speech inpainting reconstructs missing or corrupted speech segments, enabling restoration of damaged virtual lectures and online presentations. While existing audio-only approaches perform well for short gaps, their performance degrades significantly for long-duration missing regions. In this work, we introduce a novel multimodal speech inpainting framework that leverages accompanying presentation slides as semantic grounding for reconstructing both short and long missing speech segments (10 ms–1500 ms). Our approach reformulates speech inpainting from a local signal-reconstruction problem into a contextual speech generation task using a Large Audio-Language Model (LALM). To improve acoustic flexibility during generation, we propose Acoustic Similarity Groups (ASG), a token-grouping strategy that relaxes conventional hard-label training. In addition, we incorporate an auxiliary Connectionist Temporal Classification (CTC) objective to improve alignment between generated speech and target lexical content. Experiments on the M3AV-CHI benchmark demonstrate that the proposed multimodal framework substantially outperforms an audio-only baseline and consistently improves intelligibility and reconstruction quality across challenging long-duration gaps. Code and demo samples are available online.[1]

1 Introduction

Asynchronous online education has become a major component of modern learning, providing broad access to academic content through recorded lectures and virtual presentations. This transition accelerated following the COVID-19 pandemic, which drove the rapid adoption of remote learning platforms (Paudel, 2021). However, the effectiveness of such platforms depends heavily on the quality and continuity of the recorded audio. In realistic natural conditions, speech signals are often degraded by packet loss, compression artifacts, microphone failures, and environmental noise (Vincent et al., 2017; Shim and Lee, 2020), resulting in missing or unintelligible speech segments that disrupt comprehension.

Speech inpainting aims to reconstruct missing portions of an audio signal using the surrounding acoustic context. Traditional signal-processing and deep-learning approaches have demonstrated promising results for short-duration gaps, typically spanning several tens or hundreds of milliseconds (Kegler et al., 2020; Huang et al., 2022; Niizumi et al., 2022). However, their performance degrades substantially when the missing interval becomes longer, since local acoustic context alone is often insufficient to infer semantically meaningful speech content. Recent multimodal methods have attempted to address this limitation by incorporating visual cues such as lip movements or facial dynamics (Montesinos et al., 2023; Elyaderani and Shirani, 2024). While effective, these approaches rely on synchronized high-quality video streams, which are frequently unavailable or unreliable in real-world lecture recordings.

Educational presentations, however, naturally provide an additional source of semantic grounding in the form of presentation slides. Slides often contain keywords, technical terminology, and structured textual content that are strongly correlated with the speaker's utterances. Our analysis of the M3AV-CHI benchmark (Chen et al., 2024) reveals an average of 16.7 overlapping words per slide, corresponding to approximately one semantically grounded cue every 1.25 seconds of lecture audio. This high alignment density suggests that slide text can provide powerful contextual information for recovering missing speech segments, particularly in long-duration gaps where acoustic continuity alone becomes ambiguous.

Motivated by this observation, we propose a mul-

[1] https://github.com/yanir1/Speech-Inpainting.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 45455–45473
October 24-29, 2026 ©2026 Association for Computational Linguistics

timodal speech inpainting framework that leverages slide-derived text to reconstruct corrupted lecture audio. We reformulate speech inpainting as a conditional speech generation task using a Large Audio Language Model (LALM), which jointly models acoustic and linguistic information from large-scale audio-text pretraining (Arora et al., 2025; Su et al., 2025). To improve generation quality, we introduce two training enhancements: *Acoustic Similarity Groups* (ASG), which relax hard-label supervision by grouping acoustically similar tokens, and an auxiliary *Connectionist Temporal Classification* (CTC) objective that strengthens lexical alignment. We evaluate the proposed framework on the M3AV-CHI benchmark under realistic lecture-recording conditions with diverse speakers and acoustic environments. Results show that incorporating slide-derived context significantly improves long-duration speech reconstruction over an audio-only baseline, while preserving speaker consistency and acoustic quality.

Our key contributions are as follows:

- We introduce a new multimodal formulation of speech inpainting that leverages presentation slides as semantic grounding for reconstructing corrupted lecture audio.
- We propose an LALM-based generative framework enhanced with Acoustic Similarity Groups (ASG) and an auxiliary CTC objective to improve acoustic flexibility and lexical alignment.
- We demonstrate robust recovery of long-duration missing speech segments of up to 1.5 seconds under realistic natural recording conditions.
- We establish strong empirical results on the M3AV-CHI benchmark across heterogeneous speakers, accents, and acoustic environments.

2 Related Work

Audio-Only Speech Inpainting. Traditional speech restoration methods mainly relied on statistical signal processing techniques and were typically limited to very short gaps of several tens of milliseconds, as commonly addressed in Packet Loss Concealment (PLC) systems (Adler et al., 2012; Mokrý and Rajmic, 2020; Dissen et al., 2024). More recent deep learning approaches shifted toward spectrogram reconstruction using masked inpainting formulations that exploit the temporal and spectral context surrounding the missing region (Huang et al., 2022; Niizumi et al., 2022). Other works explored convolutional architectures, such as the end-to-end U-Net proposed by Kegler et al. (2020), which demonstrated recovery of time-frequency segments up to 400 ms and showed that deep feature supervision improves intelligibility compared to standard pixel-wise objectives. Despite these advances, audio-only speech inpainting methods remain largely limited to short-duration gaps. In addition, many recent approaches emphasize spectrogram reconstruction quality or downstream classification performance rather than intelligibility-oriented metrics such as Word Error Rate (WER) (Huang et al., 2022; Niizumi et al., 2022).

Audio-Visual Multimodal Grounding. To mitigate the "large gap" problem, where acoustic context alone is insufficient for reliable reconstruction, recent research has increasingly focused on multimodal integration. In particular, incorporating visual information—such as lip movements and facial dynamics—has been shown to preserve phonetic consistency across missing segments and substantially improve speech regeneration accuracy (Morrone et al., 2021; Montesinos et al., 2023). For example, ReVISE (Hsu et al., 2023) employs a self-supervised framework that enforces both acoustic coherence and synchronization with the speaker's visual cues during reconstruction. More recently, sequence-to-sequence approaches have demonstrated the ability to recover gaps of up to 1600 ms by formulating speech inpainting as a cross-modal translation problem (Elyaderani and Shirani, 2024). Despite these advances, such methods remain limited in practical deployment, as they rely on continuous, high-quality facial video streams—an assumption that does not always hold in real-world telecommunication settings.

Text-Conditioned Inpainting and Synthesis. Several recent approaches (Borsos et al., 2022; Le et al., 2023; Peng et al., 2024) condition the reconstruction process on the explicit ground-truth missing text, effectively reframing speech inpainting as a constrained Text-to-Speech (TTS) task focused on intelligibility preservation. For example, *SpeechPainter* (Borsos et al., 2022) demonstrates inpainting of gaps up to 1 second using unaligned transcripts while preserving speaker identity, prosody, and acoustic conditions. However, these methods fundamentally rely on access to the target missing words, limiting their applicability to scenarios where accurate

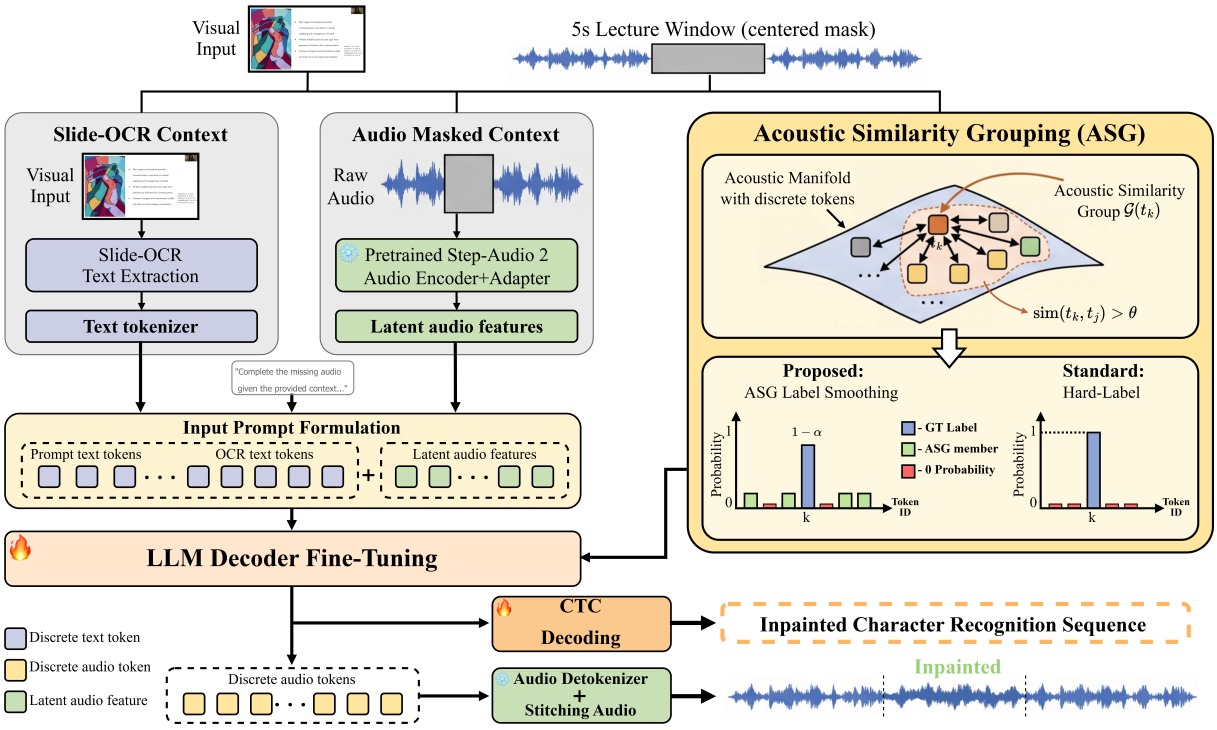


Figure 1: Overview of the proposed multimodal framework. The system integrates Slide-OCR text tokens and masked latent audio features, followed by an instruction prompt to a Large Language Model (LLM) decoder. Training is enhanced by an ASG module that modifies the Cross-Entropy (CE) loss term. During training, the framework makes use of a CTC loss as an auxiliary task. The network outputs discrete audio tokens, which are decoded to generate the inpainted audio and reconstruct the complete recovered audio waveform.

transcripts are available, a condition that is rarely satisfied in real-world speech restoration settings.

Speech Editing with LALMs. The emergence of LALMs has marked a shift toward generative speech manipulation utilizing a Language Model (Chen et al., 2025; Huang et al., 2024; LLM-Core Xiaomi, 2025; Yan et al., 2025). These frameworks treat speech synthesis as a language modeling task, performing speech manipulation directly through an LALM.

3 Problem Formulation

We introduce a new *multimodal speech inpainting* task for reconstructing a corrupted audio signal associated with a relevant, but not completely aligned, text. This can correspond to a recorded virtual lecture or webinar, which has a continuous acoustic stream paired with synchronous visual presentation slides. These types of lectures are inherently susceptible to uncontrolled acoustic degradations, ranging from network jitter and hardware malfunctions to ambient environmental interference. Such artifacts may damage the audio severely, resulting in missing temporal segments. We propose to recover these missing segments using a new *multimodal approach*, leveraging the inherent correlation between the lecturer's spoken signal and the literal content present on the accompanying slides.

Formally, let $\mathbf{x}_\mathrm{o} = (x_1, \ldots, x_N)$ denote the sample sequence of the original clean audio, and $\mathbf{c}$ a textual content extracted from an accompanying slide via Optical Character Recognition (OCR). We define the missing segment to be M samples masked out from $\mathbf{x}_\mathrm{o}$, denoted by $\mathbf{x}_\mathrm{t} = (x_{i+1}, \ldots, x_{i+M})$. The resulting masked audio is defined as $\mathbf{x}_\mathrm{m} = \mathbf{x}_\mathrm{o} \setminus \mathbf{x}_\mathrm{t} = (x_1, \ldots, \underbrace{\mathbf{0}}_{M \text{ zeros}}, \ldots, x_N)$. We assume that each missing segment $\mathbf{x}_t$ is of typical length of 10 – 1500 ms, corresponding to a complete (or partial) word, w, which appears in the given context slide, i.e. $w \in \mathbf{c}$. We wish to fill in the missing audio segment given the masked signal $\mathbf{x}_\mathrm{m}$ and the context $\mathbf{c}$.

A key challenge arises when the missing word is absent from the OCR, i.e., $w \notin \mathbf{c}$, as it may be unrelated to the available OCR context, making reliable prediction difficult. We discuss this case in § 5.2 and address it using a dedicated Hybrid-Model trained on an extended dataset, as detailed in § 6.1.

4 Methodology

We propose a multimodal generative framework for the long-duration speech inpainting task, illustrated in Fig. 1. In contrast to conventional local interpolation methods, our approach reformulates speech inpainting as a conditional generation problem.

4.1 Utilizing Large Audio Language Model

Recently, LALMs have established themselves as the leading framework for state-of-the-art results in a wide spectrum of speech and text tasks, ranging from Automatic Speech Recognition (ASR) to various speech and audio understanding tasks (Lakhotia et al., 2021; Borsos et al., 2023; Yang et al., 2026). This performance is achieved through the vast amount of speech and text training data, which ensures robustness against diverse accents and non-stationary noise inherent in real-world recordings. By effectively bridging the modality gap between speech and textual context, these models offer a more holistic approach to audio-text tasks.

Utilizing an LALM, we reframe speech inpainting from a traditional signal-reconstruction problem into a speech-understanding and generation task. This approach can capture the complex dependencies between $\mathbf{x}_{\mathrm{m}}$ and the textual context, enabling the completion of larger masked segments. Moreover, it may help to overcome generalization bottlenecks of current methods. We specifically focus on Step-Audio-2-mini-Base (Team, 2025a) based on the former Step-Audio (Team, 2025b), due to its wide range of capabilities, strong performance, and relatively small size. The model consists of an audio encoder followed by an audio adapter, a Qwen2.5-7B as an LLM decoder (Team, 2024) and an audio detokenizer.

4.2 Model Training

Let $\mathcal{M}$ denote an LALM. These models typically utilize a discrete tokenization scheme for both audio and text, forming a unified codebook $\mathcal{V}$. For Step-Audio 2, the total codebook size is $|\mathcal{V}| = 158\text{K}$ and the audio part is of size $|\mathcal{V}_{\text{audio}}| = 6.6\text{K}$. Consequently, $\mathcal{M}$ autoregressively generates both text and audio tokens from $\mathcal{V}$. From this mixed-token output during inference, we can de-interleave the audio tokens and reconstruct the final inpainted waveform using the corresponding detokenizer. We fix the final learnable token in the sequence to a special audio-end token, forcing $\mathcal{M}$ to infer when to terminate generation solely from the masked audio context, without explicit knowledge of the target length.

We employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) for parameter-efficient fine-tuning of the model. The objective is to learn a mapping function $f : (\mathbf{x}_{\mathrm{m}}, \mathbf{c}) \rightarrow \hat{\mathbf{x}}_t$ such that the inpainted speech $\hat{\mathbf{x}}_t$ should be acoustically consistent with the surrounding signal $\mathbf{x}_{\mathrm{m}}$ and semantically consistent with the context $\mathbf{c}$, where $\mathbf{x}_{\mathrm{m}}$ denotes the masked audio, $\mathbf{c}$ the slide-derived textual context, and $\hat{\mathbf{x}}_{\mathrm{t}}$ the reconstructed speech segment. The model input consists of the tuple $(\mathbf{x}_{\mathrm{m}}, \mathbf{c}, \mathbf{p})$, where $\mathbf{p}$ is a task-specific instruction prompt (Appendix A). During training, we extract $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_k)$, the sequence of one-hot ground-truth labels formed by tokenizing the missing audio segment, i.e. $\mathbf{y} = \mathrm{tokenizer}(\mathbf{x}_{\mathrm{t}})$, where k is the number of audio tokens. At each decoding step i, the model $\mathcal{M}$ produces a hidden representation $\mathbf{h}_i \in \mathbb{R}^d$, which is projected into the vocabulary space: $\mathbf{z}_i = \mathbf{W}\mathbf{h}_i + \mathbf{b}$ with logits $\mathbf{z}_i \in \mathbb{R}^{|\mathcal{V}|}$. The predicted token distribution is then obtained via $\hat{\mathbf{y}}_i = \mathrm{softmax}(\mathbf{z}_i)$, where each entry $\hat{y}_i[t] = P_{\mathcal{M}}(y_i = t \mid \mathbf{y}_{<i}, \mathbf{x}_{\mathrm{m}}, \mathbf{c})$ represents the autoregressive probability assigned by the model to token $t \in \mathcal{V}$. This yields the sequence of predicted distributions $\hat{\mathbf{y}} = (\hat{\mathbf{y}}_1, \ldots, \hat{\mathbf{y}}_k)$.

Usually, the model is trained via standard CE loss. However, our task can allow for multiple plausible completions, as long as the completed segment $\hat{\mathbf{x}}_t$ maintains the following principles:

- C1 Perceptually similar to $\mathbf{x}_t$.
- C2 Preserves the content of the missing word w.
- C3 Preserves speaker identity.

In order to achieve C1, we propose a soft version of the CE loss, as presented in § 4.3. For handling C2, we define an auxiliary task for estimating the word w via an additional prediction head, as discussed in § 4.4. Finally, for C3, the speaker identity is preserved by the detokenizer, which receives a reference audio segment of the same speaker.

4.3 Acoustic Similarity Groups (ASG)

Standard LLM frameworks utilize a CE loss that treats all tokens as independent categories. In speech inpainting, this "hard-label" approach may be too restrictive since two different audio tokens may represent similar acoustic features. Following (Yanuka et al., 2026), we propose constructing ASGs that cluster audio tokens with

strong proximity on the acoustic manifold. The model is then guided along this manifold by permitting token substitutions within the same ASG, thereby encouraging acoustically consistent transitions during generation.

Construction of ASG. Unlike text tokens that represent inherently discrete subword units, audio tokens correspond to sounds lying on a continuous acoustic manifold. In order to quantify the relations along the acoustic manifold, we utilize the similarity between the corresponding vector embeddings. Let $\mathbf{e}_i$ be the embedding vector of token $t_i \in \mathcal{V}$ from the learned embeddings codebook of $\mathcal{M}$. We define the similarity between tokens i and j using their cosine similarity:

$$\text{sim}(\mathbf{t}_i, \mathbf{t}_j) = \frac{\mathbf{e}_i \cdot \mathbf{e}_j}{\|\mathbf{e}_i\| \|\mathbf{e}_j\|}. \quad (1)$$

Accordingly, we define groups of acoustically similar tokens based on their embedding similarity. Given a token $t_i \in \mathcal{V}$, its corresponding ASG $\mathcal{G}(t_i)$ is defined as:

$$\mathcal{G}(t_i) := \{t_j \in \mathcal{V} : \text{sim}(t_i, t_j) > \theta, \;\; i \neq j\}, \quad (2)$$

where θ is a threshold that controls the group size, with lower thresholds leading to larger groups that permit more prosodic and phonetic variation, further explained in Appendix D.

ASG Label Smoothing. We adopt Label Smoothing (LS) (Szegedy et al., 2016; Vaswani et al., 2017), a regularization technique that replaces hard labels with softened targets by retaining a probability mass of $1 - \beta$ on the ground-truth label and distributing the remaining $\beta \in [0, 1]$ uniformly across all classes. We propose to integrate ASG and LS in order to improve the acoustical perception of the model during training. Based on the ASG groups, we replace standard one-hot labels with smoothed uniform weights distributed across each group. This relaxes the requirement for exact probability assignments during prediction, reducing model overconfidence and improving generalization. Specifically, we reformulate the hard-label vector $\mathbf{y}_i$ of a token $t_i \in \mathcal{V}_{\text{audio}}$ and $t_j \in \mathcal{V}_{\text{audio}}$, to take the ASG-dependent form marked as $\tilde{\mathbf{y}}_i$:

$$\tilde{y}_i[j] = (1 - \beta) \cdot y_i[j] + \frac{\beta}{|\mathcal{G}(t_i)|} \cdot \mathbb{1}_{t_j \in \mathcal{G}(t_i)} \quad (3)$$

where $\beta \in [0, 1]$ controls the weight spread uniformly across the ASG partners of t_i. Accordingly, we define a new ASG-CE loss between estimated tokens $\hat{\mathbf{y}}$ and smoothed ASG targets $\tilde{\mathbf{y}}$:

$$\mathcal{L}_{\text{ASG}} = -\frac{1}{k} \sum_{i=1}^{k} \sum_{t \in \mathcal{V}} \tilde{y}_i[t] \log \hat{y}_i[t]. \quad (4)$$

Transitioning from the hard targets $\mathbf{y}$ to the smoothed ASG targets $\tilde{\mathbf{y}}$ alleviates the ambiguity arising from the many-to-one structure of the acoustic manifold. This relaxation enables the model to explore acoustically valid phonetic alternatives within the same ASG, rather than enforcing a single deterministic target.

4.4 CTC Auxiliary Task

In addition to predicting the tokens of the missing audio segment, we introduce an auxiliary task that estimates the missing word in the language domain using a CTC module. We feed the sequence of hidden states $\mathbf{h} = (\mathbf{h}_1, \ldots, \mathbf{h}_k)$ into a CTC head comprising a linear layer followed by a softmax. This yields the probability distribution required for CTC decoding. The objective of the CTC-decoder is to accurately align the output representations of $\mathcal{M}$ with the target word w.

Finally, our proposed framework jointly optimizes both speech inpainting and speech recognition objectives. We define the overall training objective as the weighted sum of the ASG CE loss and the auxiliary CTC loss, balanced by a hyperparameter λ, as follows:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{ASG}} + \lambda \cdot \mathcal{L}_{\text{CTC}}. \quad (5)$$

5 Experimental Setup

In this section, we evaluate the proposed framework for the task of multimodal speech inpainting. To enable a realistic and reproducible evaluation, we introduce a new benchmark based on natural lecture recordings accompanied by presentation slides, incorporating varying mask durations and corresponding missing words. We further propose a set of evaluation metrics that jointly assess reconstruction fidelity, intelligibility, and semantic consistency of the restored speech.

5.1 Multimodal Dataset Creation

We derive a specialized multimodal dataset from the **CHI** (Computer Science) track of the M3AV collection (Chen et al., 2024). This subset comprises 53.2 hours of technical presentations across

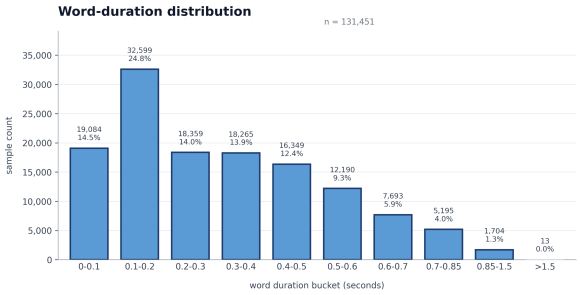


Figure 2: Dataset Distribution by $\mathbf{x}_t$ duration.

Attribute	Train	Val	Test
Total Samples	104,941	13,647	12,850
Environment	Virtual Video Conference		
Total Speaker Count	660 unique		
Slide Duration	20.8s / slide		
L_{total}	5.0s		
L_{mask}	0.01s – 1.5s		

Table 1: M3AV-CHI Corpus Statistics. Sample Duration (L_{total}). Mask Duration (L_{mask}).

9,195 slides, capturing uncontrolled natural acoustic environments typical of virtual video conferencing (VC), characterized by compression artifacts and varied background noise. To ensure signal integrity, we utilize Whisper large-v3 (Radford et al., 2023) as a cross-modal verifier; samples are retained only if the ASR transcription achieves a 100% character-level match with the given dataset's ground-truth transcriptions, effectively pruning acoustic disfluencies and misalignments of the dataset. The resulting filtered dataset consists of 131K samples out of the original 153K. The distribution over mask duration is shown in Fig. 2. The recordings are partitioned into training, validation, and test sets using an 80%/10%/10% split, while ensuring that neither speakers nor lectures overlap across the resulting subsets.

M3AV includes audio recordings, transcriptions, word-level timestamp annotations, presentation slides, and the slides' text extracted with OCR. For each lecture $\ell \in \mathcal{L}$ in the corpus, we extract a sequence of OCR-derived contexts $\mathcal{C}_\ell$, where each context $c_i \in \mathcal{C}_\ell$ corresponds to the textual content of a single slide within lecture ℓ. For each slide i, we define $\mathcal{W}_i$ as the set of spoken words that appear both in the transcription and in the slide's content. This grounding of a spoken word w in c_i is done in order to ensure that model $\mathcal{M}$ is supervised by only relevant semantic anchor words, potentially capable of inpainting back $\mathbf{x}_m$ audio. We extract 5-second audio clips centered around the matched spoken words, and construct a benchmark consisting of unique samples containing the masked audio $\mathbf{x}_m$, the target missing segment $\mathbf{x}_t$, and the corresponding missing word w. Summary of the benchmark statistics is provided in Table 1.

5.2 Word in OCR Premise

To better understand the role of OCR information, we analyzed the entire dataset and found that approximately 153K spoken words also appear in the corresponding slide OCR, whereas 329K do not. Thus, roughly one-third of the spoken words are present in the slides, suggesting that OCR provides a valuable source of contextual information in many cases. We further analyzed the relationship between the spoken transcript and the OCR text. The simplest scenario occurs when the missing word and its immediate neighboring words (one preceding and one following) appear consecutively in the OCR. This occurs in only 28% of the words that appear in the slides. In the remaining 72% of cases, the neighboring words are not adjacent in the OCR, making recovery substantially more challenging. This analysis demonstrates that the proposed framework is not merely performing local textual matching, but instead must learn to integrate information across the speech signal and OCR under much weaker contextual alignment.

5.3 Model and Training Setup

We adapt the Step-Audio 2 architecture (Team, 2025a), an end-to-end multimodal LALM, using the Step-Audio-2-mini-Base variant. To specialize the model while preserving the audio-text capabilities of the backbone, we employ LoRA (Hu et al., 2022). Specifically, we apply LoRA to the QKVO projection matrices within the attention blocks, using a high-capacity configuration with rank $r = 64$ and scaling factor $\alpha = 128$. This setup (Appendix B) achieves strong speech reconstruction performance while requiring optimization of only 40.3M parameters. In our experiments, we investigated a threshold range of $\theta \in [0.38, 0.45]$ and chose $\theta = 0.42$, and strictly upper bounded each ASG group cardinality $|\mathcal{G}| \leq 5$.

5.4 Evaluation

Evaluating long-duration speech inpainting presents a unique challenge because the task is inherently *generative* rather than reconstructive. Thus, traditional signal-level metrics, such as L1 loss or MSE, often penalize valid predictions that deviate from the exact ground-truth spectrogram

Table 2: Inpainting performance over M3AV-CHI. Original audio achieves DNSMOS=2.871 and Sim=1. Mismatch indicates the absolute duration error between the generated inpainting and the target audio segments. Since a single audio token corresponds to 40 ms, a mismatch below 40 ms indicates a high precision of sub-token alignment.

Model	ASG	CTC	Inpainting Performance					
			WER ↓	CER ↓	PER ↓	DNSMOS ↑	Sim ↑	Mismatch ↓
$\mathcal{M}_{\text{w/o OCR}}$	✓	✓	52.31%	43.75%	48.94%	2.848	0.976	22.1 ms
VOICECRAFT	-	-	5.24%	4.69%	8.93%	2.984	0.930	-
$\mathcal{M}_{\text{base}}$	✗	✗	13.34%	10.92%	11.12%	2.848	0.985	26.1 ms
$\mathcal{M}_{+\text{ASG}_{\text{rand}}}$	random	✗	13.30%	10.94%	11.00%	2.850	0.985	25.2 ms
$\mathcal{M}_{+\text{ASG}_{\text{rand}}+\text{CTC}}$	random	✓	12.68%	10.34%	10.43%	2.846	0.985	25.3 ms
$\mathcal{M}_{+\text{CTC}}$	✗	✓	12.68%	10.67%	10.54%	2.848	0.985	28.7 ms
$\mathcal{M}_{+\text{ASG}}$	✓	✗	13.06%	10.69%	11.01%	2.847	0.985	35.8 ms
$\mathcal{M}_{\text{full}}$	✓	✓	**12.15%**	**10.06%**	**10.23%**	2.845	0.985	25.6 ms

while still being perfectly good. Given our LALM-based framework, we prioritize metrics that capture semantic intelligibility and linguistic orientation over raw signal reconstruction. All metrics are computed over the inpainted segment.

Textual Content Analysis. We evaluate whether the textual content of the inpainted segment matches the missing word w. Transcription is performed using Whisper large-v3. Since ASR models rely on contextual information to produce accurate transcriptions, we do not apply the ASR model solely to the inpainted segment. Instead, we first reconstruct the full audio signal by stitching the predicted segment back into the masked audio, and then run ASR over the complete utterance. This provides a more reliable transcription and avoids ambiguity in the cases of homophones, e.g., "two" vs. "to" or "too". From the resulting transcript, we extract the text corresponding to the inpainted boundaries and compute the standard WER against the ground-truth word w. In addition, we compute the Character Error Rate (CER) to provide a more fine-grained evaluation of the inpainted sequence accuracy at the subword level.

Phonetic Error Rate (PER). PER measures phoneme-level accuracy, capturing pronunciation errors not reflected by word- or character-level metrics. We obtain IPA phoneme sequences for the ground-truth and inpainted segments using a pretrained `wav2vec2-large-xlsr-53` model (Conneau et al., 2021), and compute PER by applying standard WER calculation to the space-separated phonemes.

Acoustic Quality and Identity. To evaluate acoustic naturalness and voice consistency, we use DNSMOS (Reddy et al., 2022) and the mean WavLM-based speaker similarity (Chen et al., 2022), respectively. For DNSMOS, we report the overall quality (OVRL) score of the stitched audio.

5.5 Baselines

To the best of our knowledge, there is currently no directly comparable model for long-duration speech inpainting under our setting. Existing approaches are primarily designed for audio-only reconstruction, or alternatively rely on audio-visual cues that are not applicable in our scenario. In contrast, our framework additionally leverages contextual information from presentation slides, introducing a fundamentally different conditioning paradigm. Furthermore, prior works typically focus on short missing segments, and their performance degrades substantially for large gaps.

We assess the capabilities of our proposed framework by isolating the contributions of its individual components, considering the following variants:

- **Audio-Only (No Slide Context):** A baseline trained exclusively on masked audio sequences $\mathbf{x}_{\text{m}}$, without providing slide context $\mathbf{c}$. This quantifies the necessity of cross-modal information in long-duration gaps.
- **Standard CE:** A multimodal baseline utilizing both audio and textual context but trained with standard CE loss without ASG or CTC.
- **Ablating the ASG Method:** We introduce a randomized ASG variant, where tokens are grouped arbitrarily without considering the similarity-based proximity defined in Eq. (1). In both options, the smoothing factor is set to $\beta = 0.08$.
- **Ablating the CTC Module:** Finally, across the aforementioned configurations, we evaluate both

with and without CTC objective as an auxiliary task (Eq. (5)), choosing $\lambda = 0.25$.

In addition, we evaluate the following baselines:

- **MSM-MAE[2](Niizumi et al., 2022):** MSM-MAE reconstructs masked spectrograms using an MSE objective. We initialized the model from the released pre-trained weights and fine-tuned it on our training set. We evaluate the MSM-MAE model by fine-tuning a HiFi-GAN vocoder (Kong et al., 2020) to synthesize an audio waveform from the reconstructed mel-spectrogram. We further break down its evaluation by gap duration, as detailed in Section C.5.
- **VoiceCraft[3](Peng et al., 2024):** A state-of-the-art text-guided speech editing model. We used VOICECRAFT's speech editing pipeline, utilizing the largest publicly available checkpoint (`giga830M`). VOICECRAFT demands the complete ground-truth transcription together with the masked speech, allowing it to generate the complete audio by using explicit linguistic supervision. Therefore, it serves as an ideal baseline, whereas our model receives only slide-derived context.

6 Results

Complete test results are presented in Table 2, with bootstrap confidence intervals provided in Appendix C.6. First, the large performance gap between $\mathcal{M}_{\text{w/o OCR}}$ and all slide-conditioned variants demonstrates the significant contribution of contextual information from the presentation slides compared to audio-only inpainting. In particular, leveraging slide context substantially improves intelligibility and phonetic reconstruction, as reflected by the WER, CER, and PER metrics.

In addition, both ASG and CTC independently improve reconstruction performance over the baseline model $\mathcal{M}_{\text{base}}$. Combining the two yields the best overall configuration, $\mathcal{M}_{\text{full}}$, which achieves the lowest error rates of 12.15%, 10.06%, and 10.23% in WER, CER, and PER, respectively. This corresponds to a relative WER improvement of 76.8% over $\mathcal{M}_{\text{w/o OCR}}$ and 8.9% over $\mathcal{M}_{\text{base}}$. Performance as a function of mask length is further illustrated in Fig. 3. As expected, reconstruction becomes more challenging as the missing gap duration increases. Nevertheless, even for the longest

[2] https://github.com/nttcslab/msm-mae.

[3] https://github.com/jasonppy/voicecraft.

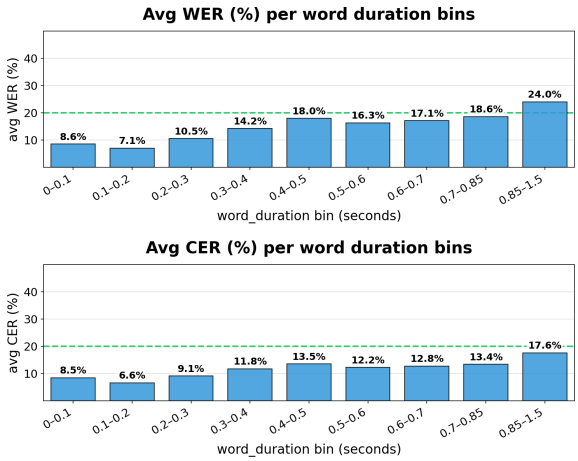


Figure 3: $\mathcal{M}_{\text{full}}$ WER(%) and CER(%) over $\mathbf{x}_t$ duration.

and most difficult gaps, the model still achieves a word recognition rate of 76% and a character recognition rate exceeding 82%.

Finally, all configurations maintain consistently high DNSMOS and speaker similarity scores. This indicates that the generated speech preserves both audio quality and speaker characteristics. Notably, these qualities remain stable across all variants and are not degraded by the additional fine-tuning process. In addition, the mismatch in generated duration remains small and relatively uniform across all methods, demonstrating stable temporal alignment.

To further demonstrate the capabilities of our framework, Fig. 4 presents the waveforms, spectrograms, and transcriptions of the original, masked, and inpainted signals alongside the corresponding contextual slide. Although the generated tokens are not always identical to the ground-truth ones, the reconstructed spectrograms remain acoustically similar. This observation further strengthens our motivation for using LS based on ASG during training.

6.1 Hybrid Model: Missing Word not in Slide

We test the robustness of our proposed method against missing words using a custom Hybrid dataset, in which 80% of the masked words are present in the slide's OCR, while the remaining 20% are absent. The non-OCR samples were selected to match the same word-duration distribution as the OCR-present samples in Fig. 2, resulting in subsets with comparable word-duration distributions. We then trained a model with the same settings as $\mathcal{M}_{\text{full}}$ on this hybrid dataset and evaluated it separately on the two test subsets.

A detailed analysis is provided in Appendix C.1. Overall, the Hybrid-Model achieves better per-

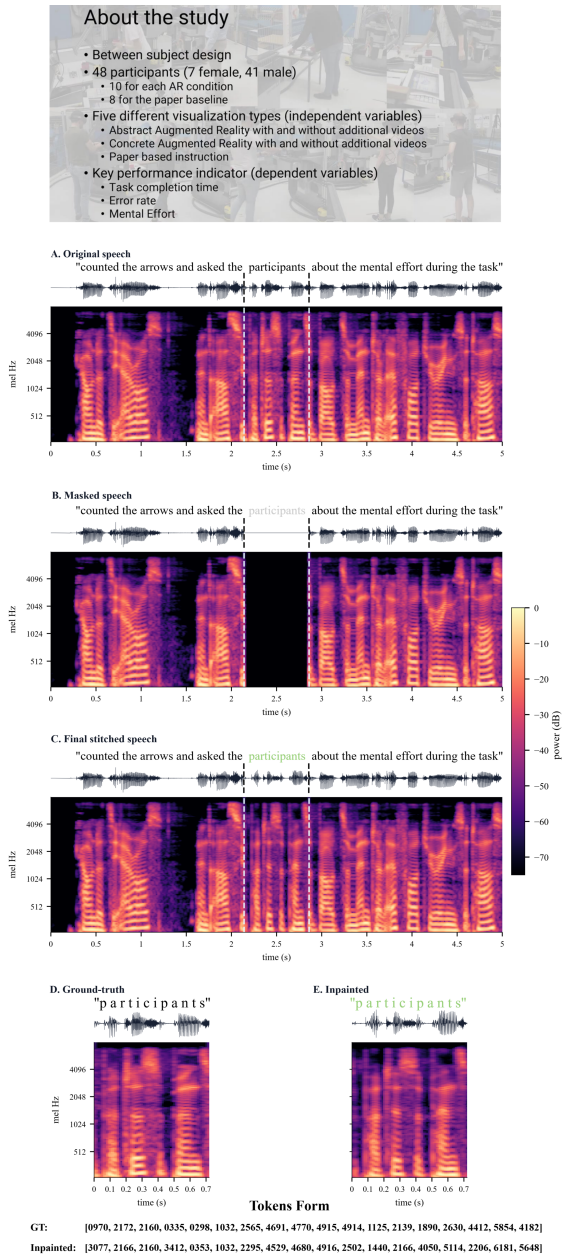


Figure 4: lecture: "CHI-24F31" | slide_idx: 3 | masked_word: "participants" | mask_duration: 720ms. Spectrogram analysis of the speech inpainting process. Utilizing context from the visual slide (top). The stages show speech waveform and mel-spectrogram of: (A) original speech, (B) masked speech ("participants") removed, and (C) final stitched output with synthesized speech. Panels (D) and (E) provide a zoom-in comparison of ground-truth versus inpainted spectrograms. Bottom row details the discrete audio token sequences.

formance (Fig. 5), largely preserving the original OCR-trained model's performance on in-OCR words (Fig. 5b) while improving performance on short out-of-OCR words (Fig. 5c). This demonstrates that a single model can effectively handle both scenarios, resulting in a more robust and practical framework for real-world applications where the availability of OCR information varies. Finally, while the current approach already achieves reasonable performance on words that are absent from the OCR, future work will investigate leveraging the language understanding capabilities of LALMs, together with the contextual information provided by the presentation slides, to further improve the reconstruction of these words.

6.2 Additional Results

We briefly highlight additional results detailed in Appendix C. Appendix C.2 presents an extensive ablation study of key architectural components and hyperparameters, including LoRA configurations, ASG thresholds, and CTC weights. Performance remains highly stable across settings, demonstrating the method's robustness and limited need for hyperparameter tuning. Additionally, we evaluated $\mathcal{M}_{\text{full}}$ in a partial-word masking setting (Appendix C.3), verifying that the model effectively reconstructs sub-word segments and yields expected improvements due to reduced overall mask durations. Furthermore, we demonstrate that our approach is robust to unseen off-center masks (Appendix C.4). Although trained exclusively on centered segments, the model achieves reasonable zero-shot performance, indicating that its learned representations generalize to new mask positions rather than simply memorizing spatial configurations.

7 Conclusion

We introduced a novel multimodal framework for long-duration speech inpainting that leverages presentation slides as semantic grounding for reconstructing corrupted lecture audio. By reframing speech inpainting from a purely signal-reconstruction problem into a contextual generation task, our approach enables robust recovery of missing segments ranging from short gaps to severe degradations of up to 1.5 seconds. Our framework combines an LALM with the proposed ASG objective and an auxiliary CTC loss, allowing the model to better navigate the acoustic manifold while preserving intelligibility and speaker consistency. Experimental results on the M3AV-CHI benchmark demonstrate that incorporating slide-derived context substantially improves reconstruction quality over an audio-only baseline under real-world recording conditions.

8 Limitations

While the proposed framework demonstrates strong performance for long-duration speech inpainting, several limitations remain. First, the model operates in a discrete latent audio space with a temporal resolution of approximately 40 ms, limiting its ability to accurately reconstruct very short acoustic events or fine-grained phonetic transitions below the token resolution. As a result, the framework is primarily optimized for semantic and perceptual continuity rather than exact waveform recovery.

Second, our approach assumes the availability of semantically relevant slide context. Although this assumption is valid for many academic presentations and webinars, the missing spoken content may not always explicitly appear in the accompanying slides. In such cases, as demonstrated, the effectiveness of contextual grounding may decrease. Future work could explore extending the receptive context to neighboring slides, lecture transcripts, or external knowledge sources in order to improve robustness when direct lexical overlap is absent.

Third, our evaluation focuses on lecture recordings within the M3AV-CHI domain, which primarily consists of English academic presentations conducted in virtual conferencing environments. Consequently, the generalization capabilities of the framework across other domains, languages, conversational settings, or highly spontaneous speech remain underexplored.

In addition, while the proposed ASG mechanism improves perceptual flexibility during generation, the grouping process is currently based solely on embedding-space cosine similarity with manually selected thresholds. More principled or adaptive approaches for constructing acoustically meaningful token neighborhoods could further improve generation quality and stability.

9 Acknowledgments

This work was supported in part by the Israel Science Foundation under Grant 2094/25, in part by the Israeli Ministry of Innovation, Science and Technology under Grant 1001818518, and in part by the Israel Innovation Authority's Magneton program under Grant 86466.

References

Amir Adler, Valentin Emiya, Maria G. Jafari, Michael Elad, Rémi Gribonval, and Mark D. Plumbley. 2012. Audio inpainting. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(3):922–932.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *Transactions on Machine Learning Research*, October-2025. Publisher Copyright: © 2025, Transactions on Machine Learning Research. All rights reserved.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533.

Zalán Borsos, Matthew Sharifi, and Marco Tagliasacchi. 2022. SpeechPainter: Text-conditioned Speech Inpainting. In *Interspeech 2022*, pages 431–435.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Sanyuan Chen, Chengyi Wang, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2025. Neural codec language models are zero-shot text to speech synthesizers. *IEEE Transactions on Audio, Speech and Language Processing*, 33:705–718.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024. M^3AV: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060. Association for Computational Linguistics.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Interspeech 2021*, pages 2426–2430.

Yehoshua Dissen, Shiry Yonash, Israel Cohen, and Joseph Keshet. 2024. Enhanced ASR Robustness to Packet Loss with a Front-End Adaptation Network. In *Interspeech 2024*, pages 5008–5012.

Mahsa Kadkhodaei Elyaderani and Shahram Shirani. 2024. Robust multi-modal speech in-painting: A sequence-to-sequence approach. *arXiv preprint arXiv:2406.00901*.

Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, and Yossi Adi. 2023. ReVISE: self-supervised speech resynthesis with visual input for universal and generalized speech regeneration. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18796–18806.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer. 2022. Masked autoencoders that listen. In *Advances in Neural Information Processing Systems*, volume 35, pages 28708–28720. Curran Associates, Inc.

Rongjie Huang, Ruofan Hu, Yongqi Wang, Zehan Wang, Xize Cheng, Ziyue Jiang, Zhenhui Ye, Dongchao Yang, Luping Liu, Peng Gao, and Zhou Zhao. 2024. InstructSpeech: Following speech editing instructions via large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 19886–19903. PMLR.

Mikolaj Kegler, Pierre Beckmann, and Milos Cernak. 2020. Deep Speech Inpainting of Time-Frequency Masks. In *Interspeech 2020*, pages 3276–3280.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*, volume 36, pages 14005–14034. Curran Associates, Inc.

LLM-Core Xiaomi. 2025. MiMo-Audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*.

Ondřej Mokrý and Pavel Rajmic. 2020. Audio inpainting: Revisited and reweighted. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2906–2918.

Juan Felipe Montesinos, Daniel Michelsanti, Gloria Haro, Zheng-Hua Tan, and Jesper Jensen. 2023. Speech inpainting: Context-based speech synthesis guided by video. In *Interspeech 2023*, pages 4459–4463.

Giovanni Morrone, Daniel Michelsanti, Zheng-Hua Tan, and Jesper Jensen. 2021. Audio-visual speech inpainting with deep learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6653–6657.

Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. 2022. Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation. In *HEAR: Holistic Evaluation of Audio Representations (NeurIPS 2021 Competition)*, volume 166 of *Proceedings of Machine Learning Research*, pages 1–24. PMLR.

Pitambar Paudel. 2021. Online education: Benefits, challenges and strategies during and after COVID-19 in higher education. *International Journal on Studies in Education (IJonSE)*, 3(2).

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. 2024. Voicecraft: Zero-shot speech editing and text-to-speech in the wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12442–12462.

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 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Chandan K A Reddy, Vishak Gopal, and Ross Cutler. 2022. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 886–890.

Tae Eun Shim and Song Yi Lee. 2020. College students' experience of emergency remote teaching due to covid-19. *Children and youth services review*, 119:105578.

Yi Su, Jisheng Bai, Qisheng Xu, Kele Xu, and Yong Dou. 2025. Audio-language models for audio-centric tasks: A survey. *arXiv preprint arXiv:2501.15177*.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2818–2826.

Qwen Team. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Step-Audio Team. 2025a. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Step-Audio Team. 2025b. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. 2017. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language*, 46:535–557.

Canxiang Yan, Chunxiang Jin, Dawei Huang, Haibing Yu, Han Peng, Hui Zhan, Jie Gao, Jing Peng, Jingdong Chen, Jun Zhou, Kaimeng Ren, Ming Yang, Mingxue Yang, Qiang Xu, Qin Zhao, Ruijie Xiong, Shaoxiong Lin, Xuezhi Wang, Yi Yuan, and 6 others. 2025. Ming-UniAudio: Speech LLM for joint understanding, generation and editing with unified representation. *arXiv preprint arXiv:2511.05516*.

Dongchao Yang, Yuanyuan Wang, Dading Chong, Songxiang Liu, Xixin Wu, and Helen Meng. 2026. UniAudio 2.0: A unified audio language model with text-aligned factorized audio tokenization. *arXiv preprint arXiv:2602.04683*.

Moran Yanuka, Paul Dixon, Eyal Finkelshtein, Daniel Rotman, and Raja Giryes. 2026. Principled coarse-grained acceptance for speculative decoding in speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 16692–16696.

A Prompt Instruction

The model was trained and tested using a specific prompt instruction, alongside the masked audio and the supplementary accompanying slide context. The original prompt was constructed as follows:

> "You are given input audio, the audio is masked exactly in the middle. The masked audio is a word, you need to output speech of the word that is missing in the audio. The missing word exists inside the text provided by the user."

To further assess the prompt significance, we replaced the aforementioned prompt with the following more generic instruction:

> "You are given input masked speech. You need to output the missing speech segment in the audio. Utilize the text context provided by the user."

The model trained with this prompt achieved performance comparable to that reported in the main part, obtaining WER $= 12.49\%$ and CER $= 10.25\%$. This indicates that the proposed approach is not sensitive to the exact wording of the prompt.

B Experimental Hyperparameters

Table 3 details the experimental configuration to facilitate reproducibility. All training and evaluation were conducted on a single NVIDIA H100 GPU using the Step-Audio-2-mini-Base backbone. A single full training run took approximately 11 hours.

Category	Value
Epochs	5
Learning Rate	5×10^{-5} (Cosine Decay)
Warmup Ratio	0.05
Effective Batch Size	16
Label Smoothing (β)	0.08
ASG-Threshold (θ)	0.42
CTC-parameter (λ)	0.25
LoRA Rank (r)	64
LoRA Alpha (α)	128
LoRA Dropout	0.05
Target Modules	$q_proj, v_proj, k_proj, o_proj$

Table 3: Training and LoRA configurations.

C Additional Results

C.1 Hybrid Training Results

We evaluated our Hybrid-Model trained on a mixture containing 80% in-OCR and 20% out-of-OCR samples. Figure 5 details the error-rate distributions for all test words, in-OCR words, and out-of-OCR words, and compares the performance of $\mathcal{M}_{\text{hybrid}}$ with that of $\mathcal{M}_{\text{full}}$.

C.2 Ablation Experiments

We further investigated the impact of the main design choices through an extensive ablation study, varying the key hyperparameters and architectural components while keeping all other settings fixed. The results are summarized in Table 4.

Across all ablations, the proposed framework exhibits stable performance under a wide range of hyperparameter settings. Most metrics vary only slightly around the best configuration, indicating that the method is robust and does not require extensive hyperparameter tuning to achieve strong performance.

C.3 Partial-Word Masking

Without any additional training, we evaluated the $\mathcal{M}_{\text{full}}$ model on a more flexible masking setting in which only sub-word segments were masked. Specifically, for each masked word, we randomly retained 50-100% of the original mask, resulting in masks covering only part of the word. The results are summarized in Table 5.

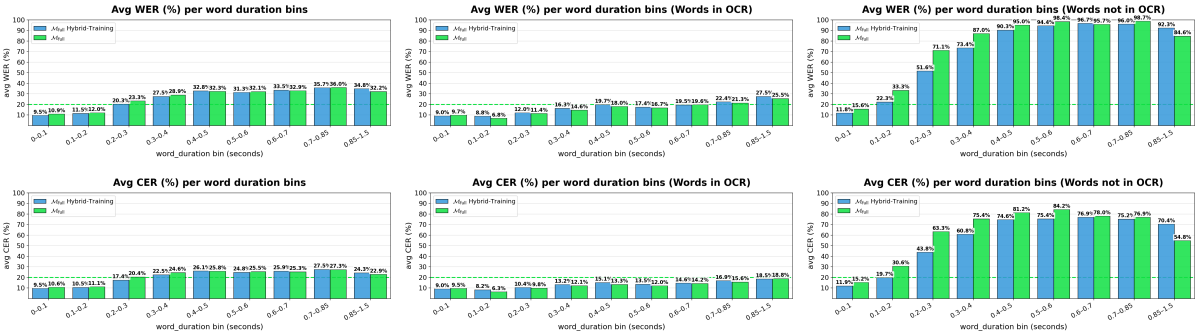
Table 5: Partial words test ($\mathcal{M}_{\text{Partial}}$) vs. full words test ($\mathcal{M}_{\text{full}}$). Both evaluations use the same model weights but different masking conditions.

Model	WER%	CER%	PER%	DNSMOS	Sim	Mismatch
$\mathcal{M}_{\text{Partial}}$	11.86	9.03	10.29	2.844	0.988	28.3 ms
$\mathcal{M}_{\text{full}}$	12.15	10.06	10.23	2.845	0.985	25.6 ms

These results demonstrate that the model can effectively reconstruct partially masked words as well. The improved performance, compared to the original main results, is expected, as partial-word masking reduces the average masked duration by approximately 25% compared to the original evaluation.

C.4 Off-Center Masking

We evaluated the same model $\mathcal{M}_{\text{full}}$ in a more challenging setting where the missing segment was randomly positioned between 1.5 s and 3.5 s within



(a) All target words (b) In-OCR target words (c) Out-of-OCR target words

Figure 5: WER (top row) and CER (bottom row) as functions of target-segment duration for M_{hybrid} (blue) and M_{full} (green): (a) all target words, (b) target words present in the slide OCR, and (c) target words absent from the slide OCR.

Table 4: Ablation Study Results

Param. Type	**Value**	**WER** ↓	**CER** ↓	**PER** ↓	**DNSMOS** ↑	**Sim** ↑	**Mismatch** ↓
LoRA	$r = 8, \alpha = 16$	15.30%	12.64%	12.78%	2.842	0.984	29.0 ms
	$r = 16, \alpha = 32$	13.98%	11.50%	11.57%	2.845	0.985	22.4 ms
	$r = 32, \alpha = 64$	12.92%	10.67%	10.97%	2.851	0.986	21.0 ms
	$r = 128, \alpha = 256$	12.44%	10.21%	10.12%	2.851	0.986	20.3 ms
	$r = 32, \alpha = 32$	13.76%	11.25%	11.55%	2.846	0.985	22.2 ms
	$r = 64, \alpha = 64$	12.73%	10.50%	10.70%	2.849	0.986	26.0 ms
	$r = 128, \alpha = 128$	12.44%	10.29%	10.15%	2.850	0.986	22.9 ms
ASG	$\lvert\mathcal{G}\rvert \leq 20$	12.60%	10.35%	10.41%	2.852	0.985	20.8 ms
	$\theta = 0.38$	12.42%	10.23%	10.20%	2.851	0.986	22.4 ms
	$\theta = 0.45$	12.78%	10.51%	10.48%	2.849	0.985	24.5 ms
CTC	$\lambda = 0.1$	12.93%	10.69%	10.90%	2.853	0.985	26.9 ms
	$\lambda = 0.4$	12.29%	10.11%	10.13%	2.853	0.985	23.0 ms
Loss (smooth)	$\beta = 0.05$	12.64%	10.34%	10.46%	2.850	0.986	22.7 ms
	$\beta = 0.15$	12.24%	10.10%	10.20%	2.852	0.986	22.8 ms
	$\beta = 0.20$	12.44%	10.20%	10.24%	2.852	0.986	23.9 ms
Prompt	More general	12.49%	10.25%	**10.07%**	2.854	0.986	27.3 ms
$\mathcal{M}_{\text{full}}$ **(original)**	*Original configuration	**12.15%**	**10.06%**	10.23%	2.845	0.985	25.6 ms

* Original configuration parameters: **LoRA:** r $= 64$, scaling $= 128$ | **ASG:** $\lvert\mathcal{G}\rvert \leq 5, \theta = 0.42$ | **CTC:** $\lambda = 0.25$ | **Loss:** $\beta = 0.08$

the 5 s audio clip. Importantly, the model was trained exclusively on centered masks and received no additional training for this evaluation. Results are presented in Table 6 and Figure 6.

These results indicate that our approach exhibits a meaningful degree of positional robustness. Although the model was trained exclusively on centered masks, it still achieves reasonable recognition performance on off-center masks (WER $= 22.95\%$). This suggests that the model does not simply memorize the spatial configuration observed during training, but instead learns representations that generalize to unseen mask positions. While some performance degradation is expected due to the distribution shift between training and testing, the results indicate that the learned representations are not tightly coupled to a specific mask location. We therefore expect that incorporating off-center mask examples, or more generally increasing positional diversity during training, would further improve robustness and overall performance.

Table 6: Off-Center words test ($\mathcal{M}_{\text{Off-Center}}$) vs. full words test ($\mathcal{M}_{\text{full}}$). Both evaluations use the same model weights but different masking conditions.

Model	WER%	CER%	DNSMOS	Sim	Mismatch
$\mathcal{M}_{\text{Off-Center}}$	22.95	20.25	2.826	0.977	71.5 ms
$\mathcal{M}_{\text{full}}$	12.15	10.06	2.845	0.985	25.6 ms

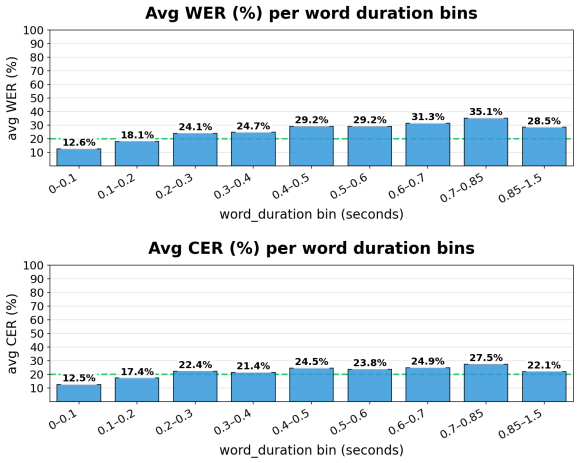


Figure 6: WER(%) and CER(%) of our $\mathcal{M}_{\text{full}}$ evaluated on off-center masks across $\mathbf{x}_t$ durations.

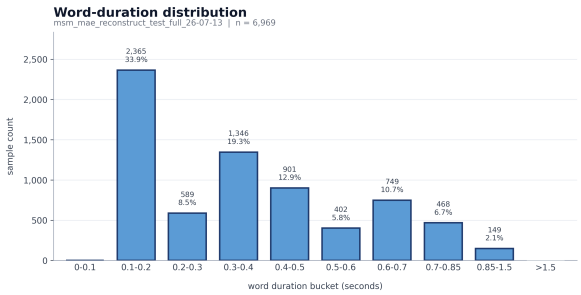


Figure 7: MSM-MAE test distribution across $\mathbf{x}_t$ durations. This represents a subset of the original dataset shown in Fig. 2.

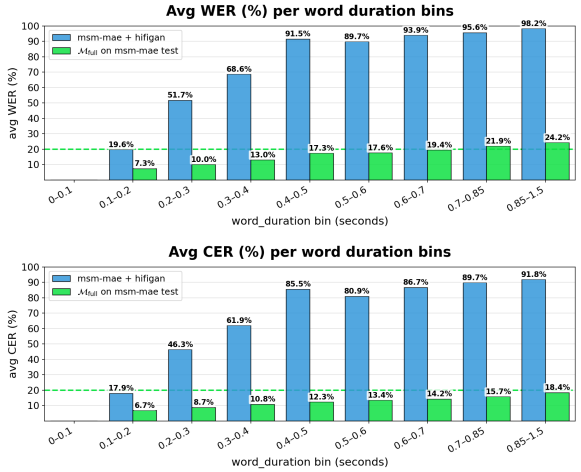


Figure 8: MSM-MAE (blue) vs. our $\mathcal{M}_{\text{full}}$ (green): WER(%) and CER(%) over $\mathbf{x}_t$ durations.

C.5 Baseline - MSM-MAE

MSM-MAE (Niizumi et al., 2022) reconstructs masked spectrograms using an MSE objective. This framework requires mask durations to be multiples of 160 ms, therefore we derived a subset from the primary dataset to meet this constraint. Specifically, we rounded mask durations to the nearest 160 ms multiple while keeping the original mask locations unchanged. By restricting the maximum boundary shift between the original and adjusted masks to 20 ms, we preserved approximately 7,000 evaluation samples (see Fig. 7) while introducing only negligible alterations to the masking configuration. Figure 8 compares the WER and CER of the MSM-MAE baseline against our proposed $\mathcal{M}_{\text{full}}$ model across various word durations. The results demonstrate that $\mathcal{M}_{\text{full}}$ consistently and significantly outperforms the baseline across all duration bins, maintaining substantially lower error rates even for the longest word durations.

C.6 Statistical Sensitivity Tests - Confidence Intervals

We have extended the evaluation by reporting 95% confidence intervals for the average WER of all models in Table 2. The intervals were estimated using 1,000 bootstrap resamples with the Confidence-Intervals Toolkit[4]. The confidence intervals are relatively narrow and comparable across models (see Table 7), indicating stable performance estimates. The proposed $\mathcal{M}_{\text{full}}$ model achieves the lowest average WER, and both its lower and upper 95% confidence bounds are lower than those of all competing variants, consistent with its improved performance.

[4]L. Ferrer, Confidence-Intervals Toolkit. https://github.com/luferrer/ConfidenceIntervals.

Table 7: 95% confidence intervals for the average WER of all models in Table 2

Model	CI Lower-Bound	Mean WER	CI Upper-Bound
$\mathcal{M}_{\text{w/o OCR}}$	49.92%	52.31%	55.01%
$\mathcal{M}_{\text{base}}$	12.07%	13.34%	14.68%
$\mathcal{M}_{+\text{ASG}_{\text{rand}}}$	12.00%	13.30%	14.74%
$\mathcal{M}_{+\text{ASG}_{\text{rand}}+\text{CTC}}$	11.32%	12.68%	14.19%
$\mathcal{M}_{+\text{CTC}}$	11.37%	12.68%	14.13%
$\mathcal{M}_{+\text{ASG}}$	11.68%	13.06%	14.40%
$\mathcal{M}_{\text{full}}$	10.73%	12.15%	13.54%

D Selection of the ASG Similarity Threshold (θ)

Audio tokens, unlike discrete text subword units, lie on a continuous acoustic manifold. To quantify relations along this manifold, we construct Acoustically Similar Groups (ASGs). Given an audio token $t_i \in \mathcal{V}$, its corresponding ASG $\mathcal{G}(t_i)$ is defined using cosine similarity between their embeddings:

$$\mathcal{G}(t_i) := \{t_j \in \mathcal{V} : \text{sim}(t_i, t_j) > \theta, i \neq j\} \quad (6)$$

where θ is a threshold that controls the group size. A lower threshold yields larger groups, capturing more prosodic and phonetic variation, whereas a higher threshold enforces stricter acoustic equivalence. Note that, in general, choosing θ too small asymptotically recovers classical label smoothing, whereas choosing θ too large asymptotically yields hard one-hot labels.

To determine an optimal value for θ, we first compute the pairwise cosine similarity matrix across the token vocabulary. A subset of 6,561 tokens (only audio ones) is visualized in Fig. 9, displaying the distribution of similarity scores which directly informs our candidate values for θ.

We empirically evaluated three candidate thresholds: $\theta = 0.38$, $\theta = 0.42$, and $\theta = 0.45$. For each threshold, we analyzed the density and distribution of the resulting ASG sizes, visualized via scatter plots (Fig. 10). A key metric in our selection is the number of groups containing multiple tokens, which ensures sufficient LS during training.

As detailed in Table 8, setting $\theta = 0.38$ results in 3,493 robust groups (with 4+ tokens), allowing maximal variation but potentially introducing acoustic noise by grouping weakly related tokens. Conversely, $\theta = 0.45$ yields only 1,145 robust groups, which strictly limits the ASG diversity. We select $\theta = 0.42$ as it offers a heuristic balance, yielding 2,255 robust groups.

To evaluate the sensitivity of the proposed method to the choice of θ, we performed a sensitivity analysis across different values of θ. These results are summarized in Table 9.

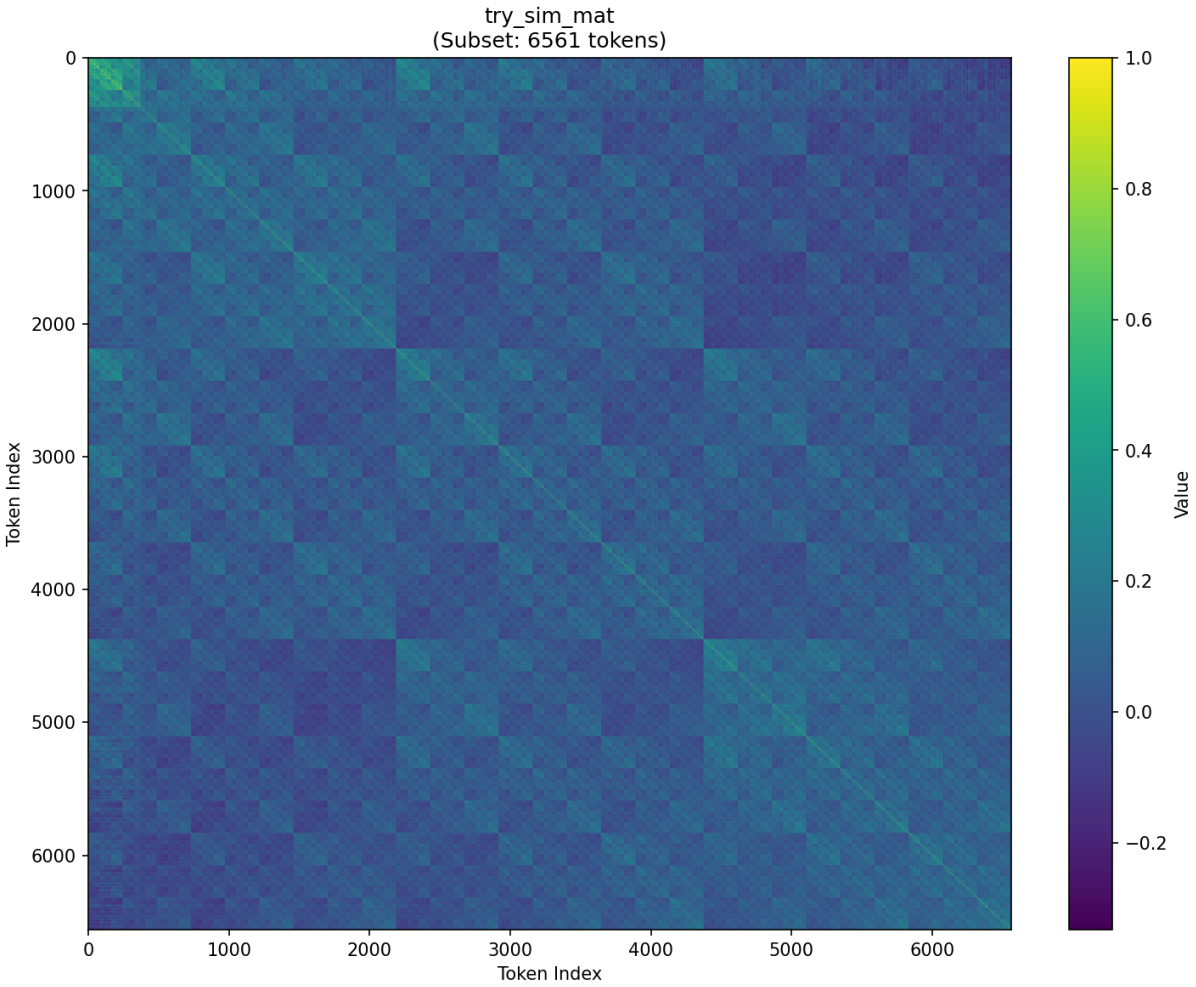


Figure 9: Pairwise cosine similarity matrix for a subset of 6,561 audio tokens.

Table 8: Distribution of Acoustic Similarity Group (ASG) sizes under different θ thresholds.

Threshold (θ)	**1 token**	**2 tokens**	**3 tokens**	**4 tokens**	**5 tokens**	$\geq$ **6 tokens**	$\geq$ **4 tokens**
0.38	1757	745	566	555	438	2500	3493
0.42	2743	928	635	409	329	1517	2255
0.45	4327	741	348	201	129	815	1145

Table 9: A sensitivity analysis over different values of θ.

Model	**WER**	**CER**	**PER**	**DNSMOS**	**Sim**	**Mismatch**
$\theta = 0.38$	12.42%	10.23%	10.20%	2.851	0.986	22.4 ms
$\theta = 0.42$ (original)	12.15%	10.06%	10.23%	2.845	0.985	25.6 ms
$\theta = 0.45$	12.78%	10.51%	10.48%	2.849	0.985	24.5 ms

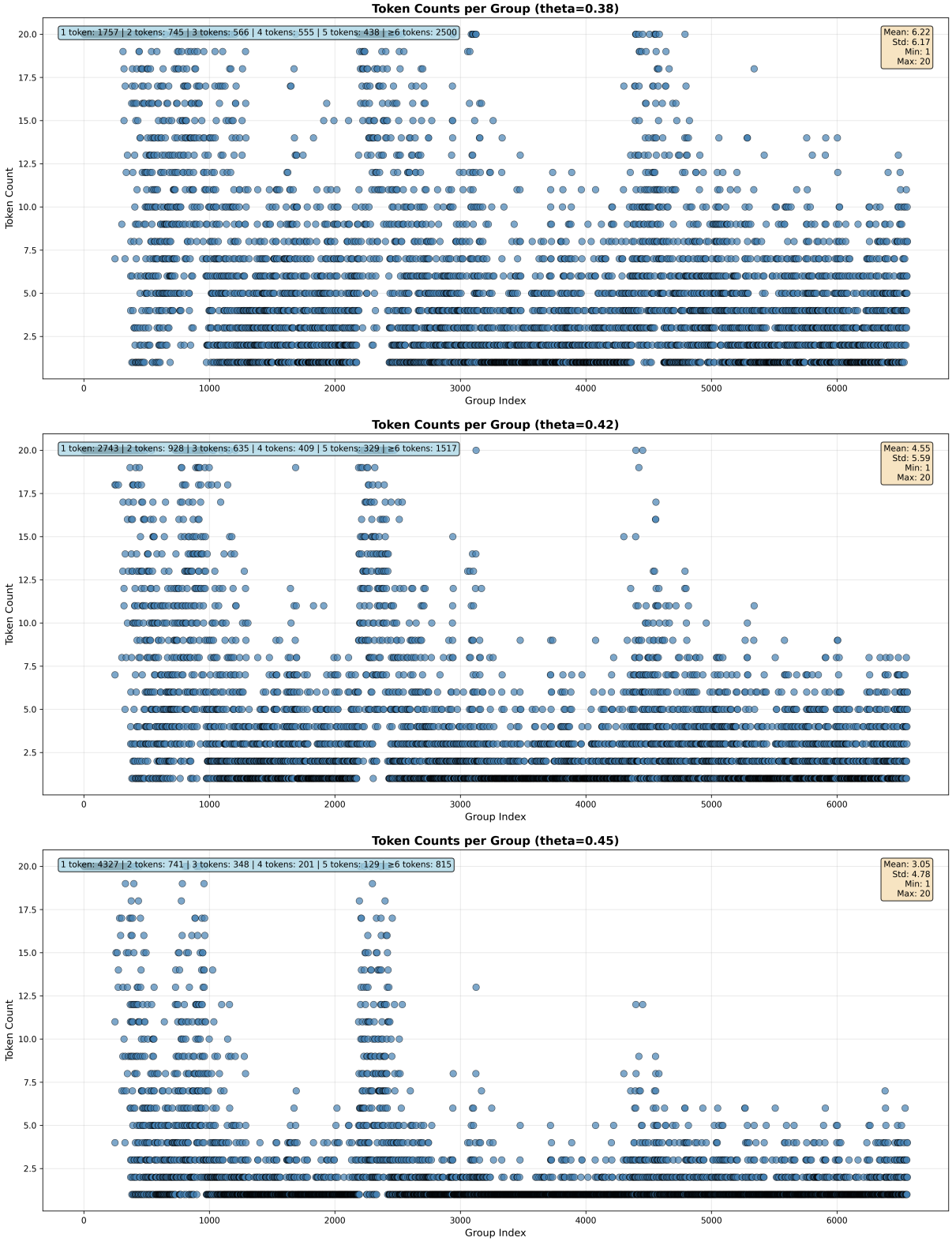


Figure 10: Scatter density plots showing token counts per group for $\theta = 0.38$ (left), $\theta = 0.42$ (center), and $\theta = 0.45$ (right).

E Glossary of Notation

Summary of primary mathematical notations and terminology used throughout this work in Table E.

Notation	Meaning
Problem Formulation & Data Sequences	
$\mathbf{x}_o, \mathbf{x}_m, \mathbf{x}_t$	Original clean audio, masked audio input, and missing target audio segment
$\hat{\mathbf{x}}_t$	The predicted inpainted speech sequence
$\mathbf{c}, w$	Slide-OCR textual context and the corresponding ground-truth target word
$\mathbf{p}$	Task-specific instruction prompt for conditional generation
$L_{\text{total}}, L_{\text{mask}}$	Duration of the total acoustic window ($5.0\,\text{s}$) and the masked interval
Model Architecture & Latent Space	
$\mathcal{M}$	The Large Audio Language Model (LALM) backbone
$\mathcal{V}, \mathcal{V}_{\text{audio}}$	The unified token vocabulary and its audio-specific subset
$\mathbf{h}_i, \mathbf{z}_i$	Hidden state vector and logit vector at token time step i
$\mathbf{e}_j$	Learned token embedding vector of token $t_j \in \mathcal{V}$
$\mathbf{y}_i, \hat{\mathbf{y}}_i$	Ground-truth one-hot label vector and predicted probability distribution over $\mathcal{V}$, at token time step i
Acoustic Similarity Groups (ASG)	
$\text{sim}(\cdot, \cdot)$	Cosine similarity function applied to token embeddings
θ	Cosine similarity threshold for ASG formation (e.g., $\theta = 0.42$)
$\mathcal{G}(t_i)$	The Acoustic Similarity Group for token t_i
$\lvert\mathcal{G}\rvert$	Cardinality of an ASG ($\lvert\mathcal{G}\rvert \leq 5$)
β	Label smoothing hyperparameter controlling probability mass spread in ASG
$\tilde{\mathbf{y}}_i$	ASG-smoothed soft-label target distribution for a token at step i
Training Objectives & Evaluation	
$\mathcal{L}_{\text{ASG}}, \mathcal{L}_{\text{CTC}}$	The ASG-smoothed Cross-Entropy loss and the auxiliary CTC loss
λ	Hyperparameter weighting the contribution of the CTC objective
WER, CER	Word/Character Error Rates computed only on the inpainted region by applying ASR over the fully stitched 5.0s sentence