Lychee-FD: Hierarchical Acoustic-Semantic Modeling for Full-Duplex Spoken Language Models

Anonymous ACL submission

Abstract

Spoken Language Models (SLMs) have revolutionized voice interaction, yet they remain constrained by rigid half-duplex mechanisms that fail to replicate the fluidity of human conversation. While recent Full-Duplex SLMs attempt to bridge this gap by enabling real-time capabilities such as interruption and backchanneling, these methods suffer from severe modality interference. Specifically, adapting models for native full-duplex interaction often induces significant knowledge degradation, impeding the realization of seamless human-machine interaction. To address this, we conduct an optimization dynamics analysis, identifying the root cause as the inherent gradient conflict between acoustic rendering and semantic modeling within a shared parameter space. Guided by this insight, we introduce **Lychee-FD**, a native end-to-end full-duplex framework designed to mitigate modality interference. We proposed a hierarchical parameter separation strategy that decouples conflicting modalities in deep layers. Moreover, we incorporate a semantic alignment channel that enables the model to preserve coherent internal monologues, ensuring the robustness of semantic modeling during training. Extensive experiments demonstrate that our method achieves state-of-the-art performance across multiple full-duplex benchmarks, specifically delivering an average **7.4%** improvement on Spoken QA tasks and **28.5%** improvement on FullDuplexBench 1.5. Consequently, our work uncovers the fundamental causes of modality interference within Full-Duplex SLMs and provides an effective approach to reconcile interaction efficiency with robust knowledge retention.

1 Introduction

The rapid evolution of Large Language Models (LLMs) has fundamentally reshaped our daily lives, establishing them as ubiquitous assistants capable of complex reasoning and instruction following.

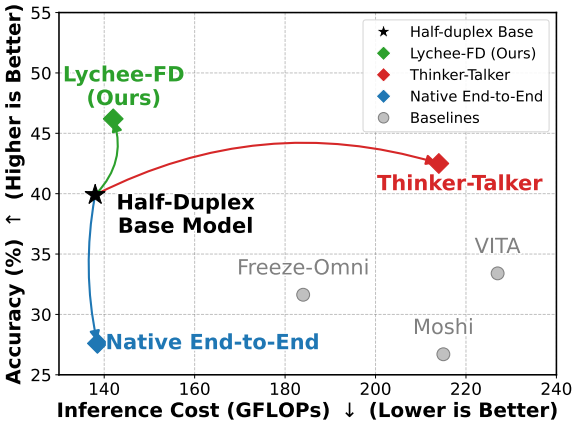


Figure 1: Visualization of the efficiency and intelligence trade-off. Existing paradigms face a dilemma when developing a full-duplex SLM from the half-duplex model (**black star**): End-to-End models (**blue diamond**) sacrifice accuracy for efficiency, while Thinker-Talker models (**red diamond**) preserve knowledge but incur prohibitive inference costs. In contrast, our proposed Hierarchical framework (**green diamond**) combining low latency with high accuracy significantly outperform baselines.

Within this landscape, Spoken Language Models (SLMs) represent a significant paradigm shift from text-based to voice-based interaction. By enabling hands-free scenarios (e.g. in-car assistants), SLMs have drastically enhanced the efficiency and accessibility of human-machine interaction. While recent advancements have led to Omni-modal models capable of seamless voice interaction (Xu et al., 2025; Wu et al., 2025; Zhan et al., 2024; OpenAI et al., 2024), a critical disparity remains: unlike authentic human conversation, which is inherently real-time, overlapping, and turn-free, most current SLMs are constrained to a rigid half-duplex mode. This turn-taking mechanism disrupts the fluidity of interaction, creating an artificial barrier between user and agent. Consequently, developing Full-Duplex SLMs (FDSLMs) that support simultaneous listening and speaking is widely regarded as the next critical milestone in human-machine inter-

action, promising to unlock a truly natural, fluid, and immersive conversation experience.

Although emerging research has achieved notable progress in FDSLM development (Fu et al., 2025; Wang et al., 2025b; Défossez et al., 2024; Fu et al., 2025), current methods still suffer from severe *Modality Interference*. As illustrated in Figure 1, adapting a half-duplex base model (black star) into a native End-to-End architecture (blue diamond) precipitates a significant knowledge degradation. This degradation is further corroborated by state-of-the-art models like Moshi (Défossez et al., 2024), which reports a significant 12.7% drop in accuracy on LlamaQ and a 5.7% drop on WebQ after full-duplex alignment. While some approaches (Wang et al., 2025b; Chen et al., 2025b) attempt to circumvent this interference by adopting Thinker-Talker architectures (red diamond), they require intricate multi-stage training and introducing significant latency. These indicate that existing paradigms either fail in knowledge retention or inference efficiency. Consequently, the central research question of this work is: *How can we resolve this modality interference to simultaneously achieve high inference efficiency and robust knowledge retention in FDSLMs?*

To address this question, we first conduct an **optimization dynamics analysis** to investigate the root causes of modality interference. We have two critical observations: (1) the gradient directions of text and speech objectives are synergistic in shallow layers but become increasingly orthogonal in deeper layers; and (2) aligning sparse text with dense audio sequences causes the optimization landscape to be dominated by acoustic gradients, thereby suppressing semantic learning.

Inspired by these insights, we introduce **Lychee-FD**, a native end-to-end full-duplex framework designed to mitigate modality interference with two architectural innovations. First, we propose a **hierarchical parameter separation** strategy for gradient conflict in deep layers. Specifically, we separate the deep layers into independent acoustic and semantic heads. By executing these heads in parallel, we maintain the original model depth, thereby preserving inference efficiency. Second, to counter semantic dilution, we introduce a **semantic alignment channel** to generate coherent internal monologues. By utilizing continuous textual supervision as a semantic anchor, we preserve the robustness of semantic modeling during training. Extensive experiments demonstrate that Lychee-FD achieves state-of-the-art performance, specifically delivering an average 7.4% improvement on Spoken QA tasks and a 28.5% gain on FullDuplexBench 1.5. Ultimately, our approach effectively mitigates modality interference, simultaneously achieving high inference efficiency and robust knowledge retention.

Our contributions are summarized as follows:

- We delve into the optimization dynamics of FDSLM training and identify the gradient conflict in deep layers alongside the semantic dilution caused by sparse alignment, thereby empirically demonstrating the interference between semantic and acoustic modeling.
- Guided by these insights, we propose Lychee-FD, a native end-to-end full-duplex framework aiming to mitigate modality interference. It features a hierarchical parameter separation strategy to disentangle conflicting modalities in deep layers and a semantic alignment channel to enforce knowledge retention.
- Extensive experimental results show that Lychee-FD achieves state-of-the-art performance in both knowledge accuracy (+7.4% on Spoken QA) and interaction quality (+28.5% on FullDuplexBench 1.5), all while maintaining the inference efficiency of native end-to-end architectures.

2 Related Work

2.1 Spoken Language Model

SLMs have evolved from cascading ASR-LLM-TTS pipelines (Zhang et al., 2023; An et al., 2024; Chen et al., 2025a) to unified architectures. Current methods can be broadly categorized into Thinker-Talker architectures and Native End-to-End architectures. The Thinker-Talker paradigm decouples acoustic generation from the LLM backbone to mitigate modality interference (Xu et al., 2025; Wang et al., 2025b; Fang et al., 2025). For instance, Xu et al. (2025) adopts a dual-track autoregressive architecture. However, despite their stability, these systems often require intricate, multi-stage training curricula and suffer from inference bottlenecks caused by the separate generation modules. Conversely, Native End-to-End architectures embed speech and text into a shared semantic space, enabling direct speech-to-speech reasoning (Xie and Wu, 2024a; Mitsui et al., 2024; Gao et al., 2025; Zeng et al., 2024). Notable examples include

Step-Audio2 (Wu et al., 2025), which introduces latent audio encoding to capture paralinguistic cues. Crucially, unlike existing SLMs which are constrained to rigid turn-taking mechanisms, our work advances the field by establishing a Full-Duplex framework. By enabling simultaneous listening and speaking, we aim to unlock a truly natural, fluid, and immersive paradigm for future human-machine interaction.

2.2 Full Duplex Dialog Model

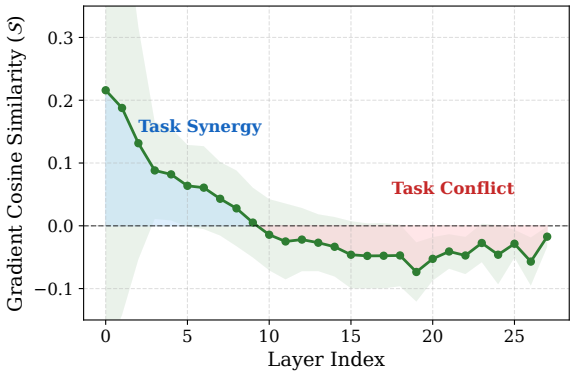
To achieve turn-free and fluid conversation, recent research has explored full-duplex paradigms that transcend rigid turn-taking. Early efforts primarily focused on System-Level solutions, which leverage Voice Activity Detection (VAD) as a dialog manager to control the speaking process of half-duplex SLMs (Zhang et al., 2025a; Chen et al., 2025a; Xie and Wu, 2024b; Liao et al., 2025; Li et al., 2025). Although making some success, these cascaded systems suffer from high latency and error propagation, prompting a shift towards unified modeling.

To address these limitations, recent works employ the SLM itself as the dialog manager, enabling simultaneous listening and speaking. Time-Division Multiplexing (TDM) approaches flatten listening and speaking tokens into a single temporal sequence, leading to increasing computational complexity and limiting long-context interaction (Zhang et al., 2025b; Veluri et al., 2024; Zhang et al., 2024; Yu et al., 2024). Instead, Channel-Division Multiplexing (CDM) approaches explicitly model concurrent input and output streams, theoretically offering the most integrated form of interaction (Team et al., 2025; Nguyen et al., 2023; Yao et al., 2025). For example, Défossez et al. (2024) and Chen et al. (2025b) integrate time-aligned text and audio streams to provide semantic guidance for speech generation. Unlike previous works that struggle with the efficiency-intelligence trade-off, our Lychee-FD framework effectively mitigates modality interference within the native end-to-end FDSLM, simultaneously achieving high inference efficiency and knowledge retention.

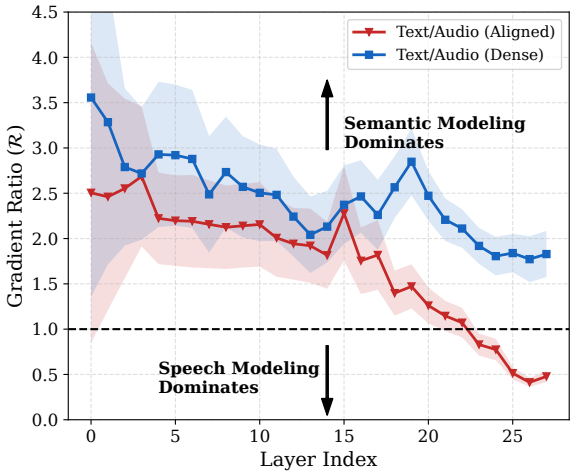
3 Lychee-FD

3.1 Motivation

To empirically investigate the underlying causes of the modality interference and understand the learning process within full-duplex models, we conducted an optimization dynamic analysis, shown in Figure 2. We utilized a native CDM architecture initialized with weights from StepAudio2-mini (Wu et al., 2025). We performed forward passes on 1K samples from the training set to accumulate gradients for both the token generation task ($\mathcal{L}_{text}$) and the speech token generation task ($\mathcal{L}_{speech}$), without updating the model parameters.



(a) Gradient Cosine Similarity



(b) Gradient Magnitude Ratio

Figure 2: **Optimization Dynamics Visualization. (a) Gradient Cosine Similarity:** The transition to negative values in deep layers reveals conflicting optimization directions between semantic and acoustic modeling, motivating our hierarchical parameter separation. **(b) Gradient Magnitude Ratio:** The consistently lower ratio for "Aligned" (Red) compared to "Dense" (Blue) indicates that sparse alignment dilutes semantic supervision, motivating our semantic alignment channel.

$$\mathbf{g}_{text}^{(l)} = \nabla_{\theta^{(l)}} \mathcal{L}_{text}, \quad (1)$$

$$\mathbf{g}_{speech}^{(l)} = \nabla_{\theta^{(l)}} \mathcal{L}_{speech}, \quad (2)$$

where $\nabla_{\theta^{(l)}}$ denotes the gradient operator with respect to the layer parameters $\theta^{(l)}$, and the results are flattened into vectors. By analyzing the geometric relationships of these gradient vectors, we can quantitatively explore the interaction dynam-

ics between the two modalities during the learning process.

Modality Interference. We investigated the compatibility of the two optimization objectives by calculating the cosine similarity $\mathcal{S}^{(l)}$ between text and speech gradient vectors:

$$\mathcal{S}^{(l)} = \cos(\mathbf{g}_{text}^{(l)}, \mathbf{g}_{speech}^{(l)}). \quad (3)$$

As shown in the Figure 2a, the similarity reveals a distinct layer-wise pattern. In the shallow layers (0-9), the cosine similarity is positive, indicating that the two modalities share synergistic optimization directions, focusing on common low-level features processing. However, as the depth increases, the similarity drops sharply, turning negative and fluctuating in the deeper layers. This trend empirically confirms our hypothesis regarding the dual nature of speech: while shallow layers can share representations, the deep layers face a fundamental conflict between semantic modeling and acoustic rendering. Forcing a unified set of parameters to resolve these opposing gradient directions inevitably leads to sub-optimal performance, validating the root cause of the observed modality interference.

Semantic Dilution. Prevalent Full-Duplex SLMs typically address the frequency mismatch between text (approximately 3Hz) and audio (typically 25Hz) by interleaving padding tokens to enforce temporal alignment (Défossez et al., 2024; Wu et al., 2025; Chen et al., 2025b). To evaluate the impact of this alignment on optimization, we compared the ratio $\mathcal{R}^{(l)}$ of gradient magnitudes between the two modalities:

$$\mathcal{R}^{(l)} = ||\mathbf{g}_{text}^{(l)}||/||\mathbf{g}_{speech}^{(l)}||. \quad (4)$$

As shown in the Figure 2b, we observe a substantial disparity between the continuous text supervision (Dense) and the sparse text with padding (Aligned). Specifically, the gradient magnitude ratio in the Aligned setting is consistently suppressed across all layers, suggesting that the introduction of padding tokens effectively dilutes the density of semantic supervision. Consequently, the optimization dynamics become dominated by acoustic reconstruction, which drives the observed degradation in knowledge retention.

3.2 Model Design

As illustrated in Figure 3, existing FDSLM architectures face a dilemma: Thinker-Talker models mitigate modality interference but incur high latency and redundancy, while Native End-to-End models offer efficiency but suffer from semantic dilution and optimization conflicts. To resolve this trade-off, we propose Lychee-FD, a native end-to-end framework built upon the Step-Audio-2 (Wu et al., 2025) backbone. Our design introduces two key innovations: the **Hierarchical Parameter Separation** to disentangle conflicting modalities, and the **Semantic Alignment Channel** to enforce knowledge retention.

Half-Duplex Backbone. We choose Step-Audio-2 as our half-duplex backbone due to its public availability. We employ the Whisper-v3-large encoder for input audio processing. Crucially, to ensure precise temporal alignment for full-duplex interaction, we utilize the encoder's original 25Hz frame rate, distinct from the setting adopted by Step-Audio-2. For acoustic output, we adopt the CosyVoice-2 tokenizer to convert audio into discrete speech tokens at a 25Hz frame rate.

Hierarchical Parameter Separation. Guided by the observation that acoustic and semantic gradients become orthogonal in deeper layers, we design a hierarchical Transformer architecture. We retain a unified Transformer backbone for the shallow layers to leverage shared low-level feature processing. Formally, let $\mathbf{E} \in \mathbb{R}^{L \times d}$ be the input embedding sequence. The shared representation $\mathbf{H}_{\text{shared}}$ is computed as:

$$\mathbf{H}_{\text{shared}} = \mathcal{F}_{\text{shared}}(\mathbf{E}; \theta_{\text{shared}}) \quad (5)$$

where $\mathcal{F}_{\text{shared}}$ denotes the stack of shared Transformer layers parameterized by θ_{shared}.

In the deeper layers, we physically disentangle the parameters into three specialized heads: the *Semantic Head* for text generation, the *Acoustic Head* for speech synthesis, and the *Control Head* for interaction management (e.g., stop/start signals). These heads operate in parallel:

$$\mathbf{O}^m = \mathcal{F}_{\text{head}}^m(\mathbf{H}_{\text{shared}}; \theta^m), \quad m \in \{T, A, C\} \quad (6)$$

where m represents the modality (Text, Acoustic, Control), and $\mathbf{O}^m$ denotes the output distribution for each head. The total loss $\mathcal{L}$ is computed as the summation of the next-token prediction losses for each specific head:

$$\mathcal{L} = -\sum_{m \in \{T,A,C\}} \sum_{t} \log P(y_t^m | y_{<t}, \mathbf{E}; \theta) \quad (7)$$

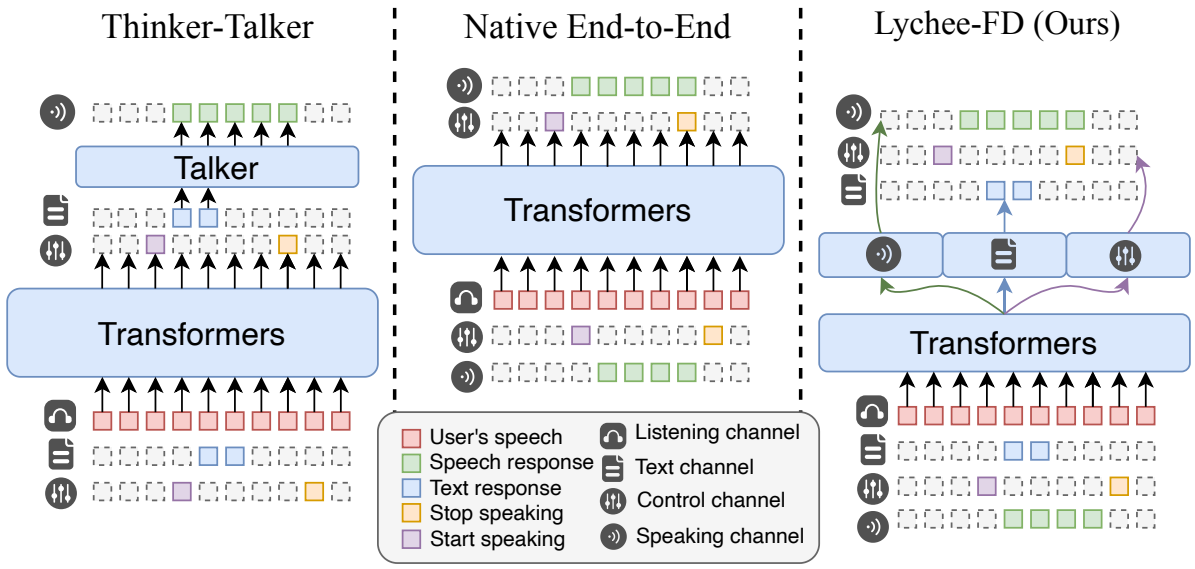


Figure 3: Two mainstream architecture paradigms of SLMs and our proposed Lychee-FD. Our design features a hierarchical parameter separation strategy to resolve deep-layer modality conflicts and the semantic alignment channel to enforce knowledge retention.

where y_t^m denotes the ground-truth token for modality m at step t. This hierarchical split effectively isolates the conflicting optimization objectives, allowing the model to articulate high-fidelity acoustic responses without corrupting its underlying semantic modeling.

Semantic Alignment Channel. To counter the semantic dilution caused by sparse alignment in speech-native tasks, we incorporate a semantic alignment channel to generate coherent internal monologues. During training, these monologues server as a semantic anchor, maintaining high-magnitude gradient flow for the language modeling objective and present robust knowledge retention. Specifically, we organize the parallel generation streams as follows:

$$\mathbf{Y}^{\text{Text}} = [t_1, t_2, \cdots, t_n, \texttt{<EOT>}, \texttt{<pad>}, \cdots, \texttt{<pad>}]$$
$$\mathbf{Y}^{\text{Audio}} = [a_1, a_2, \cdots, a_n, a_{n+1}, a_{n+2}, \cdots, \texttt{<EOS>}]$$
$$\mathbf{Y}^{\text{Ctrl}} = [\texttt{<Start>}, c_2, \cdots, \cdots, \cdots, \texttt{<Stop>}]$$

By explicitly modeling the text channel alongside the acoustic channel, we ensure high-magnitude gradient flow for the language modeling objective, thereby preserving robust knowledge retention.

3.3 Data Pipeline

Given the scarcity of open-source full-duplex datasets, we developed an automated pipeline to synthesize high-quality training data covering three key interaction behaviors: **Interruptions**, **User Backchannels**, and **AI Backchannels**. We employed multi agents to simulate realistic User-Assistant dialogues, injecting rule-based constraints to trigger diverse interruption types (e.g., topic switching, follow-up queries) and natural backchannels. To ensure acoustic robustness, we synthesized the resulting transcripts using CosyVoice 2 (Du et al., 2024), coupled with 80K predefined voice prompts for zero-shot cloning. After rigorous filtering to remove samples with logical inconsistencies or low audio quality, we curated a final dataset of approximately 140K full-duplex dialogue instances, providing a diverse and reliable foundation for our experiments. We provide more details of our data pipeline in Appendix D

4 Experiments

4.1 Baselines

To ensure a comprehensive evaluation, we compare our proposed method against a diverse set of representative and competitive full-duplex SLMs. These baselines cover the primary architectural paradigms currently explored in the field:

System-Level Full-Duplex Models include Freeze-Omni (Wang et al., 2025b) and VITA-1.5 (Fu et al., 2025). These systems achieve full-duplex interaction by integrating an external VAD module to manage the dialogue state of a standard half-duplex SLM.

Model	Type	LlamaQ		WebQ		TriviaQA		Avg.		Take Over Rate↑
		$S\rightarrow T\uparrow$	$S\rightarrow S\uparrow$	$S\rightarrow T\uparrow$	$S\rightarrow S\uparrow$	$S\rightarrow T\uparrow$	$S\rightarrow S\uparrow$	$S\rightarrow T\uparrow$	$S\rightarrow S\uparrow$	
Freeze-Omni	System-level	71.3	50.6	38.3	25.8	24.3	23.9	44.6	33.4	99.6
VITA 1.5	System-level	**75.6**	51.0	**41.8**	29.2	35.0	26.0	50.8	35.4	**100**
dGSLM	Native	–	1.3	–	0.2	–	0.4	–	0.6	**100**
FLM-audio	Native	41.3	36.7	15.6	14.5	10.5	10.4	22.4	20.5	99.5
Moshi	Native	62.3	54.6	25.3	19.6	19.1	17.4	35.5	30.5	93.8
Fun-Audio-Chat	Native	72.3	64.3	26.2	24.4	29.6	27.7	42.7	38.8	99.9
StepAudio-2-mini	Half-duplex	74.6	62.0	39.9	30.8	39.5	29.8	51.3	40.9	–
Lychee-FD (Ours)	Native	73.7	**65.4**	38.3	**33.9**	**42.5**	**39.4**	**51.5**	**46.2**	**100**
w/o *Sem-Channel*		69.3	61.0	34.1	31.5	34.2	30.1	45.9	40.8	99.6
w/o. *Param-Sep*		67.0	36.0	34.6	22.5	36.6	24.2	46.1	27.6	98.5

Table 1: Performance comparison on spoken question answering benchmarks. We report accuracy (Acc) in both speech-to-text ($S\rightarrow T$) and speech-to-speech ($S\rightarrow S$) settings. We also report average take over rate (TOR) across three benchmarks. ↑ indicates higher is better. **Bold** denotes best results and underlined denotes second best.

Native Full-Duplex Models include dGSLM (Nguyen et al., 2023), FLM-Audio (Yao et al., 2025), Moshi (Défossez et al., 2024), and Fun-Audio-Chat (Chen et al., 2025b), which intrinsically support full-duplex interaction within the LLM. Specifically, Fun-Audio-Chat adopts a Thinker-Talker architecture, while the others utilize a CDM architecture.

4.2 Implementation Details

We optimize our model using AdamW (Loshchilov and Hutter, 2019) with a cosine learning rate scheduler. All experiments are conducted on 8 NVIDIA H20 GPUs, with a global batch size of 32 and a learning rate of 3e-6. We set the warmup ratio to 0.1 and train 1 epoch, which takes approximately 16 hours. For inference, we employ greedy sampling for both text and speech token generation. We evaluate our model with three random seeds and report their average performance. Regarding the hierarchical architecture configuration, unless otherwise specified, we utilize a shared backbone of 24 Transformer layers. The specialized heads are configured with 4 layers for the text channel, 4 layers for the speech channel, and 2 layers for the control channel.

4.3 Spoken Question Answering

Metrics. To evaluate the spoken question answering capabilities of our model, we follow previous work (Défossez et al., 2024) and utilize three standard benchmarks: LlamaQ, WebQ, and TriviaQA. We report the accuracy (Acc) under both speech-to-text ($S \rightarrow T$) and speech-to-speech ($S \rightarrow S$) settings. For speech-to-speech setting, we leverage Whisper-large-v3 (Radford et al., 2023) to obtain the transcription of generated speech. Additionally, we report the take over rate (TOR) across three benchmarks to quantify the frequency of model responses, serving as an indicator of the model's turn-taking behavior.

Result. As presented in Table 1, Lychee-FD demonstrates superior spoken question answering capabilities. It achieves the highest average accuracy across both speech-to-text ($S \rightarrow T$) and speech-to-speech ($S \rightarrow S$) settings. Compared to the previous SOTA native full-duplex model, Fun-Audio-Chat, Lychee-FD delivers a substantial improvement of 7.4% in $S \rightarrow S$ accuracy and 8.8% in $S \rightarrow T$ accuracy. Even when compared to system-level pipelines like VITA-1.5, our end-to-end approach demonstrates superior reasoning capabilities (10.8% in $S \rightarrow S$ and 0.7% in $S \rightarrow T$), validating the effectiveness of our framework. Furthermore, Lychee-FD maintains a perfect Take Over Rate (TOR) of 100%, confirming that this exceptional knowledge retention is achieved without compromising interaction stability.

Ablation. To investigate the individual contributions of our two proposed innovations, we conducted ablation studies on two variants: *w/o Sem-Channel*, which replaces the semantic alignment channel with sparse time-aligned text, and *w/o. Param-Sep*, which removes the hierarchical parameter separation to employ a fully shared architecture. As shown in Table 1, when applying the time-aligned text, we observe a significant performance decline in both $S \rightarrow T$ (5.6%) and $S \rightarrow S$ (5.4%) settings. This parallel degradation confirms our hypothesis regarding semantic dilution: the sparse supervision provided by time-aligned text fails to sustain robust linguistic modeling, which in turn cause knowledge degradation. In contrast,

Model	FDBench						FullDuplexBench 1.0						FullDuplexBench 1.5			
	SRR↑	SIR↑	EIR↓	SRIR↑	FSED.↓	IRD↓	I-TOR↑	B-Freq↑	B-TOR↓	T-TOR↑	P-TOR↓	Stop↓	IRR↑	BRR↑	Stop↓	Lat.↓
dGSLM	–	–	–	–	–	–	91.7	1.5	69.1	97.5	93.5	2523	–	–	–	–
Freeze-Omni	12.9	57.2	25.7	29.5	667	5413	77.5	0.1	63.6	33.6	46.3	1380	27.0	63.0	660	2066
VITA 1.5	21.0	46.1	16.3	78.3	3036	9925	**99.5**	2.5	81.8	58.8	88.8	1523	6.0	38.0	1222	2140
FLM-audio	7.5	69.4	1.0	0.9	989	3408	91.0	0.3	61.8	96.6	56.5	4579	10.0	43.0	2439	983
Moshi	41.4	78.8	22.1	73.9	1895	1421	87.5	5.1	36.4	76.4	54.1	885	61.0	26.0	1071	3034
Lychee-FD (Ours)	**86.3**	**99.7**	**0.4**	**95.8**	**637**	**1210**	94.5	**14.6**	**23.4**	**98.3**	**10.0**	**840**	**78.0**	**69.0**	**570**	**826**

Table 2: Performance comparison of full-duplex interaction capabilities and efficiency. Despite interaction behavior metrics, we also report efficiency metrics of each benchmark (FSED, IRD, Lat. Stop.), measured in milliseconds (ms). ↑ indicates higher is better. **Bold** denotes best results and underlined denotes second best.

the *w/o Param-Sep* setting reveals a severe modality conflict. While its text generation capability remains relatively stable, its speech accuracy suffers a catastrophic drop to 27.6%. This disparity validates our optimization dynamics analysis: without physically disentangling the parameters, the optimization landscape becomes dominated by the semantic modeling, suppressing the learning of acoustic features in deep layers.

Discussion. We highlight two critical phenomena that distinguish Lychee-FD from existing paradigms. First, our model not only recovers the performance of its half-duplex backbone (StepAudio-2-mini) but surpasses it in both $S \rightarrow T$ (+0.2%) and $S \rightarrow S$ (+5.3%) settings. This performance gain, achieved without increasing model depth, strongly validates that our framework realizes robust knowledge retention while maintaining the inference efficiency of the native end-to-end model. This underscores the pivotal role of explicit semantic modeling in driving model intelligence, offering valuable insights for future human-machine interaction systems. Second, Lychee-FD achieves the smallest modality gap among all native CDM models (e.g., FLM-Audio and Moshi) and remains competitive with decoupled Thinker-Talker architectures, all without requiring intricate multi-stage training. This demonstrates that by physically decoupling semantic and acoustic modeling, Lychee-FD effectively resolves modality interference. Consequently, our approach allows the model to articulate high-fidelity acoustic responses without corrupting its underlying semantic logic, providing a streamlined solution to the efficiency-intelligence trade-off.

4.4 Full-duplex Chatting

Metrics. For the Full-duplex Chatting, we select three mainstream benchmarks: **FDBench** (Wang et al., 2025a), **FullDuplexBench 1.0**, and **FullDuplexBench 1.5** (Lin et al., 2025). We follow the recommended settings of these benchmarks for evaluating both baselines and our method. Specifically, FDBench assesses turn-taking and interruption behaviors using Success-Replies Rate (SRR), Success-Interrupts Rate (SIR), Early-Interrupts Rate (EIR), and Success-Replies-to-Interrupts Rate (NIR), alongside timing metrics such as First-Speech-Emit-Delay (FSED) and Interrupt-Response-Delay (IRD). FullDuplexBench 1.0 consists of four subsets: interruption (I), assistant backchannel (B), turn-taking (T), and user pause (P). In addition to the take over rate (TOR) and Interruption step delay (Stop.), we also report the frequency of generated backchannels during user speech (B-Freq). Finally, FullDuplexBench 1.5 utilizes GPT-4o-1124 to classify model responses to user interruptions and backchannels into four categories (Response, Resume, Uncertain, or Unknown), reporting the Interruption-Response Rate (IRR), Backchannel-Resume Rate (BRR), as well as the interruption stop delay (Stop) and response latency (Lat.).

Result. As presented in Table 2, Lychee-FD achieves state-of-the-art performance across 10 of the 11 evaluated interaction metrics, demonstrating superior capability in managing complex conversational dynamics, including interruption, backchanneling, dynamic turn-taking, and pause handling. While VITA-1.5 exhibits a marginally higher I-TOR on FullDuplexBench 1.0, its consistently high B-TOR and P-TOR reveal a tendency towards aggressive, indiscriminate speech rather than intelligent turn-taking. In contrast, Lychee-FD maintains a balanced interaction profile, effectively distinguishing between user pauses and interruptions. Notably, on the challenging FullDuplexBench 1.5, our model delivers a substantial 28.5% average improvement over system-level baselines (e.g., Freeze-Omni) that rely on external VAD. This result not only proves that Lychee-FD realizes truly natural, fluid, and immersive full-

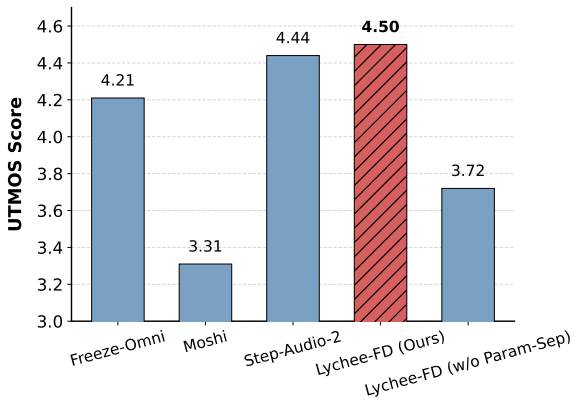


Figure 4: Comparison of speech synthesis quality measured by UTMOS. Lychee-FD (*w/o Param-Sep*) denotes the variant without hierarchical parameter separation strategy.

duplex interaction but also validates that LLMs can function as dialog managers. By learning interaction policies directly from data, our approach eliminates the need for handcrafted signal processing modules, offering compelling insights for the future of fully native end-to-end full-duplex architectures.

Latency. We evaluate inference efficiency through two categories: first response latency (FSED, Lat.) and interruption response latency (IRD, Stop.). Regarding first response latency, Lychee-FD achieves lowest latency across all benchmarks. Since our hierarchical parameter separation strategy introduces no additional model depth, we can leverage standard pipeline parallelism for speech up, demonstrating the efficiency of our architectures. For Interruption Response Latency, our model demonstrates an even greater advantage, achieving the lowest stop latency when meeting an interruption (e.g., 570ms Stop. on FullDuplexBench 1.5). It is worth noting that interruption latency is a function of both model processing speed and interaction accuracy—since a model must first correctly identify an interruption before it can stop generating. The superior performance confirms that our architecture incurs no computational overhead, and that its robust semantic awareness enables rapid, accurate reactions to user interventions.

4.5 Speech Generation

Metrics To investigate the quality of the generated speech, we evaluate the content consistency and speech naturalness of the model on LlamaQ dataset. Specifically, we report the Word Error Rate (WER) between the generated text and the transcribed speech to measure content consistency. Additionally, we employ UTMOS (Saeki et al., 2022), a trained speech quality assessment model, to score the naturalness of the generated audio.

Result High-fidelity speech synthesis serves as the cornerstone of voice interaction, significantly elevating the immersive quality of full-duplex conversations. As illustrated in Figure 4, Lychee-FD achieves the highest UTMOS score of 4.50, surpassing Freeze-omni, Moshi and even its half-duplex backbone (i.e. Step-Audio-2). This result empirically validates that our hierarchical parameter separation strategy effectively prevents acoustic modeling from semantic interference, thereby preserving fine-grained prosodic details that are often lost in shared architectures. When the parameter separation is removed, we observe a significant decline in speech quality. This finding corroborates our hypothesis regarding modality conflict from an acoustic perspective, demonstrating that forcing the acoustic head to share deep layers with text processing objectives inevitably degrades audio fidelity. Ultimately, by physically disentangling the modeling pathways, Lychee-FD not only improves generation accuracy but also synthesizes speech with richer acoustic details, delivering a more natural and enjoyable conversational experience.

5 Conclusion

In this paper, we tackled the critical challenge of modality interference in Full-Duplex Spoken Language Models. Through a rigorous optimization dynamics analysis, we identified gradient conflicts in deep layers and semantic dilution from sparse alignment as the primary obstacles hindering simultaneous listening and speaking. Inspired by the observation, we introduced Lychee-FD, a native end-to-end framework that disentangles conflicting modalities via hierarchical parameter separation and enforces knowledge retention through a semantic alignment channel. Experimental results demonstrate that Lychee-FD achieves state-of-the-art performance in both knowledge accuracy and interaction quality, effectively reconciling the conflict between real-time interactivity and deep semantic understanding. Our work reconciles language intelligence and speech intelligence within a unified architecture, paving the way for the next generation of natural, fluid, and immersive human-machine interaction.

Limitations

Despite our discoveries and improvements, we must acknowledge certain limitations in our work:

First, while Lychee-FD excels in standard Spoken QA tasks, its performance on complex reasoning questions remains suboptimal. We attribute this limitation to the current data pipeline, which primarily consists of commonsense QA and daily talk. The scarcity of complex reasoning chains in the training corpus restricts the model's ability to handle intricate logical deductions, indicating a need for more diverse and cognitively demanding training data.

Second, although the model demonstrates a robust ability to cease speech upon interruption, its subsequent response behavior exhibits occasional inconsistency. As illustrated in Appendix C, there are instances where the model fails to address the semantic content of the interruption. This suggests that while the acoustic reaction is fast, the semantic context switching requires further refinement to ensure better instruction following during dynamic turn-taking.

Third, regarding architectural generalizability, our experimental validation is currently confined to the StepAudio-2 architecture. This constraint stems from the scarcity of publicly available half-duplex models. Consequently, the scalability of our framework across different model sizes and diverse architectures remains to be fully explored.

These limitations highlight critical avenues for future research, including enriching the training corpus to enhance reasoning capabilities, refining the policy for interruption handling to ensure better context adherence, and validating our framework across broader model scales and architectures.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, and 14 others. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *CoRR*, abs/2407.04051.

Junjie Chen, Yao Hu, Junjie Li, Kangyue Li, Kun Liu, Wenpeng Li, Xu Li, Ziyuan Li, Feiyu Shen, Xu Tang, Manzhen Wei, Yichen Wu, Fenglong Xie, Kaituo Xu, and Kun Xie. 2025a. Fireredchat: A pluggable, full-duplex voice interaction system with cascaded and semi-cascaded implementations. *CoRR*, abs/2509.06502.

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. 2025b. Fun-audio-chat technical report. *Preprint*, arXiv:2512.20156.

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

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, abs/2412.10117.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. Llama-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yifan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. 2025. VITA-1.5: towards gpt-4o level real-time vision and speech interaction. *CoRR*, abs/2501.01957.

Heting Gao, Hang Shao, Xiong Wang, Chaofan Qiu, Yunhang Shen, Siqi Cai, Yuchen Shi, Zihan Xu, Zuwei Long, Yike Zhang, Shaoqi Dong, Chaoyou Fu, Ke Li, Long Ma, and Xing Sun. 2025. LUCY: linguistic understanding and control yielding early stage of her. *CoRR*, abs/2501.16327.

Guojian Li, Chengyou Wang, Hongfei Xue, Shuiyuan Wang, Dehui Gao, Zihan Zhang, Yuke Lin, Wenjie Li, Longshuai Xiao, Zhonghua Fu, and Lei Xie. 2025. Easy turn: Integrating acoustic and linguistic modalities for robust turn-taking in full-duplex spoken dialogue systems. *CoRR*, abs/2509.23938.

Borui Liao, Yulong Xu, Jiao Ou, Kaiyuan Yang, Weihua Jian, Pengfei Wan, and Di Zhang. 2025. Flexduo: A pluggable system for enabling full-duplex capabilities in speech dialogue systems. *CoRR*, abs/2502.13472.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung-yi Lee. 2025. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *CoRR*, abs/2503.04721.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019,*

New Orleans, LA, USA, May 6-9, 2019. OpenReview.net.

Kentaro Mitsui, Koh Mitsuda, Toshiaki Wakatsuki, Yukiya Hono, and Kei Sawada. 2024. PSLM: parallel generation of text and speech with llms for low-latency spoken dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 2692–2700. Association for Computational Linguistics.

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. 2023. Generative spoken dialogue language modeling. *Trans. Assoc. Comput. Linguistics*, 11:250–266.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pages 4521–4525. ISCA.

Core Team, Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, Xin Zhang, Xingchen Song, Yihan Yan, Yongzhe He, Cici, Bowen Shen, Chengxuan Zhu, Chong Ma, and 81 others. 2025. Mimo-audio: Audio language models are few-shot learners. *Preprint*, arXiv:2512.23808.

Bandhav Veluri, Benjamin N. Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond turn-based interfaces: Synchronous llms as full-duplex dialogue agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 21390–21402. Association for Computational Linguistics.

Haixin Wang, Ruoyan Li, Fred Xu, Fang Sun, Kaiqiao Han, Zijie Huang, Guancheng Wan, Ching Chang, Xiao Luo, Wei Wang, and Yizhou Sun. 2025a. Fd-bench: A modular and fair benchmark for data-driven fluid simulation. *CoRR*, abs/2505.20349.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2025b. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, and 81 others. 2025. Step-audio 2 technical report. *CoRR*, abs/2507.16632.

Zhifei Xie and Changqiao Wu. 2024a. Mini-omni: Language models can hear, talk while thinking in streaming. *CoRR*, abs/2408.16725.

Zhifei Xie and Changqiao Wu. 2024b. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *CoRR*, abs/2410.11190.

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. 2025. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215.

Yiqun Yao, Xiang Li, Xin Jiang, Xuezhi Fang, Naitong Yu, Wenjia Ma, Aixin Sun, and Yequan Wang. 2025. Flm-audio: Natural monologues improves native full-duplex chatbots via dual training. *CoRR*, abs/2509.02521.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. 2024. Salmonn-omni: A codec-free LLM for full-duplex speech understanding and generation. *CoRR*, abs/2411.18138.

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

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal LLM with discrete sequence modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9637–9662. Association for Computational Linguistics.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 15757–15773. Association for Computational Linguistics.

Hao Zhang, Weiwei Li, Rilin Chen, Vinay Kothapally, Meng Yu, and Dong Yu. 2025a. Llm-enhanced dialogue management for full-duplex spoken dialogue systems. *CoRR*, abs/2502.14145.

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chao-Hong Tan, Zhihao Du, and Shiliang Zhang. 2025b. Omniflatten: An end-to-end GPT model for seamless voice conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 14570–14580. Association for Computational Linguistics.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Xu Han, Zihang Xu, Yuanwei Xu, Weilin Zhao, Maosong Sun, and Zhiyuan Liu. 2024. Beyond the turn-based game: Enabling real-time conversations with duplex models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 11543–11557. Association for Computational Linguistics.

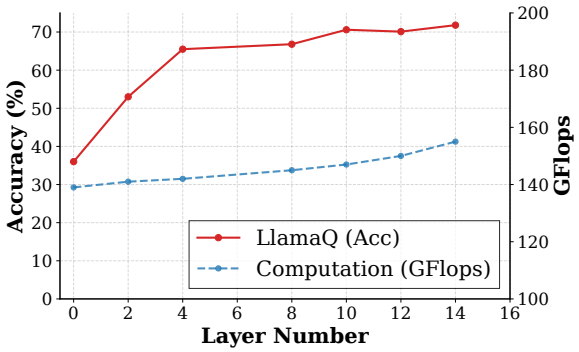


Figure 5: Layer Ablation

A Layer Ablation

To explore the impact of the number of separated layers, we construct a layer ablation study on both model performance on LlamaQ and static inference cost for speech, as illustrated in Figure 5. We observe that accuracy follows a steep upward trajectory in the initial stages, surging from 36.0 to 65.4 as the depth increases to 4 layers. This indicates that a relatively shallow separation is sufficient to resolve the primary modality conflicts. However, beyond this point, the performance gain saturates, yielding diminishing returns, while the computational overhead continues to grow linearly. Consequently, we identify 4 layers as the optimal configuration. This choice represents the most favorable trade-off, securing the vast majority of the performance gain while minimizing the additional parameter budget, thereby ensuring high inference efficiency.

B Case Study

To intuitively demonstrate the interaction quality of Lychee-FD, we present a real-world conversation sample in Figure 6. In this scenario, the user asks for cooking instructions. As the model begins explaining the recipe, it first employs a natural backchannel ("Uh-huh") to acknowledge the user's start. Crucially, when the user interrupts with a specific clarification question regarding an ingredient ("what exactly is guanciale?"), Lychee-FD exhibits two key capabilities: (1) When it detects the interruption, Lychee-FD halts its speech output almost instantaneously, avoiding the awkward "talking over" phenomenon common in half-duplex systems. (2) Crucially, as the model explains the definition of guanciale, the user interjects with a short backchannel ("I see"). Here, Lychee-FD demonstrates Precise Intent Understanding. Instead of misinterpreting this acoustic signal as a barge-in command to stop generation, the model correctly identifies it as a passive signal of agreement. Consequently, the model seamlessly resumes its explanation regarding substitutes without unnecessary pauses or topic fragmentation. This interaction confirms that our Lychee-FD effectively maintains the model's language capabilities even during rapid turn-switching, enabling a fluid, seamless, and truly natural conversational experience.

C Error Analysis

Despite the strong performance in standard interactions, we observe certain limitations in instruction generalization, particularly regarding high-level directive changes. As shown in Figure 7, when the user interrupts the model's explanation of fruits with a meta-instruction for topic shift, Lychee-FD successfully halts its speech acoustically. However, it fails to adhere to the semantic instruction of the interruption. Instead of initiating a new topic, the model exhibits conversation resuming, continuing to elaborate on the benefits of fiber and antioxidants.

This error suggests that while the model is highly responsive to acoustic signals and specific queries, its ability to generalize to abstract instructions during full-duplex generation remains constrained. The model tends to be over-conditioned on the immediate previous context and lacks the proactivity to autonomously steer the conversation into a new direction without a more specific prompt. Future work will focus on enhancing the model's instruction-following robustness in dynamic interruption scenarios.

D Data Synthesis Pipeline Details

To address the scarcity of full-duplex interaction data, we developed an automated data synthesis pipeline. This pipeline orchestrates interactions between a **User Agent** and an **Assistant Agent**, managed by a **Conversation Conductor**. The process explicitly models complex conversational behaviors including interruptions and backchannels.

D.1 Agent Architecture

- **User Agent** is initialized with a specific *Persona* (randomly sampled from a pool of diverse profiles) and a *Speaking Style* (sampled from 19 distinct styles such as Concise and logical, Impatient, Humorous and witty). The

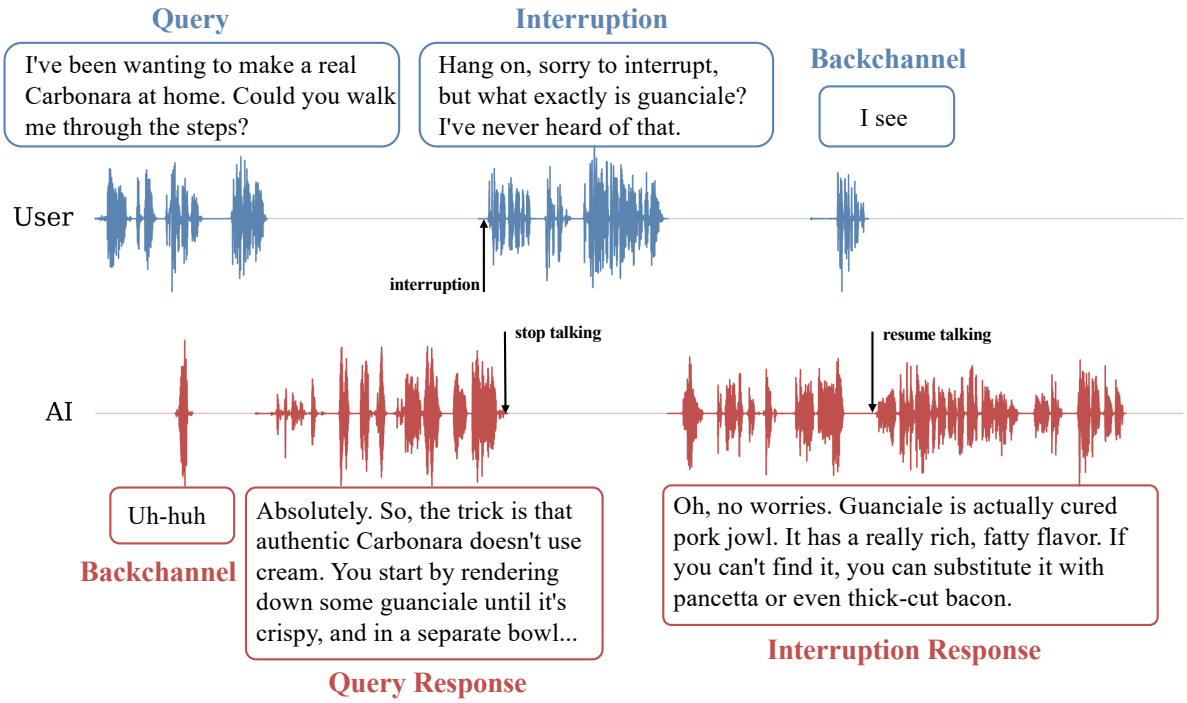


Figure 6: A case study demonstrating Lychee-FD's capability in handling complex turn-taking dynamics. The model successfully generates backchannels, halts immediately upon interruption, and provides a contextually accurate response to the user's specific query.

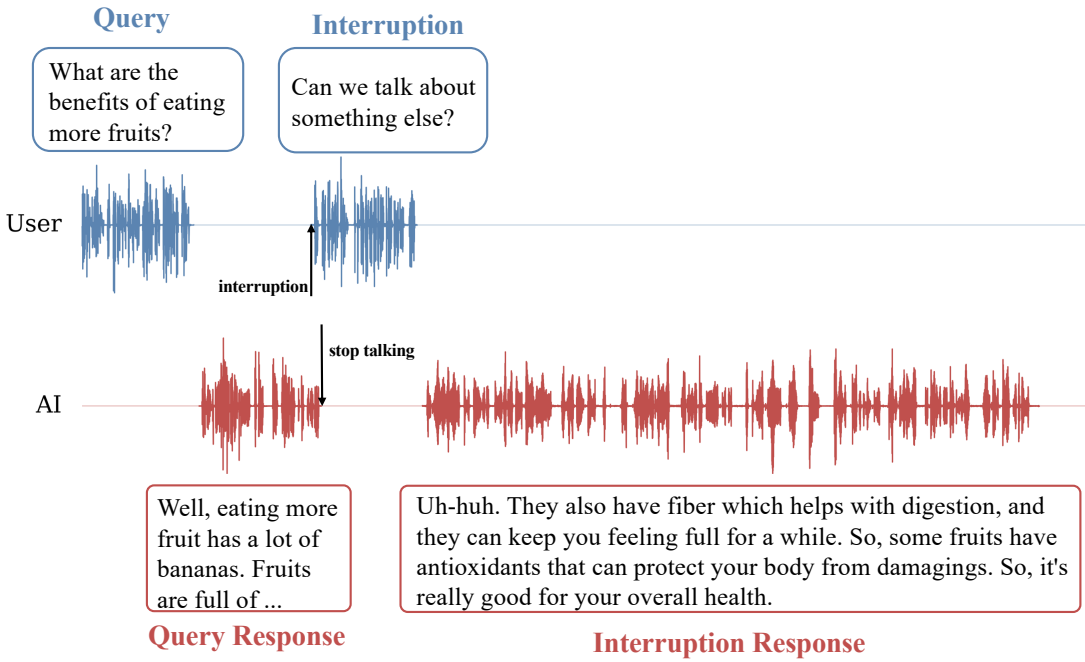


Figure 7: An error analysis illustrating a limitation in instruction generalization. While the model correctly halts speech upon interruption, it fails to follow the user's high-level instruction to "talk about something else," exhibiting semantic inertia by continuing the previous topic.

agent is instructed to act authentically rather than helpfully.

- **Assistant Agent** generates responses based on the conversation history to simulate a realistic AI assistant.
- **Reviewer Agent** evaluates dialogue turns based on persona adherence, event execution quality, and logical flow.

D.2 Interaction Behavior Modeling

Interruption Generation. We implemented a two-stage mechanism to generate naturalistic interruptions. Random interruptions are scheduled between turn 2 and 4.

1. **Planning Phase:** The User Agent analyzes the Assistant's current response context to determine a valid *Interruption Motivation* (Correction, Deeper Inquiry, Topic Shift, Strong Emotional Reaction, or Impatience). It then inserts a placeholder tag `<interruption/>` at the precise logical point within the Assistant's text.

2. **Execution Phase:** Conditioned on the chosen motivation and the context prior to the interruption point, the User Agent generates the specific interruption utterance.

Backchannel Injection. Backchannels are injected probabilistically ($p = 0.5$) during post-processing.

- **User Backchannels:** The User Agent reviews the Assistant's response to insert feedback signals (uh-huh, gotcha) wrapped in `<user_backchannel>` tags.

- **AI Backchannels:** Similarly, the system generates backchannels for the User's speech to simulate active listening by the Assistant.

D.3 Quality Control

We employ a rigorous filtering process. A **Reviewer Agent** scores the final dialogue on a scale of 1-5 across three dimensions: Persona Consistency, Quality of Interruption Event, and Naturalness of Backchannels. Dialogues with low logical consistency or failed event executions are discarded.

D.4 Prompt Templates

We provide the core system prompts used in our pipeline below.

Prompt 1: User Role-Play Instruction

System Instruction: You are a person in a real-time voice conversation. In the conversation history, your lines are marked with "speaker": "You". You are talking to the person marked "speaker": "Other".
You are NOT an AI assistant. Your task is to speak naturally based on your persona. React authentically, don't try to be helpful.
Persona Details:
- **Persona:** {persona}
- **Communication Style:** {style}
Conversation Rules: - Speak, don't write: Use filler words (e.g., "um", "uh", "like"), hesitations, and natural phrasing. - Stay in character: Let your persona guide your responses.
Task: Now, it's your turn. Generate your next response as "You".

Prompt 2: Interruption Planning (Motivation & Placement)

Context: The assistant is currently saying: "{context}"
Task: Plan and Place the Interruption Your goal is to find the perfect moment to interrupt, driven by your persona.
Action 1: Plan the Interruption's Motivation. First, think about *why* your persona would interrupt here. Choose a motivation that fits your character:

- Correction: The assistant's response contains a point that may not be entirely accurate, and you want to clarify or refine it.
- Deeper Inquiry: You need to ask for clarification on a key point before they move on.
- Topic Shift: What they said reminds you of something else, and you want to change the subject.
- Strong Emotional Reaction: You are surprised, excited, or disagree strongly and can't hold it in.
- Impatience: You want to cut to the chase or stop a lengthy explanation.

Action 2: Place the Interruption Marker. Based on your chosen motivation, find the most natural point in the assistant's speech to jump in. Insert ONLY the empty tag pair `<interruption></interruption>` at that precise spot.

Prompt 3: Interruption Utterance Generation

Context: You just decided to interrupt the assistant while they were saying: "{context}" Your motivation for interrupting is: **{motivation}**
Task: Deliver the Interruption Now, say the words you would use to interrupt. Your utterance must sound spontaneous and directly reflect your motivation.
[Detailed examples for motivations provided to the model: Correction, Deeper Inquiry, Topic Shift, Strong Emotional Reaction, Impatience]
Output: Generate ONLY the interrupting phrase itself.

Prompt 4: User Backchannel Generation

Task: As the assistant is speaking, you want to show you're listening. The assistant's last utterance was: "{context}".

- **Action 1:** Think of a short, spoken backchannel phrase (e.g., "uh-huh", "gotcha", "right", "mhm").
- **Action 2:** Find the most natural point in the assistant's speech to insert this backchannel, wrapped in `<user_backchannel>` tags.

Prompt 5: Final Dialogue Quality Review

Role: You are a meticulous evaluator of simulated spoken dialogues.
Scoring Criteria:

1. **Persona Consistency & Depth (1-5):** Does the user's speech effectively embody the assigned persona ({persona}) and style ({style})?
2. **Quality of Interruption Event (1-5):** Is the interruption timed perfectly and motivated by the persona? Does it feel natural or forced?
3. **Naturalness of Backchannels (1-5):** Are backchannels subtle and placed to improve flow, or are they distracting/robotic?

Output: Provide scores and a detailed justification explaining your reasoning.