SYNCAUDIO: PARALLEL AND SYNCHRONIZED TEXT-SPEECH DECODING FOR SPOKEN DIALOGUE

Anonymous authors
Paper under double-blind review

ABSTRACT

Extending text-based LLMs to spoken dialogue requires integrating speech generation without sacrificing the linguistic and knowledge capabilities of the underlying model. We study this problem in a shared-backbone architecture with parallel text and speech generation, where the two modalities are decoded through separate pathways but interact through a common language-model context. A central challenge is how the speech pathway interacts with the semantic structure inherited from the text LLM. Such interaction can introduce degradation in two complementary ways. First, temporal mismatch between text and speech can mix states from different stages of the response, weakening the semantic coherence of the shared context. Second, the speech pathway depends on backbone representations whose cross-modal relevance varies across depth, so conditioning on a single layer may either miss complementary information or emphasize representations that interfere with semantic preservation. We introduce **SyncAudio**, which addresses these two sources of cross-modal interference through **dynamic text-speech synchronization** and **sparse multi-depth aggregation**. For temporal coordination, SyncAudio uses an audio-guided synchronization mechanism to determine when the evolving speech state should be committed and fed back into the shared context. For representational interaction, it sparsely selects and aggregates representative hidden states from different backbone depths to provide richer conditioning for speech generation, while leaving the standard text-decoding pathway unchanged. Experiments show that dynamic synchronization substantially improves text-speech alignment while preserving natural and intelligible speech generation. More importantly, SyncAudio consistently outperforms an otherwise matched fixed-synchronization baseline on spoken question answering, suggesting better preservation of the semantic and knowledge capabilities inherited from the underlying text LLM. Further ablations show that sparse multi-depth representation aggregation provides more effective conditioning for the speech pathway than relying on the final backbone layer alone.

1 INTRODUCTION

Large language models (LLMs) provide strong linguistic and knowledge capabilities for text-based interaction, motivating growing efforts to extend them toward direct spoken dialogue (Zhang et al., 2023; Zeng et al., 2024; Défossez et al., 2024; Xu et al., 2025). Recent spoken LLMs have made substantial progress in streaming speech generation, low-latency interaction, and even full-duplex dialogue (Défossez et al., 2024; Xu et al., 2025; Wang et al., 2025b). A desirable spoken dialogue model, however, should not only generate natural speech efficiently, but also retain the linguistic and knowledge capabilities inherited from its text backbone. Recent studies have shown that adapting LLMs to speech can introduce measurable gaps between text- and speech-based capabilities, motivating explicit efforts to preserve the behavior of the original language model during speech adaptation (Wang et al., 2025b; Hu et al., 2026; Xiang et al., 2025; Wang et al., 2026a). The challenge is therefore not merely how to attach speech generation to an LLM, but how to make the acoustic generation process interact with the pretrained language model without unnecessarily disrupting the semantic structure inherited from the original text model, especially when generated speech states are recurrently fed back into the shared context.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

This issue is particularly explicit in an architecture with parallel text and speech generation over a shared backbone. Such models maintain separate text and speech generation pathways while allowing the two modalities to interact through a common language-model context. Text and speech do not progress in a fixed correspondence: the acoustic realization associated with a linguistic state varies with lexical content, pronunciation, prosody, and speaking style. Nevertheless, existing parallel systems often synchronize the two streams using fixed-size audio groups or predefined schedules (Team et al., 2025; Lu et al., 2026). Temporal mismatch can therefore cause states corresponding to different stages of the response to be combined. In a shared-backbone architecture, these mismatched states are recurrently fed into the same generation context, potentially weakening the semantic coherence maintained by the underlying language model. This motivates our first question: *when should the evolving text and speech states be synchronized?*

Temporal synchronization is only one aspect of this interaction. A second concerns how the speech pathway couples with representations inside the shared backbone. A common practice is to condition downstream generation on the hidden state from a single Transformer layer, often the final one (Xu et al., 2025; Liu et al., 2026). However, text-speech relationships can change substantially across model depth, and recent analyses associate speech–text capability gaps with changes in cross-modal representations across Transformer layers (Xiang et al., 2025; Wang et al., 2026a). Restricting the speech interface to a single depth may therefore miss complementary information distributed across the backbone. This representation choice affects the same recurrent generation loop: the backbone features used to generate speech determine the acoustic state that is subsequently returned to the shared context. From this perspective, temporal and representational interactions are two complementary aspects of the same cross-modal problem: one determines *when* speech information re-enters the shared context, while the other determines *what backbone information* is used to construct it.

We introduce **SyncAudio**, a shared-backbone spoken dialogue model with parallel and synchronized text-speech generation. SyncAudio addresses these two aspects through **dynamic text-speech synchronization** and **sparse multi-depth aggregation**. For temporal coordination, an audio-guided synchronization mechanism determines when the evolving speech state should be committed and fed back into the shared context, rather than imposing a fixed number of speech tokens at each decoding step. For representational interaction, SyncAudio sparsely selects and aggregates representative hidden states from different backbone depths to provide richer conditioning for the speech pathway, while leaving the standard final-layer text-decoding interface unchanged. Together, these designs regulate both *when* speech information re-enters the shared context and *what* backbone representations are used to construct the corresponding speech state.

Our experiments support both aspects of the design. Dynamic synchronization substantially reduces text-speech synchronization error while maintaining natural and intelligible speech generation. More importantly, SyncAudio consistently outperforms an otherwise matched fixed-synchronization baseline on spoken question answering and exhibits a smaller gap between textual and spoken response quality, suggesting better preservation of the semantic and knowledge capabilities inherited from the underlying text LLM. Ablations further show that sparse multi-depth aggregation provides more effective speech conditioning than using only the final backbone layer, while also achieving stronger overall performance than densely aggregating representations from all layers. Beyond these core evaluations, we further extend SyncAudio to real-time duplex interaction with lightweight control mechanisms, while preserving the core generation architecture.

Our main contributions are:

- We formulate cross-modal interaction in parallel text–speech generation with a shared backbone from two complementary perspectives: temporal synchronization and backbone representation, both of which can affect the semantic context inherited from the text LLM.
- We propose **SyncAudio**, combining **dynamic text-speech synchronization** to control when speech states are fed back and **sparse multi-depth aggregation** to enrich speech conditioning, while preserving the standard text-decoding interface.
- Through controlled experiments, we show that SyncAudio improves synchronization and speech generation quality, consistently improves spoken QA over fixed synchronization, and reduces the gap between textual and spoken responses. Ablations further demonstrate the benefit of sparse multi-depth conditioning for semantic and knowledge capabilities.

2 Related Work

2.1 Spoken Dialogue Models

Recent work has extended text-based LLMs toward direct speech understanding and generation, moving beyond cascaded ASR–LLM–TTS pipelines. Early systems such as SpeechGPT (Zhang et al., 2023), AudioPaLM (Rubenstein et al., 2023), SALMONN (Tang et al., 2024), and Qwen2-Audio (Chu et al., 2024) demonstrate that pretrained language models can be adapted to speech through discrete speech units or speech encoders. More recent models further support streaming speech generation and interactive dialogue. Mini-Omni (Xie & Wu, 2024), LLaMA-Omni (Fang et al., 2025), and GLM-4-Voice (Zeng et al., 2024) enable low-latency spoken responses, while Qwen2.5-Omni (Xu et al., 2025) adopts a Thinker–Talker architecture for streaming multimodal interaction. Moshi (Défossez et al., 2024), SyncLLM (Veluri et al., 2024), Freeze-Omni (Wang et al., 2025b), and OmniFlatten (Zhang et al., 2025) further explore continuous listening, overlapping speech, interruption handling, and full-duplex interaction. These studies primarily focus on equipping LLMs with increasingly natural speech interaction capabilities. Our work instead focuses on how speech generation should interact with a shared text-LLM backbone so that spoken generation can be introduced without unnecessarily degrading the linguistic and knowledge capabilities inherited from the original model.

2.2 Text-Speech Co-generation and Synchronization

Existing spoken language models differ in how text and speech are organized during generation. **Speech-centric** models, such as SpeechGPT (Zhang et al., 2023) and Moshi (Défossez et al., 2024), primarily generate responses in the speech-token space. **Interleaved** models, including Spirit-LM (Nguyen et al., 2025), Step-Audio series (Team, 2025), and GLM-4-Voice (Zeng et al., 2024), organize text and speech along a shared autoregressive trajectory. **Thinker–Talker** architectures, such as Qwen2.5-Omni (Xu et al., 2025) and Qwen3.5-Omni (Team, 2026), separate language reasoning from speech realization. Finally, **parallel** architectures, such as Fun-Audio-Chat (Team et al., 2025) and CF-Duplex (Lu et al., 2026), maintain separate text and speech generation pathways that interact throughout decoding. Across these generation paradigms, coordinating text and speech is important for stable and coherent generation, particularly for long-form responses, motivating recent adaptive alignment mechanisms such as TAIL in MiniCPM-o 4.5 (Cui et al., 2026) and ARIA in Qwen3.5-Omni (Team, 2026). SyncAudio focuses on alignment in the **parallel** setting, where text and speech are decoded separately but recurrently coupled through the shared context. Rather than relying on a fixed text-speech ratio, SyncAudio dynamically determines *when* the evolving speech state should be committed and fed back into the shared context.

2.3 Capability Preservation and Multi-Depth Representations

Adapting a text LLM to speech can also affect capabilities acquired during text pretraining. Several recent systems therefore explicitly preserve or reuse the original language model. Freeze-Omni (Wang et al., 2025b) keeps the LLM frozen during speech adaptation, while DiVA (Held et al., 2025) uses distillation to mitigate the loss of text-LLM capabilities during speech instruction tuning. PRIME-Speech (Hu et al., 2026) similarly studies capability preservation when extending speech-to-text LLMs toward speech-to-speech generation. Related analyses report persistent gaps between text- and speech-conditioned reasoning and associate them with changes in cross-modal representations across model depth (Wang et al., 2026b). The choice of backbone representation may therefore matter in addition to temporal synchronization. Prior work has long shown that representations from different network depths can provide complementary information. ELMo (Peters et al., 2018) learns task-dependent mixtures of multiple language model layers, while speech foundation models commonly combine hidden states across layers for downstream tasks (Yang et al., 2021). Recent multimodal and speech models also exploit intermediate or multi-layer representations for cross-modal modeling (Lee et al., 2025; Hu et al., 2026). Motivated by these findings, SyncAudio studies two complementary aspects of cross-modal interaction in a shared-backbone spoken dialogue model: *when* text and speech should exchange their evolving states through dynamic synchronization, and *which* backbone representations should support generation through multi-depth hidden aggregation. Both are designed toward preserving the semantic and knowledge capabilities of the underlying text LLM while enabling spoken response generation.

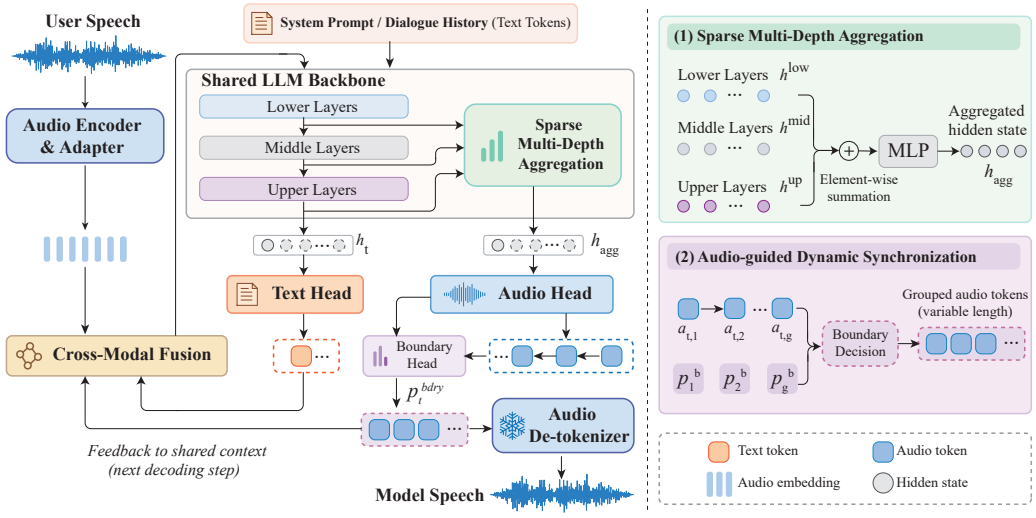


Figure 1: Overview of **SyncAudio**. It performs parallel text-speech generation with (1) **Sparse Multi-Depth Aggregation** and (2) **Audio-Guided Dynamic Synchronization**, while recurrently feeding generated text and audio groups back into the shared context.

3 METHOD

3.1 DESIGN OVERVIEW

SyncAudio extends a text-based LLM to spoken dialogue using a shared backbone with parallel text and speech generation. The model maintains *user speech*, *model text*, and *model speech* along a shared decoding timeline, with their states recurrently combined through a common language-model context. This organization also allows the same architecture to be extended to real-time duplex interaction without changing its basic generation structure.

Our design focuses on two aspects of the speech-generation interface. First, **dynamic text-speech synchronization** determines *when* the evolving speech generation should form a completed audio group and be returned to the shared context. Second, **sparse multi-depth aggregation** determines *which* backbone representations should condition speech generation. Importantly, sparse multi-depth aggregation is applied only to the audio pathway, while the text head retains the standard final-layer decoding interface of the pretrained LLM.

As illustrated in Figure 1, streaming user speech is encoded into continuous acoustic embeddings and fused with the previously generated text and speech representations. The shared LLM backbone then produces the context for two output pathways: its final-layer hidden state is used directly for text prediction, while hidden states from multiple depths are aggregated to condition the autoregressive audio head. At each shared decoding step, the model generates one text token and a variable-length group of audio tokens. The resulting text token and completed audio group are then fed back to the shared context for subsequent decoding.

3.2 PARALLEL TEXT–SPEECH GENERATION OVER A SHARED BACKBONE

Shared Context Construction. At each shared decoding step t, let $\mathbf{u}_t$, $\mathbf{e}_{t-1}^{\text{text}}$, and $\bar{\mathbf{e}}_{t-1}^{\text{audio}}$ denote the current user-speech representation, previous text-token embedding, and previous audio-group representation, respectively. Following CF-Duplex (Lu et al., 2026), we fuse the three streams through a gated cross-modal fusion module:

$$\mathbf{x}_t = \mathbf{u}_t + \mathbf{e}_{t-1}^{\text{text}} + \bar{\mathbf{e}}_{t-1}^{\text{audio}} + \sigma(W_g \mathbf{c}_t + \mathbf{b}_g) \odot \text{MLP}(\mathbf{c}_t), \tag{1}$$

where $\mathbf{c}_t = [\mathbf{u}_t; \mathbf{e}_{t-1}^{\text{text}}; \bar{\mathbf{e}}_{t-1}^{\text{audio}}]$ denotes their feature-wise concatenation. The fused representation is then processed by the shared LLM backbone, producing hidden states $\{\mathbf{h}_t^{(\ell)}\}_{\ell=1}^{L}$ across its L Transformer layers.

Parallel Text and Speech Decoding. At each shared decoding step t, the text and speech pathways decode in parallel from the same backbone context using separate generation heads. The step-wise output distribution is written as

$$p\big(y_t^{\text{text}}, \mathbf{y}_t^{\text{audio}} \mid \mathbf{h}_t\big) = \underbrace{p\Big(y_t^{\text{text}} \mid \mathbf{h}_t^{(L)}\Big)}_{\text{Text Head}} \underbrace{\prod_{g=1}^{G_t} p\Big(y_{t,g}^{\text{audio}} \mid y_{t,<g}^{\text{audio}}, \tilde{\mathbf{h}}_t^{\text{audio}}\Big)}_{\text{Audio Head}}. \tag{2}$$

Here, $\mathbf{y}_t^{\text{audio}} = (y_{t,1}^{\text{audio}}, \ldots, y_{t,G_t}^{\text{audio}})$ denotes the audio group generated at step t. The text head predicts a single text token from the final-layer state $\mathbf{h}_t^{(L)}$, while the audio head autoregressively generates speech tokens within the group. Specifically, each audio token $y_{t,g}^{\text{audio}}$ is conditioned on the previously generated tokens $y_{t,<g}^{\text{audio}}$ and the speech-specific representation $\tilde{\mathbf{h}}_t^{\text{audio}}$, and generation continues until the dynamically determined group boundary G_t is reached. The construction of $\tilde{\mathbf{h}}_t^{\text{audio}}$ is described in Section 3.4.

Recurrent Cross-Modal Feedback. After completing the audio group, we aggregate its token representations and project the group-level summary back to the shared representation space:

$$\bar{\mathbf{e}}_t^{\text{audio}} = \text{MLP}_{\text{grp}}\left(\frac{1}{G_t}\sum_{g=1}^{G_t} \mathbf{e}_{t,g}^{\text{audio}}\right). \tag{3}$$

The generated text token and audio-group representation are then fed back into the shared context for the next decoding step, together with newly arriving user speech. This recurrent feedback couples the two output pathways across decoding steps. Importantly, G_t controls both the number of generated speech tokens and the timing of audio feedback, making its dynamic determination central to text-speech synchronization.

3.3 Audio-Guided Dynamic Synchronization

The parallel generation framework in Section 3.2 requires deciding when each audio group should terminate. A fixed group size enforces the uniform text-speech ratio at every decoding step, even though the amount of speech tokens associated with a linguistic state may vary. SyncAudio therefore models the audio-group boundary dynamically. We also consider two alternative dynamic strategies: *Text-Guided*, which predicts the group size before audio generation, and *Native*, which uses an end-of-group token for termination; details are provided in Appendix B.

Boundary Prediction from the Audio State. Instead of determining the group size based on the linguistic state before the group generation, SyncAudio predicts the boundary from the evolving hidden state of the autoregressive audio decoder. Let $\mathbf{z}_{t,g}$ denote the audio-decoder hidden state after generating the g-th speech token at shared decoding step t. A lightweight boundary head predicts

$$p_{t,g}^{\text{bdry}} = \sigma(f_{\text{bdry}}(\mathbf{z}_{t,g})), \qquad G_t = \min\Big\{g : p_{t,g}^{\text{bdry}} > \tau\Big\}, \tag{4}$$

where τ is the boundary threshold. Thus, G_t determines both the number of generated speech tokens and when the completed audio group is returned to the shared context. Because the decision is updated throughout audio generation, it can adapt to the speech realization produced so far rather than relying only on the pre-generation LLM state.

Boundary Supervision. During training, the boundary predictor is supervised with reference audio-group boundaries. Let $b_{t,g}^* \in \{0, 1\}$ indicate whether position g is the reference boundary at decoding step t. We optimize the boundary head with binary cross-entropy:

$$\mathcal{L}_{\text{sync}} = -\sum_{t,g}\Big[b_{t,g}^* \log p_{t,g}^{\text{bdry}} + (1 - b_{t,g}^*)\log\Big(1 - p_{t,g}^{\text{bdry}}\Big)\Big]. \tag{5}$$

At inference time, the audio decoder generates autoregressively until the predicted boundary is reached, after which the completed group is aggregated and fed back into the shared context. This enables adaptive synchronization within the speech generation process while preserving recurrent coupling between the text and speech pathways.

3.4 MULTI-DEPTH REPRESENTATION AGGREGATION

Dynamic synchronization determines *when* generated speech information is returned to the shared context. We further ask *what backbone representation* should condition speech generation. Transformer representations evolve substantially with depth, and prior analyses suggest that text-speech relationships are distributed across different layers (Xiang et al., 2025; Wang et al., 2026a). Relying only on the final layer may overlook complementary representations useful for speech generation.

Sparse Multi-Depth Aggregation. We introduce a compact depth-stratified aggregation scheme that sparsely samples the representation hierarchy rather than densely combining all Transformer layers. Specifically, we select representative hidden states from the lower, middle, and upper regions of the shared LLM. Let $\mathbf{h}_t^{\text{low}}$, $\mathbf{h}_t^{\text{mid}}$, and $\mathbf{h}_t^{\text{up}}$ denote the corresponding representations at decoding step t. They are fused as

$$\mathbf{s}_t^{\text{audio}} = \mathbf{h}_t^{\text{low}} + \mathbf{h}_t^{\text{mid}} + \mathbf{h}_t^{\text{up}}, \qquad \tilde{\mathbf{h}}_t^{\text{audio}} = \text{MLP}_{\text{depth}}\big(\mathbf{s}_t^{\text{audio}}\big). \tag{6}$$

This sparse design captures complementary information across the representation hierarchy while avoiding dense aggregation over every backbone layer. The resulting $\tilde{\mathbf{h}}_t^{\text{audio}}$ provides a compact conditioning representation for the autoregressive audio head.

Asymmetric Cross-Modal Interface. Multi-depth aggregation is applied only to the speech pathway, while the text head continues to decode from the final-layer state $\mathbf{h}_t^{(L)}$. This asymmetric design enriches speech conditioning without altering the standard text-decoding interface of LLM.

3.5 TRAINING

Training Objective. SyncAudio is trained jointly for text generation, speech generation, and dynamic synchronization. The text and audio pathways use autoregressive cross-entropy objectives, while the boundary predictor is supervised by Eq. 5. The overall objective is

$$\mathcal{L} = \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{audio}}\mathcal{L}_{\text{audio}} + \lambda_{\text{sync}}\mathcal{L}_{\text{sync}}, \tag{7}$$

where $\mathcal{L}_{\text{text}}$ and $\mathcal{L}_{\text{audio}}$ are the autoregressive cross-entropy losses for text and speech generation, respectively. The synchronization objective jointly teaches the audio pathway when to terminate the current group and return its state to the shared context.

Staged Training. Training follows a two-stage curriculum. We first train on **ASR** and **streaming TTS** to establish basic speech understanding and generation, optimizing all trainable components except the de-tokenizer. We then introduce speech-to-text dialogue (**S2TD**) and speech-to-text-and-speech dialogue (**S2TSD**) for conversational and joint text-speech generation, while freezing the speech encoder and continuing to optimize the remaining modules. Dynamic synchronization and multi-depth audio conditioning are used throughout both stages.

Real-Time Extension. We also further extend SyncAudio to **SyncAudio-Realtime** for continuous duplex interaction by introducing lightweight wait and interruption control, while preserving the core generation architecture. Details are provided in Appendix G.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETUP

Datasets We train SyncAudio on four task families: ASR, streaming TTS, S2TD, and S2TSD. The ASR and TTS training data comprises over 300k hours of speech, with ASR data collected from large-scale public corpora such as LibriSpeech (Panayotov et al., 2015), GigaSpeech (Chen et al., 2021), and VoxPopuli (Wang et al., 2021), while TTS data is primarily drawn from VoxBox (Wang et al., 2025a). For S2TD and S2TSD, we construct approximately 1.9M spoken dialogue samples from public QA datasets through spoken-style rewriting and speech synthesis. Detailed data sources, statistics, and construction procedures are provided in Appendix C.

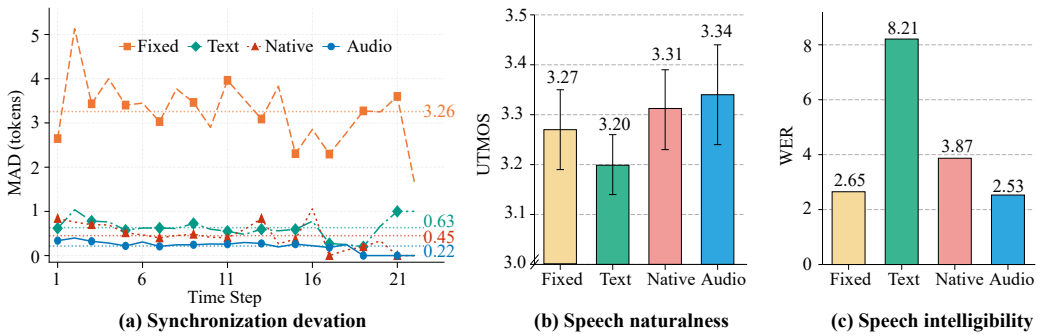


Figure 2: Comparison of different synchronization mechanisms from three complementary perspectives: synchronization accuracy, speech naturalness, and intelligibility.

Implementation Details We use Qwen3-4B as the backbone LLM and keep it frozen throughout training. We train LoRA adapter (Hu et al., 2021) on top of the backbone with rank $r = 16$ and scaling factor $\alpha = 32$. The audio head is initialized from Qwen3-0.6B. The audio tokenizer follows CosyVoice 2 (Du et al., 2024) at 25Hz, with waveform reconstruction via the pretrained token2wav model. Following CF-Duplex (Lu et al., 2026), the audio stream is prefixed with two padding tokens to improve text-speech consistency and speech naturalness. All models are trained on 16 NVIDIA H200 GPUs using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 0.01. Stage 1 is trained for 350k steps with a learning rate of 1×10^{-4} and 8k warm-up steps, while Stage 2 is trained for 20k steps with a learning rate of 5×10^{-5} and 4k warm-up steps.

Baselines We compare SyncAudio with representative spoken dialogue models spanning speech-to-speech, interleaved, and parallel generation, including SpeechGPT (Zhang et al., 2023), GLM-4-Voice (Zeng et al., 2024), Moshi (Défossez et al., 2024), Freeze-Omni (Wang et al., 2025b), and CF-Duplex (Lu et al., 2026). To isolate the effect of dynamic text-speech synchronization, we construct a controlled baseline, termed **FixedAudio**, using the same architecture, training data, and training recipe as SyncAudio, but with a fixed audio group size of $G = 4$ instead of dynamic grouping.

Evaluation Benchmarks and Metrics. We primarily evaluate SyncAudio on four spoken question answering (QA) benchmarks to assess different aspects of spoken dialogue performance: **LLaMAQ**[1], **TriviaQ** (Joshi et al., 2017), **WebQ**[2], and **AlpacaEval** (Li et al., 2023), using audio questions from OpenAudioBench[3]. These benchmarks are used to evaluate response quality, text-speech decoding consistency, and the effect of multi-depth aggregation. For LLaMAQ, TriviaQA, and WebQ, we report binary response correctness, while AlpacaEval uses GPT-based scoring on a 0-5 scale and is rescaled to 0-100 when computing cross-benchmark averages. We also evaluate synchronization mechanisms using synchronization deviation, speech naturalness, and intelligibility. Additional ASR/TTS and real-time full-duplex interaction evaluations are provided in Appendices D and G, respectively.

4.2 Comparison of Synchronization Mechanisms

Setup. We first examine how the synchronization boundary should be determined during parallel text-speech generation. SyncAudio adopts **Audio-Guided synchronization**, where the boundary is predicted from the evolving hidden state of the audio decoder, allowing the decision to depend on the speech realization generated so far. To analyze alternative synchronization signals, we compare Audio-Guided with three variants: **Fixed**, which uses a constant group size $G = 4$; **Text-Guided**, which predicts the group size from the shared LLM state before audio generation; and **Native**, which relies on a dedicated end-of-group token in the autoregressive audio sequence. All variants share the same backbone, training data, and generation architecture, differing only in how the audio-group boundary is determined. We compare four synchronization strategies using three metrics:

[1] https://github.com/google-research-datasets/LLAMA1-Test-Set
[2] https://huggingface.co/datasets/stanfordnlp/web_questions
[3] https://huggingface.co/datasets/baichuan-inc/OpenAudioBench

synchronization deviation, measured by the mean absolute deviation (MAD) between predicted and ground-truth audio-group sizes for each text token; speech naturalness, measured by UTMOSv2 (Saeki et al., 2022); and speech intelligibility, measured by transcription word error rate (WER).

Results. As shown in Figure 2, Fixed grouping yields an MAD of 3.26, while Text-Guided, Native, and Audio-Guided reduce it to 0.63, 0.45, and 0.22, respectively. However, lower synchronization error does not always lead to better speech generation. Text-Guided increases WER from 2.65 to 8.21 and slightly lowers UTMOS from 3.27 to 3.20, while Native obtains a WER of 3.87. In contrast, Audio-Guided achieves the lowest MAD (0.22), the highest UTMOS (3.34), and the lowest WER (2.53). These results show that the synchronization signal matters in addition to boundary accuracy itself. Conditioning on the evolving audio-generation state provides a more reliable basis for determining when the current speech group should terminate. We therefore use **Audio-Guided synchronization** as the default configuration of SyncAudio in all subsequent experiments.

4.3 Spoken Question Answering

Setup. We evaluate whether SyncAudio preserves the semantic and knowledge capabilities of the underlying LLM under joint text-speech generation. We focus on **speech-to-text-and-speech dialogue (S2TSD)**, where both text and audio heads are active and jointly generate textual and spoken responses. We compare SyncAudio with FixedAudio and representative spoken dialogue systems, and evaluate textual (T) and spoken (S) response quality separately on the four QA benchmarks.

Results. As shown in Table 1, SyncAudio achieves stronger overall performance than FixedAudio, with higher average scores for both textual and spoken responses, indicating better preservation of the semantic and knowledge capabilities of the underlying LLM during joint generation. We further evaluate **speech-to-text dialogue (S2TD)** with the audio pathway disabled at inference time, where SyncAudio again consistently outperforms FixedAudio. This suggests that the benefit is not limited to generated speech, but also extends to text-only decoding. Complete S2TD results, together with the original Qwen3-4B text model as a reference, are provided in Appendix F.

Table 1: Spoken question answering performance on S2TSD. Higher is better.

Model	Backbone LLM	LLaMAQ		TriviaQ		WebQ		AlpacaEval		Avg.	
		T	S	T	S	T	S	T	S	T	S
SpeechGPT	LLaMA-13B	–	21.0	–	14.8	–	6.5	–	–	–	–
GLM-4-Voice	GLM4-9B	64.7	50.7	**39.1**	26.5	32.2	15.9	3.8	3.6	<u>53.1</u>	41.2
BayLing-Duplex	GLM4-9B	–	46.0	–	–	–	18.1	–	–	–	–
Moshi	Helium-7B	60.8	54.5	25.6	16.7	23.4	22.1	1.8	1.8	36.4	32.1
SALM-Duplex	TinyLlama-1.1B	–	51.3	–	16.9	–	25.0	–	3.0	–	–
CF-Duplex	Qwen3-1.7B	57.3	50.7	19.6	18.1	30.3	28.0	<u>4.2</u>	<u>3.9</u>	47.6	43.9
FixedAudio	Qwen3-4B	<u>66.0</u>	<u>56.0</u>	<u>32.4</u>	<u>28.7</u>	<u>34.5</u>	**34.9**	3.9	3.4	52.8	<u>47.0</u>
SyncAudio	Qwen3-4B	**66.7**	**58.5**	31.9	**28.9**	**36.0**	<u>33.1</u>	**4.3**	**4.0**	**55.2**	**49.9**

4.4 Text-Speech Decoding Consistency

Setup. We evaluate whether the parallel textual and spoken outputs convey consistent content from two complementary perspectives. For **lexical consistency**, we compute the *word error rate* (WER) between the generated text and the ASR transcription of the generated speech. For **response-level consistency**, we measure the *Text-Speech Response Gap* (<u>TSG</u>), defined as the discrepancy between the textual and spoken responses correctness. Lower indicates better consistency for both metrics.

Results. As shown in Table 2, SyncAudio consistently improves text–speech consistency over FixedAudio across all four benchmarks. Overall, WER decreases from 13.2% to 11.9%, while TSG decreases from 8.9% to 7.9%. The improvement is more pronounced on the long-response subset: SyncAudio reduces average TSG from 9.3% to 5.9%, with consistent gains on three of the four benchmarks, while also reducing average WER from 17.7% to 16.6%. This suggests that dynamic

synchronization is particularly beneficial when text-speech mismatch is more likely to accumulate over longer responses. Complete results across different response-length ranges are provided in Appendix E.

Table 2: Text-speech consistency across four benchmarks. WER and TSG are reported in percentage (%); lower is better for both. AlpacaEval TSG is rescaled from 0–5 to 0–100. Positive Δ (FixedAudio − SyncAudio) indicates improvement.

		LLaMAQ		**TriviaQ**		**WebQ**		**AlpacaEval**		**Avg.**	
	Model	WER	TSG	WER	TSG	WER	TSG	WER	TSG	WER	TSG
Overall	FixedAudio	6.5	10.7	8.9	4.4	17.3	5.3	16.2	15.0	13.2	8.9
	SyncAudio	**5.8**	**9.7**	**7.7**	**3.6**	**15.8**	**4.5**	**14.8**	**13.6**	**11.9**	**7.9**
	$\Delta_{\mathrm{F-S}}$	+0.7	+1.0	+1.2	+0.8	+2.1	+0.8	+1.5	+1.4	+1.3	+1.0
Long-set	FixedAudio	6.8	8.0	8.9	4.4	23.4	7.2	28.1	17.6	17.7	9.3
	SyncAudio	**5.5**	**2.7**	**6.7**	**2.0**	23.9	**3.2**	**25.0**	**15.6**	**16.6**	**5.9**
	$\Delta_{\mathrm{F-S}}$	**+1.3**	**+5.3**	**+2.2**	+2.4	−0.5	**+4.0**	**+3.1**	+2.0	+1.1	**+3.4**

4.5 Ablation on Multi-Depth Representation Aggregation

Setup. We study how the depth coverage of backbone representations affects audio conditioning. We compare three variants: *Final-Layer*, which uses only the final backbone representation; *Dense Multi-Depth Aggregation (D-MDA)*, which aggregates representations from all Transformer layers; and *Sparse Multi-Depth Aggregation (S-MDA)*, which selects representative layers from the lower, middle, and upper backbone regions. All other components are kept unchanged.

Results. As shown in Table 3, S-MDA consistently outperforms Final-Layer and achieves the strongest overall performance. D-MDA provides no consistent additional benefit, suggesting that selectively covering the representation hierarchy is more effective than densely aggregating all backbone layers. Notably, S-MDA also improves the textual output under joint text–speech generation despite being applied only to the audio pathway, consistent with the recurrent cross-modal feedback in SyncAudio. Overall, S-MDA provides more effective speech conditioning than either Final-Layer or D-MDA. This confirms that sparse, depth-stratified aggregation is sufficient to capture the complementary backbone information needed for effective text-speech interaction.

Table 3: Effect of multi-depth representations for audio conditioning. T^* denotes text-only decoding, while T and S denote textual and spoken outputs under joint generation, respectively.

Aggregation	**LLaMAQ**			**TriviaQ**			**WebQ**			**AlpacaEval**			**Avg.**		
	T^*	T	S	T^*	T	S	T^*	T	S	T^*	T	S	T^*	T	S
Final-Layer	66.3	65.7	55.7	34.4	28.9	23.6	37.2	34.2	30.2	4.14	3.92	3.50	55.18	51.80	44.88
D-MDA	66.0	**67.0**	**59.7**	33.7	31.6	26.0	37.0	34.6	31.5	4.27	4.20	**4.05**	55.53	54.30	49.55
S-MDA	**69.3**	66.7	58.5	**36.1**	**31.9**	**28.9**	**38.3**	**36.0**	**33.1**	**4.58**	**4.31**	3.96	**58.83**	**55.20**	**49.93**

5 Conclusion

We introduced **SyncAudio**, a shared-backbone spoken dialogue model for parallel text-speech generation. SyncAudio combines audio-guided dynamic synchronization to determine *when* generated speech states should return to the shared context, with sparse multi-depth aggregation to provide richer speech conditioning while preserving the standard final-layer text-decoding interface. Experiments show that SyncAudio improves text-speech synchronization, speech quality, and spoken question answering over a matched fixed-synchronization baseline. Ablations further show that sparse multi-depth aggregation provides more effective conditioning than either final-layer-only or dense all-layer aggregation. Together, these results highlight the importance of coordinating both *when* speech states interact with the shared context and *which* backbone representations are used for speech generation.

REFERENCES

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520/.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Interspeech 2021*, pp. 3670–3674, 2021. doi: 10.21437/Interspeech.2021-1965.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *CoRR*, abs/2407.10759, 2024. doi: 10.48550/ARXIV.2407.10759. URL https://doi.org/10.48550/arXiv.2407.10759.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *CoRR*, abs/2410.00037, 2024. doi: 10.48550/ARXIV.2410.00037. URL https://doi.org/10.48550/arXiv.2410.00037.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3029–3051, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.183. URL https://aclanthology.org/2023.emnlp-main.183/.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. In *International Conference on Learning Representations*, volume 2025, pp. 57607–57624, 2025.

Daniel Galvez, Greg Diamos, Juan Torres, Keith Achorn, Juan Cerón, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/202cb962ac59075b964b07152d234b70-Paper-round1.pdf.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: A large-scale, extensive, multilingual, and diverse dataset for speech generation. *CoRR*, abs/2501.15907, 2025. doi: 10.48550/ARXIV.2501.15907. URL https://doi.org/10.48550/arXiv.2501.15907.

William Held, Yanzhe Zhang, Minzhi Li, Weiyan Shi, Michael J Ryan, and Diyi Yang. Distilling an end-to-end voice assistant without instruction training data. In *Proceedings of the 63rd Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 7876–7891, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Yuxuan Hu, Heng Lu, Ruchao Fan, Yao Qian, Xiaofei Wang, Jian Xue, Heming Wang, Shuohang Wang, Young Jin Kim, Yelong Shen, et al. Preserving speech-to-text llm capabilities in speech-to-speech generation. *arXiv preprint arXiv:2606.30944*, 2026.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1601–1611. Association for Computational Linguistics, 2017. doi: 10.18653/V1/P17-1147. URL https://doi.org/10.18653/v1/P17-1147.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Hyunji Lee, Danni Liu, Supriti Sinhamahapatra, and Jan Niehues. How do multimodal foundation models encode text and speech? an analysis of cross-lingual and cross-modal representations. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 600–610, 2025.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung-Yi Lee. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2025, Honolulu, HI, USA, December 6-10, 2025*, pp. 1–8. IEEE, 2025. doi: 10.1109/ASRU65441.2025.11433838. URL https://doi.org/10.1109/ASRU65441.2025.11433838.

Zhenyu Liu, Xuanyu Zhang, Yunxin Li, Qixun Teng, Shenyuan Jiang, Haolan Chen, Mingjun Zhao, Fanbo Meng, Yu Xu, Yancheng He, et al. Hierarchical acoustic-semantic modeling: Modality separation and semantic coherence for full-duplex slms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9264–9280, 2026.

Hui Lu, Xueyuan Chen, Huimeng Wang, Shuhai Peng, Shiyin Kang, Xixin Wu, and Zhiyong Wu. How should llms listen while speaking? a study of user-stream routing in full-duplex spoken dialogue. *arXiv preprint arXiv:2605.10199*, 2026.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne (eds.), *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, CEUR Workshop Proceedings. CEUR-WS.org, 2016. URL https://ceur-ws.org/Vol-1773/CoCoNIPS_2016_paper9.pdf.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. Generative spoken dialogue language modeling. *Trans. Assoc. Comput. Linguistics*, 11:250–266, 2023. doi: 10.1162/TACL_A_00545. URL https://doi.org/10.1162/tacl_a_00545.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, et al. Spirit-lm: Interleaved spoken and written language model. *Transactions of the Association for Computational Linguistics*, 13:30–52, 2025.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1202. URL https://aclanthology.org/N18-1202/.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. MLS: A large-scale multilingual dataset for speech research. In Helen Meng, Bo Xu, and Thomas Fang Zheng (eds.), *21st Annual Conference of the International Speech Communication Association, Interspeech 2020, Virtual Event, Shanghai, China, October 25-29, 2020*, pp. 2757–2761. ISCA, 2020. doi: 10.21437/INTERSPEECH.2020-2826. URL https://doi.org/10.21437/Interspeech.2020-2826.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264/.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=piLPYqxtWuA.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In Hanseok Ko and John H. L. Hansen (eds.), *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pp. 4521–4525. ISCA, 2022. doi: 10.21437/INTERSPEECH.2022-439. URL https://doi.org/10.21437/Interspeech.2022-439.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 16607–16629, 2024.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

StepFun Audio Team. Step-audio 2 technical report. *CoRR*, abs/2507.16632, 2025. doi: 10.48550/ARXIV.2507.16632. URL https://doi.org/10.48550/arXiv.2507.16632.

Tongyi Fun Team, Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, Qinglin Zhang, Qiquan Zhang, and Jingren Zhou. Fun-audio-chat technical report. *CoRR*, abs/2512.20156, 2025. doi: 10.48550/ARXIV.2512.20156. URL https://doi.org/10.48550/arXiv.2512.20156.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21390–21402, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1192. URL https://aclanthology.org/2024.emnlp-main.1192/.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 993–1003, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.80. URL https://aclanthology.org/2021.acl-long.80/.

Chaoren Wang, Heng Lu, Xueyao Zhang, Shujie Liu, Yan Lu, Jinyu Li, and Zhizheng Wu. Closing the modality reasoning gap for speech large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18821–18835, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.857. URL https://aclanthology.org/2026.acl-long.857/.

Chaoren Wang, Heng Lu, Xueyao Zhang, Shujie Liu, Yan Lu, Jinyu Li, and Zhizheng Wu. Closing the modality reasoning gap for speech large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18821–18835, 2026b.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *CoRR*, abs/2503.01710, 2025a. doi: 10.48550/ARXIV.2503.01710. URL https://doi.org/10.48550/arXiv.2503.01710.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025b. URL https://proceedings.mlr.press/v267/wang25aw.html.

Bajian Xiang, Shuaijiang Zhao, Tingwei Guo, and Wei Zou. Understanding the modality gap: An empirical study on the speech-text alignment mechanism of large speech language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5187–5202, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.262. URL https://aclanthology.org/2025.emnlp-main.262/.

Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*, 2024.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical

report. *CoRR*, abs/2503.20215, 2025. doi: 10.48550/ARXIV.2503.20215. URL https://doi.org/10.48550/arXiv.2503.20215.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *CoRR*, abs/2412.02612, 2024. doi: 10.48550/ARXIV.2412.02612. URL https://doi.org/10.48550/arXiv.2412.02612.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1055. URL https://aclanthology.org/2023.findings-emnlp.1055/.

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chao-Hong Tan, Zhihao Du, and ShiLiang Zhang. OmniFlatten: An end-to-end GPT model for seamless voice conversation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14570–14580, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.709. URL https://aclanthology.org/2025.acl-long.709/.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 35139–35148. AAAI Press, 2026. doi: 10.1609/AAAI.V40I41.40820. URL https://doi.org/10.1609/aaai.v40i41.40820.

A AI USE STATEMENT

In accordance with ICLR policy, we disclose that large language models (LLMs) were used as assistive tools in this work. During manuscript preparation, LLMs were used for language polishing and refinement, including improving grammar, wording, readability, and fluency. In addition, LLM-based evaluators were used as evaluation metrics for assessing experimental results, as described in the corresponding evaluation protocols. The LLMs were not used for research ideation, methodological or experimental design, conducting experiments, or drawing scientific conclusions. All scientific contributions, analyses, and conclusions are the work of the human authors, who take full responsibility for the content of this paper.

B DETAILS OF SYNCHRONIZATION VARIANTS

We provide additional implementation details for the four synchronization mechanisms compared in Section 4.2. All variants use the same backbone, audio decoder, training data, and parallel text–speech generation framework. They differ only in how the audio-group size G_t is determined at each shared decoding step.

Fixed. The Fixed variant follows a pre-defined and fixed text–speech token ratio and generates a constant number of audio tokens at every shared decoding step. Specifically, we set the audio-group size to $G_t = 4$ for all t. After four audio tokens are generated, their token embeddings are aggregated into the group-level audio state and returned to the shared context. This provides a controlled baseline without adaptive boundary prediction.

Text-Guided. Text-Guided variant determines the audio-group size using the shared LLM representation at the current decoding step, before starting the generation of the corresponding audio group. A lightweight prediction head takes the final backbone hidden state $h_t^{(L)}$ as input and predicts the size of the upcoming audio group. The audio decoder auto-regressively generates the corresponding number of tokens before returning the completed group to the shared context. Thus, Text-Guided variant allows the group size to vary across decoding steps, but its synchronization decision is made solely from the linguistic context available before the current speech realization is generated. It is worth noting that this strategy is similar to duration prediction utilized in non-autoregressive text-to-speech modeling Ren et al. (2021).

Native. Native variant makes the synchronization decision implicitly during the autoregressive audio-token sequence generation. We introduce a dedicated end-of-group token, denoted $\langle \mathrm{EOG} \rangle$, into the audio vocabulary. During training, $\langle \mathrm{EOG} \rangle$ is appended at each reference audio-group boundary and is optimized together with the ordinary audio tokens using the autoregressive audio-generation objective. At inference time, the audio decoder continues generation until $\langle \mathrm{EOG} \rangle$ is emitted, at which point the current audio group is terminated and fed back into the shared context. Unlike Text-Guided, Native variant can therefore determine the boundary based on the generated audio history, but the boundary decision is coupled with audio-token prediction.

Audio-Guided. Audio-Guided variant, used by SyncAudio, separates boundary prediction from audio-token prediction while conditioning both on the evolving audio generation process. After generating the g-th audio token at step t, a lightweight boundary head operates on the corresponding audio-decoder hidden state $z_{t,g}$:

$$p_{t,g}^{\mathrm{bdry}} = \sigma(f_{\mathrm{bdry}}(z_{t,g})) \,. \tag{8}$$

Generation of the current group terminates when the predicted boundary probability exceeds a threshold τ. The boundary head is trained with binary cross-entropy using the reference audio-group boundaries. Because the decision is updated after every generated audio token, Audio-Guided variant can adapt G_t to the speech realization generated so far without introducing a boundary symbol into the audio-token sequence.

Comparison. The four variants differ primarily in both the source and timing of their synchronization decisions. Fixed variant imposes a predefined boundary, Text-Guided variants predicts the boundary from the shared linguistic state before speech generation, while Native and Audio-Guided variants determine it during autoregressive speech generation. Native variant represents the boundary as an additional audio token, whereas Audio-Guided uses a separate boundary predictor over the evolving audio-decoder state. This controlled comparison isolates whether adaptive synchronization is beneficial and, more importantly, which signal provides an effective basis for determining the text–speech boundary.

C Dataset Details

ASR. We utilize several openly available English speech corpora, including LibriSpeech (Panayotov et al., 2015), GigaSpeech (Chen et al., 2021), PeopleSpeech (Galvez et al., 2021), MLS (Pratap et al., 2020), CommonVoice (Ardila et al., 2020), VoxPopuli (Wang et al., 2021), and Emilia-Large (He et al., 2025), yielding 217k hours of English ASR data in total. The user stream consists of the audio input and a text prompt, while the model text stream contains the transcription.

Streaming TTS. We use VoxBox (Wang et al., 2025a), which provides 104k hours of English and Chinese text-to-speech data. The model is trained to emit G_t audio tokens per text token position after a delay of $\mathcal{D}$ text tokens, with the delay positions filled by $\langle$AUDIO_WAIT$\rangle$ tokens.

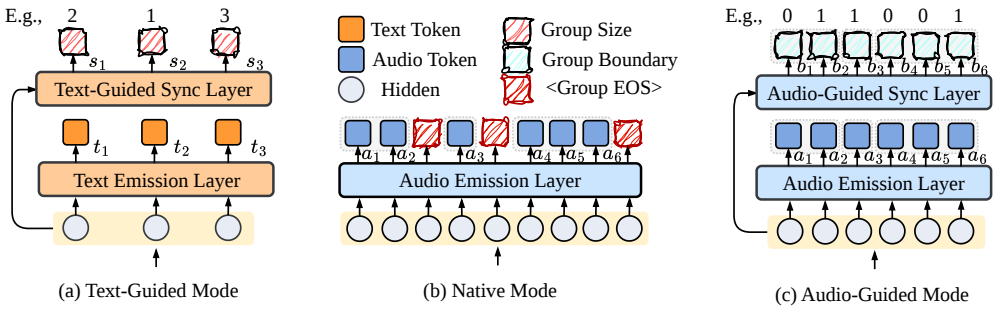


Figure 3: Diagram of different sync mode.

Turn-Based Spoken Dialogue. We follow Lu et al. (2026) to construct turn-based spoken dialogue data for S2TD and S2TSD training from publicly available textual QA and dialogue datasets, including SQuAD (Rajpurkar et al., 2016), MS-MARCO (Nguyen et al., 2016), HotpotQA (Yang et al., 2018), Natural Questions (Kwiatkowski et al., 2019), UltraChat (Ding et al., 2023), and `I_Wonder_Why-Chinese` [4]. We use `Qwen3-30B-A3B-Instruct` [5] to rewrite these samples into English spoken-style question–answer pairs and synthesize them into speech using IndexTTS-2 (Zhou et al., 2026). We filter out samples with overly long audio or poor text–audio consistency. In total we obtain 16.8k hours of turn-based spoken dialogue data.

Full-Duplex Spoken Dialogue. Following Lu et al. (2026), we construct full-duplex spoken dialogue by composing turn-based spoken dialogues into two-channel audio interactions. We synthesized full-duplex spoken dialogues with two scenarios: user interruptions and backchanneling. We formulate full-duplex samples on the fly during training to randomize insertion timing of user interruptions and backchannels. In total we construct 36.4k hours of full-duplex spoken dialogue data.

Table 4: Overview of training data for different training tasks.

Task	Source	Size (k hours)	Construction
ASR	Public	217	Combined from openly available English corpora
TTS	Public	102	VoxBox
Dialogue (Turn)	Public Derived	16.8	QA datasets rewritten into spoken style, synthesized via TTS, filtered for quality
Dialogue (FD)	Public Derived	36.4	Turn-based dialogues composed into two-channel interactions with simulated interruptions and backchannels

D Basic Speech Understanding and Generation

Setup. We evaluate the fundamental speech capabilities of SyncAudio on ASR and streaming TTS. ASR is evaluated by word error rate (WER) on LibriSpeech (Panayotov et al., 2015) test-clean and test-other sets, while streaming TTS is evaluated by WER on the English subset of Seed-TTS-Eval [6]. We compare against representative spoken dialogue systems as well as task-specific ASR and TTS models.

Results. As shown in Table 5, SyncAudio achieves WERs of 3.08 and 8.17 on LibriSpeech test-clean and test-other sets, respectively, improving over FixedAudio (3.30/9.27) and remaining competitive with existing spoken dialogue systems. On streaming TTS, SyncAudio achieves a WER of 2.53, compared with 2.65 for FixedAudio, and obtains the lowest WER among the compared

[4] `https://huggingface.co/datasets/Mxode/I_Wonder_Why-Chinese`
[5] `https://huggingface.co/Qwen/Qwen3-30B-A3B-Instruct-2507`
[6] `https://github.com/BytedanceSpeech/seed-tts-eval`

systems. These results show that the proposed dynamic synchronization text-speech decoding strategy preserves strong speech understanding while supporting highly intelligible streaming speech generation.

Table 5: ASR performance (WER ↓) on LibriSpeech and streaming TTS performance (WER ↓) on Seed-TTS-Eval (English). Best results in bold, second best underlined.

(a) ASR: LibriSpeech WER (↓)

Model	test-clean	test-other
Whisper-large-v3	**2.01**	**3.91**
OmniFlatten	7.91	19.21
Mini-Omni2	4.80	9.80
GLM-4-Voice	<u>2.82</u>	<u>7.66</u>
FLM-Audio	3.20	–
CF-Duplex	3.32	8.93
FixedAudio	3.30	9.27
SyncAudio	3.08	8.17

(b) Streaming TTS: Seed-TTS-Eval WER (↓)

Model	Streaming	eval-en
CosyVoice2	Half	2.57
FlashTTS	Full	<u>2.55</u>
VoXtream	Full	3.81
GLM-4-Voice	Full	2.91
FLM-Audio	Full	2.95
CF-Duplex	Full	2.89
FixedAudio	Full	2.65
SyncAudio	Full	**2.53**

E Text-speech Decoding Consistency Across Different Length Groups

We further analyze the text-speech decoding consistency of **FixedAudio** and **SyncAudio** across different response length groups. For each benchmark, we group the samples into four equal-size windows by the word count of the textual response, where W_1–W_4 run from the shortest to the longest responses.

SyncAudio yields overall better text-speech decoding consistency. The advantage is most pronounced on longer textual responses (W_3–W_4), where it attains the largest gain versus FixedAudio on 8 of 10 WER and TSG metrics. This is expected: as the textual sequence grows, FixedAudio's fixed text–speech token ratio is more likely to desynchronize the two streams, whereas SyncAudio keeps token-level alignment through adaptive group-size prediction, independent of response length. On shorter groups (W_1–W_2), the two models are comparable, and SyncAudio is often as good or better, particularly on WER.

Table 6: Text-speech decoding consistency by generated text length.

		LLaMAQ		**TriviaQ**		**WebQ**		**AlpacaEval**		**Avg.**	
	Model	WER	TSG	WER	TSG	WER	TSG	WER	TSG	WER	TSG
Overall	FixedAudio	6.5	10.7	8.9	4.4	17.3	5.3	16.2	15.0	13.2	8.9
	SyncAudio	**5.8**	**9.7**	**7.7**	**3.6**	**15.8**	**4.5**	**14.8**	**13.6**	**11.9**	**7.9**
	$\Delta_{\mathrm{F-S}}$	+0.7	+1.0	+1.2	+0.8	+2.1	+0.8	+1.5	+1.4	+1.3	+1.0
W_1	FixedAudio	9.0	25.7	7.5	3.6	11.1	3.6	7.6	13.1	8.5	11.5
	SyncAudio	**8.3**	25.7	8.0	4.8	**8.6**	4.0	**6.6**	13.5	**7.7**	12.0
	$\Delta_{\mathrm{F-S}}$	+0.7	+0.0	−0.5	−1.2	**+2.5**	−0.4	+1.0	−0.4	+0.8	−0.5
W_2	FixedAudio	5.3	5.3	9.5	3.2	8.6	5.6	9.6	13.6	9.0	6.9
	SyncAudio	7.2	**4.0**	**8.8**	4.0	**6.7**	**5.2**	9.6	**10.8**	**8.4**	**6.0**
	$\Delta_{\mathrm{F-S}}$	−1.9	+1.3	+0.7	−0.8	+1.9	+0.4	+0.0	**+2.8**	+0.6	+0.9
W_3	FixedAudio	5.4	4.0	9.2	6.4	7.3	4.8	13.2	15.6	9.4	7.7
	SyncAudio	**4.9**	6.7	**8.3**	**3.2**	**6.4**	5.6	**12.0**	**14.4**	**8.1**	**7.5**
	$\Delta_{\mathrm{F-S}}$	+0.5	−2.7	+0.9	**+3.2**	+0.9	−0.8	+1.2	+1.2	**+1.3**	+0.2
W_4	FixedAudio	6.8	8.0	8.9	4.4	23.4	7.2	28.1	17.6	17.7	9.3
	SyncAudio	**5.5**	**2.7**	**6.7**	**2.0**	23.9	**3.2**	**25.0**	**15.6**	**16.6**	**5.9**
	$\Delta_{\mathrm{F-S}}$	**+1.3**	**+5.3**	**+2.2**	+2.4	−0.5	**+4.0**	**+3.1**	+2.0	+1.1	**+3.4**

F SPEECH-TO-TEXT DIALOGUE EVALUATION

We additionally evaluate SyncAudio under the **speech-to-text dialogue (S2TD)** setting, where the audio head is disabled and the model generates only a textual response to a spoken query. This setting isolates the model's text-generation capability from the text-speech joint decoding mode and provides a complementary view of how well the knowledge and semantic capabilities of the underlying LLM are preserved.

We evaluate on the same four spoken QA benchmarks used in the main paper: LLaMAQ, TriviaQA, WebQ, and AlpacaEval. In addition to SyncAudio and FixedAudio, we report the original Qwen3-4B text model as an upper reference and CF-Duplex as a representative spoken dialogue baseline.

Table 7: S2TD question answering performance.

Method	Mode	LLaMAQ	TriviaQ	WebQ	AlpacaEval
Qwen3-4B	T→T	74.0	57.7	55.2	5.01
CF-Duplex	S→T	59.3	21.3	26.6	2.91
FixedAudio	S→T	68.3	34.3	36.7	4.10
SyncAudio	S→T	**69.3**	**36.1**	**38.3**	**4.60**

As shown in Table 7, SyncAudio consistently improves over FixedAudio across all four benchmarks, including gains from 68.3 to 69.3 on LLaMAQ, 34.3 to 36.1 on TriviaQA, 36.7 to 38.3 on WebQ, and 4.10 to 4.60 on AlpacaEval. Although a gap remains relative to the original text-only Qwen3-4B model, SyncAudio retains substantially stronger QA performance than the spoken dialogue baselines considered here. These results complement the S2TSD evaluation in the main paper and indicate that SyncAudio preserves strong text-generation capabilities even when operating in text-only response mode.

G REAL-TIME DUPLEX INTERACTION

Real-Time Extension. SyncAudio operates on user speech, model text, and model speech within the same recurrent generation framework, making it straightforward to extend the model from turn-based dialogue to continuous full-duplex speech interaction, as shown in Figure 4. We introduce modality-specific control tokens that allow the text and speech pathways to wait when no output should be produced and to react to newly arriving user speech during model generation. The resulting model, termed **SyncAudio-Realtime**, preserves the same shared LLM backbone, cross-modal fusion module, text head, audio head, and synchronization mechanism as the turn-based model. The corresponding fixed-synchronization variant is denoted as **FixedAudio-Realtime**.

During real-time full-duplex interaction, streaming user-speech representations continue to enter the shared context while the model is generating its response. The model can therefore update its output behavior according to newly observed user speech without switching to a separate interaction architecture. For training, full-duplex dialogue data with interruption and backchannel behaviors are incorporated so that the model learns when to wait, take turn, or respond to user interruptions and backchannels.

Evaluation Protocol. We evaluate real-time full-duplex spoken interaction using Full-Duplex Bench v1.0 (Lin et al., 2025). The benchmark considers three representative conversational behaviors: *interruption handling*, *smooth turn-taking*, and *backchanneling*. Following the benchmark protocol, we report turn-over rate (TOR), response latency, and the benchmark's GPT-based score for interruption handling; TOR and latency for smooth turn-taking; and TOR, backchannel frequency, and Jensen–Shannon divergence (JSD) for backchanneling. We compare SyncAudio-Realtime and FixedAudio-Realtime with representative real-time spoken dialogue systems including dGSLM (Nguyen et al., 2023), Freeze-Omni (Wang et al., 2025b), Moshi Défossez et al. (2024), and Gemini Live.

Results. Table 8 reports the evaluation results. For interruption handling, SyncAudio-Realtime achieves a TOR of 1.000 with a response latency of 0.346 s, compared with 0.990 and 0.445 s for

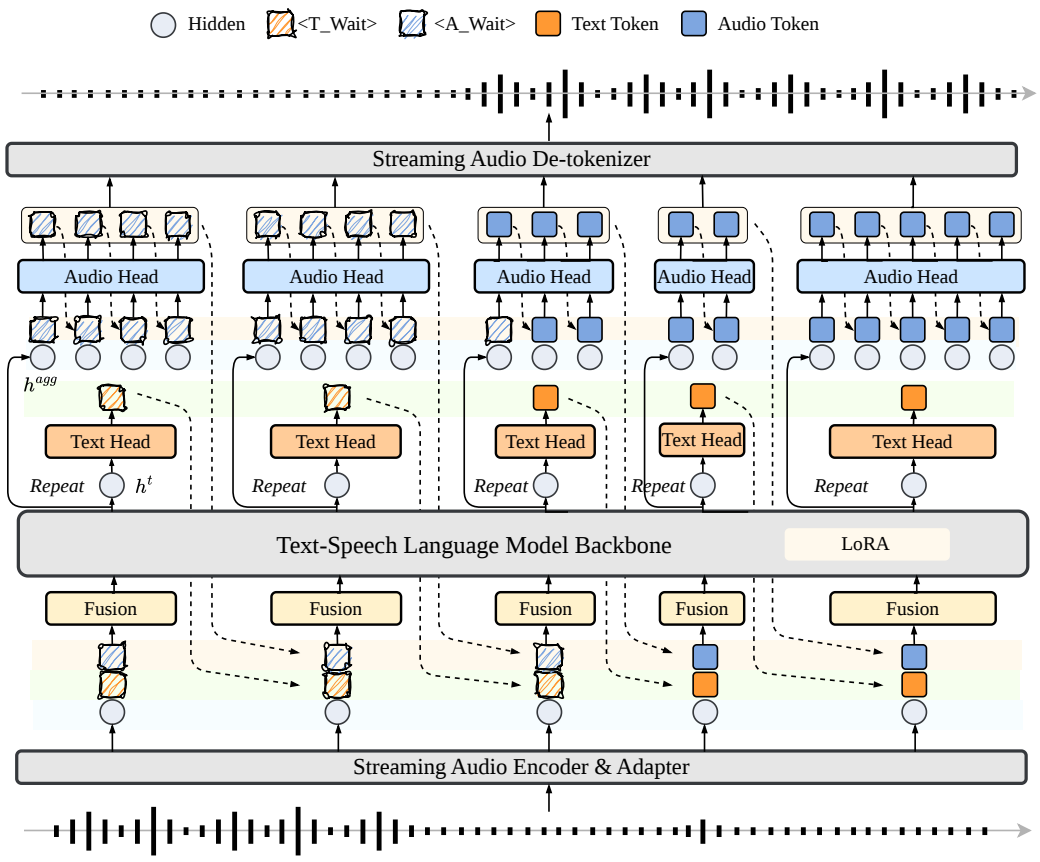


Figure 4: Diagram of the extended SyncAudio-Realtime system.

FixedAudio-Realtime. The GPT-based score is slightly worse than FixedAudio-Realtime, indicating that the benefit of dynamic synchronization is not uniform across all interruption metrics.

The advantage of SyncAudio-Realtime over FixedAudio-Realtime is clearer for smooth turn-taking. SyncAudio-Realtime realizes better TOR and latency than FixedAudio-Realtime while also being competitive to other baselines. For backchanneling, though SyncAudio-Realtime is worse than FixedAudio-Realtime on take over rate, it realizes best performance for Freq and JSD across all compared models.

Overall, these results show that the same SyncAudio generation framework can support full-duplex speech interaction scenarios such as interruption, turn-taking, and backchannel behaviors without requiring a separate duplex architecture. While SyncAudio-Realtime does not uniformly improve every benchmark metric, it provides substantially better smooth turn-taking and lower latency than its FixedAudio-Realtime.

Table 8: Results on Full-Duplex Bench v1.0.

Models	User Interruption			Smooth Turn Taking		Backchannel		
	TOR (↑)	GPT-4o (↑)	Latency (↓)	TOR (↑)	Latency (↓)	TOR (↓)	Freq (↑)	JSD (↓)
dGSLM	0.917	0.201	2.531	0.975	0.352	0.691	0.015	0.934
Freeze-Omni	0.867	3.615	1.409	0.336	0.953	0.636	0.001	0.997
Moshi	**1.000**	0.765	**0.257**	0.941	0.265	1.000	0.001	0.997
Gemini Live	0.891	3.376	1.183	0.655	1.301	0.091	0.012	0.896
FixedAudio-Realtime	0.990	3.850	0.445	0.479	0.537	0.836	0.039	0.798
SyncAudio-Realtime	**1.000**	3.570	<u>0.346</u>	<u>0.958</u>	<u>0.272</u>	1.000	**0.192**	**0.650**