DUPLEXTTS: A SCALABLE FRAMEWORK FOR FULL-DUPLEX CONVERSATIONAL SPEECH SYNTHESIS AND EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The scarcity of full-duplex conversational speech data limits the development of natural spoken interaction models. Conventional text-to-speech systems primarily generate isolated utterances or alternating turns, leaving the joint generation of conversational content and interaction timing underexplored. Motivated by advances in native full-duplex speech generation, we introduce DuplexTTS, a scalable framework for synthesizing and evaluating full-duplex conversational speech. DuplexTTS supports two synthesis routes: native autonomous generation, where a single speech model generates two role-conditioned streams through shared acoustic feedback, and rule-based orchestration, where scripted utterances are synthesized and arranged into turn exchanges, backchannels, interruptions, and side talk. A unified pipeline connects synthesis with dialogue-level quality screening, additional source-utterance checks for rule synthesis, interaction analysis, and downstream evaluation. Using this framework, we construct over 1,800 hours of screened Chinese and English dialogue. Human listeners rate the generated dialogues favorably for naturalness and meaningfulness. Training Moshi on the two corpora yields different conversational behaviors. The mixed-data model achieves 16.60 percentage points higher macro expected-behavior accuracy on Full-Duplex-Bench than the original model. These results establish DuplexTTS as a practical framework for expanding full-duplex training data and studying how synthesis and curation shape spoken interaction. Audio demos are available at `https://duplextts.github.io`.

1 INTRODUCTION

Natural conversation requires knowing when to speak, acknowledge a partner, yield to an interruption, and continue through unrelated speech. These choices are expressed through pauses and overlaps as well as words (Stivers et al., 2009). Training full-duplex speech systems therefore requires conversations that capture this behavior across varied situations and roles. Natural corpora such as Fisher provide real two-channel exchanges (Cieri et al., 2004), while the supply of scenario-conditioned full-duplex training data remains limited (Guo et al., 2026). Synthesis offers a way to expand that supply with control over the participants and interactions.

Conventional TTS converts supplied text into speech. Dialogue TTS extends this capability to multiple speakers and longer exchanges (Zhang et al., 2024; Peng et al., 2026; Chen et al., 2026). Generating full-duplex training data also calls for control over how an exchange develops and how its interaction patterns enter the audio. dGSLM showed that a speech model can directly generate synchronized dialogue streams and learn their turn-taking patterns (Nguyen et al., 2023). This motivates our central question: how can native dialogue generation and explicit event composition be developed into a scalable source of full-duplex training data?

We present DuplexTTS, a synthesis and screening pipeline with two routes. The native route adapts a single PersonaPlex model (Roy et al., 2026) to generate both sides of a conversation. Its temporal backbone and acoustic decoder are shared across two role-conditioned states. We train the backbone using rewards on complete autonomous dialogues, then distill the acoustic decoder with the backbone fixed. Given a scenario and role descriptions, the resulting model develops words and

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

speech timing through shared acoustic feedback, producing separate synchronized tracks. The rule route starts from a structured script, synthesizes and aligns each utterance, then uses text anchors and event-specific rules to arrange backchannels, interruptions, and side talk. These routes supply autonomous interaction and explicit event control within one corpus construction process.

DuplexTTS screens the generated speech for quality and records the interaction patterns in the resulting conversations. We assess the synthesized conversations through interaction analysis and human listening, and evaluate their usefulness for training a full-duplex speech model.

Our contributions are threefold:

1. We develop two synthesis routes: a single speech model adapted on complete autonomous dialogues to generate both participants, and an alignment-based composition method that controls conversational events.
2. We construct 107,072 dialogues totaling 2,189.1 hours in Chinese and English. Final-audio screening, with additional source-utterance checks for the rule route, retains 92,940 dialogues and 1,835.3 hours, accompanied by quality decisions and interaction statistics.
3. We connect data construction to interaction quality and training utility through full-corpus analysis, human listening, and full-parameter Moshi training. The mixed-data model improves macro expected-behavior accuracy on Full-Duplex-Bench by 16.60 percentage points over the original model.

2 Related work

Dialogue text-to-speech. CoVoMix converts dialogue text into speaker-specific semantic streams and synthesizes a mixture containing overlapping speech (Zhang et al., 2024). VibeVoice generates long-form multi-speaker speech with next-token diffusion (Peng et al., 2026). Text-Audiobox learns text–speech alignment through cross-attention and generates stereo full-duplex dialogue conditioned on a transcript (Chen et al., 2026). These models advance the realization of supplied dialogue text. DuplexTTS studies data construction through both scripted event composition and native generation of future words during acoustic interaction.

Generative spoken dialogue. dGSLM models two speech channels with cross-attending Transformer towers (Nguyen et al., 2023); NTPP uses next-token-pair prediction with decoder-only models (Wang et al., 2025). Moshi combines parallel speech streams with text-guided hierarchical codec generation and describes multi-stream TTS (Défossez et al., 2024). PersonaPlex adds text role prompts and audio voice prompts (Roy et al., 2026). DuplexTTS builds its native synthesis route on PersonaPlex, using scenario and role conditions to drive autonomous interaction while retaining the directly generated audio.

Interaction alignment. Multi-Faceted Interactivity Alignment applies group relative policy optimization (GRPO) to text tokens, using recorded interlocutors and rewards for interaction and response quality (Ohashi et al., 2026). Our method learns from complete autonomous dialogues and adapts the temporal/text backbone and acoustic modules in sequence.

Dialogue synthesis. Agentic-DuplexGen supplies dialogue content in advance, lets full-duplex models determine timing through online interaction, and re-renders the result with TTS (Wang et al., 2026). DuplexGen calibrates scenario-dependent turn-taking synthesis using human preferences and evaluates a model trained on its generated data; its speech construction uses TTS, word alignment, and scheduled overlap (Kim et al., 2026). DuplexDrama combines persona and scenario scripts, expressive TTS, overlap assembly, sound events, and quality validation to construct bilingual dialogue data (Guo et al., 2026). DuplexTTS studies shared-model autonomous generation alongside alignment-based event composition, connecting these construction choices to measured interaction and downstream model behavior.

Acoustic assessment. SCOREQ learns speech-quality representations through contrastive regression and is evaluated against human mean opinion score (MOS) datasets (Ragano et al., 2024). We

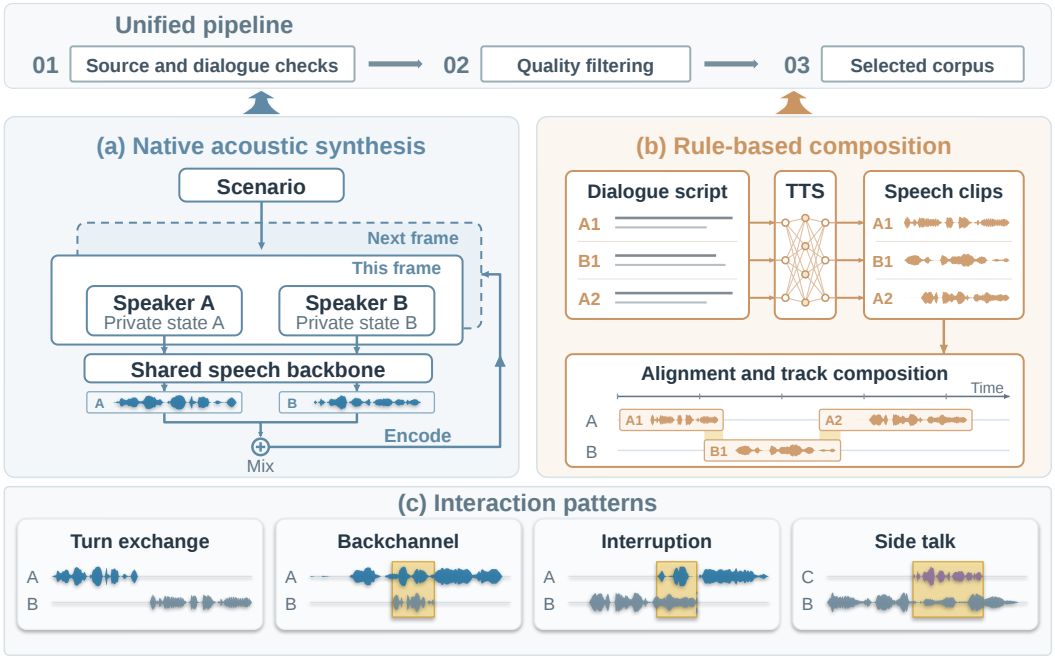


Figure 1: DuplexTTS generates synchronized dialogue through two routes. (a) One speech model generates both participants using shared backbone and acoustic-decoder weights. A and B retain separate role states and receive shared acoustic feedback. (b) A script supplies utterances and events; alignment maps text anchors to a schedule. Both routes feed a common screening and export process. (c) Schematic waveforms illustrate turn exchange, backchannel, interruption, and side talk. Rule output places A and C on the left channel and B on the right.

use its no-reference predictions alongside speech activity, transcription, and repetition measurements to screen dialogues and analyze how selection changes their pauses and overlaps.

3 DuplexTTS: Synthesis and Screening

Figure 1 summarizes DuplexTTS: native acoustic interaction and rule-based event placement produce synchronized dialogue tracks. Both routes screen final audio; the rule route also checks source utterances. The pipeline retains audio, generation conditions, and quality decisions for training preparation.

3.1 A Shared Model for Native Dialogue Synthesis

We use one speech model $p_{\theta,\phi}$ for both participants, where θ denotes the temporal/text backbone and ϕ the acoustic Depth modules, which predict audio tokens across codebooks within each frame. The model maintains two role-conditioned states. Each state has its own role profile, voice prompt, and causal history H_t^i, for $i \in \{A, B\}$, while both receive the shared acoustic observation u_t. At each frame, the same model predicts each role's text action y_t^i and eight own-audio codebook tokens c_t^i:

$$(y_t^i, c_t^i) \sim p_{\theta,\phi}(\cdot \mid H_t^i, u_t), \qquad i \in \{A, B\}. \tag{1}$$

Role and voice prompts distinguish the participants; both use the same θ and ϕ throughout training and synthesis. Their actions are sampled separately from their respective histories on a synchronized frame clock. Future words and turn boundaries develop during generation.

For completed decoded chunks x_t^A, x_t^B, the next public observation is

$$u_{t+1} = E_{\text{Mimi}}\big(0.5(x_t^A + x_t^B)\big)\,. \tag{2}$$

Here E_{Mimi} is the streaming Mimi audio encoder. Its state is retained, and identical encoded observations are delivered to both roles. Concatenating each role's chunks gives its output track. Generation uses a 12.5-Hz frame clock and 24 kHz floating-point waveforms. Codec and model delays determine the offset between text decisions and synthesized audio.

The shared frame clock lets either role speak while the other is speaking. Feeding back the completed mixture makes the resulting overlap part of both roles' next observation, while separate output tracks preserve speaker identity. Appendix A.4 examines how removing speech from this feedback changes subsequent dialogue. In a private-fact test, one role receives a fact and the other is asked to confirm it. With normal feedback, the receiver mentions the target without its alternative in eleven of sixteen inputs; removing the informed role's audio reduces this to two. Opposite facts then produce identical receiver waveforms in all eight paired tests. This verifies that the shared observation carries information between the role-conditioned states.

Dialogue-level backbone training. We adapt the PersonaPlex backbone on complete generated dialogues, sampling a group of candidates for each scenario. A large language model (LLM) judges timestamped transcripts of the generated audio on scenario adherence, role adherence, coherence, progression, responsiveness, and timing. We average the six scores and standardize rewards within each group. Both role trajectories receive the same dialogue-level advantage, so their updates are guided by the quality of the complete exchange. Scoring audio transcripts ties the reward to what was actually spoken. Clipped text-token GRPO (Shao et al., 2024) updates the full temporal/text backbone, with reference-model Kullback–Leibler (KL) constraints on text and acoustic predictions. Acoustic modules and Mimi remain fixed.

Acoustic adaptation with the backbone fixed. The backbone also conditions acoustic decoding, so changing it changes the representations from which speech is rendered. We freeze the adapted backbone and train the acoustic Depth modules to match the original model's predictions across all output codebooks. Teacher and student receive the same sampled histories and text/codebook prefixes, with their respective backbone states. We reuse the generated trajectories and select the acoustic checkpoint by KL on a scenario-disjoint development split. The resulting shared backbone and acoustic decoder generate both participants; the judge and distillation teacher are used during training. Appendix A.1 gives the objectives, training configurations, and checkpoint-selection procedure.

Effect of the two training stages. Table 1 compares the original, backbone-adapted, and acoustically distilled models on twenty development scenarios, using the same roles, voices, seeds, and generation settings. Backbone adaptation improves coherence, responsiveness, and timing while reducing SCOREQ from 3.831 to 3.571. Acoustic distillation restores it to 3.765 and retains gains in those three dialogue dimensions over the original. Scenario and role adherence do not improve. The two stages thus adjust dialogue behavior and acoustic rendering in sequence. Appendix A.3 details the evaluation.

Table 1: Sequential adaptation on twenty shared development scenarios. Dialogue scores use 1–5 scales; SCOREQ averages per-track medians. Acoustic distillation yields the final native synthesis model.

Model	SCOREQ	Scene	Role	Coherence	Progress	Response	Timing
Original	3.831	3.40	2.45	3.20	3.05	3.35	2.40
Backbone adaptation	3.571	3.30	2.40	3.90	3.00	4.10	3.60
Acoustic distillation	3.765	3.05	2.20	3.45	3.00	3.70	3.10

3.2 Rule-based synthesis from aligned utterances

The rule route constructs speech interactions from topic-conditioned scripts. Topics span everyday life, social and cultural activities, and science and technology, including family, work, education, travel, health, finance, and the arts. For a chosen topic and language, an LLM authors a coherent script of 10–30 utterances, organized into 5–15 two-utterance groups and covering the four interaction patterns in Figure 1. A and B sustain the main exchange; C contributes removable background

side talk. Each non-normal group specifies a *text anchor*: a unique contiguous span of the source utterance used to locate the insertion. Structural checks enforce event coverage, speaker fields, group counts, and anchor occurrence. The script provides separate control over content, interaction type, and insertion location. The rule corpus covers Chinese and English.

Each utterance is synthesized independently using local Fun-CosyVoice3-0.5B (CosyVoice 3; Du et al., 2025). Reference voices are held fixed within a dialogue; Chinese synthesis uses one reference per role, while English synthesis samples references from voice pools. Qwen3-ForcedAligner-0.6B (Shi et al., 2026) aligns the supplied text to each utterance. Text anchors locate backchannels and interruptions; separate placement rules arrange background speech. TTS realizes the scripted utterances and voices, while clip composition determines overlap and stopping.

For an interruption, the host continues briefly after the insertion and then fades out. English source clips are trimmed to retain at most 0.2 seconds before the first aligned word; the event schedule uses these trimmed clips. The output mixes A and the background participant C on the left channel and places B on the right. Each event remains linked to its script, source clips, word alignments, and final placement. Appendix B.1 specifies the event rules and language-specific voice and background-placement settings.

3.3 Preparing dialogue supervision

Both routes retain synchronized audio, channel identities, and quality decisions. Downstream training requires matching text to the speech present in this audio. For native dialogue, we transcribe detected speech with Whisper and align the transcripts used for training. Rule synthesis retains the script, source clips, word alignments, and event placements. The English training export maps source words onto the composed timeline, accounting for leading-silence trimming. Words crossing an interruption cutoff or an ambiguous same-channel mixture are excluded from supervision. Appendix B.2 gives the projection and word-selection procedure.

3.4 Selecting complete and clear dialogues

Screening addresses two questions: does the dialogue contain usable speech, and does it preserve the utterances required by its construction? We inspect each final channel for sufficient speech, transcription coverage, excessive repetition, and predicted acoustic quality. Voice activity detection (VAD) locates speech, automatic speech recognition (ASR) provides its transcript, and SCOREQ estimates audio quality. A dialogue is retained when both channels pass these checks; identical participant tracks are rejected.

Rule synthesis also checks each source utterance before accepting the assembled dialogue. This matters because a missing or silent short reply can be hidden by the other speech in a channel. We therefore require every expected utterance to be present and contain signal, as well as requiring the final dialogue to pass the audio checks. Each record stores the measurements and rejection reasons, enabling us to analyze what screening removes and how the selected corpus changes. Appendix C.1 gives the thresholds and language-specific settings.

4 Corpus construction and screening

4.1 Corpus size and language coverage

DuplexTTS yields a screened corpus of 92,940 dialogues and 1,835.3 hours. Native synthesis contributes English conversations generated from scenarios and roles; rule synthesis contributes Chinese and English conversations with scripted interaction events. Table 2 summarizes the generated and selected data by language and route. We count each dialogue's shared timeline once, using one canonical rendering per source dialogue.

The native corpus uses 208 scenarios, 32 roles, and 32 voices, with a fixed model and generation configuration for all one-minute dialogues. The rule corpus covers 73 topics in each language. Screening retains 98.31% of native dialogues, 74.11% of English rule dialogues, and 68.66% of Chinese rule dialogues. These selected sets provide the basis for the interaction analysis in Section 5.

Table 2: Corpus size and screening by language and synthesis route, counting one canonical rendering per source dialogue. Hours measure the shared timeline once. Retention is the fraction of dialogues passing each route's screen.

		Generated		Selected		
Language	Route	Dialogues	Hours	Dialogues	Hours	Kept (%)
English	Native	60,000	1,000.0	58,985	983.1	98.31
Chinese	Rule	17,073	366.0	11,723	248.1	68.66
English	Rule	29,999	823.1	22,232	604.2	74.11
Total		107,072	2,189.1	92,940	1,835.3	86.80

Independent dialogue requests allow synthesis to scale across GPUs. In a completed run, eight A100 GPUs generated and structurally validated 50 hours of native dialogue in 3.96 hours, producing 12.63 audio hours per wall-clock hour, including waveform writing. ASR and quality screening run afterward. This throughput supports corpus expansion with a fixed model. Appendix A.1 gives the configuration and timing scope.

4.2 Checking every utterance improves dialogue completeness

An assembled dialogue can contain clear, fluent speech even when one of its short utterances is missing. Across all 47,072 rule dialogues, 12,977 pass the final-channel screen despite containing a missing or near-silent source clip. Checking each expected utterance catches these failures. Conversely, 118 dialogues contain usable source clips but fail the final-channel screen. Requiring both checks retains 33,955 rule dialogues (852.2 hours). The two checks thus address utterance completeness and final audio quality at their respective levels. Appendix Table 9 gives the complete breakdown; Appendix B.2 describes how retained speech is converted into training labels.

5 Corpus quality and interaction analysis

5.1 Interaction in the final corpus

We analyze every dialogue in the final screened corpus, grouping results by language and synthesis route. The two routes are developed and run independently; their distributions describe the data each contributes to the corpus.

Following the event definitions of dGSLM (Nguyen et al., 2023), we join same-channel speech intervals separated by at most 200 ms into inter-pausal units (IPUs). We measure speech segments, shared silences bounded by activity on the same channel (pauses) or different channels (gaps), and simultaneous channel activity (overlap). Native channels represent A and B; rule channels represent A+C and B, so rule statistics describe the physical channel activity. Figure 2 summarizes every dialogue in the final corpus. Event frequencies divide counts by total dialogue minutes; dots and labels show event-duration medians, and lines show the middle 50% of durations.

Native English contains 25.31 speech segments and 4.75 overlaps per minute; rule English contains 18.95 and 3.84, and rule Chinese contains 17.97 and 4.78. Median speech-segment durations are 1.60, 2.68, and 3.10 seconds, respectively, while median overlap durations are 0.92, 0.54, and 0.54 seconds. The native corpus thus combines more frequent, shorter speech segments with longer typical overlaps. The two rule corpora provide distinct English and Chinese interaction profiles. Appendix C.3 gives event definitions and additional screening statistics.

5.2 Screening and event realization

Screening preserves all 208 native scenarios and all 73 rule topics in each language. Acoustic quality and repetition account for native rejections, while mean overlap changes little after selection. Among retained native dialogues, 29.29% contain an internal shared silence of at least five seconds. A full-corpus threshold sweep quantifies how stricter acoustic selection changes retention and temporal composition (Appendix C.1).

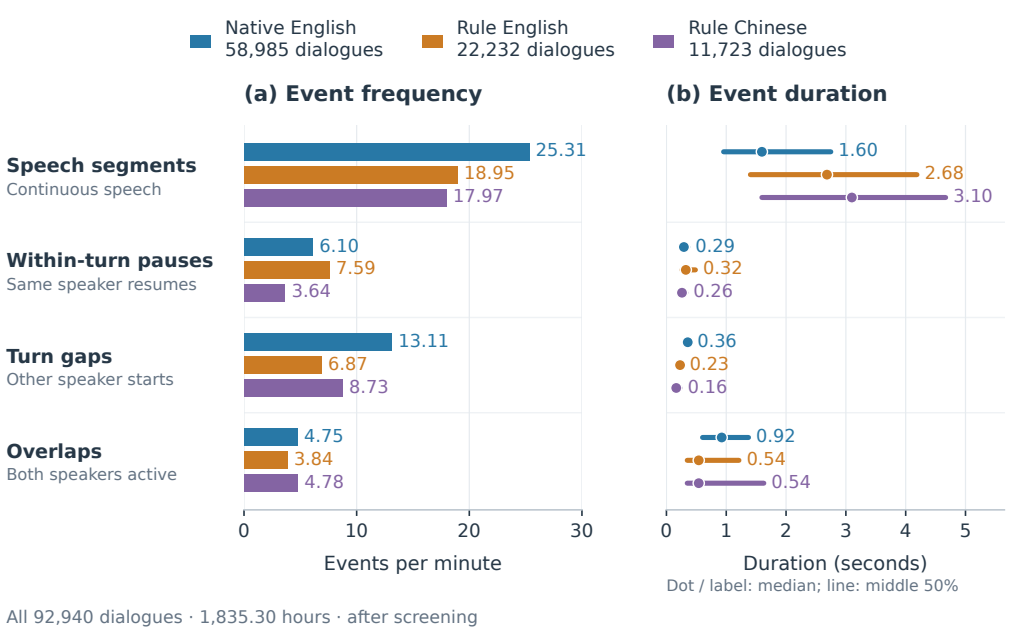


Figure 2: Interaction frequency and duration across all 92,940 dialogues in the final screened corpus (1,835.3 hours). (a) Event counts divided by total dialogue minutes. (b) Event-duration medians (dots and labels) and 25th–75th percentiles (lines). Groups represent language and synthesis route; rule channels contain A+C and B.

Extending generation also changes the amount of actual conversation. Across eight five-minute dialogues, mean speech-active time falls from 45.00 seconds in the first minute to 8.22 in the fifth. Asking explicitly for continued discussion raises the fifth-minute value to 11.93 seconds, with early closure and repeated farewells in some examples (Appendix A.5). This distinction between recording length and active exchange motivates reporting speech activity alongside corpus hours and identifying sustained interaction as a target for native-model adaptation.

In the final rule corpus, 50.26% of Chinese and 82.91% of English backchannel windows contain dual-channel VAD overlap. These measurements connect scheduled events to simultaneous speech in the final audio (Appendix C.3).

A separate fixed-source experiment tests onset compensation by advancing inserted clips while keeping their waveforms and the interruption cutoff unchanged. The corpus statistics above use the original corpus renderings; the onset-advanced outputs are used only in this controlled experiment. Measured post-anchor overlap increases from 66/78 to 77/78 observable backchannels and from 11/39 to 37/39 interruptions. These results support scheduling against actual speech onset. Appendix B.3 reports the paired design and full event counts.

5.3 Human Listening Evaluation

Figure 3 summarizes two listeners' ratings of 100 dialogues: 40 native English, 40 rule English, and 20 rule Chinese. We use category-stratified random sampling, grouping candidates by scenario or rule category to cover varied interaction settings. The native set spans 40 scenarios. Listeners score dialogue playback and can inspect channels and available transcripts. On five-point scales, interaction naturalness (N-MOS) is 4.10, 4.56, and 4.50, respectively; content meaningfulness (M-MOS) is 4.16, 4.94, and 4.83. These ratings support natural and coherent dialogue from both routes, with higher means for the rule English set. Listeners' annotations identify pauses and competing speech as the main opportunities to improve native interaction. Appendix C.5 reports the protocol, channel-quality scores, and annotations.

Listening places the temporal measurements in their conversational context. Both native channels receive mean audio-quality scores above 4.5, while interaction naturalness averages 4.10. Listeners

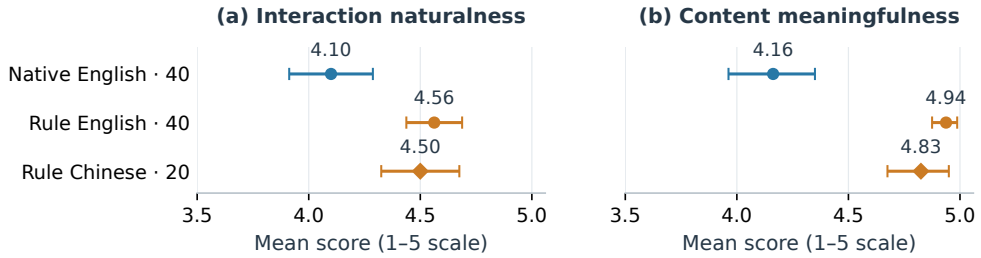


Figure 3: Human assessment of dialogue quality. Points show mean interaction naturalness and content meaningfulness on a five-point scale; bars show 95% intervals from dialogue-level bootstrap resampling. Each dialogue has two ratings. Group labels give language, synthesis route, and dialogue count.

mark nineteen inappropriate overlaps and thirteen long silences in the native set. This separation between clear speech and natural coordination explains why we assess whole dialogues as well as individual channels: a well-formed utterance can still enter at an unsuitable moment or be followed by a disruptive pause.

6 Downstream Training with Synthesized Dialogue

We test how synthetic dialogues change a downstream model's interaction behavior. Starting from the same Moshi checkpoint, we update all language-model parameters on native English, rule English, or their combination, with the codec fixed. Native transcripts are aligned after audio screening; rule labels are projected from source utterances after signal checks and removal of cut or ambiguous words. Single-route experiments use fixed training subsets. Joint training combines native and rule-based English data. We reserve disjoint scenarios or topics from the synthetic data for checkpoint selection using text and audio validation losses, and evaluate the selected models on the same 498 Full-Duplex-Bench v1.5 cases (Lin et al., 2025). The benchmark tests two interaction behaviors: responding to a user interruption, and continuing the preceding task through backchannels, other-directed speech, or background speech. Table 3 reports each behavior and their unweighted macro average. Appendix D.1 gives training configurations, data splits, and checkpoint selection.

The two data routes produce distinct response patterns. Native-data adaptation raises interruption response from 59.50% to 72.00%, a paired gain of 12.50 percentage points (95% interval: 4.00–21.00). The model also responds more often to non-interruptions: 98/298 cases, compared with 40/298 for the original. This broader tendency to respond explains why the interruption gain coexists with a macro score of 25.31%. Rule-data adaptation shifts behavior toward continuing the preceding task. Continuation reaches 28.57% for backchannels, 44.00% for other-directed speech, and 43.00% for background speech, while interruption response falls to 36.00%. Under the tested configurations, the native-data model responds more readily, while the rule-data model more often maintains the ongoing task.

Downstream behavior after mixed-data training. With the full English training pool and a 9,500-update schedule, the mixed model reaches 43.03% macro expected behavior, compared with 26.43% for the original: a gain of 16.60 percentage points (95% interval: 10.98–22.00). Its interruption response remains close to the original (58.50% versus 59.50%), while continuation improves in all three non-interruption tasks. The mixed model has the highest backchannel and other-directed continuation scores among the four models; the rule model retains the highest background continuation score. Responses to non-interruptions decrease to 13/298 cases (4.36%), compared with 40/298 (13.42%) for the original. Response precision—the fraction of all response-labeled cases that are actual interruptions—rises from 74.84% to 90.00% (Appendix D.3). Together with improved continuation, this identifies the mixed model's gain as more selective engagement: it better maintains the current task through unrelated speech while preserving near-original responsiveness to interruptions.

Table 3: Full-Duplex-Bench expected behavior (%, higher is better). Interruptions require a response; the other tasks require continuation.

Training data	Interruption	Backchannel	Other-directed	Background	Macro
Cases	200	98	100	100	
Original	59.50	11.22	21.00	14.00	26.43
Native	72.00	12.24	11.00	6.00	25.31
Rule	36.00	28.57	44.00	43.00	37.89
Mixed	58.50	31.63	47.00	35.00	43.03

Behavioral improvements and acoustic quality. We also use UTMOSv2, an automatic predictor of perceived speech naturalness (Baba et al., 2024), to assess benchmark outputs under clean conditions. Rule-data adaptation raises its scores in all four tasks, whereas native-data adaptation lowers them. The mixed model scores between these two models in every task. Evaluating both behavior and speech quality therefore distinguishes successful interaction from the quality of the response it produces. Appendix D.1 reports full acoustic results, behavior-label counts, paired intervals, and checkpoint-selection details.

7 Discussion

Control over content and timing. Native synthesis develops an exchange from roles and a shared situation, while rule synthesis specifies utterances and their placement. The former supports studying coordination as a conversation unfolds; the latter permits holding speech fixed while changing its timing, as in the onset experiment. These construction choices produce different segment and overlap patterns. The downstream models also show distinct response tendencies: native-data training encourages responding, while rule-data training favors maintaining the ongoing task. Joint training with the larger English pool and longer schedule achieves the strongest aggregate benchmark result.

Using audio together with its measurements. Source checks establish whether the expected utterances are present; final-channel screening measures speech activity, transcription coverage, repetition, and acoustic quality. Keeping these decisions and temporal measurements alongside the audio lets users select examples for a training objective: overlap statistics identify simultaneous channel activity, while internal-silence measurements distinguish recording length from sustained exchange. Source alignments and placements support constructing supervision for retained words and varying rule-event timing. Native activity patterns and listener annotations identify sustained exchange and context-appropriate overlap as targets for further model adaptation.

Preserving coverage during corpus curation. Screening preserves all 208 native scenarios, 32 roles, and 32 voices, and all 73 rule topics in each language. Reporting this coverage alongside temporal distributions describes both the range of generation conditions and the interaction realized under them. These two views support choosing training examples by setting and interaction pattern, and guide further synthesis toward broader conditions or different timing within a condition.

8 Conclusion

DuplexTTS generates and screens full-duplex dialogue for downstream training. Native acoustic interaction develops conversations from scenarios and roles; rule-based synthesis controls dialogue events. The routes produce 107,072 dialogues totaling 2,189.1 hours, with 1,835.3 hours retained after final-audio screening and the rule route's source-utterance checks. Experiments identify speech-onset placement and source checks as effective construction choices. Moshi training produces distinct response and continuation patterns; the mixed-data model achieves the highest macro accuracy.

AI USE STATEMENT

AI assisted implementation, experiment organization, scenario/script preparation, literature retrieval, analysis, and drafting. An LLM supplied dialogue rewards and evaluation scores. ASR, VAD, and SCOREQ provided automatic measurements. The authors retain responsibility for the manuscript and submission disclosure.

REFERENCES

Kaito Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the VoiceMOS Challenge 2024: Transfer learning from deep image classifier to naturalness MOS prediction of high-quality synthetic speech. In *IEEE Spoken Language Technology Workshop (SLT)*, pp. 818–824, 2024. doi: 10.1109/SLT61566.2024.10832315.

Sanyuan Chen, Min-Jae Hwang, Sho Inoue, Anna Sun, Bokai Yu, David Kant, Dongmin Hyun, Dorian Desblancs, Gregory Antonovsky, Oleg Repin, Peng-Jen Chen, Xutai Ma, Zehai Tu, Juan Pino, and Wei-Ning Hsu. Alignment-free Text-Audiobox for voice dubbing and full-duplex dialogue synthesis, 2026. URL https://arxiv.org/abs/2609.03992v1.

Christopher Cieri, David Miller, and Kevin Walker. The Fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004. European Language Resources Association (ELRA). URL https://aclanthology.org/L04-1500/.

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, 2024. URL https://arxiv.org/abs/2410.00037.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training, 2025. URL https://arxiv.org/abs/2505.17589v2.

Qingxiang Guo, Wenke Fan, Shuofeng Zhao, Dawei Yang, Zhiyang Zhou, Yingxin Shang, Hongwei Cai, Zhou Wang, Weixu Wang, Lin Yang, Shuran Zhou, and Yang Song. DuplexDrama: A synthesized dialogue dataset with scenarios, full-duplex behaviors, expressive speech, and sound events, 2026. URL https://arxiv.org/abs/2609.12872v1.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Takyoung Kim, Kang-wook Kim, Sang Hoon Woo, Julia Hirschberg, Gunhee Kim, and Dilek Hakkani-Tür. DuplexGen: Adaptive synthesis of human-AI turn-taking dialogues, 2026. URL https://arxiv.org/abs/2607.26178v2. Version 2.

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, 2025. URL https://arxiv.org/abs/2507.23159v4. Version 4, revised April 2026.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266, 2023. doi: 10.1162/tacl_a_00545. URL https://aclanthology.org/2023.tacl-1.15/.

Atsumoto Ohashi, Neil Zeghidour, Alexandre Défossez, and Eugene Kharitonov. Multi-faceted interactivity alignment in full-duplex speech models, 2026. URL https://arxiv.org/abs/2606.11167v2. Version 2.

Zhiliang Peng, Jianwei Yu, Wenhui Wang, Yaoyao Chang, Yutao Sun, Li Dong, Yi Zhu, Weijiang Xu, Hangbo Bao, Zehua Wang, Shaohan Huang, Yan Xia, and Furu Wei. VibeVoice: Expressive podcast generation with next-token diffusion. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=FihSkzyxdv.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Alessandro Ragano, Jan Skoglund, and Andrew Hines. SCOREQ: Speech quality assessment with contrastive regression. In *Advances in Neural Information Processing Systems*, volume 37, pp. 105702–105729. Curran Associates, Inc., 2024. doi: 10.52202/079017-3353. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bece7e02455a628b770e49fcfa791147-Abstract-Conference.html.

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. In *ICASSP 2026 – 2026 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2026. doi: 10.1109/ICASSP55912.2026.11463413. URL https://research.nvidia.com/labs/adlr/files/personaplex/personaplex_preprint.pdf.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-ASR technical report, 2026. URL https://arxiv.org/abs/2601.21337v2.

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

Tanya Stivers, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter de Ruiter, Kyung-Eun Yoon, and Stephen C. Levinson. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(26):10587–10592, 2009. doi: 10.1073/pnas.0903616106. URL https://www.mpi.nl/publications/item66202/universals-and-cultural-variation-turn-taking-conversation.

Pengcheng Wang, Sheng Li, Jiyi Li, and Takahiro Shinozaki. Agentic-DuplexGen: Decoupling content, timing, and acoustics for synthetic dialogue speech, 2026. URL https://arxiv.org/abs/2608.16053v2. Version 2.

Qichao Wang, Ziqiao Meng, Wenqian Cui, Yifei Zhang, Pengcheng Wu, Bingzhe Wu, Irwin King, Liang Chen, and Peilin Zhao. NTPP: Generative speech language modeling for dual-channel spoken dialogue via next-token-pair prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63912–63928. PMLR, 2025. URL https://proceedings.mlr.press/v267/wang25by.html.

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. In *Interspeech 2019*, pp. 1526–1530, 2019. doi: 10.21437/Interspeech.2019-2441. URL https://www.isca-archive.org/interspeech_2019/zen19_interspeech.html.

Leying Zhang, Yao Qian, Long Zhou, Shujie Liu, Dongmei Wang, Xiaofei Wang, Midia Yousefi, Yanmin Qian, Jinyu Li, Lei He, Sheng Zhao, and Michael Zeng. CoVoMix: Advancing zero-shot speech generation for human-like multi-talker conversations. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3183. URL https://papers.neurips.cc/paper_files/paper/2024/hash/b5fd95d6b16d3172e307103a97f19e1b-Abstract-Conference.html.

Supplementary material follows the construction process: native synthesis and adaptation (A), rule synthesis and event realization (B), corpus screening and assessment (C), and downstream training and evaluation (D).

A NATIVE SYNTHESIS AND ADAPTATION

A.1 TRAINING OBJECTIVES AND CONFIGURATION

The native route uses one PersonaPlex parameter set for both participants, with separate role histories and a shared acoustic observation. Training first adjusts dialogue behavior, then fits acoustic decoding to the adapted backbone. Table 4 presents the two stages in training order.

Backbone adaptation. Starting from the original temporal/text backbone θ_0 and acoustic modules ϕ_0, we first optimize complete-dialogue rewards with ϕ_0 and Mimi fixed. For each of 64 training scenarios, eight 60-second candidates yield 512 trajectories. The reward is the equal mean of six dialogue scores (Appendix A.2), standardized within each candidate group. Both roles receive the same episode advantage.

Clipped text-token GRPO (Shao et al., 2024) averages each role's valid text actions, including padding actions, before averaging the two roles. An eight-group rollout batch is optimized in eight successive minibatches: each takes the same candidate index from every group. Eight rollout batches therefore yield 64 updates. Rewards, behavior probabilities, and advantages are fixed before optimizing each batch.

The original model constrains the update through three KL terms: current-to- reference text-trajectory KL (weight 0.01), reference-to-current KL on clean audio/text anchors (0.1/0.01), and reference-to-current KL on generated-history audio (0.1 over codebooks 1–7). Clean audio anchors also constrain codebook 0; all KL temperatures are one. Clean anchors are eight fixed 256-frame rollouts from the original model with clean voice prompts; optimization cycles through them. The full temporal/text backbone is updated; Depth modules, audio embeddings, audio heads, and Mimi remain fixed.

Acoustic distillation. We then freeze the adapted backbone θ_D and fit the acoustic Depth modules by knowledge distillation (KD) (Hinton et al., 2015). The original model supplies every output-codebook teacher distribution p_T:

$$\mathcal{L}_{\mathrm{KD}} = \mathbb{E}_{z,q}\,\mathrm{KL}\Big(p_T^{(q)}(c_q \mid z, c_{<q}) \,\|\, p_{\theta_D,\phi}^{(q)}(c_q \mid z, c_{<q})\Big)\,. \tag{3}$$

Here z is a sampled role-frame history with its text prefix, q indexes the eight audio codebooks, and $c_{<q}$ contains earlier codebook tokens. Teacher and student use the same histories and prefixes, with hidden states from their respective backbones. Valid role-frame-codebook targets receive equal weight. Training updates Depth conditioning projections, text/audio embeddings, and supervised own-audio heads; temporal input-audio embeddings and unused peer-only decoder parameters stay fixed.

Table 4: Two successive training stages for the native synthesis model. Columns identify the parameters, data, and optimization used in each stage. The acoustic stage reuses backbone-training trajectories; its selected checkpoint minimizes development-set distillation loss (KL).

Setting	Backbone adaptation	Acoustic distillation
Trainable scope	Full temporal/text backbone	Acoustic Depth modules
Data	64 scenarios; 512 trajectories	448 train / 64 development
Optimizer	Adafactor	AdamW
Learning rate	Relative-step cap 0.0002	0.00002
Weight decay	0.1	0
Gradient clipping	2	1
Policy clipping	0.2	—
Updates	64	1,313
Selection	Completed backbone training	Minimum dev. KL; step 768

Backbone optimization accumulates FP32 CPU master-weight updates before casting to the forward dtypes. Acoustic parameters use FP32 with BF16 autocast. The acoustic split holds out eight of the 64 rollout scenarios, giving 56 training scenarios and eight development scenarios. Development KL uses 128 sampled role-frames per trajectory (8,192 total), evaluated every 256 updates and at the end. The rollout pool includes trajectories from successive backbone-training batches. The original-teacher acoustic model selected at update 768 is the model used to construct the native corpus.

Generation. Role prompts combine identity, profile, shared scene setting, and private scenario context. Voice conditioning uses the native 24 kHz frontend, loudness normalization, and prefix construction. Both roles share parameters and receive the completed mixture at gain 0.5 per track, with independent causal states. Generation uses a 12.5-Hz clock, audio/text temperatures 0.8/0.7, and top-k values 250/25. Waveforms are saved as FLOAT without post-generation normalization. The corpus consists of 60-second dialogues.

Production throughput. Offline synthesis distributes independent dialogue requests across GPUs. In one completed production run, eight NVIDIA A100 GPUs generated and structurally validated 3,000 one-minute dialogues (50 hours) in 3.96 hours of wall-clock time, yielding 12.63 dialogue-audio hours per wall-clock hour. This measurement covers generation, waveform writing, and structural validation. ASR and acoustic-quality screening run as subsequent stages.

A.2 Dialogue reward and model assessment

The training reward and native-model comparison use the same six-axis rubric with GPT-5.6 Sol at medium reasoning. Inputs contain the scenario, both roles, and timestamped transcripts of the generated waveforms from Whisper large-v3-turbo (Radford et al., 2023), together with Silero speech intervals (Silero Team, 2024). Each dimension receives an integer score from one to five, a confidence level, and evidence identifying a track and time span. The judge considers the complete exchange and evaluates the two roles symmetrically. Table 5 summarizes the actual questions. The first two dimensions ask whether the exchange follows its setting; the remaining four ask how well the participants develop and coordinate it.

Table 5: Six dimensions used for native-model rewards and comparison.

Dimension	Assessment question
Scenario adherence	Does the dialogue remain grounded in the supplied situation, facts, relationship, and task?
Role adherence	Do both participants behave consistently with their roles while reacting naturally?
Coherence	Is the exchange understandable without unexplained contradictions or topic breaks?
Progression	Does it add relevant information, clarification, decisions, or natural closure?
Responsiveness	Do participants answer, acknowledge, repair, or develop each other's meaning?
Timing	Are gaps, overlaps, floor transfers, interruptions, and backchannels appropriate in context?

The shared scale anchors are: 1, pervasive or severe failure; 2, major defects over a substantial part of the sample; 3, meaningful strengths and defects; 4, mostly strong with minor local defects; and 5, consistently strong without material defects. Role assessment considers both speakers beyond name mentions; coherence spans early, middle, and late dialogue. Brief backchannels can support progression while the other participant holds the floor. Overlap alone does not establish responsiveness, and more overlap is not inherently better timing. These judgments concern the supplied transcript and timeline; acoustic quality is measured separately with SCOREQ.

A.3 Model comparison protocol

The model-study library contains 144 scenarios, 32 role profiles, and 32 voice prompts. The twenty development scenarios in Table 1 are disjoint from the 64 backbone-training scenarios. Each model

generates all twenty 60-second dialogues with matched scenario, role, voice, and seed settings. Dialogue scores use the six-axis rubric above. SCOREQ first takes the median of active-window scores within each track, then averages the two track medians. This comparison follows the training sequence used to construct the native corpus.

A.4 SHARED ACOUSTIC FEEDBACK

Four private-fact tasks use alternative facts (north/south, blue/green, Monday/Friday, blankets/pillows), each assigned to either role: sixteen inputs and eight task–direction pairs. Only the informed role receives the fact; the receiver is asked to confirm it. Its prompt, voice, and positions are fixed across alternatives. We compare normal feedback with removing the informed role from the public audio after the first completed frame. Original and backbone-only models provide normal-feedback controls.

Table 6 counts recognized target and alternative words. Target-only requires the target without its alternative; a successful pair requires target-only output for both alternatives. Questions and negations count as mentions, so the endpoint measures lexical response to the private fact. The first four counts partition sixteen inputs; the last uses eight pairs to test whether the receiver changes its answer with the private fact.

Table 6: Receiver word mentions on 16 private-fact inputs per row. Target-only means the target appears without its alternative, including in questions or negations. A pair succeeds when both fact versions are target-only. Removing the informed speaker yields identical receiver speech across opposite facts.

Weights	Public observation	Target only	Alternative only	Both	Neither	Successful pairs
Original	Normal	11/16	3/16	1/16	1/16	4/8
Backbone only	Normal	10/16	1/16	1/16	4/16	3/8
Native model	Normal	11/16	1/16	0/16	4/16	4/8
Native model	Holder removed	2/16	2/16	0/16	12/16	0/8

The final model yields eleven target-only transcripts with normal feedback and two with the informed role removed. In all eight masked pairs, opposite facts produce identical receiver tokens, waveforms, and public inputs, confirming that this manipulation removes the acoustic information path. The original model also yields eleven target-only transcripts and four successful pairs: adaptation does not improve these lexical endpoints.

A.5 CONVERSATION OVER LONGER GENERATION WINDOWS

We extend four scenarios with two seeds each from 60 to 300 seconds, keeping the native model, voices, sampling, and shared feedback fixed. The temporal context remains 3,000 frames (240 seconds before prefix positions), and the longer outputs reproduce their 60-second prefixes. A paired prompt asks for five minutes of continued discussion through follow-up questions, explanations, and alternatives, without repeating settled points or closing early. Figure 4 tracks what happens in each successive minute: speech union counts seconds with either participant speaking, while SCOREQ tracks predicted audio quality. The two curves differ only in the duration instruction.

Figure 4 includes all eight dialogues per prompt. Speech union falls from 45.00 to 8.22 seconds between minutes one and five with the original prompt, and from 45.32 to 11.93 with the duration instruction. The instruction raises activity most in minute four but does not sustain exchange throughout. Both animal-shelter seeds contain early closure, repeated farewells, and a later unrelated topic under the added prompt. The corpus uses one-minute generations; maintaining the ongoing situation is a target for longer native synthesis. Minute-wise SCOREQ uses four-second windows contained within each minute, averaging the two track means.

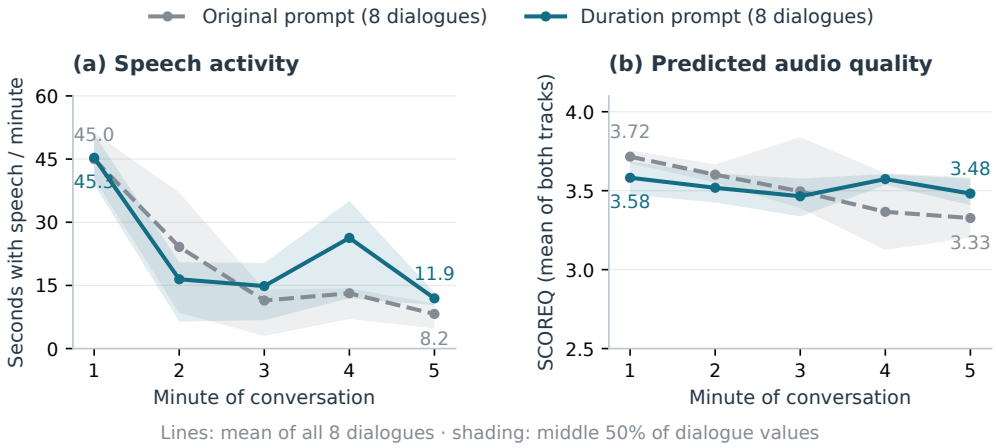


Figure 4: Original versus added duration instruction on eight paired conditions. Gray dashed and teal solid lines show the mean of all eight dialogues per prompt; shading spans the 25th–75th percentiles of dialogue values. Labels give the first- and fifth-minute means. SCOREQ averages the two track means. Minute-wise observations are repeated measurements within each dialogue.

B Rule synthesis and event realization

B.1 Scripts, voices, and alignment

Rule scripts cover 73 topics per language across everyday life, society, culture, science, and technology. Qwen and DeepSeek language models author 10–30 utterance fields in 5–15 two-utterance groups. Every script contains normal exchange, backchannel, interruption, and side talk. A and B sustain the main exchange; C supplies unrelated background speech that can be removed without breaking the main dialogue. Structural validation checks group counts, event coverage, speaker fields, and unique contiguous text anchors; failed validation requests a new complete script. Table 7 shows how a text anchor specifies where another speaker may enter. The italic span belongs to the ongoing utterance; its alignment locates the brief acknowledgment or interrupting question on the audio timeline.

Table 7: From a text anchor to an inserted utterance. Each row pairs an ongoing utterance with the speech to insert. Italics identify the anchor; alignment and the event-specific scheduling rule determine the insertion time. Examples are translated from Chinese.

Event	Source utterance	Inserting utterance
Backchannel	A: I sit in an office and hardly sweat; *forcing down that much water* makes my stomach uncomfortable.	B: Right.
Interruption	B: Replacing electrolytes after heavy sweating is an exception; *otherwise, plain water is* the best choice.	A: Does an electrolyte drink after exercise count as a special case?

Fun-CosyVoice3-0.5B synthesizes each utterance independently, using zero-shot voice references, non-streaming decoding, and speed 1.0. Chinese uses one fixed reference per role. English samples one male and one female reference from ten of each, randomly assigns them to A/B, and samples C independently from twenty background references. These forty LibriTTS test-clean clips (Zen et al., 2019) represent twenty-nine speakers; eleven identities overlap between the main and background pools. A dialogue keeps its assigned references for all utterances.

Source waveforms are mono 24 kHz PCM16. Qwen3-ForcedAligner aligns the supplied text in the requested language, internally resampling to mono 16 kHz and leaving the source waveform unchanged. An anchor is matched after removing punctuation and whitespace. Its normalized characters inherit their alignment token's endpoint. If text normalization changes an anchor, uniquely

matched neighboring context locates the corresponding interval; unresolved or ambiguous matches prevent event placement. English clips retain up to 0.2 seconds of context before the first aligned word.

B.2 Event placement

Scheduling the four events. Let T denote a group's start, d_i a source clip's duration after configured leading trim, and α the event's insertion offset in that trimmed clip. The final corpus uses:

1. **Normal exchange:** A starts at T, followed by B at $T + d_A$. The next group starts when B ends.
2. **Backchannel:** B starts at A's anchor endpoint while A continues. The next group starts at A's end, allowing an unfinished B tail to continue.
3. **Interruption:** A starts at the endpoint of the normalized character immediately after B's anchor, clamped to the final character. B is retained to the endpoint below; the last at most 0.15 seconds fade linearly. The next group starts after both retained utterances end.
4. **Side talk:** B continues and C is placed by the language-specific background policy; B's end advances the main timeline.

$$t_{\mathrm{B,end}} = T + \min(d_B, \alpha + r), \qquad r \sim \mathcal{U}(0.3, 0.6)\ \mathrm{s}. \tag{4}$$

The Chinese background policy advances C relative to its B-anchor so that it ends by B's end when possible. The English policy samples C's start between neighboring A turns, fitting the full clip into the interval when possible. Long clips can overlap the next A, and the English final output extends to preserve C's tail. Chinese output ends at the final main-timeline cursor. A and C mix on the left channel; B occupies the right. Samples add with PCM16 clipping. There is no global loudness normalization or background attenuation.

Projecting words onto the final audio. Rule records retain source clips, word alignments, event placements, and screening decisions. To construct training labels, a clip placed at time p_i after leading trim δ_i maps an aligned word $[s_j, e_j]$ to $[p_i + s_j - \delta_i, p_i + e_j - \delta_i]$. The projected interval is checked against the retained source extent, final audio length, and same-channel contributions. Words crossing a cut or an ambiguous mixture receive masks. The downstream English export determines placement and leading trim by matching source PCM to the final waveform, and uses B as output with A+C as listening input.

Content at an interruption boundary. In a Chinese example, an interruption about the price of hairy crabs starts at 40.800 seconds, while the host's source mention aligns to 41.600–42.160 seconds, after its retained endpoint of 41.177 seconds. Projecting labels removes that inaudible host phrase. The example also shows why conversational appropriateness is assessed against the speech available at insertion, in addition to checking the alignment of individual words.

B.3 Aligning insertion to speech onset

The controlled experiment uses twenty English scenarios with two seeds each. Each script has fourteen utterances, including two backchannels and one interruption. CosyVoice synthesizes the source utterances with fixed references; all 560 calls pass the source-signal gate on the first attempt. One condition fails because its last aligned word extends 40 ms beyond a 4.600-second clip. The remaining 39 conditions supply 78 backchannels and 39 interruptions; all forty planned conditions remain in completion accounting.

We compare the original placement with advancing each inserted file by its first positive-duration aligned token onset τ. If the host-relative anchor is α, the file starts at $T + \alpha - \min(\tau, \alpha)$. Source waveforms are identical across conditions. In this experiment the host interruption cutoff remains $\min(d_B, \alpha + 0.45)$ with the same at-most-0.15-second fade. Full outputs are cropped or padded to a 60-second observation window.

Event overlap is dual-channel VAD activity after the host anchor, ending at the earliest of anchor plus one second, host end, insert end, and window end. Zero overlap remains zero; an absent observation

remains missing. Both events must be observable for a condition's backchannel mean; both seeds must be present for a scenario mean. Nineteen complete scenarios enter paired comparisons, with 2,000 scenario-bootstrap resamples. Table 8 reports overlap occurrence and duration. "Planned" additionally counts the failed condition; "observable" does not. The final rows require at least 100 milliseconds of overlap.

Table 8: Advancing inserted speech increases realized overlap. Each count is events meeting the overlap criterion / events in that denominator. "Original rule" places the file at the anchor; "onset advance" compensates for its leading silence. Duration means use the 39 observable conditions.

Measurement	Original rule	Onset advance
Backchannels with overlap / observable	66/78	77/78
Backchannels with overlap / planned	66/80	77/80
Interruptions with overlap / observable	11/39	37/39
Interruptions with overlap / planned	11/40	37/40
Mean backchannel overlap (s)	0.39208	0.81428
Mean interruption overlap (s)	0.03836	0.32723
Backchannels with $\geq$100-ms overlap / observable	62/78	77/78
Interruptions with $\geq$100-ms overlap / observable	6/39	29/39

Advancing the insert increases paired mean backchannel overlap by 0.426 seconds (95% interval: 0.375–0.476) and interruption overlap by 0.289 seconds (0.219–0.347). Excluding the initial discovery scenario preserves both increases across eighteen scenarios. A post-hoc 100-ms minimum-overlap check retains the direction for both event types (Table 8). The SCOREQ difference is 0.010 $[-0.002, 0.022]$. Earlier insert endings can advance later groups within the observation window; the source samples and host-relative cutoff remain fixed. This experiment supports scheduling the audible speech onset when realizing scripted events.

C Corpus screening and assessment

C.1 Screening criteria and completeness

Every native dialogue is structurally validated and screened without altering its waveform. A dialogue is rejected if either channel has fewer than three seconds of detected speech, fewer than eight ASR words, an untranscribed VAD fraction above 0.5, a repeated four-gram fraction above 0.6 with at least twenty words, more than three consecutive identical nonempty IPUs, or mean active-window SCOREQ below 3.0. Identical A/B waveforms are rejected. SCOREQ uses four-second windows with two-second hop at 16 kHz; windows below -55 dBFS RMS are excluded, and a channel with no eligible window fails. The threshold applies to each channel's mean. FLOAT excursions beyond $[-1, 1]$ are flagged and their original values preserved.

Rule screening combines these final-channel checks with inspection of every source utterance. A source must be finite, nonempty, decodable mono 24 kHz PCM16 with RMS at least -55 dBFS and peak above one quantization unit. Chinese text gates count normalized Unicode alphanumeric characters; English gates count words. The minimum-eight-unit and four-gram rules use these language-specific units. Missing sources retain their rejection status even when a composer has substituted silence in the assembled dialogue. Table 9 separates source completeness from final audio quality across the entire rule corpus. Rows show source checks and columns show assembled-audio checks. Only the top-left cell is selected; the bottom-left cell passes the audio screen despite missing or near-silent source speech.

Across the rule corpus, 3,016 missing fields and 11,602 near-silent clips affect 12,999 dialogues. Joint selection retains the 22,232 English and 11,723 Chinese dialogues reported in Table 2.

C.2 Screening components and threshold sensitivity

We evaluate gate omissions on all 60,000 native dialogues using the same fixed measurements. Low SCOREQ and repetition cover all 1,015 rejected dialogues: 929 involve SCOREQ alone, 73 repetition alone, and thirteen both. Table 10 includes all gates, including those with zero exclusive

Table 9: Source-signal and final-channel checks on all 47,072 Chinese and English rule dialogues. Combined selection retains their intersection (33,955); source checks identify 12,977 additional failures among final-channel passes.

Source-signal gate	Audio pass	Audio reject	Total
Pass	33,955	118	34,073
Reject	12,977	22	12,999
Total	46,932	140	47,072

contribution in this corpus. The columns distinguish all failures from failures unique to one check. For example, low SCOREQ flags 942 dialogues, of which 929 would pass all the other gates.

Table 10: Which checks reject native dialogues? The first count includes every dialogue failing that check; the second includes only those passing all other checks. A dialogue can fail several checks, so the first column is not additive. All 60,000 native dialogues are evaluated.

Gate	Rejected by this check	Rejected only by this check
Identical tracks	0	0
Insufficient speech	2	0
Insufficient ASR words	1	0
Untranscribed speech	0	0
Repeated four-grams	75	64
Repeated IPUs	11	9
No active SCOREQ windows	0	0
Low SCOREQ	942	929

Figure 5 asks how a stricter audio-quality threshold changes the selected native corpus. Panels (a) and (b) share the same threshold rows: read retained dialogue count and mean overlap together. Panel (c) fixes the chosen threshold at 3.0 and groups the 32 voices by their retention rate.

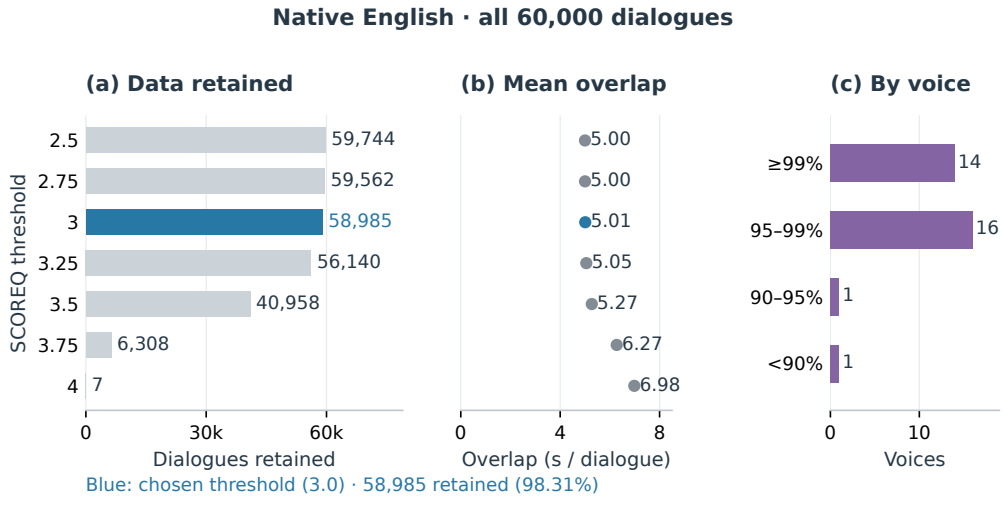


Figure 5: Screening effects across all 60,000 native dialogues (1,000 hours). (a) Dialogue counts retained at each SCOREQ threshold; blue marks the chosen threshold of 3.0. (b) Mean overlap within each corresponding retained set. (c) Number of voices in each retention range at threshold 3.0, with upper endpoints excluded below 99%. Voice-associated retention groups whole-dialogue decisions. Other screening rules remain fixed; threshold 4.0 retains seven dialogues.

With the other gates fixed, increasing the SCOREQ threshold from 3.0 to 3.5 reduces retention from 58,985 (98.31%) to 40,958 (68.26%) dialogues. Mean raw-VAD overlap changes from 5.009 to 5.271 seconds in the retained sets. Thus, the higher threshold removes many more dialogues while changing mean overlap only modestly. The corpus uses 3.0. Voice-associated retention counts whole-dialogue decisions; its lowest values are 89.71% and 91.30%. All 208 scenarios, 32 roles, and 32 voices remain represented, together with 27,443 of 27,450 input conditions. The corpus scenarios comprise 64 from policy training and 144 newly authored ones, distinct from the initial 144-scenario model-study library.

C.3 Interaction measurements in the final corpus

Temporal analysis covers all 107,072 generated dialogues and their 92,940 selected outputs. Silero VAD 6.2.1 runs at 16 kHz with threshold 0.5, minimum speech 250 ms, minimum silence 100 ms, and 30 ms padding. Following dGSLM (Nguyen et al., 2023), within-channel gaps of at most 200 ms are bridged to form IPUs. A pause is shared silence bounded by the same channel; a gap is bounded by different channels. Overlap is simultaneous channel IPU activity. Leading, trailing, and ambiguous intervals are counted separately. The rule left channel includes both A and C, so these measures characterize A+C versus B. Raw-VAD overlap used in screening omits the IPU bridge.

Event frequencies divide counts by total shared-timeline minutes. Duration percentiles weight events equally. The complete selected-corpus distributions appear in Figure 2. The overlap share of shared time changes from 8.55% to 8.57% for native English, from 5.69% to 5.84% for rule English, and from 7.79% to 7.94% for rule Chinese after screening. Internal shared silences of at least five seconds occur in 29.29%, 0.07%, and 0% of the selected dialogues, respectively. All 73 rule topics in each language remain represented.

Scheduled events and measured channel overlap are recorded separately. Within selected rule backchannel windows, positive dual-channel VAD overlap occurs in 12,820/25,509 Chinese events (50.26%) and 39,017/47,058 English events (82.91%). Each window is the scheduled host/insert intersection clipped to the final waveform. These measurements describe actual event realization alongside the intended script labels.

C.4 Dialogue lengths and common expressions

The final corpus contains 58,985 native English, 22,232 rule English, and 11,723 rule Chinese dialogues. Native generation uses a fixed 60-second window. Rule dialogue lengths instead follow the script and event schedule. Figure 6 shows their full distributions: English dialogues have a median duration of 96.00 seconds, with the middle 50% spanning 84.52–109.12 seconds; Chinese dialogues have a median of 74.84 seconds and a corresponding range of 65.00–86.56 seconds.

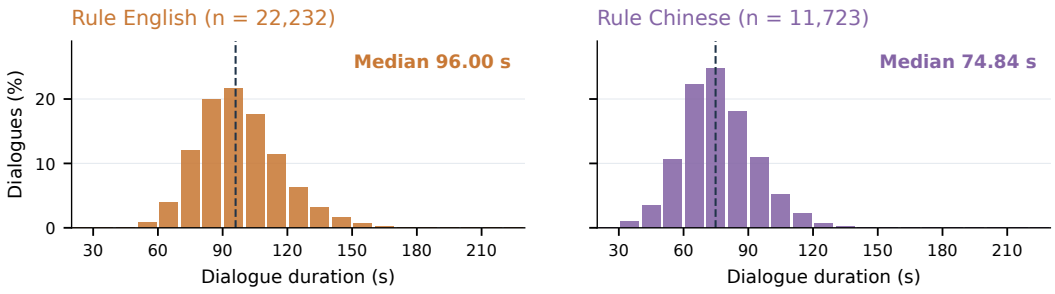


Figure 6: Duration of all retained rule dialogues, normalized within each language. Bars use 10-second bins; dashed lines mark medians. Both panels use the same axes and include the full duration range. Native dialogues use a fixed 60-second generation window. Durations count the shared timeline once.

Figure 7 summarizes common expressions in the final-audio ASR transcripts. Each term is counted once per dialogue, so repeated occurrences within a dialogue do not increase its weight. The clouds highlight frequent conversational expressions: “exactly” appears in 90.5% of native English dialogues and 63.1% of rule English dialogues; the name “Sarah” appears in 50.0% of rule English

dialogues. These word-level summaries describe recurring expressions, while the scenario and topic counts describe the supplied content categories.



Figure 7: The 32 most frequent non-stopword terms in each retained corpus, weighted by the number of dialogues containing the term. English words are lowercased; Chinese uses jieba segmentation with terms of at least two Han characters. Size reflects within-panel frequency and layout, with independent scaling across panels. All retained dialogues contribute, using the same fixed stopword lists as the lexical statistics.

C.5 Human listening evaluation

Two listeners rate 40 native English, 40 rule English, and 20 rule Chinese dialogues, giving 200 reviews. Sampling is stratified by scenario or rule category, with random selection within categories to cover different interaction settings. The native set spans forty distinct scenarios from the backbone-training scenario library. Its generation conditions were selected using input metadata, before synthesis, and each output lasts one minute. Rule English and Chinese dialogues last 57.96–134.56 and 47.32–117.20 seconds, respectively. Native ratings from one listener were carried forward from an earlier questionnaire, as recorded in the evaluation export.

The interface provides playback, channel inspection, transcripts, and timestamped issue annotation. Integer scores range from one (lowest) to five (highest). Channel MOS measures audio quality; following dGSLM (Nguyen et al., 2023), N-MOS measures natural interaction and M-MOS meaningful, coherent content. Native dialogues also receive a scenario-and-role adherence score. All applicable dimensions have two completed-listen ratings per dialogue (880 scores). We average listeners within each dialogue, then dialogues within each set. The 95% percentile intervals use 10,000 dialogue-bootstrap resamples, keeping both listeners together; they describe variation for these two listeners. Table 11 gives means and 95% intervals. Its first two score rows assess individual channels; the next two assess the conversation.

Table 11: Human listening scores (1–5, higher is better): dialogue means with 95% bootstrap intervals in brackets. Each dialogue has two listeners. In rule synthesis, the left channel contains A and C; the right contains B. Setting adherence uses the native dialogues' supplied scenarios and roles.

Measurement	Native English	Rule English	Rule Chinese
Dialogues	40	40	20
	4.538	4.938	4.825
Left-channel MOS	[4.388, 4.675]	[4.888, 4.988]	[4.675, 4.950]
	4.550	4.975	4.500
Right-channel MOS	[4.425, 4.663]	[4.938, 5.000]	[4.350, 4.650]
	4.100	4.563	4.500
Interaction naturalness (N-MOS)	[3.913, 4.288]	[4.438, 4.688]	[4.325, 4.675]
	4.163	4.938	4.825
Content meaningfulness (M-MOS)	[3.963, 4.350]	[4.875, 4.988]	[4.675, 4.950]
	4.125		
Scenario and role adherence	[3.963, 4.275]	—	—

Channel-quality means exceed 4.5 in both English sets. Native dialogue-level naturalness and meaningfulness are lower than its channel-quality ratings, locating the remaining improvement opportunity in the exchange as a whole. Listeners mark 38 issues across 25 native dialogues: thirteen long

silences, nineteen inappropriate overlaps, four setting deviations, and two other issues. Rule English has one voice-drift annotation; rule Chinese has one repetition annotation. Multiple marks may concern one dialogue or overlapping intervals. The native annotations identify interaction timing as an improvement target.

Agreement between listeners. Table 12 gives the proportions of dialogues with identical ratings or ratings within one point. For example, 45.0 / 92.5 means 45.0% match exactly and 92.5% differ by at most one point. Content ratings agree exactly in 32.5% of native English, 92.5% of rule English, and 85.0% of rule Chinese dialogues; naturalness ratings vary more between listeners in the rule sets.

Table 12: Listener agreement (%). Each cell gives exact agreement / agreement within one point. Denominators are 40, 40, and 20 dialogues, respectively.

Measurement	Native English	Rule English	Rule Chinese
Left-channel MOS	45.0 / 92.5	87.5 / 100.0	80.0 / 95.0
Right-channel MOS	45.0 / 85.0	95.0 / 100.0	35.0 / 85.0
Interaction naturalness (N-MOS)	50.0 / 87.5	42.5 / 80.0	35.0 / 95.0
Content meaningfulness (M-MOS)	32.5 / 72.5	92.5 / 100.0	85.0 / 100.0
Scenario and role adherence	22.5 / 70.0	—	—

D Downstream training and evaluation

D.1 Data, optimization, and checkpoint selection

All adaptation conditions initialize the same official Moshiko checkpoint (Défossez et al., 2024). Training updates the temporal and Depth transformers, embeddings, and prediction heads; Mimi stays fixed. Native dialogues supply both speaker orientations. English rule data use B as output and A+C as listening input, with source-signal checks and PCM-verified word projection. Loss masks exclude interruption fades, words crossing windows, same-channel ambiguity, and code positions beyond observed duration. Table 13 separates the data used for training from the fixed examples used to choose a checkpoint. "Training updates" is the total optimization run; "selected update" identifies the checkpoint evaluated on Full-Duplex-Bench. Validation views are used for selection, not benchmark scoring.

Table 13: Data and checkpoint selection for the three downstream training conditions. Hours count unique training audio once. A validation view is one speaker orientation or one rule window. The selected checkpoint can precede the final training update.

Setting	Native	Rule	Mixed
Training dialogues	36,848	6,148	73,816
Training hours	614.13	167.97	1,425.61
Fixed validation views	48	48	144
Training updates	2,000	2,000	9,500
Selected update	2,000	700	9,100

The native-only split has 196 training scenarios and twelve held-out scenarios containing 2,427 dialogues (40.45 hours). Two fixed dialogues per held-out scenario provide 48 speaker views. The rule-only split has 61 training topics and twelve held-out topics; its 1,190 validation dialogues yield 2,464 windows (32.31 hours), with four fixed windows per topic used for selection. Training uses 12,683 rule windows. Rule windows span at most 60 seconds; native inputs span 60 seconds.

The mixed split combines 55,202 native and 18,614 rule dialogues, providing 148,670 speaker/window views. Uniform view sampling yields 74.26% native and 25.74% rule data. Validation fixes 48 native and 96 rule views from held-out scenarios/topics, with equal weight per view. These split-specific hours and label requirements define the data used by each training condition.

All conditions use eight-GPU fully sharded data parallelism, BF16 parameters, FP32 master parameters and AdamW states, and gradient checkpointing. One view per GPU with two accumulation steps gives sixteen views per update. OneCycle peaks at 2×10^{-6} after the first 5% of the scheduled updates; weight decay is 0.1 and gradient clipping is 1.0. Validation every 100 updates selects the minimum sum of text and audio losses on the fixed held-out views. Model selection precedes Full-Duplex-Bench evaluation.

D.2 Benchmark protocol and behavior labels

Full-Duplex-Bench v1.5 (Lin et al., 2025) contains 200 interruption, 98 backchannel, 100 talking-to-others, and 100 background-speech cases. Each checkpoint produces 996 clean/overlap generations. Original Moshiko and the selected native, rule, and mixed models share Mimi, tokenizer, BF16 inference, audio/text temperatures 0.8/0.7, and top-k values 250/25. Case-derived seeds pair generations; each session starts a fresh streaming state and receives 1,920-sample frames at 24 kHz every 80 ms. Parakeet-TDT-0.6B-v2 transcribes outputs, and `gpt-4o-2024-08-06` applies the benchmark's fixed behavior instruction.

The judge assigns exactly one of four categories to each overlap case: `C_RESPOND` addresses the overlap's content; `C_RESUME` ignores that content and continues or completes the model's preceding task; `C_UNCERTAIN_HANDLING` requests clarification without addressing the content; and `C_UNKNOWN` covers off-target or low-information output. Table 14 retains all four labels. The expected category is `C_RESPOND` for user interruptions and `C_RESUME` for the other three tasks. Each row of Table 14 partitions its n cases. Task success means Respond for interruption and Continue for the other tasks.

Table 14: How each model handles the incoming speech. Counts in the four behavior columns sum to n. Continue means resume or complete the preceding task; Uncertain means request clarification; Unknown covers off-target or low-information output. Each model covers all 498 cases.

Task and model	n	Respond	Continue	Uncertain	Unknown
Interruption, original	200	119	35	15	31
Interruption, native	200	144	17	3	36
Interruption, rule	200	72	108	2	18
Interruption, mixed	200	117	48	4	31
Backchannel, original	98	4	11	1	82
Backchannel, native	98	1	12	2	83
Backchannel, rule	98	2	28	0	68
Backchannel, mixed	98	0	31	0	67
Talking to others, original	100	21	21	6	52
Talking to others, native	100	40	11	5	44
Talking to others, rule	100	9	44	1	46
Talking to others, mixed	100	8	47	0	45
Background speech, original	100	15	14	1	70
Background speech, native	100	57	6	4	33
Background speech, rule	100	9	43	0	48
Background speech, mixed	100	5	35	0	60

D.3 Response selectivity and paired differences

Across all 298 non-interruption cases, the original, native, rule, and mixed models respond in 40, 98, 20, and 13 cases. Response precision is the fraction of all response-labeled cases that are actual interruptions. The corresponding ratios are 119/159 (74.84%), 144/242 (59.50%), 72/92 (78.26%), and 117/130 (90.00%). Non-interruption response rates are 13.42%, 32.89%, 6.71%, and 4.36%, respectively. Unknown and uncertain outcomes remain in the task denominators for expected-behavior accuracy. Table 15 compares each trained model with the original on the same cases. Positive changes mean more task-appropriate behavior; intervals spanning zero include changes in either direction. For mixed training, all three continuation-task intervals lie above zero, whereas the interruption interval spans zero.

Table 15: Changes in task-appropriate behavior relative to the original model. Cells give percentage-point changes and paired 95% intervals. Positive is better: responding for interruption, continuing for the other three tasks.

Task	Native minus original	Rule minus original	Mixed minus original
Interruption	12.50 [4.00, 21.00]	−23.50 [−32.00, −15.00]	−1.00 [−11.00, 8.50]
Backchannel	1.02 [−7.14, 9.18]	17.35 [6.12, 28.57]	20.41 [9.18, 32.65]
Talking to others	−10.00 [−21.00, 0.00]	23.00 [12.00, 34.00]	26.00 [14.00, 38.00]
Background speech	−8.00 [−17.00, 1.00]	29.00 [19.00, 39.00]	21.00 [11.00, 31.00]

Unknown outcomes remain frequent for backchannels: 67/98 for mixed and 82/98 for original (Table 14). The mixed model's macro gain is 16.60 percentage points (95% interval: 10.98–22.00).

D.4 Predicted audio quality and uncertainty

Table 16 compares speech quality in the clean condition; higher UTMOSv2 predicts more natural-sounding speech. Each row compares models on the same task.

Table 16: Clean-condition UTMOSv2 by task. Means use the official inclusive 5th–95th-percentile trim.

Task	n	Original	Native	Rule	Mixed
Interruption	200	2.733	2.364	2.938	2.542
Backchannel	98	2.426	2.366	2.687	2.580
Talking to others	100	2.463	2.307	2.804	2.634
Background speech	100	2.492	2.298	2.716	2.548

Table 16 follows the official 5th–95th-percentile trim, retaining 180/88/90/90 clean-condition scores per task and model. Paired comparisons use all 498 finite scores. Rule-model quality gains have 95% intervals above zero in every task; native-model intervals are below zero except for backchannels, whose interval spans zero. Mixed-model interruption quality is lower than the original, while other-directed quality is higher.

Downstream intervals use 5,000 paired bootstrap resamples of benchmark cases. Pooled ratios and macro scores additionally preserve task counts (200/98/100/100). Replicates use the same case indices across models; central 95% intervals describe case variation for these fixed checkpoints and judge assignments.