Decoupling Conversational Dynamics in Full-Duplex Spoken Models through Reinforcement Learning

Yuxin Li[1] **Donghang Wu**[1] **Guan-Ting Lin**[2] **Hung-yi Lee**[2] **Chengwei Qin**[3]
Zhehuai Chen[4] **Chen Chen**[4]

[1]Nanyang Technological University [2]National Taiwan University
[3]The Hong Kong University of Science and Technology [4]NVIDIA

Abstract

Recent full-duplex spoken dialogue models have demonstrated compelling progress toward human-like interaction, enabling agents to respond with low latency, produce backchannels, and handle user barge-ins. Yet these improvements in conversational dynamics often come with weaker reasoning and instruction-following abilities, revealing a potential tension between interactive dynamics and intelligence capability. In this paper, we argue that such an **intelligence–dynamics** trade-off is not fundamental: conversational dynamics can instead be optimized as a distinct objective, within the same model, using human dialogue data. To this end, we propose DuplexPO, a reinforcement learning (RL) framework that decouples when to speak from what to say. It preserves the semantic response capability of an instruction-tuned assistant, while optimizing its temporal interaction behavior over selected high-impact windows from long human conversations. To quantitatively optimize these dynamics, we formulate the Factorized Conversational Dynamics Reward (FCDR) to enable fine-grained temporal credit assignment for turn initiation, backchanneling, and termination. The policy is then optimized with a GRPO-style objective. Experiments show that DuplexPO substantially improves full-duplex behaviors, including timely backchannels, smooth turn-taking, and barge-in handling, while maintaining strong reasoning and instruction-following performance. Moreover, improvements in dynamics-oriented metrics are reflected in better user experience, suggesting that optimizing conversational timing as a standalone objective can promote more natural full-duplex interaction.

1 Introduction

Speech is widely regarded as one of the most natural modalities for human-computer interaction, making spoken dialogue systems a long-standing focus in both academia and industry. In recent years, advances in neural modeling and large-scale data have substantially accelerated progress in this area [Cui et al., 2024]. In particular, end-to-end spoken dialogue models have emerged as an increasingly important paradigm for building interactive and helpful voice agents [Nguyen et al., 2023].

More recently, full-duplex spoken dialogue models have gained increasing attention in the research community due to their ability to support simultaneous listening and speaking [Défossez et al., 2024, Yu et al., 2025]. By allowing the model to process incoming user speech while generating responses, this paradigm enables real-time agent behaviors such as handling user barge-ins and producing timely backchannels. As a result, full-duplex modeling offers a more responsive and human-like conversational experience compared to traditional turn-based approaches [Wu et al., 2025b].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

However, some full-duplex models underperform their half-duplex counterparts on instruction-following and reasoning benchmarks [Hu et al., 2025, Jain et al., 2025], suggesting a potential **intelligence–dynamics** trade-off. We consider two possible sources of this tension. At the modeling level, multi-step reasoning requires sustained inference [Wei et al., 2022, Xie et al., 2025], whereas natural full-duplex interaction requires local, real-time decisions about whether to speak, pause, or remain silent as new input arrives. At the data level, natural dialogue corpora such as Fisher [Cieri et al., 2004] provide useful interaction patterns, but their casual exchanges are not necessarily aligned with helpful, instruction-following behavior.

These considerations motivate treating "*what to say*" and "*when to speak*" as distinct optimization objectives. The former concerns semantic response generation, while the latter governs turn-taking, backchanneling, and yielding [Lin et al., 2025a]. Rather than imitating complete human dialogue responses, an instruction-tuned spoken language model can adapt its policy at interaction-critical moments. Our central hypothesis is that timing-focused policy optimization can improve conversational dynamics while preserving instruction-following and reasoning abilities. This follows the broader principle of targeted speech-specific alignment, where auxiliary speech capabilities are learned without replacing general language capability [Chen et al., 2025b].

To operationalize this separation, we propose DuplexPO, an RL framework that treats conversational dynamics as an independent optimization target for full-duplex spoken dialogue models. Concretely, DuplexPO samples dynamics-critical windows from long multi-turn human conversations, focusing optimization on informative full-duplex events such as turn transitions, backchannels, and barge-ins. Since trajectories are sampled from the model itself, the policy is improved under its own response distribution rather than forced to imitate heterogeneous conversational corpora. It then computes a Factorized Conversational Dynamics Reward (FCDR) from the sampled windows and uses it as the reward signal for RL. The automatic metric improvements are accompanied by higher human preference in a blinded listening study. Finally, the GRPO-style objective includes KL regularization toward the frozen SFT reference policy to constrain policy updates [Schulman et al., 2017, Ouyang et al., 2022, Shao et al., 2024].

Empirically, DuplexPO improves turn-taking, backchanneling, and barge-in handling across multiple conversational-dynamics evaluations, while maintaining comparable performance on factual QA, instruction following, speech understanding, and reasoning benchmarks. We provide a demo page at `https://liyuxin44.github.io/DuplexPO/` and codes are available at `https://github.com/LiYuxin44/DuplexPO_official`.

2 Related Work

2.1 Full-Duplex Dialogue: Systems vs. Models

Traditional spoken dialogue systems usually decompose real-time interaction into recognition, dialogue management, incremental processing, and synthesis modules [Moritz et al., 2020, He et al., 2019, Schlangen and Skantze, 2011, Chiba and Higashinaka, 2025, Skerry-Ryan et al., 2018]. Full-duplexity is therefore best understood as a functional property rather than a single architecture. A system can obtain it through external orchestration, such as VAD or dialogue-management control [Yamamoto et al., 2025, Shi et al., 2023, Zhang et al., 2025], while a model can internalize it as part of its own generation process [Yan et al., 2026a, Veluri et al., 2024, Wang et al., 2024a]. In system-level designs, auxiliary control can determine *when to speak* while the core model focuses on response generation.

Model-level full duplex follows a different formulation. Models such as Moshi, SALMONN-omni, SALM-Duplex, and related latent-reasoning variants expose turn-taking, backchanneling, and barge-in behavior directly to the model [Défossez et al., 2024, Yu et al., 2025, Hu et al., 2025, Wu et al., 2026]. Recent full-duplex benchmarks similarly treat these behaviors as explicit model capabilities rather than peripheral interface features [Lin et al., 2025a, 2026b,a]. This paper therefore focuses on model-level full-duplex dialogue, where the key challenge is not only generating an appropriate answer, but learning a temporally coordinated policy for deciding when that answer should enter, pause, or leave the shared speech stream.

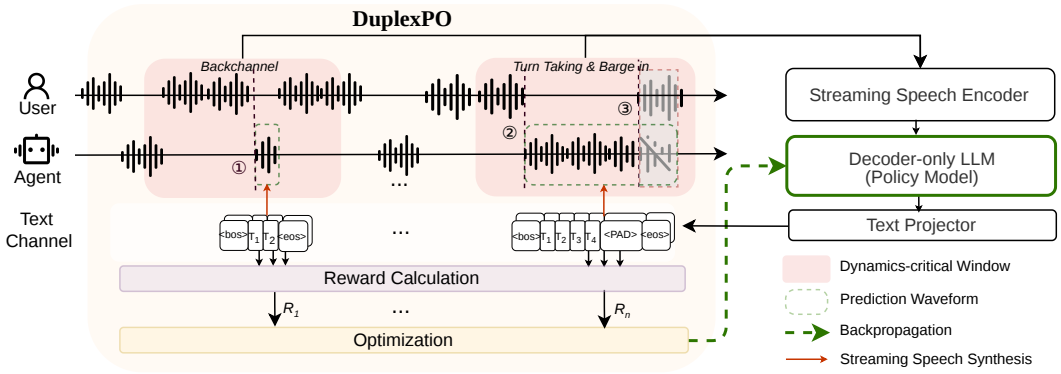


Figure 1: Overview of DuplexPO for full-duplex spoken dialogue models. The red shadows indicate dynamics-critical windows that may include important conversational behaviors, such as: ① backchanneling, ② turn-taking, and ③ user barge-in. For each window, the policy model generates multiple rollouts for reward calculation and is optimized with RL.

2.2 Reinforcement Learning for Dialogue Agents

RL is well suited to dialogue policy optimization, where actions often have delayed effects beyond next-token likelihood [Sutton and Barto, 2018]. Prior work has applied RL to task-oriented dialogue, turn-taking, conversational reasoning, retrieval-augmented QA, and AI-feedback-based response optimization [Raux and Eskenazi, 2012, Khouzaimi et al., 2016, Acikgoz et al., 2026, Lupart et al., 2025, Arora et al., 2026, Rakib et al., 2026]. This direction has also begun to reach end-to-end spoken language models. Align-SLM uses AI-feedback preference optimization to improve semantic coherence in textless SLMs, while user-interaction alignment constructs large-scale preference pairs from raw multi-turn speech conversations for full-duplex speech-to-speech models [Lin et al., 2025b, Wu et al., 2025a]. These studies show that reward-based learning is useful when the desired behavior is sequential, context-dependent, or only weakly specified by supervised targets.

Full-duplex dynamics pushes this view into a more local temporal regime. Yielding after user barge-in, producing a brief backchannel, or delaying a turn start are decisions whose quality depends on a narrow acoustic and conversational context, yet whole-dialogue objectives can blur their credit assignment [Zhao et al., 2025]. Recent spoken-dialogue RL and reward-modeling work has begun to explicitly optimize interaction behaviors such as turn-taking and backchanneling by learning dialogue-level policies or reward models over entire conversations [Chen et al., 2025a, 2026]. In contrast, our approach focuses on fine-grained temporal credit assignment, restricting policy updates to short, interaction-critical windows rather than optimizing over full dialogues. ASPIRin shows that raw-token RL for full-duplex timing can degrade semantic quality, and addresses this by projecting the action space into active-speech versus inactive-silence states before GRPO-style optimization [Hsiao et al., 2026]. DuplexPO instead keeps the original action space and changes the optimization unit, updating only dynamics-critical windows rather than the full dialogue.

3 Methodology

3.1 Overview

We propose DuplexPO, a policy optimization method that decouples when and how to engage in conversation (turn-taking, backchanneling, and yielding to user barge-ins) from what to say (semantic content). This separation is implemented through the optimization objective and its temporal support within the same full-duplex model, while retaining the original action space. As shown in Figure 1, the framework consists of three main components. First, dynamics-critical window sampling selects local regions around annotated conversational events. Second, the Factorized Conversational Dynamics Reward (FCDR) scores the timing and coordination of sampled responses within each window. Third, group-based policy optimization uses these rewards to improve the policy by comparing continuations generated from the same context.

3.2 Problem Formulation

We formulate full-duplex conversational dynamics learning as frame-level policy optimization over streamed dialogue. Let $x_{1:T}$ denote the user audio stream and $y_{1:T}$ denote the agent's real-time interaction decisions on a uniform temporal grid with frame duration Δ. At each frame, the policy predicts

$$\pi_\theta(y_t \mid y_{<t}, x_{\leq t}), \tag{1}$$

where $x_{\leq t}$ is the causal user context and y_t is the token the agent emits on its text stream at frame t: a text token while speaking, the control tokens <BOS> and <EOS> that initiate and yield a turn, or <PAD> while silent. The agent's interaction state (silent, speaking, initiating, yielding) is read off from these ordinary generated tokens; the action space is the model's full vocabulary, not a reduced set of states.

3.3 Dynamics-Critical Window Sampling

Conversational turn transfer often depends on local timing and prosodic cues near possible response points [Ward and Tsukahara, 2000, Sacks et al., 1974]. Motivated by this, DuplexPO selects dynamics-critical windows around annotated turn transitions, backchannels, and user barge-ins.

Let (a_i, b_i) denote the start and end times of the i-th annotated human agent segment, and let c_i denote its event metadata, including whether the segment is a backchannel and whether the corresponding window contains user barge-in. Each window is defined as

$$W_i = (s_i, e_i, g_i, h_i, c_i), \tag{2}$$

where $[g_i, h_i)$ is the discretized reference speaking interval and $[s_i, e_i)$ is the sampled window. The reference interval is

$$g_i = \lfloor a_i/\Delta \rfloor, \qquad h_i = \lfloor b_i/\Delta \rfloor. \tag{3}$$

Candidate windows extend L seconds before each annotated segment and B seconds after it. Segments are indexed chronologically, before filtering or sampling. With $q_0 = 0$, we set

$$s_i = \max(q_{i-1}, \lfloor (a_i - L)/\Delta \rfloor), \qquad \bar{e}_i = \lfloor (b_i + B)/\Delta \rfloor. \tag{4}$$

For a non-final segment, the next segment limits the boundary:

$$q_i = \min(\bar{e}_i, \max(q_{i-1}, \lfloor (a_{i+1} - L)/\Delta \rfloor)). \tag{5}$$

For the final segment, $q_i = \min(\bar{e}_i, T)$, where T is the conversation length in frames. To allow observation of stopping behavior after the reference offset h_i, we extend the window where possible:

$$\tilde{e}_i = \min(\bar{e}_i, T, \max(q_i, h_i + 3)). \tag{6}$$

Here, three frames correspond to 0.24 s and the next window's start still uses q_i.

We discard candidates with $\tilde{e}_i \leq s_i$ or $\tilde{e}_i \leq g_i$, and retain only those starting at least 2 s into the conversation. Full-turn segments must additionally last at least 2 s. Using a fixed seed, we sample at most three full-turn windows and one backchannel window per conversation, then cap each rollout at $H = 200$ frames:

$$e_i = \min(\tilde{e}_i, s_i + H).$$

Within $[s_i, e_i)$, the policy samples a continuation conditioned on the teacher-forced history before s_i. The policy loss is applied only to sampled tokens within this window.

3.4 Factorized Conversational Dynamics Reward (FCDR)

ORISE [Chen et al., 2025a] demonstrates the potential of rollout-level rewards for improving spoken interaction. To target specific conversational events, we introduce FCDR, which combines event-specific onset scores with conditional duration and termination corrections. The reward design is motivated by findings on human turn-taking and backchannel behavior [Sacks et al., 1974, Stivers et al., 2009, Levinson and Torreira, 2015]. We compare FCDR with an ORISE-inspired rollout-level reward under the same backbone and approximately matched RL update counts in Section 6.3.

For each window W_i, FCDR evaluates a sampled continuation with a scalar reward. We omit the continuation index here for clarity:

$$R_i = R(W_i, y_{[s_i, e_i)}), \tag{7}$$

Table 1: FCDR scoring rules. A reference is present when $h_i > g_i$. The last four rows apply when both a reference and a predicted onset are present. For a predicted interval, $d_i = \Delta(\hat{h}_i - \hat{g}_i)$ if a stop is observed, and $d_i = \infty$ otherwise.

Condition	Final reward R_i
No reference, no predicted onset	0
Reference present, no predicted onset	-0.5
No reference, predicted onset present	-0.5
Full turn, stop observed	r_i
Full turn, no stop observed	$0.75r_i$
Backchannel, $r_i > 0$ and $d_i > 1.5\,\mathrm{s}$	-0.5
Backchannel, otherwise	r_i

where $y_{[s_i,e_i)}$ denotes the agent's sampled tokens within the window and R_i is the final reward.

Within $[s_i, e_i)$, we score the first predicted speech interval, whose onset $\hat{g}_i$ is the first <BOS> position and whose offset $\hat{h}_i$ is the first subsequent <EOS> or <PAD> position. All positions use the same conversation-level frame grid as g_i and h_i. If no onset is generated, the prediction is absent. If an onset is generated without a subsequent stop, its duration is treated as infinite. The missing-stop rule also applies when the reference offset falls beyond the rollout horizon. The reference speaking interval is $[g_i, h_i)$.

When both a reference and a predicted onset are present, define the onset error in seconds as

$$\tau_i = \Delta(\hat{g}_i - g_i). \tag{8}$$

The backchannel branch is used when the event is labeled as a backchannel or its reference duration $(h_i - g_i)\Delta$ is shorter than 1 second. The onset score is

$$r_i = \begin{cases} \exp\left(-\dfrac{\tau_i^2}{2\sigma(\tau_i)^2}\right), & \text{full turn,} \\ \max\left(-0.5,\, 0.5 - \dfrac{\max(0, |\tau_i| - 0.5\,\mathrm{s})}{1\,\mathrm{s}}\right), & \text{backchannel,} \end{cases} \tag{9}$$

where $\sigma(\tau_i) = 0.2\,\mathrm{s}$ for $\tau_i < 0$ and $0.5\,\mathrm{s}$ otherwise.

The final reward combines this event-specific onset score with conditional duration and termination corrections, as summarized in Table 1. Full-turn onset uses an asymmetric Gaussian to penalize early starts more strongly than late starts. Backchannel onset receives equal credit within a tolerance interval around the reference onset, followed by a linear decrease with timing error. For positively scored predictions, an overlong or unterminated backchannel receives a penalty, while an unterminated full turn receives a reduced score. These are behavior-informed heuristics, not uniquely optimal functions. The resulting scalar window reward is used for group-based optimization. We also compare FCDR with a neural reward model based on temporal state prediction in Appendix I.

3.5 Group-based Optimization

DuplexPO uses a GRPO-style objective [Shao et al., 2024] to optimize sampled continuations within each dynamics-critical window. Continuations in the same group share the dialogue history and target event, allowing their rewards to be compared under a common conversational context. Group-normalized advantages use the rewards of all sampled continuations, whereas our DPO-style baseline constructs a preference pair from the highest- and lowest-reward continuations.

For window i, we normalize rewards across its sampled continuations to obtain a clipped group advantage:

$$A_{i,k} = \mathrm{clip}\left(\frac{R_{i,k} - \mu_i}{\sigma_i + \epsilon}, -5, 5\right), \tag{10}$$

where $R_{i,k}$ is the final reward of the k-th continuation in window i, and μ_i, σ_i are the mean and standard deviation of these final rewards within the group. We apply the policy-gradient loss only to sampled tokens inside the window, together with a KL penalty toward a frozen reference policy,

namely the SFT model before RL. The reference log-probabilities of the sampled tokens are computed with the frozen model and used in the KL term:

$$\mathcal{L}_{\mathrm{GRPO}} = \mathcal{L}_{\mathrm{policy}} + \beta \mathcal{L}_{\mathrm{KL}}. \tag{11}$$

We also implement a DPO-style baseline by constructing pairwise preferences among continuations from the same window, and report the ablation in Appendix J.

4 Data and Model Training

4.1 Conversational Dynamics-Aware Dialogue Data Generation

Fisher [Cieri et al., 2004] and Seamless-Naturalistic-HQ (Seamless) [Agrawal et al., 2025] provide natural two-party conversations with word-level timing, making them suitable sources for learning conversational dynamics. The two corpora cover complementary interaction regimes: Fisher is closer to symmetric peer-to-peer conversation, whereas Seamless contains more role-asymmetric, question-answer-oriented interactions. This combination enables us to test whether DuplexPO can handle both balanced conversational exchange and assistant-like spoken interaction. All Fisher and Seamless evaluations use held-out conversations that are split before reconstruction and window annotation. The training conversations, derived windows, and timing annotations are disjoint from the evaluation conversations and evaluation annotations. Reconstruction, filtering, and sampling details for data preprocessing are given in Appendix C.

4.2 Training Pipeline

The training pipeline consists of four stages: pre-training, SFT, RL with DuplexPO, and speech-synthesis SFT. The pre-training and SFT stages follow the setup of Nemotron-VoiceChat [NVIDIA, 2026] and we provide more training and data details in Appendix C. Then, DuplexPO is applied to optimize conversational dynamics with explicit rewards over dynamics-critical windows. Finally, in speech-synthesis SFT, we freeze all parameters except the speech generation module and train the streaming TTS model on the same speech dataset [Du et al., 2024]. The implementation, optimization, and stage-wise hyperparameters are in Appendix D.

5 Evaluation and Benchmark

We evaluate Spoken Dialogue Language Models (SDLMs) from two complementary aspects, namely conversational dynamics and model intelligence. The former assesses full-duplex conversational dynamics, such as turn-taking and backchanneling. The latter examines general capabilities, including factual knowledge, instruction following, speech understanding, and reasoning.

5.1 Conversational Dynamics Evaluation

Conversational dynamics are evaluated on Fisher [Cieri et al., 2004], Seamless [Agrawal et al., 2025], and Full-Duplex-Bench v3 (FDB-v3) [Lin et al., 2026a]. For each conversation, we evaluate the event windows of Section 3.3 across the dialogue, each scored once and conditioned independently on the teacher-forced history before its own start. Consecutive windows share at most the post-offset extension $\tilde{e}_i - q_i$ of Eq. (6). On Fisher and Seamless, we report window-level initiation rate, onset MAE, and yield rate for both full-turn and backchannel events, measuring <BOS> occurrence, <BOS> timing error, and turn termination, respectively. Here, <BOS> and <EOS> denote the beginning and end of the agent's speaking turn, respectively. Onset MAE is computed only for initiated events. Exact rules, denominators, and time thresholds for these and for the rhythm accuracy, interruption yield, and overlong-backchannel metrics used in Section 6.3 and Appendix B are given in Appendix E. FDB-v3 tests robustness to premature agent speech during disfluent user turns and user re-entry. We also report turn-taking latency, voiced interrupt rate (VIR), and yield rate. VIR measures the percentage of agent speech onsets that overlap with voiced user speech. We exclude onsets during internal user pauses and within the final 0.5s of a voiced user segment, where overlap may reflect turn-transition ambiguity rather than interruption. Yield rate measures whether the agent releases the floor before the user resumes speaking.

Table 2: Window-level turn-taking and backchannel performance on Fisher and Seamless. Rates (except Onset MAE) are reported in %. Best values in each column are bolded.

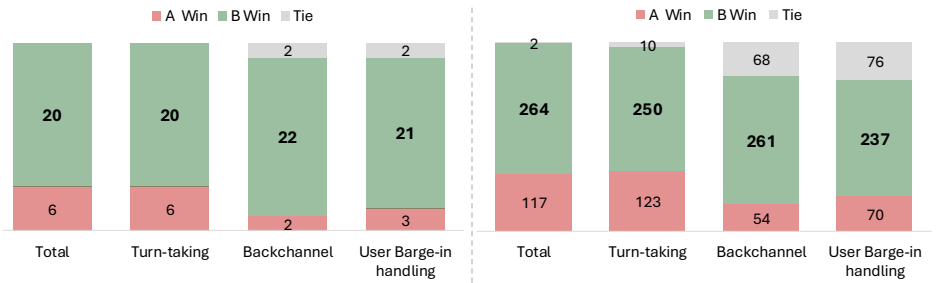
Method	Dataset	Onset MAE (↓)	Turn-taking		Backchannel	
			Init Rate (↑)	Yield Rate (↑)	Init Rate (↑)	Yield Rate (↑)
Moshi [Défossez et al., 2024]	Fisher	1.99	87.7	76.5	83.9	66.7
PersonaPlex [Roy et al., 2026]		1.50	98.6	80.3	91.4	69.2
SFT Baseline		0.98	97.8	92.1	95.7	57.1
SFT Dynamics		1.14	97.1	78.9	91.4	57.1
DuplexPO		**0.69**	**100.0**	**98.7**	**97.8**	**100.0**
Moshi [Défossez et al., 2024]	Seamless	2.16	80.1	69.0	76.2	71.3
PersonaPlex [Roy et al., 2026]		1.77	82.8	79.6	82.3	84.0
SFT Baseline		1.22	91.8	78.5	92.2	79.4
SFT Dynamics		1.34	79.7	63.0	83.2	63.4
DuplexPO		**1.03**	**98.0**	**93.6**	**99.5**	**93.3**

Following prior pairwise evaluation protocols [Zheng et al., 2023, Chiang et al., 2024], we use Gemini 3.0 Pro to compare conversational dynamics between DuplexPO and the *SFT Baseline*. The judge receives anonymized timelines constructed from the same sampled windows for both models, together with aggregate dynamics statistics. Prompts and dimension-level results are provided in Appendix H.

Human evaluation. We conduct a blinded A/B listening study on 96 paired continuous excerpts from Fisher, Seamless, and FDB-v3. Models use identical inputs, TTS voices, and fixed user tracks, with full-sequence inference and randomized candidate order. Each pair receives three independent ratings from a pool of five evaluators; pair-level majority decisions assess overall dynamics, turn-taking, backchannels, and barge-in handling. Details are provided in Appendix A.

5.2 Model Intelligence Evaluation

To verify that conversational dynamics optimization preserves task-level capability, we evaluate SDLMs on factual knowledge, instruction following, speech understanding, and reasoning benchmarks. Factual knowledge is measured by accuracy on four quick QA sets: Llama Questions [Nachmani et al., 2023], WebQuestions [Berant et al., 2013], TriviaQA [Joshi et al., 2017], and SDQA [Faisal et al., 2021]. Instruction following is evaluated with AlpacaEval and CommonEval from VoiceBench [Chen et al., 2024]; the latter uses real human speech and better reflects spoken instruction-following scenarios. Following the official VoiceBench protocol, open-ended responses are scored by a GPT judge on a 1–5 scale. Speech understanding and reasoning are measured by multiple-choice accuracy on OpenBookQA and MMSU [Wang et al., 2025]. For all benchmarks, generated speech is transcribed with Whisper large v3 [Radford et al., 2022] before computing text-based metrics under the official VoiceBench protocol.



(a) Pairwise Judge Evaluation on Fisher test set (n=26) (b) Pairwise Judge Evaluation on Seamless test set (n=383)

Figure 2: Conversation-level Gemini pairwise evaluation comparing the *SFT Baseline* (A) with DuplexPO (B). The judge observes timestamps, transcripts, and aggregate dynamics statistics to assess turn-taking, backchanneling, and barge-in handling.

Table 3: Turn-taking dynamics, VIR, and yield rate on FDB-v3. Rates are reported in %, and Latency is reported in seconds. The results of GPT-Realtime, Gemini Live 2.5, Gemini Live 3.1, Grok and Ultravox are taken from [Lin et al., 2026a]. Best values in each column are bolded.

Method	Turn-Taking Dynamics		VIR (↓)	Yield (↑)
	Turn-Take (↑)	Latency (↓)		
GPT-Realtime [OpenAI, 2026]	96.0	2.65	–	–
Gemini Live 2.5 [Google, 2025]	92.0	3.84	–	–
Gemini Live 3.1 [Google, 2026]	78.0	1.59	–	–
Grok [x.ai, 2026]	94.0	3.13	–	–
Ultravox [fixie-ai, 2026]	96.0	1.90	–	–
Moshi [Défossez et al., 2024]	**100.0**	0.44	12.0	71.9
PersonaPlex [Roy et al., 2026]	98.0	1.41	15.0	66.7
SFT Baseline	99.0	7.33	8.0	64.8
SFT Dynamics	**100.0**	14.24	**4.0**	93.3
DuplexPO	**100.0**	**0.24**	5.0	**100.0**

Table 4: Model intelligence evaluation. QA, OpenBookQA, and MMSU metrics are accuracy in %; AlpacaEval and CommonEval are GPT-scores on a 1–5 scale. Baseline results are taken from [Yu et al., 2025, Chen et al., 2024]. FD denotes full duplex. LlamaQ, WebQ, TriQA, OBQA, AlpacaE, and ComE denote Llama Questions, WebQuestions, TriviaQA, OpenBookQA, AlpacaEval, and CommonEval. Bold values indicate improvements of DuplexPO over the SFT Baseline.

Method	FD	LlamaQ	WebQ	TriQA	SDQA	AlpacaE	ComE	OBQA	MMSU
Moshi [Défossez et al., 2024]	✓	54.5	22.1	16.7	15.6	2.01	1.60	25.9	24.0
Freeze-Omni [Wang et al., 2024b]	✓	56.2	27.9	28.5	53.5	4.03	3.46	31.0	28.1
SALMONN-omni [Yu et al., 2025]	✓	73.6	43.7	56.0	-	3.22	-	-	30.0
SALM-Duplex [Hu et al., 2025]	✓	51.3	25.0	16.9	26.0	2.99	2.50	39.6	26.3
GLM-4-Voice [Zeng et al., 2024]	✗	65.7	37.0	47.5	37.0	3.97	3.42	53.4	39.8
Qwen2-Audio [Chu et al., 2024]	✗	69.7	45.2	40.3	35.7	3.74	3.43	49.5	35.7
Kimi-Audio [Ding et al., 2025]	✗	68.3	37.3	51.2	63.1	4.46	3.97	83.5	62.2
Baichuan-Audio [Li et al., 2025]	✗	74.0	40.7	53.0	45.8	4.41	4.08	71.7	53.2
STITCH-R [Chiang et al., 2025]	✗	70.0	50.3	49.6	-	2.70	-	-	-
SFT Baseline	✓	72.0	44.3	48.1	47.2	3.43	3.48	72.2	54.9
DuplexPO	✓	**75.3**	**44.5**	**49.9**	**49.8**	**3.68**	**3.74**	**73.7**	**56.2**

6 Results and Analysis

6.1 Conversational Dynamics

We compare DuplexPO with two SFT-only baselines. *SFT Baseline* is trained without dynamics-aware dialogue data and is the checkpoint from which DuplexPO is initialized, whereas *SFT Dynamics* keeps the same SFT recipe and additionally includes the reconstructed dynamics-aware dialogue data, i.e., it is exposed to the same dynamics data as DuplexPO but through imitation rather than RL.

Table 2 shows that DuplexPO consistently improves window-level turn-taking and backchannel behavior on both Fisher and Seamless. *SFT Dynamics* regresses on most reported dynamics metrics relative to the *SFT Baseline*, while Fisher backchannel yield remains unchanged. This indicates that adding natural dialogue imitation alone does not reliably improve conversational coordination in our setting. DuplexPO achieves the highest initiation and yield rates for both turn-taking and backchannels, as well as the lowest onset MAE on both datasets.

Table 3 presents a latency–interruption trade-off on FDB-v3: commercial models and Ultravox have higher latency, while Moshi and PersonaPlex respond faster but exhibit higher VIR. *SFT Dynamics* reduces VIR at the cost of long latency. DuplexPO achieves 100% turn-taking and yield rates, the lowest latency among the evaluated models, and a competitively low VIR, indicating improved coordination between responsiveness and floor control.

Figure 2 evaluates whether these metric gains are reflected at the conversation level. The pairwise judge prefers DuplexPO in 76.9% of Fisher comparisons and 69.3% of non-tie Seamless comparisons. Dimension-level preferences are consistent with the automatic metrics for turn-taking, backchanneling, and user barge-in handling. A timing-perturbation sanity check of the judge is reported in Appendix G.

Table 5: Controlled reward comparison using the same backbone, initialization, and approximately matched RL update counts. Metrics are defined in Appendix E.

Dataset	Reward	Rhythm accuracy (%) ↑	Interruption yield (%) ↑
Fisher	ORISE-inspired	87.88	82.05
	FCDR	**88.31**	**87.18**
Seamless	ORISE-inspired	77.52	85.15
	FCDR	**78.49**	**92.16**

Human evaluation. DuplexPO is preferred in 61 of the 86 non-tied overall comparisons (70.9%; 25 baseline wins and 10 ties among 96 pairs; 95% Wilson CI: 60.6%–79.5%; two-sided exact binomial $p = 1.3 \times 10^{-4}$). Its preference rates among non-tied comparisons are 68.6% for turn-taking, 75.9% for backchannel appropriateness, and 72.5% for barge-in handling. Inter-rater agreement is moderate (Krippendorff's $\alpha = 0.44$). On the same 96 pairs, Gemini agrees with human-majority decisions on 70 pairs (72.9%; Cohen's $\kappa = 0.54$).

6.2 Model Intelligence

Table 4 shows that DuplexPO maintains the *SFT Baseline*'s performance, with small improvements across the evaluated factual QA, instruction-following, speech-understanding, and reasoning benchmarks.

6.3 Reward and Window Ablations

We compare FCDR with an ORISE-inspired rollout-level reward using the same backbone, joint training data, seed, and approximately equal numbers of RL updates. As shown in Table 5, FCDR achieves higher rhythm accuracy and interruption yield on both Fisher and Seamless. These results support the benefits of event-structured rewards for aggregate rhythm coordination and interruption handling. Relative to the shared pre-RL checkpoint, joint-corpus FCDR improves rhythm accuracy and interruption yield on both Fisher and Seamless. This supports joint training as a practical way to optimize a single policy across complementary interaction regimes, while single-corpus transfer remains asymmetric (Appendix B).

We separately compare *SFT Dynamics*, full-dialogue FCDR, and windowed FCDR to examine the effect of localizing optimization. On Fisher, their turn-taking yields are 78.9%, 94.2%, and 98.7%, respectively, while their backchannel yields are 57.1%, 87.1%, and 100.0%. Windowed FCDR achieves higher yield rates in this comparison, consistent with the benefit of localizing optimization.

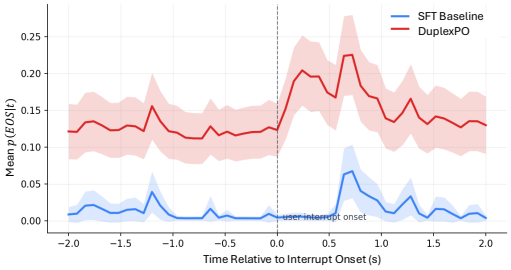
6.4 Transfer and Training Stability

Cross-corpus transfer is asymmetric: Seamless-only RL improves Fisher dynamics, while Fisher-only RL yields mixed results on Seamless. Three independent RL training seeds show limited variation in the evaluated dynamics metrics. Detailed transfer results and mean $\pm$ standard deviation across seeds are reported in Appendix B.

6.5 Discussion and Visualization

We examine boundary-token behavior on shared SFT-generated trajectories. Suppressed Intent Rate (SIR) measures how often the SFT Baseline assigns a boundary token probability above a fixed threshold without selecting it; Suppression Release Ratio (SRR) measures how often DuplexPO selects that token on the same frames (Appendix F). Figure 3(b) shows that DuplexPO selects many of these previously unselected boundary actions. Figure 3(a) compares <EOS> probabilities around user barge-ins. These diagnostics describe changes in boundary decisions on shared contexts; they do not establish an internal intent mechanism or characterize changes outside the analyzed frames. Appendix K provides a qualitative token-level example.

Additional comparisons evaluate the neural reward model and DPO-style optimization alternatives (Appendices I and J). FCDR and GRPO achieve better results on most reported dynamics metrics, with some metric-specific exceptions.



(a) <EOS> impulse around user interruption onset.

Dataset	<BOS> Impulses		<EOS> Impulses	
	SIR	SRR	SIR	SRR
Fisher	11.0	70.6	4.9	100.0
FDB-v3	10.8	71.9	0.9	100.0

(b) Suppressed-intent analysis. SIR: share of frames at which the *SFT Baseline* holds a latent <BOS>/<EOS> intent but does not act on it. SRR: share of exactly those frames at which DuplexPO emits the token. Values in %.

Figure 3: Analysis of suppressed boundary intent in *SFT Baseline* and DuplexPO. (a) <EOS> impulse around user interruption onset on Fisher test set (n=148 events from 20 conversations). Both policies are evaluated on the SFT trajectory; shaded regions show 95% CI. (b) Suppressed-intent analysis (Appendix F): SIR is measured on the *SFT Baseline*, SRR on DuplexPO over the same suppressed-intent frames; both are evaluated on the SFT trajectory.

6.6 Limitations

DuplexPO's factorized rewards enable interpretable event-level credit assignment, but may miss subtle pragmatic factors such as user intent, discourse content, speaker style, and culture-specific timing preferences. Human timing annotations are also not unique ground truth, since turn initiation, backchanneling, and yielding can vary across speakers, languages, and conversational settings. Window restriction preserves the available dialogue history as conditioning context but localizes the rollout and reward horizon. Consequently, delayed effects and error accumulation under self-generated histories remain outside the present evaluation: Fisher and Seamless dynamics metrics use teacher-forced histories, the listening study retains a fixed user stream, and the intelligence benchmarks primarily assess single-turn capabilities. The finite rollout horizon can affect termination scoring for long turns. Cross-corpus transfer is also asymmetric. The matched-window Gemini evaluation (Appendix H) uses a fixed candidate assignment, so a position bias cannot be excluded there; the human study and the matched-pair Gemini evaluation on the same 96 pairs randomize candidate order. Future work could extend evaluation to autonomous multi-turn interaction, multilingual settings, and user-adaptive timing preferences.

7 Conclusion

We propose DuplexPO, an RL framework that optimizes conversational timing over dynamics-critical windows while retaining the original model action space. Experiments show improved turn-taking, backchanneling, and barge-in handling, accompanied by higher human preference and maintained performance on the evaluated instruction-following, factual QA, speech-understanding, and reasoning benchmarks. These results support treating conversational dynamics as a distinct optimization objective within an instruction-tuned spoken dialogue model.

Acknowledgements and Disclosures

The authors have no funding or competing interests to disclose.

References

Emre Can Acikgoz, Jinoh Oh, Jie Hao, Joo Hyuk Jeon, Heng Ji, Dilek Hakkani-Tur, Gokhan Tur, Xiang Li, Chengyuan Ma, and Xing Fan. Speakrl: Synergizing reasoning, speaking, and acting in language models with reinforcement learning. In *Proceedings of the 16th International Workshop on Spoken Dialogue System Technology*, pages 312–325, 2026.

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

Siddhant Arora, Jinchuan Tian, Jiatong Shi, Hayato Futami, Yosuke Kashiwagi, Emiru Tsunoo, and Shinji Watanabe. Optimizing conversational quality in spoken dialogue systems with reinforcement learning from ai feedback. *arXiv preprint arXiv:2601.19063*, 2026.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. Hi-fi multi-speaker english tts dataset. *arXiv preprint arXiv:2104.01497*, 2021.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544, 2013.

Chen Chen, Ke Hu, Chao-Han Huck Yang, Ankita Pasad, Edresson Casanova, Weiqing Wang, Szu-Wei Fu, Jason Li, Zhehuai Chen, Jagadeesh Balam, et al. Reinforcement learning enhanced full-duplex spoken dialogue language models for conversational interactions. In *Second Conference on Language Modeling*, 2025a.

Chen Chen, Yuchen Hu, Siyin Wang, Helin Wang, Zhehuai Chen, Chao Zhang, Chao-Han Huck Yang, and Eng Siong Chng. Audio large language models can be descriptive speech quality evaluators. *arXiv preprint arXiv:2501.17202*, 2025b.

Yifu Chen, Shengpeng Ji, Zhengqing Liu, Qian Chen, Wen Wang, Ziqing Wang, Yangzhuo Li, Tianle Liang, and Zhou Zhao. Dual-axis generative reward model toward semantic and turn-taking robustness in interactive spoken dialogue models. *arXiv preprint arXiv:2604.14920*, 2026.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024.

Cheng-Han Chiang, Xiaofei Wang, Linjie Li, Chung-Ching Lin, Kevin Lin, Shujie Liu, Zhendong Wang, Zhengyuan Yang, Hung-yi Lee, and Lijuan Wang. Stitch: Simultaneous thinking and talking with chunked reasoning for spoken language models. *arXiv preprint arXiv:2507.15375*, 2025.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Yuya Chiba and Ryuichiro Higashinaka. Investigating the impact of incremental processing and voice activity projection on spoken dialogue systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3687–3696, 2025.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

C. Cieri, D. Miller, and K. Walker. The fisher corpus: A resource for conversational speech. *LDC*, 2004.

Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Yiwen Guo, and Irwin King. Recent advances in speech language models: A survey. *arXiv preprint arXiv:2410.03751*, 2024.

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

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

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

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. Sd-qa: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

fixie-ai. ultravox. https://github.com/fixie-ai/ultravox, 2026. Accessed: 2026-05-07.

Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. Freesound datasets: a platform for the creation of open audio datasets. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, pages 486–493, Suzhou, China, 2017.

Google. gemini2.5-flash-native-audio-preview-12-2025. https://ai.google.dev/gemini-api/docs/models/gemini2.5-flash-native-audio-preview-12-2025, 2025. Accessed: 2026-05-07.

Google. gemini3.1-flash-live-preview. https://ai.google.dev/gemini-api/docs/models/gemini3.1-flash-live-preview, 2026. Accessed: 2026-05-07.

Yanzhang He, Tara N Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, et al. Streaming end-to-end speech recognition for mobile devices. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6381–6385. IEEE, 2019.

Chi-Yuan Hsiao, Ke-Han Lu, Yu-Kuan Fu, Guan-Ting Lin, Hsiao-Tsung Hung, and Hung-yi Lee. Aspirin: Action space projection for interactivity-optimized reinforcement learning in full-duplex speech language models. *arXiv preprint arXiv:2604.10065*, 2026.

Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Żelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, and Boris Ginsburg. Efficient and direct duplex modeling for speech-to-speech language model. *arXiv preprint arXiv:2505.15670*, 2025.

Dhruv Jain, Harshit Shukla, Gautam Rajeev, Ashish Kulkarni, Chandra Khatri, and Shubham Agarwal. Voiceagentbench: Are voice assistants ready for agentic tasks? *arXiv preprint arXiv:2510.07978*, 2025.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Hatim Khouzaimi, Romain Laroche, and Fabrice Lefèvre. Reinforcement learning for turn-taking management in incremental spoken dialogue systems. In *IJCAI*, pages 2831–2837, 2016.

Nithin Rao Koluguri, Monica Sekoyan, George Zelenfroynd, Sasha Meister, Shuoyang Ding, Sofia Kostandian, He Huang, Nikolay Karpov, Jagadeesh Balam, Vitaly Lavrukhin, et al. Granary: Speech recognition and translation dataset in 25 european languages. *arXiv preprint arXiv:2505.13404*, 2025.

Oleksii Kuchaiev, Jason Li, Huyen Nguyen, Oleksii Hrinchuk, Ryan Leary, Boris Ginsburg, Samuel Kriman, Stanislav Beliaev, Vitaly Lavrukhin, Jack Cook, et al. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019.

Stephen C Levinson and Francisco Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:136034, 2015.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, et al. Baichuan-audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*, 2025.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. Yodas: Youtube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

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

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Aditya Gourav, Yile Gu, Ankur Gandhe, Hung-yi Lee, and Ivan Bulyko. Align-slm: Textless spoken language models with reinforcement learning from ai feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20395–20411, 2025b.

Guan-Ting Lin, Chen Chen, Zhehuai Chen, and Hung-yi Lee. Full-duplex-bench-v3: Benchmarking tool use for full-duplex voice agents under real-world disfluency. *arXiv preprint arXiv:2604.04847*, 2026a.

Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, Shinji Watanabe, and Hung-yi Lee. Full-duplex-bench v1. 5: Evaluating overlap handling for full-duplex speech models. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 19447–19451. IEEE, 2026b.

Simon Lupart, Mohammad Aliannejadi, and Evangelos Kanoulas. Chatr1: Reinforcement learning for conversational reasoning and retrieval augmented question answering. *arXiv preprint arXiv:2510.13312*, 2025.

Niko Moritz, Takaaki Hori, and Jonathan Le. Streaming automatic speech recognition with the transformer model. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6074–6078. IEEE, 2020.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm. *arXiv preprint arXiv:2305.15255*, 2023.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11: 250–266, 2023.

Vahid Noroozi, Zhehuai Chen, Somshubra Majumdar, Steve Huang, Jagadeesh Balam, and Boris Ginsburg. Instruction data generation and unsupervised adaptation for speech language models. *arXiv preprint arXiv:2406.12946*, 2024a.

Vahid Noroozi, Somshubra Majumdar, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. Stateful conformer with cache-based inference for streaming automatic speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12041–12045. IEEE, 2024b.

NVIDIA. Nemotron 3 VoiceChat Model Card. `https://build.nvidia.com/nvidia/nemotron-voicechat/modelcard`, 2026. Accessed: 2026-05-07.

OpenAI. gptrealtime-1.5. `https://developers.openai.com/api/docs/models/gptrealtime-1.5`, 2026. Accessed: 2026-05-07.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL `https://arxiv.org/abs/2212.04356`.

Tazeek Bin Abdur Rakib, Ambuj Mehrish, Lay-Ki Soon, Wern Han Lim, and Soujanya Poria. Dialogxpert: Driving intelligent and emotion-aware conversations through online value-based reinforcement learning with llm priors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 29967–29975, 2026.

Antoine Raux and Maxine Eskenazi. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing (TSLP)*, 9(1):1–23, 2012.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, et al. Fast conformer with linearly scalable attention for efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

Rajarshi Roy, Jonathan Raiman, Sang-gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. Personaplex: Voice and role control for full duplex conversational speech models. *arXiv preprint arXiv:2602.06053*, 2026.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *language*, 50(4):696–735, 1974.

David Schlangen and Gabriel Skantze. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111, 2011.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Mohan Shi, Yuchun Shu, Lingyun Zuo, Qian Chen, Shiliang Zhang, Jie Zhang, and Li-Rong Dai. Semantic vad: Low-latency voice activity detection for speech interaction. *arXiv preprint arXiv:2305.12450*, 2023.

RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR, 2018.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*, 2015.

Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592, 2009.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2459–2475, 2025.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, second edition, 2018.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL `https://qwenlm.github.io/blog/qwen2.5/`.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. Beyond turn-based interfaces: Synchronous llms as full-duplex dialogue agents. *arXiv preprint arXiv:2409.15594*, 2024.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*, 2025.

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

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

Nigel Ward and Wataru Tsukahara. Prosodic features which cue back-channel responses in english and japanese. *Journal of pragmatics*, 32(8):1177–1207, 2000.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Anne Wu, Laurent Mazaré, Neil Zeghidour, and Alexandre Défossez. Aligning spoken dialogue models from user interactions. *arXiv preprint arXiv:2506.21463*, 2025a.

Donghang Wu, Haoyang Zhang, Chen Chen, Tianyu Zhang, Fei Tian, Xuerui Yang, Gang Yu, Hexin Liu, Nana Hou, Yuchen Hu, et al. Chronological thinking in full-duplex spoken dialogue language models. *arXiv preprint arXiv:2510.05150*, 2025b.

Donghang Wu, Tianyu Zhang, Yuxin Li, Hexin Liu, Chen Chen, Eng Siong Chng, and Yoshua Bengio. The silent thought: Modeling internal cognition in full-duplex spoken dialogue models via latent reasoning. *arXiv preprint arXiv:2603.17837*, 2026.

x.ai. grok-voice-agent-api. `https://x.ai/news/grok-voice-agent-api`, 2026. Accessed: 2026-05-07.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*, 2025.

Kenta Yamamoto, Ryu Takeda, and Kazunori Komatani. Analysis of voice activity detection errors in api-based streaming asr for human-robot dialogue. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 245–253, 2025.

Ruiqi Yan, Wenxi Chen, Zhanxun Liu, Ziyang Ma, Haopeng Lin, Hanlin Wen, Hanke Xie, Jun Wu, Yuzhe Liang, Yuxiang Zhao, et al. Soulx-duplug: Plug-and-play streaming state prediction module for realtime full-duplex speech conversation. *arXiv preprint arXiv:2603.14877*, 2026a.

Ruiqi Yan, Wenxi Chen, Zhanxun Liu, Ziyang Ma, Haopeng Lin, Hanlin Wen, Hanke Xie, Jun Wu, Yuzhe Liang, Yuxiang Zhao, et al. Soulx-duplug: Plug-and-play streaming state prediction module for realtime full-duplex speech conversation. *arXiv preprint arXiv:2603.14877*, 2026b.

Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. Salmonn-omni: A standalone speech llm without codec injection for full-duplex conversation. *arXiv preprint arXiv:2505.17060*, 2025.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv preprint arXiv:1904.02882*, 2019.

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

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

Yue Zhao, Xiaoyu Wang, Dan Wang, Zhonglin Jiang, Qingqing Gu, Teng Chen, Ningyuan Xi, Jinxian Qu, Yong Chen, and Luo Ji. Dream to chat: Model-based reinforcement learning on dialogues with user belief modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4764–4781, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A Human Evaluation

We evaluate 96 paired continuous excerpts from Fisher, Seamless, and FDB-v3. Each pair compares DuplexPO with the *SFT Baseline* under the same user input, TTS system, and voice. Both models perform full-sequence inference over each continuous excerpt; the samples are not formed by concatenating independently generated windows. Model identities are concealed and candidate order is randomized.

Five evaluators provide three independent ratings per pair, yielding 288 judgments. We use the pair-level majority decision as the unit of analysis. Evaluators assess overall conversational dynamics, turn-taking, backchannel appropriateness, and barge-in handling. Backchannel appropriateness is judged from its timing, content, duration, and degree of floor competition in the local context.

DuplexPO wins 61 overall comparisons, the baseline wins 25, and 10 are ties. The preference rate is calculated over non-tied comparisons, giving $61/86 = 70.9\%$. We report a 95% Wilson confidence interval and a two-sided exact binomial test against an equal-preference null hypothesis. The corresponding interval is 60.6%–79.5%, with $p = 1.3 \times 10^{-4}$. Dimension-level non-tie preference rates are 68.6% for turn-taking, 75.9% for backchannel appropriateness, and 72.5% for barge-in handling. Inter-rater agreement is Krippendorff's $\alpha = 0.44$.

We also evaluate Gemini on the same 96 pairs using anonymized candidates, randomized candidate order, and no aggregate dynamics statistics. Gemini agrees with human-majority decisions on 70 pairs (72.9%), with Cohen's $\kappa = 0.54$. This matched-pair evaluation is separate from the matched-window Gemini protocol reported in Appendix H.

Materials and presentation. Each test item consists of one user audio track paired with two candidate system responses, presented as two complete two-channel interactions labeled A and B. The two candidates share the identical user track; only the system channel differs. The assignment of systems to the labels A and B was randomized per item, item order was randomized per evaluator, and evaluators were blind to system identity throughout. Evaluators listened over headphones, could replay and scrub freely, and were required to listen to both interactions in full at least once before answering.

Evaluation instructions. For each item, evaluators answered four questions in the fixed order below, selecting *A*, *B*, *Tie*, or *N/A* for each. They were instructed to judge only the system's behavior—the user's turns are identical across A and B and are not being rated—and to confine each criterion to its own aspect: voice quality, content correctness, and naturalness of wording are not considered under criteria 1–3 and may inform only the overall judgment (criterion 4).

1. **Turn-taking.** Which interaction starts and stops speaking at more appropriate moments? Prefer the system that (i) less often starts a turn while the user is still speaking, (ii) leaves fewer gaps that are noticeably longer than a natural pause after the user has finished, and (iii) does not cut off its own turn before its contribution is complete. Short overlaps that function as listener feedback are assessed under criterion 2, not here. Overlap caused by the user interrupting the system is assessed under criterion 3, not here. This criterion applies to every item.
2. **Backchannel appropriateness.** Backchannels are brief listener signals (e.g., "mm-hm", "right", "I see") produced while the user holds the floor and not intended to take it. Which interaction provides backchannels whose timing (at phrase or clause boundaries rather than mid-word), content (consistent with what the user just said), and duration and frequency (brief, not so frequent as to distract) fit the context, and which remain backchannels rather than growing into a full turn that competes for the floor? The absence of backchannels is not by itself a defect: a system that produces none may be preferred over one that produces ill-timed ones. Select N/A only if the user's turns in the item are too short to afford any backchannel opportunity.
3. **Barge-in handling.** Applicable only when the user begins speaking while the system is speaking. Which interaction responds more appropriately: does the system stop promptly (within roughly a syllable or two), refrain from resuming over the user, and, when it resumes, address what the user said rather than restarting its earlier content? A system should not be penalized for continuing when the user's interjection is itself a backchannel (e.g., "uh-huh")

rather than an attempt to take the floor. Select N/A when no user barge-in occurs in the item; because the user track is shared, N/A applies to the item as a whole.

4. **Overall preference.** Answered last. Which interaction would you rather have been the user in? This judgment integrates criteria 1–3 together with any other aspect of the experience (response content, voice, naturalness). It is not a tally of the preceding criteria: a single severe failure (e.g., talking over the user for several seconds) may outweigh small advantages elsewhere.

Tie indicates that, after listening, the evaluator perceives no difference between the candidates on the criterion, or a difference too small to support a confident preference; both candidates performing equally well or equally poorly are both ties. A tie does not mean the evaluator is unsure how to apply the criterion. Such cases were raised with the experimenters and resolved before continuing. *N/A* indicates that the criterion did not arise in the item and is kept distinct from a tie so that preference rates are computed only over items where the criterion applies. For any non-tie judgment on criteria 1–3, evaluators were encouraged to note a time stamp and a short description of the decisive event (e.g., "A starts over the user at 0:14"); these notes were used for adjudicating disagreements.

B Cross-Corpus Transfer and Training Stability

For cross-corpus evaluation, we train Fisher-only and Seamless-only RL variants and evaluate each on the corpus excluded from its RL training. All variants in Tables 5 and 6 use the same pre-RL checkpoint and RL budget; the *None* rows denote this shared checkpoint. Table 6 summarizes the reported results.

Table 6: Cross-corpus transfer. The evaluation corpus is excluded from RL training, and *None* denotes the shared pre-RL checkpoint. BC denotes backchannel. Metrics are defined in Appendix E.

Evaluation	RL data	Rhythm accuracy (%) ↑	Interruption yield (%) ↑	Overlong BC (%) ↓
Fisher	None	83.98	70.51	21.51
Fisher	Seamless	91.34	92.21	7.53
Seamless	None	74.22	89.22	24.84
Seamless	Fisher	75.00	87.25	25.48

Seamless-only RL improves all three reported metrics on Fisher. Fisher-only RL improves onset MAE and turn-level yield on Seamless, but produces mixed changes in the metrics shown in Table 6. This asymmetry may reflect differences in corpus scale, interaction structure, event distributions, alignment quality, or preprocessing, which are not independently controlled.

We repeat DuplexPO training with three independent RL seeds initialized from the same final *SFT Baseline* checkpoint as the main run, using the same checkpoint-selection and evaluation protocols. Table 7 reports the mean and standard deviation across these runs. On FDB-v3, the direction of improvement relative to the pre-RL checkpoint is consistent across all three seeds.

Table 7: Mean $\pm$ standard deviation across three independent RL training seeds. Time values are in seconds and rates are in percentages.

Dataset	Metric	Mean $\pm$ SD
Fisher	Onset MAE ↓	0.71 ± 0.03
	Turn-taking yield ↑	97.4 ± 1.4
	Backchannel yield ↑	96.2 ± 3.5
Seamless	Onset MAE ↓	1.05 ± 0.03
	Turn-taking yield ↑	92.1 ± 1.4
	Backchannel yield ↑	92.4 ± 1.2
FDB-v3	Latency ↓	0.29 ± 0.05
	VIR ↓	5.3 ± 1.2
	Yield ↑	98.7 ± 1.2

C Data Details

In this appendix, we illustrate all the synthetic and real data we used in this work. In pre-training, we use the speech-continuation data to make the model understand speech input under full-duplex modeling. Then, all other data is used in SFT to improve the reasoning ability and build a helpful speech agent. The broader data-construction recipe follows the released Nemotron-VoiceChat materials [NVIDIA, 2026].

1. Speech-continuation data. Speech-continuation pre-training uses 530K hours of examples derived from large text corpora [Su et al., 2025]. Each continuous passage is converted into a two-party pseudo-dialogue by assigning successive sentences to the user and agent streams. Turn lengths are randomized: a turn stops after one sentence with probability 0.8, while longer turns are formed by appending additional sentences with a decaying continuation probability; turns longer than 200 words are handed to the other speaker. The resulting turns are synthesized with two speaker profiles, aligned in time, and packed into synchronized two-stream full-duplex examples. We also swap user and agent roles to increase coverage of both conversational directions.

2. Instruction-following QA data. The instruction-following mixture contributes 70K hours of spoken QA data. Within this set, 10K hours are single-turn examples spanning broad topics, and the remaining multi-turn examples train the assistant to maintain conversations lasting up to four minutes. Dialogue generation uses a helpful-assistant prompt and is conditioned on heterogeneous textual contexts, including Wikipedia pages.[1] The generated turns are then rendered with the multi-speaker TTS pipeline.

3. Synthetic user interruption. We augment multi-turn examples with simulated barge-ins. When an agent utterance lasts more than four seconds, a user interruption is inserted with probability 0.1, and its onset is sampled uniformly from the 20%–80% span of the agent utterance. After the interruption begins, the agent stream is given an 8-token reaction delay (≈ 0.64 s) before it is forced to emit an end-of-sequence (<EOS>) token. At inference time, an <EOS> on the text channel masks the corresponding audio channel to silence.

4. ASR-QA data. ASR-QA examples expose the model to real recording conditions. We use speech segments from the open-source corpora aggregated in NeMo ASRSET [Noroozi et al., 2024b] as acoustic contexts. Following Noroozi et al. [2024a], an LLM writes questions grounded in the ASR transcripts. These clips are shorter and less knowledge-dense than the synthetic QA conversations, but they add real speaker variation, channel effects, and background noise to the full-duplex training mixture.

Model diversity and prompt speaker pool. The synthetic portions of our training mixture are generated with several text and acoustic models rather than a single generator. Text responses are sampled from a pool of instruction-tuned LLMs, including GPT-OSS-120B,[2] Qwen2.5-72B Instruct,[3] and Llama3.1-70B-Instruct.[4] Speech rendering is performed with multiple TTS backends, including Chatterbox,[5] Magpie-TTS,[6] and Mooncast.[7] For voice conditioning, we build a prompt pool from all 5–10 s speech segments in LibriTTS [Zen et al., 2019], YODAS [Li et al., 2023], and Hifi-TTS [Bakhturina et al., 2021]. This pool contains more than 100K prompt segments and covers over 20K speakers.

Dynamics-aware dialogue reconstruction. Fisher [Cieri et al., 2004] and Seamless-Naturalistic-HQ (Seamless) [Agrawal et al., 2025] provide the natural timing signals used for rhythm-aware training. For Fisher, starting from word-level alignments, we recover utterance segments by merging nearby words while keeping short lexical acknowledgments as standalone backchannels. For

[1] https://huggingface.co/datasets/wikimedia/wikipedia
[2] https://huggingface.co/openai/gpt-oss-120b
[3] https://huggingface.co/Qwen/Qwen2.5-72B-Instruct
[4] https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct
[5] https://github.com/resemble-ai/chatterbox
[6] https://docs.nvidia.com/nemo-framework/user-guide/latest/speech_ai/magpietts.html
[7] https://github.com/jzq2000/MoonCast

Seamless, we start from utterance-level transcripts, remove non-dialogue sessions and low-quality utterance fragments, and merge turns across short gaps before sampling. The processed training set contains 24.6K Fisher samples and 43.1K Seamless samples. Splits are fixed at the conversation or session level before reconstruction, so no evaluation conversation contributes SFT examples, RL windows, reward targets, or timing annotations. Fisher evaluation uses a held-out Fisher test SHAR view, while Seamless evaluation uses a held-out eval-clean view rather than the naturalistic training split. During training, a randomized round-robin sampler balances the two sources, and each sample is truncated at 210 seconds.

Acoustic augmentation. Acoustic robustness is improved with both feature-level and waveform-level augmentation. User speech features are augmented with SpecAugment, and the waveform stream is mixed with a curated set of 10K noise clips from Freesound [Fonseca et al., 2017] and MUSAN [Snyder et al., 2015]. Noise is added with probability 0.5, with the signal-to-noise ratio sampled uniformly between 0 dB and 60 dB.

D Training Details

Implementation. The implementation builds on the released Nemotron-VoiceChat recipe [NVIDIA, 2026] and is trained with the NeMo Toolkit [Kuchaiev et al., 2019]. We select a smaller language backbone Qwen2.5-7B-Instruct [Team, 2024] than Nemotron-VoiceChat. Streaming speech is encoded by a 600M-parameter Parakeet-based encoder with causal convolutional context,[8] followed by a 1024-dimensional Transformer modality adapter that maps acoustic features into the LLM embedding space [Koluguri et al., 2025, Rekesh et al., 2023]. The speech codec and the streaming flow-matching generator follow CosyVoice2 [Du et al., 2024]. The speech encoder and codec remain frozen during training. The speech encoder and LLM advance at 12.5 Hz, whereas the audio codec runs at 25 Hz; each LLM frame therefore predicts two speech tokens to keep the text and audio streams temporally aligned. All runs use 64 A800 80 GB GPUs.

Supervised fine-tuning. SFT is applied after speech-continuation pre-training to adapt the model to assistant-style spoken interaction. The base SFT mixture uses the instruction-following QA data in Appendix C, together with ASR-QA and synthetic-interruption examples that expose the model to real acoustics and basic barge-in handling. *SFT Baseline* uses this mixture and excludes dynamics-aware dialogue data. *SFT Dynamics* keeps the same schedule but adds the reconstructed dynamics-aware dialogue data from Section 4.1, making it the data-matched SFT control for DuplexPO. DuplexPO is initialized from the *SFT Baseline* checkpoint, which also serves as the frozen reference policy for the KL term and as the comparison policy in the suppressed-intent analysis of Appendix F; DuplexPO therefore sees the dynamics-aware dialogue data only through RL on dynamics-critical windows, whereas *SFT Dynamics* sees the same data through supervised training.

Optimization. Pre-training, SFT, and RL use AdamW. For pre-training and SFT, we use $\beta = (0.9, 0.98)$, a weight decay of 0, and an inverse-square-root learning-rate schedule. Pre-training uses a peak learning rate of 5×10^{-4} after a 2,500-step warm-up, while the SFT stages use a peak learning rate of 5×10^{-5}. These stages are trained with bfloat16 mixed precision, and gradients are clipped to a maximum norm of 1.0. Additional RL and reward hyperparameters are summarized in Table 8.

E Definitions of Conversational Dynamics Metrics

All Fisher and Seamless metrics are computed on the evaluation windows of Section 5.1, built with the validation lead and buffer of 2.0 s (Table 8) and conditioned on the teacher-forced history before each window. For a window W_i with reference interval $[g_i, h_i)$ and sampled span $[s_i, e_i)$, the predicted onset $\hat{g}_i$ is the first `<BOS>` frame in $[s_i, e_i)$ and the predicted stop $\hat{h}_i$ is the first subsequent `<EOS>` or `<PAD>` frame, exactly as in Section 3.4. Decoding is truncated at e_i: a stop that would occur after the window end is treated as *not observed*, and an onset after e_i as *not initiated*. Windows are grouped into full-turn and backchannel events by the reference label of Section 3.3, and all rates are macro-averaged over windows (not over conversations). Let $\mathcal{E}$ be the set of windows of a given event type and $\mathcal{I} \subseteq \mathcal{E}$ the subset with a predicted onset.

[8] `https://huggingface.co/nvidia/nemotron-speech-streaming-en-0.6b`

Table 8: DuplexPO RL and reward hyperparameters used for the reported run. Time values are in seconds unless otherwise noted.

Group	Hyperparameter	Value	Role
Windowing	Frame duration Δ	0.08	LLM decision grid.
Windowing	Training lead time L	1.0	Context before annotated agent onset.
Windowing	Training buffer B	2.0	Rollout region after annotated agent offset.
Windowing	Validation lead / buffer	2.0 / 2.0	Wider evaluation context for testing early starts and yielding behavior.
Windowing	Minimum window start	2.0	Windows starting earlier in the conversation are not sampled.
Windowing	Minimum full-turn duration	2.0	Shorter non-backchannel segments are not eligible as full-turn windows.
Windowing	Full-turn windows per conversation	3	Maximum sampled non-backchannel segments.
Windowing	Backchannel windows per conversation	1	Maximum sampled backchannel segments.
Rollout	Samples per window	4	GRPO group size for reward normalization.
Rollout	Temperature / top-p	1.0 / 0.9	Sampling policy for continuations.
Rollout	Maximum rollout steps	200	Maximum generated frames inside a window.
Rollout	Rollout chunk size	128	Chunk size for rollout processing.
Optimization	RL learning rate	1×10^{-5}	AdamW learning rate.
Optimization	Warm-up / minimum LR	100 / 1×10^{-6}	Inverse-square-root schedule.
Optimization	KL coefficient β	0.2	Regularization strength toward the frozen SFT reference policy.
Optimization	Advantage clipping	$[-5, 5]$	Clip range for group-normalized advantages.
Reward	Turn onset widths (early / late)	0.2 / 0.5	Gaussian widths for full-turn onset error.
Reward	Backchannel onset tolerance	0.5	Half-width of the constant-score onset region.
Reward	Maximum backchannel duration	1.5	Duration threshold for positively scored backchannels.
Reward	Missed-event penalty	-0.5	Reference present but no predicted onset.
Reward	False-alarm penalty	-0.5	Predicted onset without a reference event.
Reward	Backchannel overlong penalty	-0.5	Final score for a positively scored backchannel exceeding the duration threshold or lacking a stop.
Reward	Missing-stop scale	0.75	Multiplier for full-turn onset reward when neither <EOS> nor <PAD> is observed after onset.

Initiation rate. $|\mathcal{I}|/|\mathcal{E}|$: the fraction of reference events for which the model starts speaking anywhere inside the window.

Onset MAE. $\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}|\Delta(\hat{g}_i - g_i)|$ in seconds. Non-initiated windows do not contribute (they are counted by the initiation rate instead).

Yield rate (Tables 2, 13, 14). The fraction of initiated windows with an observed stop satisfying the event-specific criterion. For a full turn, the stop must occur within the evaluation window, $\hat{h}_i < e_i$; this measures window-level termination rather than alignment with the reference offset. For a backchannel, the stop must be observed and the predicted duration $\Delta(\hat{h}_i - \hat{g}_i)$ must not exceed 1.5 s. A window without an observed stop counts as not yielded. The denominator is the set of initiated windows $\mathcal{I}$, so yield rate can exceed initiation rate.

Rhythm accuracy (Tables 5 and 6). A frame-level agreement score over both event types. For each frame $t \in [s_i, e_i)$ of every window in $\mathcal{E}_{\text{turn}} \cup \mathcal{E}_{\text{bc}}$, the predicted speaking state (1 if $\hat{g}_i \le t < \hat{h}_i$, else 0) is compared with the reference state (1 if $g_i \le t < h_i$, else 0); rhythm accuracy is the fraction of matching frames, pooled over all frames of all windows. A non-initiated window contributes all-silent predictions, and a window without an observed stop contributes speaking predictions up to e_i.

Interruption yield (Tables 5 and 6). Unlike the yield rates above, which concern stopping at the end of the agent's *own* event, interruption yield concerns stopping because the *user* takes the floor. It is computed only on windows whose event metadata c_i carries the user barge-in flag (Section 3.3), i.e. the user starts a segment at time $u_i < h_i$ while the reference agent segment is still active. Among these, the denominator is the set of windows in which the model itself is speaking at u_i ($\hat{g}_i < u_i$ and no stop observed before u_i); windows in which the model never started, or had already stopped before u_i, are excluded. The model yields if a stop is observed within 0.5 s after u_i; a later stop or no observed stop counts as not yielded. The FDB-v3 Yield Rate (Table 3) follows the benchmark's own definition [Lin et al., 2026a] and is not interchangeable with this metric.

Overlong backchannel (Table 6) and no-stop errors (Section 6.3). Overlong BC is the fraction of initiated backchannel windows whose predicted duration exceeds 1.5 s or whose stop is not observed inside the window. The Fisher no-stop error rate is the analogous fraction of initiated full-turn windows without an observed stop. Both are lower-is-better.

Resumption yield (Section 6.3). The fraction of initiated backchannel windows in which the model has stopped by the time the user resumes speaking, $\hat{h}_i \le u_i^{\text{res}}$, where u_i^{res} is the onset of the first user segment after $\hat{g}_i$. Unlike overlong BC, it depends on the user's resumption time rather than on a fixed duration ceiling, so the two can move in opposite directions.

FDB-v3. Turn-taking latency, VIR, and Yield Rate are computed with the FDB-v3 protocol [Lin et al., 2026a], with the VIR exclusions described in Section 5.1.

F Definition of Suppressed Intent Rate (SIR) and Suppression Release Ratio (SRR)

To quantify a full-duplex agent's ability to *start* speaking when appropriate and to *stop* speaking when the user takes the floor, we introduce two complementary metrics. Both are computed at the frame level on the agent's text head and are defined symmetrically for the two control actions $a^\star \in \{\texttt{<BOS>}, \texttt{<EOS>}\}$.

Let π_{SFT} and π_{RL} denote the *SFT Baseline* and DuplexPO policies. Let $p_t^{\text{SFT}}(a)$ and $p_t^{\text{RL}}(a)$ denote the probability each policy assigns to token a at frame t. To isolate differences attributable to the policy head rather than to trajectory divergence, both quantities are evaluated on a single shared context, namely the SFT-generated trajectory. Let $\hat{y}_t^{\text{SFT}}$ and $\hat{y}_t^{\text{RL}}$ denote the corresponding $\arg\max$ tokens. We say that the SFT policy has an *intent* for action $a^\star$ at frame t when $p_t^{\text{SFT}}(a^\star) > \tau$, with τ a fixed threshold (we use $\tau = 0.1$ throughout).

Table 9: Judge calibration under synthetic timing perturbations on Fisher. The transcript content is identical across candidates; only the agent timestamps are shifted. Candidate order is randomized. Win rates are computed over non-tie judgments. p-values are computed using a two-sided binomial sign test.

Perturbation	# Conv.	Original Wins	Perturbed Wins	Ties	Original Win Rate	p-value
Delayed by $+1.0$s	26	18	8	0	69.2%	0.076
Advanced by -1.0s	26	21	5	0	80.8%	0.002

For each action we restrict attention to a frame set $\mathcal{F}_{a^\star}$ in which the corresponding intent is meaningful. We define $\mathcal{F}_{\texttt{<BOS>}} = \{t \mid \hat{y}_t^{\text{SFT}} = \text{PAD}\}$ as the set of frames at which the SFT model remains silent, where the intent is the impulse to *start* speaking. We define $\mathcal{F}_{\texttt{<EOS>}} = \{t \mid \hat{y}_t^{\text{SFT}} \neq \texttt{<EOS>}$ and the agent is mid-utterance at $t\}$ as the set of frames at which the SFT model continues speaking, where the intent is the impulse to *stop* speaking.

Suppressed Intent Rate (SIR). SIR is the fraction of frames in $\mathcal{F}_{a^\star}$ at which the SFT policy holds a latent intent for $a^\star$ but is overridden by a different action,

$$\text{SIR}_{a^\star} \;=\; \frac{\left|\{t \in \mathcal{F}_{a^\star} \;\mid\; p_t^{\text{SFT}}(a^\star) > \tau\}\right|}{\left|\mathcal{F}_{a^\star}\right|}. \tag{12}$$

A higher $\text{SIR}_{a^\star}$ indicates that the SFT model frequently assigns non-negligible probability to the action without taking it. We use SIR as an operational measure of latent boundary propensity; it does not by itself establish how this propensity is represented internally.

Suppression Release Ratio (SRR). SRR is a paired metric defined on exactly the suppressed-intent frames identified by SIR. Let

$$\mathcal{S}_{a^\star} \;=\; \{t \in \mathcal{F}_{a^\star} \;\mid\; p_t^{\text{SFT}}(a^\star) > \tau\}. \tag{13}$$

On this same frame set, SRR measures how often the RL policy actually takes action $a^\star$,

$$\text{SRR}_{a^\star} \;=\; \frac{\left|\{t \in \mathcal{S}_{a^\star} \;\mid\; \hat{y}_t^{\text{RL}} = a^\star\}\right|}{\left|\mathcal{S}_{a^\star}\right|}. \tag{14}$$

A higher $\text{SRR}_{a^\star}$ indicates that RL takes the action on a large share of the frames where the SFT model already assigned it non-negligible probability. SRR is computed only on $\mathcal{S}_{a^\star}$ and therefore does not by itself exclude additional action-taking outside this set. .

G Judge Calibration with Synthetic Timing Perturbations

To assess whether the pairwise LLM judge is sensitive to conversational timing rather than transcript content alone, we conduct a controlled timing-perturbation sanity check on Fisher. For each conversation, we keep the user timeline and the agent transcript unchanged, but shift the agent timestamps by ± 1.0 seconds. A $+1.0$s shift makes agent responses and backchannels systematically late, while a -1.0s shift makes agent speech more likely to occur prematurely during ongoing user speech. The judge is then asked to compare the original timeline with the perturbed timeline using the same blinded pairwise protocol as in Appendix H. Candidate order is randomized across examples.

Importantly, in this calibration experiment, we remove aggregate dynamics statistics from the judge input. Thus, the judge must rely on the aligned timestamps and transcripts rather than precomputed automatic metrics. Win rates are computed over non-tie judgments, and statistical significance is evaluated with a two-sided binomial sign test.

As shown in Table 9, the judge prefers the original timing over both delayed and advanced perturbations. The preference is statistically significant for advanced perturbations ($p = 0.002$), but not for delayed perturbations at the 0.05 level ($p = 0.076$). Because transcript content is identical across candidates, these results demonstrate sensitivity to timing, with stronger evidence for detecting premature turn entry than comparable response delays.

H LLM Pairwise Judge for Full-Duplex Dialogue Naturalness

This section describes the matched-window evaluation using fixed candidate assignments and aggregate rhythm statistics. The separate evaluation on the 96 human-rated pairs uses randomized candidate order and excludes aggregate dynamics statistics, as described in Appendix A.

The pairwise judge is designed to evaluate conversational dynamics rather than semantic answer quality. Table 11 gives the system instruction, and Table 12 gives the user-prompt template used to present aligned candidate timelines. We used Gemini 3.0 Pro and counted a result only when the returned JSON matched the requested schema and contained no error. Candidate labels were anonymized in the prompt, and in these runs candidate A is the *SFT Baseline* and candidate B is DuplexPO. Table 10 reports both overall and dimension-level preferences. Win rates are computed over non-tie judgments, and the p-values use a two-sided binomial sign test over the non-tie preferences.

Table 10: Conversation-level LLM pairwise judge results. Candidate B is DuplexPO; candidate A is the *SFT Baseline*. Win rates exclude ties.

Dataset	Criterion	Total	B wins	A wins	Ties	B win rate	p-value
Fisher	Overall	26	20	6	0	76.9	0.009
Fisher	Turn-taking	26	20	6	0	76.9	0.009
Fisher	Backchannel	26	22	2	2	91.7	$< 10^{-4}$
Fisher	Barge-in handling	26	21	3	2	87.5	$< 10^{-3}$
Seamless	Overall	383	264	117	2	69.3	$< 10^{-13}$
Seamless	Turn-taking	383	250	123	10	67.0	$< 10^{-10}$
Seamless	Backchannel	383	261	54	68	82.9	$< 10^{-32}$
Seamless	Barge-in handling	383	237	70	76	77.2	$< 10^{-21}$

Table 11: System prompt for LLM pairwise judging of full-duplex dialogue dynamics.

System prompt

```
You are an expert evaluator for full-duplex spoken dialogue rhythm.

Your role:
- Compare two candidate system turns from the same local dialogue context.
- Evaluate ONLY three interaction abilities: turn-taking, backchanneling,
  and user barge-in handling.
- Use only the observable local dialogue context and candidate turns
  provided in the prompt. Do NOT assume access to any hidden annotations,
  target response, target timing interval, or dataset label.
- Do NOT evaluate factual correctness, knowledge, helpfulness,
  informativeness, or whether the answer fully solves the user's question.
- Do NOT prefer a candidate merely because it is more detailed, more
  informative, or more semantically complete.
- If both candidates are similar on these three interaction abilities,
  output "tie".

Evaluation dimensions:
1. turn_taking: Which candidate starts at a more natural time and yields
   smoothly for the local conversation state. Do not reward a candidate
   for being longer or more substantive.
2. backchannel: Which candidate better uses short listener feedback when
   appropriate, without being stale, too long, or floor-grabbing.
3. user_barge_in_handling: Which candidate better avoids talking over
   the user or better yields/gets out of the way when the user starts or
   continues speaking.

Important judgment rules:
- Judge from timing and local dialogue behaviour, not answer quality.
- A short response can be better if it functions as a natural backchannel.
- A longer response should not be preferred just because it holds the
  floor or gives more content.
- Penalise competitive interruption, awkward delayed feedback, excessive
  overlap with the user, and failure to yield when the user appears to
  take the floor.
- If one dimension is hard to determine from the visible context and
  candidate timestamps, mark that dimension as "tie".

You MUST:
1. Read the shared local context first.
2. Infer the local turn-taking state: user_holding_floor,
   user_yielding_floor, backchannel_opportunity, full_turn_opportunity,
   or uncertain.
3. Compare candidate A and candidate B on the three dimensions.
4. State preferred_candidate as "A", "B", or "tie".
5. Give concise reasons focused only on turn-taking, backchanneling,
```

```
   and user barge-in handling.
6. Give confidence from 0 to 1.

Output MUST be valid JSON matching this schema, with no markdown and no
extra text:
{
  "interaction_id": "string",
  "pair_id": "string",
  "candidate_a_id": "string",
  "candidate_b_id": "string",
  "inferred_turn_state": "user_holding_floor|user_yielding_floor|
                          backchannel_opportunity|full_turn_opportunity|
                          uncertain",
  "preferred_candidate": "A|B|tie",
  "preference_reason": "string",
  "dimension_preferences": {
    "turn_taking":            {"preferred_candidate": "A|B|tie",
                               "reason": "string"},
    "backchannel":            {"preferred_candidate": "A|B|tie",
                               "reason": "string"},
    "user_barge_in_handling": {"preferred_candidate": "A|B|tie",
                               "reason": "string"}
  },
  "major_rhythm_issues": ["string"],
  "content_quality_ignored_note": "string",
  "confidence": 0.0-1.0
}
```

Table 12: User-prompt template for conversation-level pairwise dynamics judging.

User prompt template

```
Compare candidate A and candidate B for overall full-duplex
spoken-dialogue rhythm across this conversation sample. Ignore factual
correctness, answer completeness, and informativeness. Output only valid
JSON matching the schema.

## Interaction ID
{interaction_id}

## Shared user timeline
- [user] ({user_turn_1_start_s:.2f}s--{user_turn_1_end_s:.2f}s)
  {user_turn_1_text}
- [user] ({user_turn_2_start_s:.2f}s--{user_turn_2_end_s:.2f}s)
  {user_turn_2_text}
...

## Candidate A assistant timeline
- [candidate] ({candidate_a_turn_1_start_s:.2f}s--
               {candidate_a_turn_1_end_s:.2f}s)
  {candidate_a_turn_1_text}
- [candidate] ({candidate_a_turn_2_start_s:.2f}s--
               {candidate_a_turn_2_end_s:.2f}s)
  {candidate_a_turn_2_text}
...

## Candidate A observable rhythm summary
- n_turns:                      {candidate_a_n_turns}
- total_speech_ms:              {candidate_a_total_speech_ms}
- total_user_overlap_ms:        {candidate_a_total_user_overlap_ms}
- n_user_overlaps:              {candidate_a_n_user_overlaps}
- n_interrupted_by_user:        {candidate_a_n_interrupted_by_user}
- median_pause_after_user_ms:   {candidate_a_median_pause_after_user_ms}

## Candidate B assistant timeline
- [candidate] ({candidate_b_turn_1_start_s:.2f}s--
               {candidate_b_turn_1_end_s:.2f}s)
  {candidate_b_turn_1_text}
- [candidate] ({candidate_b_turn_2_start_s:.2f}s--
               {candidate_b_turn_2_end_s:.2f}s)
  {candidate_b_turn_2_text}
...

## Candidate B observable rhythm summary
- n_turns:                      {candidate_b_n_turns}
- total_speech_ms:              {candidate_b_total_speech_ms}
- total_user_overlap_ms:        {candidate_b_total_user_overlap_ms}
- n_user_overlaps:              {candidate_b_n_user_overlaps}
- n_interrupted_by_user:        {candidate_b_n_interrupted_by_user}
- median_pause_after_user_ms:   {candidate_b_median_pause_after_user_ms}
```

I Ablation Study: Neural Reward Model

The neural reward model teacher is SoulX-Duplug [Yan et al., 2026b], denoted f_ϕ. This method does not use assistant ground-truth. For each training sample, the user waveform u and its aligned word

sequence $\{(w_j, t_j^s, t_j^e, t_j^{emit})\}_j$ are fed to f_ϕ, producing a chunk-level teacher timeline

$$\mathcal{T} = \{(a_m, b_m, s_m)\}_{m=1}^{M}, \qquad s_m \in \{\texttt{idle}, \texttt{nonidle}, \texttt{speak}, \texttt{blank}\}. \tag{15}$$

The neural reward model first converts the SoulX-Duplug timeline into a per-frame state: frame t receives the state s_m of the chunk $[a_m, b_m)$ that contains it, and frames covered by no chunk are `blank`. Agent decisions are binary, $d_t \in \{\texttt{speak}, \texttt{silent}\}$, and are obtained from a token sequence z by the same decoding as FCDR: $g(z)$ marks a frame `speak` from the first `<BOS>` up to (excluding) the first subsequent `<EOS>` or `<PAD>`, and `silent` otherwise. The transition type $c_i \in \{\text{yield}, \text{interrupt}\}$ is taken from the window metadata of Section 3.3: $c_i =$ interrupt for windows flagged as containing a user barge-in and $c_i =$ yield otherwise. A prescription prior $p_T(d \mid s, c)$ scores whether decision d_t is appropriate under teacher state s_t and transition type c_i. It is a fixed table over the three non-blank teacher states and the two transition types, with $p_T(\texttt{silent} \mid s, c) = 1 - p_T(\texttt{speak} \mid s, c)$: speaking is prescribed when the teacher state is `speak`, silence when it is `idle`, `nonidle` lies in between, and the `interrupt` column lowers the speaking prior so that continuing to talk over a user barge-in is penalized. For a predicted decision sequence $d = g(z)$ in window W_i with $n_i = e_i - s_i$ frames, we compute the salience-weighted teacher score

$$A_i(d) = \frac{\sum_{t=1}^{n_i} w_{i,t} \log p_T(d_t \mid s_{i,t}, c_i)}{\sum_{t=1}^{n_i} w_{i,t}}, \tag{16}$$

where blank frames have $w_{i,t} = 0$, frames within $\sim$160 ms (two frames) of a SoulX-Duplug state transition are weighted 1.0 and other non-blank frames 0.5. Let $z_{i,k}$ be the token sequence of the k-th sampled continuation of the policy in W_i ($k = 1, \ldots, 4$, the GRPO group), and let $z_{i,l}^{\text{ref}}$, $l = 1, \ldots, K_0$, be $K_0 = 4$ continuations of the same window sampled from the frozen SFT reference policy with the same decoding settings; they only serve as a baseline and receive no gradient. The reward subtracts the per-window reference baseline:

$$r_{i,k}^{\text{core}} = A_i(g(z_{i,k})) - \frac{1}{K_0} \sum_{l=1}^{K_0} A_i(g(z_{i,l}^{\text{ref}})). \tag{17}$$

$r_{i,k}^{\text{core}}$ is the final NRM reward, with no additional terms; it replaces the FCDR reward $R_{i,k}$ in the group-normalized advantage in Eq. (10), with all other components (window sampler, group size, KL, optimizer) unchanged.

Table 13: Comparison with SFT controls and reward calculation ablation. SFT Baseline and SFT Dynamics are non-RL controls without reward optimization. NRM and FCDR use the same dynamics-critical window sampler and GRPO-style optimization, and differ only in the reward calculation method. Rates are in %; Onset MAE is in seconds. Best values in each column are bolded.

Dataset	Training setting	Onset MAE (↓)	Turn-taking		Backchannel		Barge-in	
			Init (↑)	Yield (↑)	Init (↑)	Yield (↑)	VIR (↓)	Yield (↑)
Fisher	*SFT Baseline*	0.98	97.8	92.1	95.7	57.1	–	–
	SFT Dynamics	1.14	97.1	78.9	91.4	57.1	–	–
	NRM reward	1.22	96.4	81.6	89.2	50.0	–	–
	FCDR reward	**0.69**	**100.0**	**98.7**	**97.8**	**100.0**	–	–
Seamless	*SFT Baseline*	1.22	91.8	78.5	92.2	79.4	–	–
	SFT Dynamics	1.34	79.7	63.0	83.2	63.4	–	–
	NRM reward	1.32	84.5	67.6	93.5	74.0	–	–
	FCDR reward	**1.03**	**98.0**	**93.6**	**99.5**	**93.3**	–	–
FDB-v3	*SFT Baseline*	–	–	–	–	–	8.0	64.8
	SFT Dynamics	–	–	–	–	–	**4.0**	93.3
	NRM reward	–	–	–	–	–	**4.0**	93.3
	FCDR reward	–	–	–	–	–	5.0	**100.0**

Table 13 shows that FCDR outperforms the evaluated NRM variant on most dynamics metrics, while NRM achieves lower VIR on FDB-v3. NRM improves over SFT Dynamics on several metrics but does not consistently surpass SFT Baseline. These results favor the explicit event-based reward in this setting; they do not establish a general advantage over neural reward models. The comparison changes the source and formulation of the reward together, so it does not isolate which difference accounts for the observed gains.

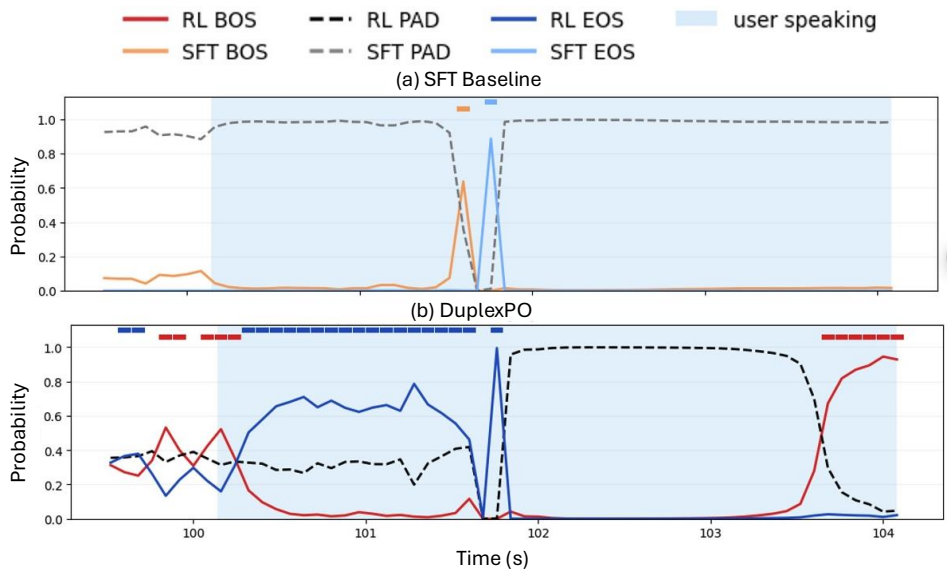


Figure 4: Example token-level conversational pattern. Compared with the *SFT baseline*, DuplexPO maintains and releases <BOS>/<EOS> impulses around user speech instead of staying near PAD.

J Ablation Study: Optimization Methods

Holding the FCDR-based reward calculation and the dynamics-critical window sampler fixed, we ablate the preference-optimization objective by comparing GRPO with a DPO-style variant. In the DPO-style variant, the highest- and lowest-reward continuations within each sampled window are used as the preferred and dispreferred pair. Table 14 reports the resulting dynamics metrics on Fisher, Seamless, and FDB-v3.

Table 14: Optimization method ablation. Both variants use the same FCDR and dynamics-critical window sampler, and differ only in the preference-optimization objective. Rates are reported in %; Onset MAE is reported in seconds. The best result in each column is bolded.

Dataset	Optimization	Onset MAE (↓)	Turn-taking		Backchannel		Barge-in	
			Init (↑)	Yield (↑)	Init (↑)	Yield (↑)	VIR (↓)	Yield (↑)
Fisher	DPO	0.81	**100.0**	93.6	91.4	71.4	–	–
	GRPO	**0.69**	**100.0**	**98.7**	**97.8**	**100.0**	–	–
Seamless	DPO	**1.00**	95.1	77.8	95.8	84.7	–	–
	GRPO	1.03	**98.0**	**93.6**	**99.5**	**93.3**	–	–
FDB-v3	DPO	–	–	–	–	–	**4.0**	93.3
	GRPO	–	–	–	–	–	5.0	**100.0**

GRPO achieves higher yield rates than the DPO-style variant across the evaluated datasets and improves most reported dynamics metrics. DPO achieves slightly lower onset MAE on Seamless and lower VIR on FDB-v3. These results support the selected GRPO-style objective in this setting without establishing the mechanism behind its advantage.

K Example of a Conversational Pattern

Figure 4 contrasts the token-level conversational pattern of DuplexPO with that of the *SFT baseline*. DuplexPO maintains and releases <BOS>/<EOS> impulses around user speech instead of remaining near PAD, providing a qualitative illustration alongside the shared-context diagnostics in Appendix F.

L Broader Impacts

This work improves the naturalness of full-duplex spoken dialogue models. Positively, this can significantly enhance accessible interfaces and conversational assistants. Negatively, highly human-like voice dynamics (e.g., natural backchanneling and interruption) could be misused for deceptive practices like realistic vishing (voice phishing). Future deployments should consider appropriate safeguards, such as watermarking or identity disclosure, to mitigate these risks.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction describe timing-focused policy optimization and the evaluated dynamics and intelligence results. Section 6.6 specifies the scope of the long-horizon and transfer evidence.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6.6 discusses heuristic rewards, variability in human timing, fixed-history evaluation, autonomous interaction, and asymmetric cross-corpus transfer.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper is empirical and does not present theoretical results requiring proofs.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the DuplexPO algorithm, data sources, preprocessing procedures, training hyperparameters, and evaluation protocols. The accompanying public repository provides the data-processing pipeline and DuplexPO implementation.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The public repository linked in the introduction provides the data-processing pipeline and DuplexPO algorithm. The paper identifies the source datasets and reports the preprocessing procedures and training hyperparameters.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 and Appendices C and D describe the data construction, held-out evaluation separation, optimizer, and principal training hyperparameters. Reward and window ablations are described in Section 6.3.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix A reports the human preference counts, a Wilson confidence interval, and an exact binomial test. Table 7 reports mean and standard deviation across three independent RL training seeds. Table 10 reports LLM-judge preference tests.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix D reports sufficient information on the computer resources.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors affirm that the research conforms to the NeurIPS Code of Ethics. The paper credits the research assets used and discusses potential misuse and evaluation limitations.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix L discusses benefits for accessible interfaces and conversational assistants, as well as risks of deceptive voice applications.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The accompanying release consists of algorithm and preprocessing code; it does not include model weights or redistributed source speech corpora. Potential misuse of full-duplex voice systems is discussed in the broader impacts section.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original sources of the datasets, models, and software used in this work. Their licenses and applicable terms of use are documented in the supplementary material and respected in our use of these assets.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The public repository linked in the introduction documents the released FCDR implementation and data-processing pipeline. This statement concerns the released code, not unreleased model checkpoints or source speech corpora.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Appendix A reports the instructions and evaluation criteria for the blinded A/B listening study, as well as the evaluators' affiliation and compensation status.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The listening study was limited to evaluating model-generated conversational outputs. The evaluation did not solicit sensitive personal information or require participants to disclose personal experiences.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper describes the language-model backbone, synthetic data generation, and use of Gemini for complementary pairwise evaluation. The matched-window judge protocol is documented in Appendix H.