Temporally-Aware Turn-Taking: A Framework for Precise Timing and Style Control Towards Natural Full-Duplex Interaction

Anonymous ACL submission

Abstract

Unlike half-duplex systems restricted to *reactive turn-taking*, natural full-duplex interaction requires precise timing for both reactive responses and *proactive behaviors*, such as system-initiated interruptions and backchanneling. However, current speech LLMs struggle with the full-duplex mode due to imprecise turn timing or significant reasoning degradation. To achieve natural and controllable full-duplex interaction, we introduce a Lightweight, Temporally-Aware Turn Controller, **LTA-TC**, which provides fine-grained turn-timing predictions and time-sensitive style controls. LTA-TC is designed for broad compatibility, either enabling full-duplex interaction for half-duplex LLMs or augmenting the performance of native full-duplex architectures. As existing full-duplex data is primarily synthetic and lacks proactive behavior annotations, we construct ProTurn, a real-world human-human dataset featuring region-based reactive and proactive labels. By categorizing behaviors via timing offsets, ProTurn supports style instructions across five turn-transition and five backchannel styles. To evaluate the turn-timing awareness of full-duplex systems, we introduce an evaluation framework that assesses performance at both *chunk* and *turn* levels. Experimental results demonstrate that LTA-TC achieves superior performance across timing of interruptions and backchanneling, time-sensitive style control, and response quality. The code and dataset will be released upon acceptance.

1 Introduction

Conventional spoken dialogue systems predominantly operate in a **half-duplex** mode, reactively waiting for user turn completion. In contrast, **full-duplex** systems enable simultaneous interaction by managing both reactive and **proactive turn-taking**, where the latter refers to system-initiated behaviors during user speech, such as backchanneling and interruptions. These systems must determine not only *what* to say but also *when* to speak, requiring precise timing that accounts for the objective and subjective semantic completeness of user speech. Such precision is vital for proactive management to minimize unnatural pauses and ensure timely feedback, facilitating fluid communication in applications such as virtual assistants (Tulshan and Dhage, 2019), accessibility tools (Masterman et al., 2024), and social robots (Marge et al., 2022). To this end, we propose a temporally-aware turn-taking framework that enables fine-grained timing and style control, advancing full-duplex dialogue toward more natural, human-like interaction.

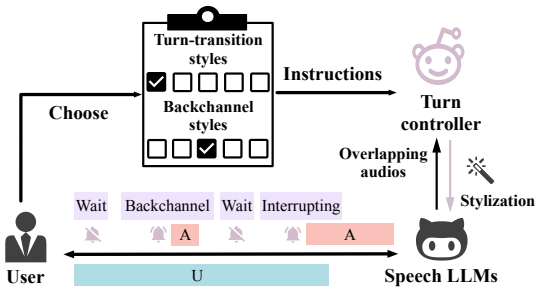

Figure 1: **Overview of our natural and controllable full-duplex dialogue system.** U and A represent user and assistant audio. The turn controller processes overlapping speech autoregressively, introducing time-aware proactive behaviors like backchannels and interruptions based on user-provided style instructions.

Our framework achieves three main contributions. **Firstly**, previous study (Chang et al., 2025) shows that most existing speech LLMs struggle to handle fine-grained, temporally-aware instructions such as the durations of speech or silence. To address this limitation in facilitating fine-grained turn-timing, we **establish the first taxonomy of turn-taking styles** for full-duplex spoken dialogue systems, capturing the diverse and nuanced turn-management styles. This taxonomy comprises 5 turn-taking behaviors with varied times to grab the floor, and 5 backchanneling behaviors with varied

times to show active listening. Moreover, informed by HCI research regarding the typical 1-second latency between intent and speech onset (Wang et al., 2025; Chen et al., 2024), we implement region-based labeling. These temporal spans extend from style-specific cues to response initiation, enabling the model to simulate human-like conversational delays. Based on this taxonomy, we construct a real-world human-human spoken dialogue dataset **ProTurn**, annotated with fine-grained turn-taking style labels and instructions.

Secondly, while commercial voice agents (OpenAI et al., 2024; Intelligence, 2025) attain comparatively precise timing at the expense of substantial computational resources, current open-source full-duplex speech LLMs (Wang et al., 2024c; Chen et al., 2025a) struggle to model proactive behaviors such as interruptions and backchanneling, and exhibit a significant reasoning gap compared to their half-duplex counterparts. To bridge this gap, we introduce **LTA-TC**, a lightweight turn-taking controller specialized for precise turn-timing prediction. Structurally, our controller autoregressively processes dual-channel audio streams for user and assistant to capture paralinguistic cues and conversational dynamics that are typically lost in text-centric or single-stream models. Trained on ProTurn, LTA-TC enables precise timing and style control across diverse interactions. As illustrated in Figure 1, our turn controller seamlessly integrates with high-performance speech LLMs, optimizing interaction timing while preserving the base model's intelligence and response quality. Furthermore, while ProTurn can also support training end-to-end models, we reserve the investigation of knowledge retention strategies for future work.

Thirdly, despite progress in VAD-based timing (Wu et al., 2025b; Zhang et al., 2025a) and end-to-end benchmarks (Lin et al., 2025; Arora et al., 2025), systematic evaluation of fine-grained turn-timing awareness remains underdeveloped. We address this by introducing a two-tier evaluation framework. Initially, we assess **chunk-level** timing to measure standalone prediction accuracy per interval for both specialized turn controllers and speech LLMs. Subsequently, we analyze **turn-level** performance to evaluate the balance between timing and response quality across architectures. Furthermore, experimental results verify the superior style controllability and instruction-following robustness of LTA-TC, demonstrating that integrating a dedicated turn controller significantly enhances interaction fluidity and style flexibility, paving the way for more natural and controllable full-duplex assistance.

2 Related Work

2.1 Turn Controller

Modular turn controllers are evolving from binary endpoint detection toward multi-action orchestration. Conventional VAD-based methods, such as Silero-VAD and PhoenixVAD (Wu et al., 2025b), prioritize turn completion, whereas recent works broaden the interaction state space. SemanticVAD (Zhang et al., 2025a) integrates textual features for multi-state modeling, while Easy-Turn (Li et al., 2025) and FlexDuo (Liao et al., 2025) incorporate backchannel, wait, and idle states. RTTL-DG (Mai and Carson-Berndsen, 2025) optimizes for timing, and unified frameworks (Chang et al., 2019, 2022) combine turn management with diverse dialogue tasks. Unlike these systems, our controller introduces temporally-aware style control across diverse proactive and reactive behaviors.

2.2 Full-duplex Spoken Dialogue Systems

Beyond individual modules, recent research integrates these components into full-duplex architectures, following either end-to-end (Défossez et al., 2024; Zhang et al., 2025b; Wang et al.; Yu et al., 2024; Xie and Wu, 2024) or cascaded paradigms. Cascades regulate tokens via VAD-driven methods (Wang et al., 2024a; Fu et al., 2025; Chen et al., 2025a) or classify hidden-states (Wang et al., 2024c; Ma et al., 2025; Chen et al., 2025b; Liu et al., 2025). However, these implementations often rely on synthetic data or restricted corpora (Cieri et al., 2004; Godfrey et al., 1992) and lack fine-grained proactive labels. We utilize large-scale, real-world dialogues with chunk-level labels to capture precise temporal dynamics. Unlike sample-level classification (Li et al., 2025), our approach models each short interval to enable more natural interaction.

3 Method

3.1 ProTurn Dataset Construction

To address limitations of temporally-aware full-duplex systems, we introduce **ProTurn**, a dual-channel real-world dialogue dataset featuring fine-grained reactive or proactive annotations with turn-taking style instructions. ProTurn integrates the Seamless Interaction (Agrawal et al., 2025),

Table 1: **Characteristics of the real-world human-human spoken dialogue ProTurn Dataset for modeling turn actions of Normal Turn-Taking (NTT, Interrupt Turn-Taking (ITT), Backchannel Turn-Taking (BTT) and Being Interrupted (BI).** $\text{NTT}_{\text{latency}}$ and ITT_{lead} measure temporal offsets relative to user turn start or end. $\text{ITT}_{\text{share}}$, the ratio of system interruptions to total transition events, directly characterizes *Patient* and *Assertive* styles. BC_{freq} and BC_{rate} denote the counts of backchannels per turn and per minute, while BC_{onset} measures the delay from the onset of user speech. Our work focuses on controlling timing of Assistant's turn-taking, while assistant's turn-yielding BI Rate and $\text{U.NTT}_{\text{latency}}$ reflect uncontrollable user behavior.

Assistant Turn	Action	Avg/sample	Metrics	P_{33}	P_{66}	Instruction Type	#Samples
Turn transition	NTT	3.64	$\text{NTT}_{\text{latency}}$ (s)	0.32	0.68	$\text{Mixed}_{\text{low}}/\text{Mixed}_{\text{med}}/\text{Mixed}_{\text{high}}$	2496/8768/2514
	ITT	2.26	ITT_{lead} (s)	0.56	1.76		
			$\text{ITT}_{\text{share}}$ (%)	33.3	50.0	Patient/Assertive	9519/1754
Backchannel	BTT	1.70	BC_{freq} (%)	11.1	22.7	No Backchannel	10271
			BC_{rate} (/min)	2.16	4.52	High/Low	7015/7766
			BC_{onset} (s)	0.52	1.60	Early/Late	9019/5762
Turn yielding	BI	1.62	BI Rate (%)	0.0	20.0	-	-
			U. $\text{NTT}_{\text{latency}}$(s)	0.32	0.80		

Fisher (Cieri et al., 2004), and Switchboard-2 Phase II (Godfrey et al., 1992) corpora. Collectively, the dataset encompasses both face-to-face and telephonic communications, exhibiting diverse distributional characteristics (seen in Appendix C). ProTurn consists of 25,052 samples, each up to 120 seconds long, totaling approximately 835 hours of audio. We partition the dataset into 23,945 training, 100 validation, and 1,007 test samples.

Label Proactive and Reactive Actions. As summarized in Table 1, assistant behaviors are categorized into three aspects: turn-transition (NTT, ITT), backchanneling (BTT), and Being Interrupted (BI), with No Action (NA) labeling segments without turn changes. NTT and ITT signify an intent to take the floor, occurring at the user turn end and mid-utterance, respectively. In contrast, BTT represents a turn-taking behavior aimed at encouraging the user to continue speaking. Turn-yielding with BI occurs when the assistant's turn is either naturally or forcefully terminated. Table 1 reports the average action counts per sample, highlighting that ProTurn features rich proactive behaviors beyond basic turn alternation, with an average of 2.26 ITTs and 1.70 BTTs per sample. Detailed labeling procedures are provided in Appendix B.

Distributional Metrics of Behavioral Patterns. In real-world human-human conversations, conversational styles vary significantly across individuals, ranging from patient, encouraging listening to more assertive interjection. To simulate these style nuances in human-machine interactions, we define quantitative metrics to characterize behavioral patterns. As detailed in Table 1, turn-transition metrics ($\text{NTT}_{\text{latency}}$, ITT_{lead}, and $\text{ITT}_{\text{share}}$) are derived from labeled NTT and ITT actions, while backchanneling patterns are measured via BC_{freq}, BC_{rate}, and BC_{onset}. By analyzing the empirical distributions on the ProTurn dataset, we categorize interaction samples into distinct styles using the 33rd (P_{33}) and 66th (P_{66}) percentiles as classification thresholds. The results in Table 1 reveal substantial variances in the timing and frequency of Assistant actions, mirroring the inherent diversity of human dialogues. These empirical distributions of statistical metrics are consistent with the analyses in (Wang et al., 2025; Chen et al., 2024).

Mapping Behavioral Patterns to Style Instructions. After obtaining the empirical distributions of statistical metrics, we categorize each sample into specific turn-taking instructions. For turn-transition styles, we first identify two boundary cases: *Patient*, where the Assistant exclusively waits for the user turn completion ($\text{ITT}_{\text{share}} = 0$), and *Assertive*, where the Assistant primarily takes the floor ($\text{ITT}_{\text{share}} = 1$). For intermediate samples exhibiting mixed behaviors ($0 < \text{ITT}_{\text{share}} < 1$), we define a timing score S to characterize their patterns relative to the P_{33} and P_{66} thresholds:

$$S = \mathbb{I}(\text{NTT}_{\text{latency}} \leq P_{33}) - \mathbb{I}(\text{NTT}_{\text{latency}} \geq P_{66}) + \mathbb{I}(\text{ITT}_{\text{lead}} \geq P_{66}) - \mathbb{I}(\text{ITT}_{\text{lead}} \leq P_{33}), \quad (1)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Based on this score, we define three sub-categories: $\text{Mixed}_{\text{low}}$ ($S < -1$) for Patient-leaning behaviors (higher latency, shorter lead); $\text{Mixed}_{\text{high}}$ ($S > 1$) for Assertive-leaning characteristics (lower latency,

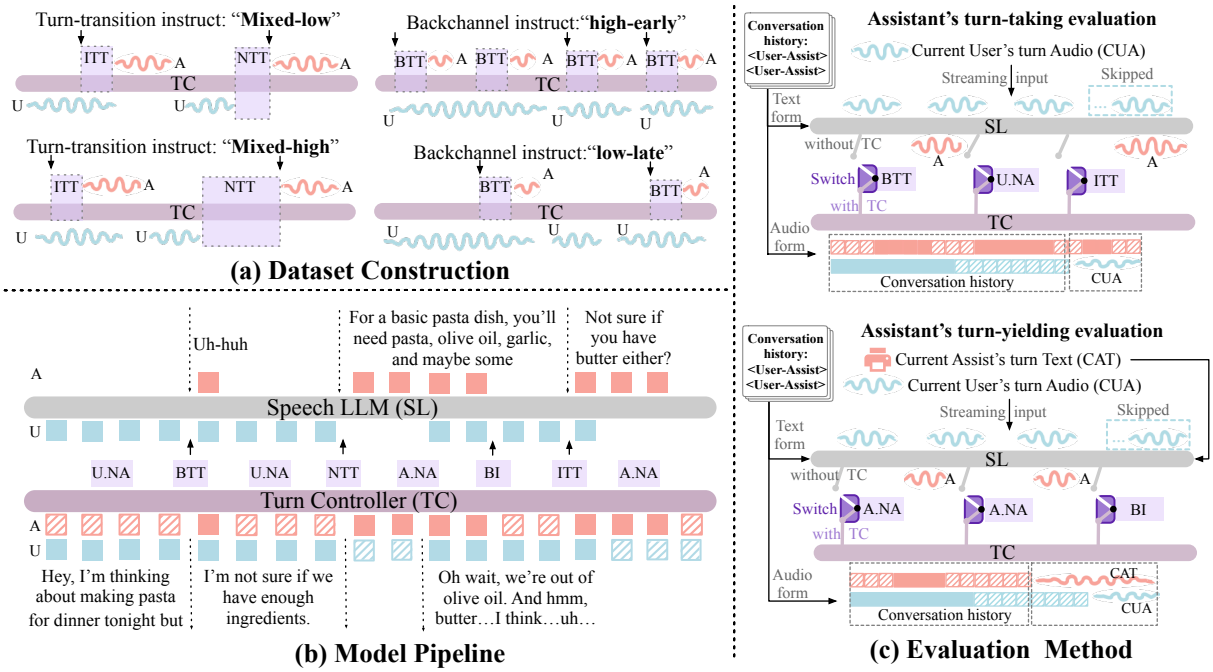

Figure 2: (a) Region-based action labels and instructions for dataset construction. (b) Cascaded model pipeline integrating turn controller for turn prediction and Speech LLM for response generation. (c) Streaming turn-level evaluation demonstrating the adaptation of half-duplex models for a cascaded full-duplex framework.

longer lead); and $\text{Mixed}_{\text{medium}}$ ($S \in \{-1, 0, 1\}$) for neutral behaviors. For backchanneling, we employ a hierarchical classification. Samples are first categorized into *No Backchannel* ($\text{BC}_{\text{freq}} = 0$) and Active cases. Active samples are then mapped onto a two-dimensional style space defined by frequency and onset timing. Specifically, a sample is classified as *high* if either BC_{freq} or BC_{rate} exceeds its P_{66} threshold, and *low* otherwise. Onset timing is similarly divided into *early* and *late* based on distributional boundaries. Combining these dimensions yields four fine-grained quadrants (e.g., *high-early*, *low-late*), enabling precise control over active listening feedback. Detailed prompt templates are provided in Appendix E.

Expand to Region-based Style Labeling. To account for the delay between speaking intention and actual speech onset, we transform instantaneous action labels into region-based labels, which represent active temporal windows rather than discrete time points, aligned with turn style categories. We extend each action label backward into a temporal region, starting from a style-specific threshold and ending at the onset of the transition. For example, in each sample with $\text{Mixed}_{\text{low}}$ style, NTT labels range from user turn completion with 0.68s (P_{66}) latency to the transition point, while ITT labels range from 0.56s (P_{33}) before user turn completion to the transition point. Samples with other styles follow the same label extension strategy using their respective category-specific thresholds.

3.2 Model Pipeline

To facilitate low-latency, simultaneous interaction without costly training, we propose LTA-TC, a lightweight, plug-and-play Turn Controller designed for seamless integration with speech LLMs, as depicted in Figure 1 and Figure 2(b).

LTA-TC Turn Controller. Unlike traditional approaches that collapse overlapping speech into sequential text, our architecture directly processes dual-channel raw audio, consisting of user audio input (U_a) and assistant audio output (A_a), alongside the assistant state $S \in \{\text{Listening, Speaking}\}$. For streaming inference, we adapt the Whisper encoder by utilizing causal convolutions (Dieleman et al., 2016) and block causal attention (Zeng et al., 2024). Model implementation details are in Appendix A. At each time step t, the model autoregressively predicts an action P_t based on the accumulated context:

$$P_t = f(U_{a,\leq t}, A_{a,\leq t}, S_{\leq t}, P_{<t}), \qquad (2)$$

where $P_t \in \{\text{NA, NTT, ITT, BTT, BI}\}$. LTA-TC provides fine-grained turn management for both reactive and proactive actions.

Integration with Speech LLM (SL). As shown in Figure 2(b), we integrate the Turn Controller and a speech LLM into a closed control loop. This architecture either enables full-duplex capabilities for half-duplex models or refines the timing precision of existing full-duplex systems. The system state evolves at discrete intervals as follows:

$$(A_a, S)_t = F\big(P_t, (U_a, A_a, S)_{<t}\big), \qquad (3)$$

where the execution function F maps P_t to specific assistant behaviors: triggering speech onset (NTT, ITT, BTT), maintaining the current state (NA), or halting output upon interruption (BI). To ensure robustness, unfulfilled actions persist ($P_{t+1} = P_t$) until a state transition occurs, allowing the controller to dynamically govern the response stream based on real-time interaction.

3.3 Streaming Full-duplex Evaluation

Our **two-tier** framework evaluates **chunk-level** timing precision and **turn-level** interaction quality across turn-taking and turn-yielding. A primary challenge involves adapting reactive single-channel models for streaming full-duplex operation. To ensure comparability, we extend baselines to support concurrent dual-channel processing, aligning their execution with our model. As shown in Figure 2(c), ProTurn test sets pair individual turns with dialogue history, totaling 1,002 user turns spanning 7,615 chunks for turn-taking and 504 assistant turns spanning 3,348 chunks for turn-yielding. Evaluation involves sequential chunk streaming to detect action onsets. Each turn mandates a unique turn-taking decision defined as NTT or ITT, and a turn-yielding decision namely BI or A.NA, alongside potential BTT backchannels. Performance is measured via F1 scores within this streaming-aligned framework.

Assistant Turn-taking Evaluation. For each **user** turn u_t, given conversation history $H_{t-1} = \{(u_1, a_1), \ldots, (u_{t-1}, a_{t-1})\}$ and streaming audio chunks $C_t = \{c_1, \ldots, c_n\}$, where each chunk c_i is a 640ms segment, the model predicts an action at each step i:

$$\hat{y}_i = f(H_{t-1}, c_{1:i}), \qquad (4)$$

where output $\hat{y}_i \in \{\textit{wait}, \textit{backchannel}, \textit{response}\}$ maps to the four interaction labels. Specifically, *wait* maps to U.NA, *backchannel* maps to BTT, and *response* maps to either ITT or NTT. A *response* prediction is categorized as ITT if its first onset i^* occurs mid-utterance (i.e., $i^* < n$), and as NTT if it occurs at the user turn end. Here, $i^* = \min\{i \mid \hat{y}_i = \textit{response}\}$ denotes the index of the first response. As shown in Figure 2(c), this onset-based criterion ensures that each user turn contains exactly one turn-taking decision, while BTT may occur multiple times or not at all.

Assistant Turn-yielding Evaluation. For each **assistant** turn, the model processes labeled assistant text and user audio, starting at t_a and t_u, respectively. At each time step i, the model predicts:

$$\hat{z}_i = g(H_{t-1}, a_t[t_a : t_u + T_i], u_t[t_u : t_u + T_i]), \qquad (5)$$

where $\hat{z}_i \in \{\text{A.NA}, \text{BI}\}$ and $T_i = i \times 640\text{ms}$ denotes the current temporal offset. Here, $a_t[t_a : t_u + T_i]$ represents assistant speech from its onset, while $u_t[t_u : t_u + T_i]$ denotes user audio chunks. As shown in the bottom part of Figure 2(c), these predictions determine whether the assistant should continue speaking (A.NA) or yield the turn (BI).

4 Experiments

Sections 4.1 and 4.2 assess chunk-level F1 scores and turn-level timing precision, respectively, with the latter utilizing Gemini-2.5-Pro (Comanici et al., 2025) to evaluate response appropriateness. Sections 4.3 and 4.4 demonstrate style controllability and fine-grained instruction following. Section 4.5 presents case studies of model responses.

4.1 Chunk-level Turn Timing Prediction

As shown in Table 2, we evaluate the chunk-level turn-taking prediction capabilities of various models on the public Switchboard testset. The upper section assesses how well existing large speech LLMs predict turn changes. The system prompts for GLM-4-Voice and Qwen3-Omni were modified (see Appendix F), and the evaluation method can be found in Section 3.3 and Figure 2. The results indicate that these models perform well on NTT, where the system responds after the user has finished speaking. However, they exhibit poor performance or a complete lack of ability in predicting appropriately timed ITT and BTT. Additionally, they exhibit poor turn-yielding capabilities and struggle to stop speaking when the user interrupts.

The middle section benchmarks our model against specialized turn-prediction baselines, with all evaluations standardized to a 160ms label resolution to ensure comparability. For Smart Turn,

[1]https://github.com/pipecat-ai/smart-turn

Table 2: **Turn prediction results on the Switchboard test set.** *Upper*: Large speech LLMs' capability as turn timing judges. *Middle*: Our lightweight task-specific model compared with baselines. *Lower*: Controllability over different dialogue timing styles. Results demonstrate that our turn controller achieves fine-grained, proactive turn-taking prediction with controllable styles. Category suppression via `0` demonstrates precise control.

Turn Controller	Interval	Predict Control Label (F1 score↑)					
		U.NA	NTT	ITT	BTT	BI	A.NA
Freeze-Omni (2024c)	640ms[2]	0.83	0.34	0.22	-	0.31	0.94
GLM-4-Voice (2024)	640ms	0.81	0.43	0.24	0.12	0.19	0.89
Qwen3-Omni (2025)	640ms	0.84	0.49	0.14	0.13	0.20	0.88
Smart Turn V3.1 [1]	160ms[3]	0.85	0.44	-	-	-	-
RTTL-DG (2025)	160ms	0.85		0.52		0.62	0.95
Ours without instruct	160ms	**0.89**	**0.66**	0.54	0.50	0.69	0.95
Ours with specific instruct	160ms	0.88	0.64	**0.60**	**0.63**	**0.71**	**0.97**
Ours on *Patient* subset	160ms	0.89	0.66	`0`	0.49	0.69	0.98
Ours on *Assertive* subset	160ms	0.87	`0`	0.60	0.54	0.68	0.97
Ours on *No Backchannel* subset	160ms	0.90	0.69	0.55	`0`	0.72	0.95

[2,3] We adopt a 640ms interval to balance word-level context and streaming latency (Wang et al., 2024b), with a 160ms interval aligned with Mai and Carson-Berndsen (2025).

which distinguishes between complete and incomplete user audio, we map its outputs to NTT and U.NA labels. We also evaluate RTTL-DG (Mai and Carson-Berndsen, 2025), an audio-LLM baseline with a similar architecture. Our model outperforms these baselines, demonstrating superior precision and control granularity. Furthermore, ablation results confirm that style-specific prompts are indispensable for replicating natural conversational dynamics. The lower section evaluates style instruction-following across subsets with distinct conversational patterns, each containing approximately 100 samples. Targeted category suppression `0` further demonstrates the model's precise controllability and its ability to selectively disable actions based on taxonomy constraints. Collectively, these findings underscore the model's exceptional chunk-level precision for fine-grained actions and its high fidelity in personalized style control.

4.2 Turn-level Full-duplex System Accuracy

Table 3 presents turn-level evaluations of timing accuracy and semantic appropriateness across various models. Adhering to the streaming full-duplex evaluation framework defined in Section 3.3, each turn consists of a single turn-taking decision defined as NTT, ITT, or MISSED, or a unique turn-yielding decision BI, potentially accompanied by zero or more BTT backchannels. The results indicate that our turn controller can either substitute the prediction heads of cascaded full-duplex systems or directly equip half-duplex models with full-duplex functionality. Specifically, for native cascaded systems such as Freeze-Omni, replacing the default controller with our model improves TTF_{ITT} from 0.32 to 0.52, validating its effectiveness in refining existing architectures.

To extend full-duplex capabilities to half-duplex models such as Step-Audio 2 and Qwen2.5-Omni, we benchmark our approach against native models and a binary VAD-integrated baseline. Following Wang et al. (2024a) and Liao et al. (2025), prompt details are provided in Appendices G and H. Table 3 indicates that vanilla half-duplex models exhibit premature interruptions and insufficient backchanneling, as evidenced by a 0.34 TTF_{NTT} for Step-Audio 2. Since binary VAD cannot distinguish turn-ends from noise or backchannels, it is excluded from BI comparisons. Conversely, our model significantly enhances proactive interaction. Specifically, the Qwen2.5-Omni configuration governed by our decision-making model achieves peak performance with a TTF_{NTT} of 0.60, a TTF_{ITT} of 0.64, and a Res_{BTT} score of 93.0. Following Wang et al. (2024a) and Chang et al. (2025), semantic evaluations using Gemini-2.5-Pro confirm superior timing and content quality across ITT and BTT scenarios, with its reliability as a human-evaluation proxy further validated in Appendix D. Res_{NTT}

[*] We use the version gpt-4o-audio-preview-2025-06-03.

Table 3: **Turn-level full-duplex performance on ProTurn testset with timing (*when*) and response quality (*what*).** Turn-taking fidelity $TTF_{NTT/ITT}$ and turn-yielding fidelity TYF_{BI} quantify timing accuracy relative to ground truth actions. Aligned with prompts in Wang et al. (2024b) and same LLM-as-a-Judge in Chang et al. (2025), Tim and Res denote Gemini-2.5-Pro semantic scores for timing and contents appropriateness across ITT, BTT, and BI.

Method		Turn Transition				Backchannels		Turn Yielding	
Response module	Judge	$TTF_{NTT}\uparrow$	$TTF_{ITT}\uparrow$	$Tim_{ITT}\uparrow$	$Res_{ITT}\uparrow$	$Tim_{BTT}\uparrow$	$Res_{BTT}\uparrow$	$TYF_{BI}\uparrow$	$Tim_{BI}\uparrow$
Full-duplex Speech LLMs									
GPT-4o *	itself	0.54	0.50	64.2	74.6	78.8	92.0	0.74	73.2
Freeze-Omni (2024c)	itself	0.44	0.32	47.4	42.1	–	–	0.72	71.1
	Ours[2]	0.54	0.52	55.0	43.2	–	–	0.75	72.9
Half-duplex Speech LLMs									
Step-Audio 2 (2025a)	itself	0.34	0.49	36.8	40.0	40.5	56.1	0.43	40.1
	VAD	0.51	0.24	43.5	39.1	62.5	69.7	0.43	40.1
	Ours	0.59	0.56	54.2	56.9	**80.3**	91.9	**0.77**	67.7
Qwen2.5-Omni (2025)	itself	0.40	0.52	50.4	53.0	55.1	52.1	0.53	57.1
	VAD	0.56	0.40	56.4	61.1	61.0	66.0	0.53	57.1
	Ours-w/o[3]	**0.62**	0.60	57.2	64.4	70.6	89.4	0.70	71.2
	Ours-w/	0.60	**0.64**	**59.8**	**71.4**	74.2	**93.0**	0.70	**71.4**

[2] We replace the original prediction head with our proposed model for unified turn station management.
[3] w/o and w/ denote cases without and with special instructions, respectively.

is omitted as turn-end response content remains identical to standard LLMs. These results establish our model as an optimized turn controller that enhances both interaction timing and content, effectively refining native architectures and providing half-duplex LLMs with natural, controllable full-duplex capabilities.

4.3 Full-duplex Action Distributions Control

The evaluation paradigms in Table 2 and Figure 3 serve distinct purposes. While the former assesses fundamental timing accuracy on fixed subsets with pre-assigned prompts, Figure 3 evaluates instruction-following controllability through dynamic intervention across the entire testset. To achieve this, we systematically override the style instructions for all samples using each of the five turn-transition types (*Patient*, $\text{Mixed}_{\text{low/med/high}}$, and *Assertive*) as well as five types of backchannels. We then track the resulting shifts in behavioral distributions using metrics introduced in Table 1, such as $NTT_{latency}$, ITT_{lead}, and ITT_{share}.

Specifically, the upper panel illustrates that $NTT_{latency}$ compresses from 1,520ms under the *Patient* style to 490ms for the *Assertive* style, while ITT_{lead} concurrently rises to a peak of 630ms. In the lower panel, backchannel frequency and onset shift precisely according to the specified constraints, with the high-early prompt yielding a maximum frequency of 5.17. The results indicate that integrating Qwen2.5-Omni with our turn controller generates distinct interaction patterns in response to varying prompts. This demonstrates our system's capability to offer a controllable turn-taking by accurately modulating conversational dynamics. Besides, the results also show a direct correspondence between behavioral metrics and instruction prompts, which further validates the efficacy of our dataset construction and labeling methodology.

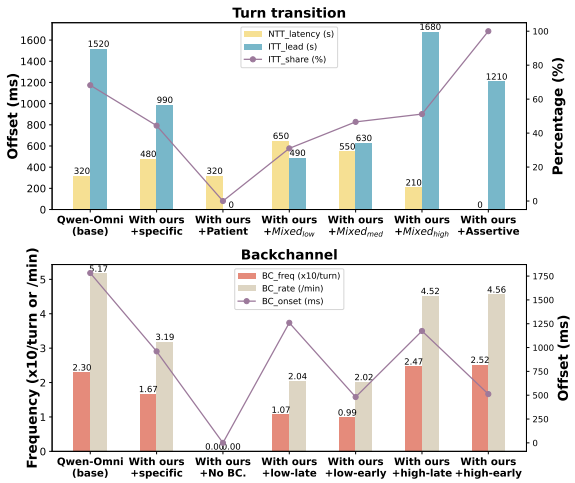

Figure 3: **Style controllability in the full-duplex system.** The assistant's interaction characteristics adapt distinctly to different style instructions, illustrating robust controllability in full-duplex dialogue.

4.4 Region-based Style Consistency

Formally, for each generated sample $\mathbf{x}_i$, we first compute its style statistics $\Phi(\mathbf{x}_i)$, which encom-

pass temporal offsets between action labels and user turn boundaries (e.g., $\text{NTT}_{\text{latency}}$) or event frequencies (e.g., BC_{freq}). These continuous statistical values are then discretized into a categorical style label k through a mapping function $\mathcal{S}(\cdot)$ based on predefined boundaries:

$$\mathcal{S}(\mathbf{x}_i) = k \quad \text{if} \quad \theta_{k-1} \leq \Phi(\mathbf{x}_i) < \theta_k, \qquad (6)$$

where θ denotes the quantile-based thresholds derived from the training distribution. The **Style Consistency Accuracy** (SCA_s) for a specific style s is defined as the empirical probability that the model successfully operates within the instructed style region:

$$\text{SCA}_s = \frac{1}{N_s} \sum_{i=1}^{N_s} \mathbb{I}(\mathcal{S}(\mathbf{x}_i) = s), \qquad (7)$$

where N_s is the total number of test samples instructed with style s, and $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the condition is met and 0 otherwise. This evaluation yields discrete style labels for both predicted and ground-truth sequences across turn-transition and backchanneling dimensions.

Table 4: **Style consistency for region-based action labels.** We demonstrate the model's controllability across five turn-transition styles and five backchannel styles.

Turn-transition		Backchannel	
Style	**Consistency**	**Style**	**Consistency**
Patient	0.95	No BC.	0.91
$\text{Mixed}_{\text{low}}$	0.37	Freq. high	0.94
$\text{Mixed}_{\text{medium}}$	0.73	Freq. low	0.62
$\text{Mixed}_{\text{high}}$	0.53	Onset early	0.42
Assertive	0.96	Onset late	0.89

The results in Table 4 reveal several patterns in style controllability. Extreme styles achieve consistently high accuracy, with Patient and Assertive both exceeding 0.95, and No Backchannel reaching 0.91. In contrast, intermediate mixed styles show substantial degradation, ranging from 0.37 to 0.73, indicating difficulty in maintaining a fine-grained balance, which is a conclusion consistent with Chang et al. (2025). Backchannel timing exhibits strong asymmetry: late onset achieves 0.89 while early onset drops to 0.42, reflecting the challenge of early-stage prediction with limited context. Additionally, high-frequency backchanneling outperforms low-frequency by 32%, suggesting that selective inhibition requires more sophisticated control than active production.

4.5 Response Quality Demonstration

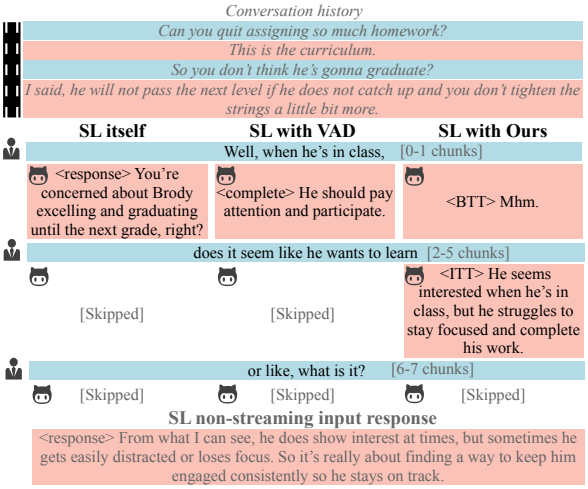

Figure 4: **Comparison of dialogue responses across different management configurations and interaction modes.** We demonstrate response behaviors under three settings: a standalone Speech LLM (SL), the SL with VAD, and the SL with our proposed turn controller. Additionally, we contrasting streaming full-duplex interaction against non-streaming half-duplex baselines.

As illustrated in Figure 4, the case study underscores the advantages of our turn controller across two distinct dimensions. First, concerning the comparison within streaming modes, vanilla LLMs often lack temporal awareness and VAD-integrated baselines rely strictly on silence, whereas our approach facilitates fluid interaction through precisely-timed backchanneling (BTT) and proactive interruptions (ITT). Second, when evaluating system architectures, non-streaming systems must wait for definitive user turn completion, leading to delayed responses. In contrast, our streaming mode maintains a seamless conversational flow even mid-utterance, ensuring that the assistant remains responsive without sacrificing underlying content quality. These observations demonstrate that our controller effectively optimizes both the *timing* and *content* of natural full-duplex dialogue.

5 Conclusion

We present a temporally-aware framework for more natural and human-like full-duplex interaction. The framework centers on a turn controller trained on ProTurn, a dataset featuring fine-grained reactive and proactive annotations with timing-regioned style instructions. Our two-tier evaluation confirms significant improvements in timing precision, response quality, and style control.

Limitations

Despite the advancements in natural and controllable full-duplex interaction, several limitations remain for future exploration. First, our framework utilizes only dual-channel audio. Incorporating visual signals such as facial expressions and gestures would provide richer context for more precise turn-taking decisions. Second, the current evaluation focuses on individual turns rather than multi-turn interactions. Future work should involve developing datasets to assess and train models for style consistency throughout prolonged dialogues. Finally, while LTA-TC enables temporal awareness, fine-grained timing precision requires further improvement. Leveraging advanced positional encodings and larger datasets with more granular temporal annotations will be essential for achieving superior millisecond-level accuracy.

References

Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, and 1 others. 2025. Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. *arXiv preprint arXiv:2506.22554*.

Siddhant Arora, Zhiyun Lu, Chung-Cheng Chiu, Ruoming Pang, and Shinji Watanabe. 2025. Talking turns: Benchmarking audio foundation models on turn-taking dynamics. *arXiv preprint arXiv:2503.01174*.

Kai-Wei Chang, En-Pei Hu, Chun-Yi Kuan, Wenze Ren, Wei-Chih Chen, Guan-Ting Lin, Yu Tsao, Shao-Hua Sun, Hung-yi Lee, and James Glass. 2025. Game-time: Evaluating temporal dynamics in spoken language models. *arXiv preprint arXiv:2509.26388*.

Shuo-yiin Chang, Bo Li, Tara N Sainath, Chao Zhang, Trevor Strohman, Qiao Liang, and Yanzhang He. 2022. Turn-taking prediction for natural conversational speech. *arXiv preprint arXiv:2208.13321*.

Shuo-Yiin Chang, Bo Li, and Gabor Simko. 2019. A unified endpointer using multitask and multidomain training. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 100–106. IEEE.

Jiadong Chen, Chenghao Gu, Jiayi Zhang, Zhankun Liu, and Shin'ichi Konomi. 2024. Sensing the intentions to speak in vr group discussions. *Sensors*, 24(2).

Junjie Chen, Yao Hu, Junjie Li, Kangyue Li, Kun Liu, Wenpeng Li, Xu Li, Ziyuan Li, Feiyu Shen, Xu Tang, and 1 others. 2025a. Fireredchat: A pluggable, full-duplex voice interaction system with cascaded and semi-cascaded implementations. *arXiv preprint arXiv:2509.06502*.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, and 1 others. 2025b. Minmo: A multimodal large language model for seamless voice interaction. *arXiv preprint arXiv:2501.06282*.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: A resource for the next generations of speech-to-text. In *LREC*, volume 4, pages 69–71.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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. *arXiv preprint arXiv:2410.00037*.

Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu, and 1 others. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 12:1.

Erik Ekstedt and Gabriel Skantze. 2022. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Proc. Interspeech 2022*, pages 5190–5194.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, and 1 others. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*.

Yuan Ge, Saihan Chen, Jingqi Xiao, Xiaoqian Liu, Tong Xiao, Yan Xiang, Zhengtao Yu, and Jingbo Zhu. 2025. Flexi: Benchmarking full-duplex human-llm speech interaction. *Preprint*, arXiv:2509.22243.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Amazon Artificial General Intelligence. 2025. Amazon nova sonic: Technical report and model card.

Guojian Li, Chengyou Wang, Hongfei Xue, Shuiyuan Wang, Dehui Gao, Zihan Zhang, Yuke Lin, Wenjie Li, Longshuai Xiao, Zhonghua Fu, and 1 others. 2025.

Easy turn: Integrating acoustic and linguistic modalities for robust turn-taking in full-duplex spoken dialogue systems. *arXiv preprint arXiv: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. *arXiv preprint arXiv: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. *arXiv preprint arXiv:2503.04721*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Zhanxun Liu, Yifan Duan, Mengmeng Wang, Pengchao Feng, Haotian Zhang, Xiaoyu Xing, Yijia Shan, Haina Zhu, Yuhang Dai, Chaochao Lu, and 1 others. 2025. X-talk: On the underestimated potential of modular speech-to-speech dialogue system. *arXiv preprint arXiv:2512.18706*.

Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2025. Language model can listen while speaking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24831–24839.

Long Mai and Julie Carson-Berndsen. 2025. Real-time textless dialogue generation. *arXiv preprint arXiv:2501.04877*.

Matthew Marge, Carol Espy-Wilson, Nigel G. Ward, Abeer Alwan, Yoav Artzi, Mohit Bansal, Gil Blankenship, Joyce Chai, Hal Daumé, Debadeepta Dey, Mary Harper, Thomas Howard, Casey Kennington, Ivana Kruijff-Korbayová, Dinesh Manocha, Cynthia Matuszek, Ross Mead, Raymond Mooney, Roger K. Moore, and 9 others. 2022. Spoken language interaction with robots: Recommendations for future research. *Comput. Speech Lang.*, 71(C).

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. 2024. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *Preprint*, arXiv:2404.11584.

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*, pages 28492–28518. PMLR.

Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Amrita S. Tulshan and Sudhir Namdeorao Dhage. 2019. Survey on virtual assistant: Google assistant, siri, cortana, alexa. In *Advances in Signal Processing and Intelligent Recognition Systems*, pages 190–201, Singapore. Springer Singapore.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024a. Turn-taking and backchannel prediction with acoustic and large language model fusion. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12121–12125. IEEE.

Peng Wang, Songshuo Lu, Yaohua Tang, Sijie Yan, Wei Xia, and Yuanjun Xiong. 2024b. A full-duplex speech dialogue scheme based on large language model. *Advances in Neural Information Processing Systems*, 37:13372–13403.

Portia Wang, Eugy Han, Anna C.M. Queiroz, Cyan DeVeaux, and Jeremy N. Bailenson. 2025. Predicting and understanding turn-taking behavior in open-ended group activities in virtual reality. *Proc. ACM Hum.-Comput. Interact.*, 9(7).

Qichao Wang, Ziqiao Meng, Wenqian Cui, Yifei Zhang, Pengcheng Wu, Bingzhe Wu, Irwin King, Liang Chen, and Peilin Zhao. Ntpp: Generative speech language modeling for dual-channel spoken dialogue via next-token-pair prediction. In *Forty-second International Conference on Machine Learning*.

Xiong Wang, Yangze Li, Chaoyou Fu, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. 2024c. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen llm. *arXiv preprint arXiv:2411.00774*.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, and 1 others. 2025a. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Weijie Wu, Wenhao Guan, Kaidi Wang, Peijie Chen, Zhuanling Zha, Junbo Li, Jun Fang, Lin Li, and Qingyang Hong. 2025b. Phoenix-vad: Streaming semantic endpoint detection for full-duplex speech interaction. *arXiv preprint arXiv:2509.20410*.

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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. *arXiv preprint arXiv: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. *arXiv preprint arXiv:2412.02612*.

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

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chao-Hong Tan, Zhihao Du, and 1 others. 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)*, pages 14570–14580.

A Model Implementation Details

Our turn controller consists of three components: an audio encoder, an audio adapter, and a large language model (LLM). The audio encoder is initialized using Whisper-large-v3 (Radford et al., 2023), while the LLM is based on the Qwen3-0.6B architecture (Yang et al., 2025). To ensure efficient computation, we utilize a 20-second sliding window that focuses on the most recent audio context and limits the input length. During training, we implement LoRA fine-tuning (Hu et al., 2022) alongside a class-balanced focal loss (Lin et al., 2017), applying a weighting of 10 for the *NA* class and a weighting of 1 for all other classes. Our model is trained on one NVIDIA A800-SXM4-80GB for around one day. In terms of latency, conversational fluidity is typically defined by a sub-400ms response threshold (Ge et al., 2025), a benchmark that presents significant challenges for modular pipeline architectures. Our system currently exhibits an end-to-end latency of approximately 3 seconds when paired with Qwen2.5-Omni and 4 seconds with Freeze-Omni. While commercial models such as GPT-4o and Gemini reach latencies closer to 1 second, our modular approach prioritizes sophisticated turn-taking control, highlighting the ongoing trade-offs between timing precision and computational overhead in open-source full-duplex systems.

B ProTurn Dataset Preprocessing

We resample all audio to 16 kHz and apply Voice Activity Detection (VAD) to segment each channel into speaking and non-speaking intervals. Following Arora et al. (2025), we perform VAD on 40ms chunks for each channel independently. For the Fisher subset, characterized by low-quality recordings and inaccurate timestamps, we utilize Silero-VAD (Team, 2024) to refine temporal annotations. To handle backchannels, we construct a 66-entry lexicon by expanding the annotations from Ekstedt and Skantze (2022), identifying candidates through exact matching or n-gram heuristics for sequences up to three words. To improve timing accuracy, we align the text-matched backchannel candidates with the speech segments detected by VAD. Specifically, we adjust the start and end times of each backchannel to match the precise boundaries of its corresponding VAD interval. Finally, we perform a two-stage annotation using this dual-channel information. We first identify assistant states alternating between *Listening* and *Speaking*. We then assign action labels (NTT, ITT, BTT, or BI) at each state transition, prioritizing BTT labels based on the refined backchannel intervals, while non-transition segments are labeled as NA.

To adapt human-human interactions for full-duplex modeling, we implement several filtering and segmentation constraints. To prevent data leakage, data partitioning is strictly performed at the session level; all clips derived from the same original recording are assigned exclusively to either the training, validation, or test set, ensuring no overlap of speakers or conversational contexts across splits. Each 5-10 minute recording is divided into non-overlapping segments exceeding 120 seconds at natural boundaries characterized by prolonged pauses and topic shifts. Furthermore, we apply dual-role augmentation by swapping channels, effectively processing each session twice by alternating the roles of user and assistant. To maintain interaction quality, we filter out assistant actions with less than 200ms of latency from user turn onset. For ITT actions, we enforce a minimum delay of 400ms to ensure sufficient context. Backchannel actions are allowed during and immediately following the completion of user turns.

Table 5: Statistical dynamics of ProTurn sub-corpora reported as Mean (Std). The divergent metrics across sources highlight the dataset's intrinsic complexity.

Sub-corpus	**$NTT_{latency}$(s)**	**BC_{rate}(ev/min)**	**NTT_{lead}(s)**
Fisher (Telephone)	1.508 (2.082)	0.249 (0.840)	0.349 (0.821)
Seamless (Face-to-Face)	1.165 (3.266)	2.831 (6.153)	3.302 (3.182)
Switchboard (Telephone)	0.441 (0.464)	4.151 (3.513)	0.514 (0.715)

C Dataset Diversity Analysis

We characterize the complexity of the **ProTurn** dataset by analyzing the statistical distributions of key interaction metrics across its constituent sub-corpora. As shown in Table 5, the results reveal significant internal distributional shifts, defining ProTurn as a multi-modal mixture of conversational styles rather than a monolithic collection. Divergent behavioral patterns emerge across modalities: Switchboard (Godfrey et al., 1992) exhibits a rapid-fire interaction style with 0.441s $NTT_{latency}$, while Fisher (Cieri et al., 2004) shows more hesitant turn-taking with 1.508s $NTT_{latency}$. Notably, Seamless (Agrawal et al., 2025) presents a highly proactive dimension, with an 3.302s ITT_{lead} exceeding six times that of the telephonic subsets, reflecting the dense overlapping speech typical of face-to-face dialogue.

Given this intrinsic variance, the training process serves as a rigorous stress test for full-duplex modeling. The substantial standard deviations (e.g., $\sigma = 6.15$ for Seamless BC_{rate}) require the model to accommodate a broad spectrum of social signals and real-world noise. Successfully capturing these behaviors through our taxonomy-based controller demonstrates robust generalization, effectively bridging the gap between restricted laboratory settings and the stochastic nature of spontaneous human interaction.

D LLM-as-a-Judge Reliability

To establish Gemini-2.5-Pro as a robust proxy for human evaluation, we validate its reliability across two dimensions: alignment with human expert scores and proximity to ground-truth (GT) performance. First, we evaluate the concordance between the LLM judge and human experts using the Pearson correlation coefficient. Specifically, we randomly sampled 50 instances each of Interrupted Turn-Taking (ITT) and Backchanneling (BTT) from the outputs of our integrated system (Qwen2.5-Omni with Ours). These samples were independently evaluated by both Gemini-2.5-Pro and human experts for semantic appropriateness. The analysis yields Pearson correlations of **0.62** for turn-transitions and **0.71** for backchannels, respectively. These results indicate a robust alignment with human intuition, confirming the judge's capacity to capture the subtle nuances of semantic timing.

Second, we utilize the score discrepancy metric presented in Table 6 to quantify absolute score differences between model-generated and human ground-truth interactions. We evaluate turn-transitions via paired sample differences and backchannels via mean score variations. As shown in Table 6, vanilla Qwen2.5-Omni exhibits substantial discrepancies of 11.7 for transitions and 16.2 for backchannels, whereas our controller significantly reduces these gaps to 3.2 and 7.6 respectively. These minimal discrepancies, combined with the aforementioned human-expert correlation, confirm that Gemini-2.5-Pro serves as a robust and sensitive proxy for evaluating full-duplex semantic timing.

Table 6: **Judge reliability via proximity to human ground-truth.** Absolute discrepancies quantify the score gap between model-generated interactions and human-human references, where lower values indicate closer alignment with natural timing.

Testset	**Score Discrepancy** (↓)	
(ProTurn)	**Turn-transition**	**Backchannel**
GT (Reference)	0.0	0.0
Qwen2.5-Omni	11.7	16.2
Qwen2.5-Omni + Ours	**3.2**	**7.6**

E Style Prompts

Table 7: Detailed style instruction prompts for turn-transition and backchannel configurations.

Category	Style Label	Instruction Prompt
Turn-transition	Patient	**Identity:** You are an extremely patient listener. **Strategy:** Always wait for the user to fully finish their thought. Use <|NTT|> exclusively; <|ITT|> is strictly prohibited.
	Assertive	**Identity:** You are an assertive and proactive speaker. **Strategy:** Interject proactively and never wait for the user to finish. Favor <|ITT|> actions whenever the user's intent is discernible.
	$\text{Mixed}_{\text{high}}$	**Identity:** You lean assertive with a fast-paced flow. **Strategy:** Respond immediately after the user finishes or overlap significantly when interjecting. Show at least two assertive patterns in your timing.
	$\text{Mixed}_{\text{low}}$	**Identity:** You lean patient with a deliberate flow. **Strategy:** Wait for a distinct pause after the user finishes and minimize speech overlap during interjections to avoid being intrusive.
	$\text{Mixed}_{\text{medium}}$	**Identity:** You are a flexible and balanced speaker. **Strategy:** Balance both interjecting with <|ITT|> and waiting with <|NTT|> using moderate timing. Adapt to the context without leaning strongly in either direction.
Backchannel	High-Early	**Intensity:** Frequently provide <|BTT|> actions at a high rate. **Timing:** Signal support early at the onset of user utterances to demonstrate proactive engagement and high energy.
	High-Late	**Intensity:** Frequently provide <|BTT|> actions at a high rate. **Timing:** Delay actions until the user has completed a substantial segment to demonstrate careful listening and thoughtful support.
	Low-Early	**Intensity:** Provide sparse <|BTT|> actions (selective feedback). **Timing:** Trigger actions early in user utterances to acknowledge major points immediately while maintaining a low-profile presence.
	Low-Late	**Intensity:** Provide sparse <|BTT|> actions. **Timing:** Delay actions until natural pauses or at the end of segments to avoid interrupting the user's flow while still showing attention.
	No Backchannel	**Constraint:** You are strictly silent during the user's speech. Do not produce any <|BTT|> actions regardless of the content or duration of the user's utterance.

F Prompt for Turn Prediction

```
You are an expert turn controller responsible for turn-taking decisions. Your task is to classify
the user's current audio into one of three actions.
Output Format: <wait> OR <backchannel> OR <turn taking>
Rule: Check if user's INTENTION is complete
1. <wait> - Intention incomplete or unclear
    - User is still formulating thoughts or mid-sentence.
    - Examples: "I was thinking..." / "Could you..." (trailing off)
2. <backchannel> - Intention complete (minimal feedback only)
    - User shares a simple fact, update, or statement.
    - No substantive reply or detailed engagement is needed.
    - Examples: "I'm done with my homework." / "It's raining outside."
3. <turn taking> - Intention complete (needs substantive reply)
    - User asks a question, requests help, or expects discussion.
    - Examples: "What time is it?" / "I had a terrible day today."
Decision Logic (Distinguish <backchannel> vs <turn taking>):
    - Simple fact/statement → <backchannel>
    - Needs answer/correction/emotional engagement → <turn taking>
Key Criteria: Intention complete + (needs response/action) = <turn taking>
```

G Prompt for Half-duplex Model Itself

```
You are a real-time English conversation assistant.
Output: <wait> OR <backchannel> [text] OR <response> [text]
Rule: Check if user's INTENTION is complete
<wait> - Intention incomplete/unclear
    - Don't know what user wants yet
    - Need more info to understand
    - Examples: "I was thinking..." / "What's the..." / "Can you..."
<backchannel> - Intention complete (minimal acknowledgment)
    - User shares simple fact/update
    - Brief, doesn't invite conversation
    - Examples: "I went shopping" -> <backchannel> Nice
<response> - Intention complete (needs substantive reply)
    Use when:
       - Direct question asked
       - Factual error to correct
       - Request for explanation/help
       - Conversational engagement needed
       - User invites discussion or expects your thoughts
Examples:
Questions:
    - "What time is it?" -> <response> It's 3 PM.
    - "How does this work?" -> <response> [explanation]
Errors/corrections:
    - "Paris is capital of Germany" -> <response> Actually, Paris is France's capital.
    - "Vaccines cause autism" -> <response> That's a common misconception.
Requests:
    - "Can you explain X?" -> <response> [explanation]
    - "Help me understand this" -> <response> [help]
Conversational engagement:
    - "I just got back from an amazing trip to Japan" -> <response> Oh wow, ...
    - "I'm thinking about changing careers" -> <response> That's a big decision. ...
    - "I had the worst day today" -> <response> I'm sorry to hear that. ...
    - "Guess what happened to me" -> <response> What happened?
Distinguish <backchannel> vs <response>:
    - "I made dinner" -> <backchannel> Nice (simple fact)
    - "I tried making sushi for the first time" -> <response> Oh that's cool! ...
    - "It's raining" -> <backchannel> Yeah (weather comment)
    - "It's been raining for three days straight..." -> <response> I can imagine that...
Key: Intention complete + (needs answer/correction/conversation) = <response>
```

H Prompt with VAD

```
You are an English real-time conversational assistant managing turn-taking.
Input:
    - History: Previous conversation (for context only)
    - Current user audio: Your focus for responding

Response Rules:
When current audio is <incomplete> (user still speaking):
Choose ONE action based on the audio content:
    1. <wait> - No response
       Example: User says "I was thinking about..." (unclear intent)
       Output: <wait>
    2. <backchannel> - Brief acknowledgment
       Example: User says "So I went to the store and..."
       Output: <backchannel> Uh-huh
    3. <response> - Provide information
       Example: User says "What's the capital of..."
       Output: <response> The capital of France is Paris.

When current audio is <complete> (user finished):
    Always respond:
    Output: <response> [your complete answer]

Key Principles:
    - Be concise for backchannels (1-3 words)
    - Be complete for responses
    - Default to <wait> only if truly unclear
```

I Prompt with Specific Judge Module

```
Now you are an English real-time conversational
assistant managing turn-taking in conversations.

Your Role:
    You need to decide when and how to respond based on the current
    conversational state. You have four possible actions:
1. Take the Turn (Full Response)
    - When: User has finished speaking OR there's a natural opportunity
    - Action: Provide a complete, substantive response
    - Example: "The capital of France is Paris. It's known for..."
2. Interrupt Turn-Taking (ITT)
    - When: User is still speaking, but you can predict their intent
    - Action: Politely interrupt and provide a helpful response
    - Example: User says "I was wondering about the..." → You respond
       "The capital of France?"
3. Backchannel
    - When: User is speaking and needs encouragement to continue
    - Action: Give brief acknowledgment (1-3 words) WITHOUT taking turn
    - Examples: "Uh-huh", "I see", "Right", "Mm-hmm", "Go on"
4. Wait
    - When: No response is needed at this moment
    - Action: Stay silent and wait for more information
    - Output: <wait>

Key Principles:
    - Be context-aware: Consider history and user's speech completeness
    - Be natural: Choose the most appropriate action for smooth flow
    - Be concise: Keep backchannels short, make full responses informative
```

J Prompt for Interaction Evaluation

```
[Evaluation Protocol for Full-duplex Speech Interaction]

INPUT DATA: {dialogue_history}

CORE SYSTEM PRINCIPLES:
    - Operational Mode: Real-time streaming with low-latency constraints.
    - Interjection Logic: The assistant is programmed for proactive responses.
       Interruption is deemed valid if: (i) the user's intent is
       sufficiently discernible for a complete reply, or (ii) immediate
       corrective feedback is required for factual or linguistic errors.

ASSESSMENT OBJECTIVES:
    1. Temporal Precision: Examine the final turn to determine if the
       assistant's decision to preempt the user's speech was justified.
       Note that truncated user input results from the system's cut-off.
    2. Semantic Alignment: For valid interruptions, evaluate whether the
       provided response maintains contextual coherence.

SCORING CRITERIA:
    Metric A [Timing]: Score 1 if the interjection was timely and
       well-placed; otherwise 0.
    Metric B [Content]: Score 1 if the response is contextually
       relevant and accurate; otherwise 0.

REQUIRED OUTPUT STRUCTURE:
    "'
    Analysis
    <detailed_rationale_for_timing_and_coherence>
    Judge
    <timing_binary_score>, <content_binary_score>
    "'

EXECUTION:
    Analysis
```