

Behavior-SD: Behaviorally Aware Spoken Dialogue Generation with Large Language Models

Sehun Lee* Kang-wook Kim* Gunhee Kim

Seoul National University

shlee@vision.snu.ac.kr full1324@snu.ac.kr gunhee@snu.ac.kr

<https://yhytoto12.github.io/Behavior-SD>

Abstract

Spoken dialogue involves behaviors like turn-taking, interruptions, filler words, and backchannels, which make interactions more natural and engaging but are often overlooked in language models. These models struggle to explicitly model these behavioral traits, resulting in a less natural and personalized communication style that aligns with user needs. To address this challenge, we make two key contributions. First, we introduce **Behavior-SD**, a large-scale dataset containing over 100K spoken dialogues (2,164 hours) annotated with various conversational behaviors, synthesized via LLMs to model diverse full-duplex interactions. Second, we propose **BeDLM**, the first dialogue model capable of generating natural conversations conditioned on specific behavioral and narrative contexts, supporting simultaneous contributions from both speakers. Through human evaluations and behavior-adherence metrics, we demonstrate that BeDLM outperforms baseline models in generating natural, coherent, and behaviorally rich dialogues. Our work opens new possibilities for developing behaviorally-aware dialogue systems that more closely mimic human conversational dynamics, enhancing user engagement and communication effectiveness.

1 Introduction

Large language models (LLMs) have significantly improved their ability to generate coherent text; however, producing human-like spoken dialogues remains a considerable challenge. Spoken conversations are inherently dynamic, involving elements like prosody, emotion (Osman, 2022), turn-taking, and interruptions (Nguyen et al., 2023). Capturing these nuances is crucial for generating life-like dialogue, yet current TTS systems (Guo et al., 2023; Du et al., 2024) and spoken language mod-

els (Zhang et al., 2023; Li et al., 2024) struggle to replicate this complexity fully.

A particularly challenging aspect of real-world conversations is managing *full-duplex* communication, where speakers frequently overlap or interrupt each other. Full-duplex conversations, unlike traditional systems that enforce strict turn-taking, enable more fluid interactions, which can be beneficial in applications such as customer service, education, or voice assistant applications (Si et al., 2023; Google, 2024; OpenAI, 2024). Additionally, conversational dynamics are influenced by personal traits, social contexts, and factors like politeness (Bevacqua et al., 2012; Yamamoto et al., 2018), making this task even more complex.

Despite the prevalence of these dynamics in human conversations, they are often underexplored in dialogue generation. While existing spoken dialogue datasets (Cieri et al., 2004; Godfrey et al., 1992; Reece et al., 2023) include conversational behaviors such as interruptions and backchannels, these behaviors are not labeled at the speaker level, limiting the ability to model nuanced interaction patterns. Recent datasets (Lee et al., 2023; Lin et al., 2024) focus on sequential turn-taking, overlooking critical aspects of natural conversations like overlapping speech and interruptions.

To address these gaps, we present **Behavior-SD**, a large-scale dataset of over 100K spoken dialogues (SD), totaling 2,164 hours, and covering a wide range of conversational behaviors and social situations. Unlike existing datasets constrained by narrow contexts (Cieri et al., 2004; Godfrey et al., 1992; Reece et al., 2023), Behavior-SD explicitly models key conversational traits such as verbosity, backchannels, and interruptions, allowing for more realistic modeling of human-like dialogues. This dataset serves as a foundation for improving both the scale and behavioral diversity of dialogue generation. Specifically, it captures four key speaker behavior traits, as outlined in Table 1.

*Equal contribution.

Behavior types	Definition
(V) Verbosity	The length of utterances during a speaker’s turn.
(F) Filler words	The occurrence of filler words (<i>e.g.</i> , "like," "um", "you know") during a speaker’s turn.
(B) Backchannels	The frequency of backchannel responses (<i>e.g.</i> , "yeah," "uh-huh") during the other speaker’s turn.
(I) Interruptions	The frequency of interrupting the other speaker’s turn to begin speaking.

Table 1: Definition of behavior types for full-duplex interactions in conversational dialogues. Each behavior type is classified into three levels: none (0), moderate (1), and frequent (2).

Moreover, we introduce **Behavior-conditioned spoken Dialogue Language Model (BeDLM)**, the first model capable of generating full-duplex spoken dialogues by incorporating behavior traits. By integrating these traits through control tokens, BeDLM produces more dynamic and lifelike conversations, adapting naturally to speaker personalities and narrative context. This novel approach enables customization of dialogue behaviors in a way that previous models, which typically focus on static turn-taking, cannot achieve.

In our experiments, BeDLM demonstrates improvements in dialogue naturalness and adherence to conditions compared to state-of-the-art models. The dataset, model, and code will be made publicly available to encourage further research in this area.

2 Related Works

Automatic spoken dialogue generation. The automatic synthesis of dialogues using LLMs has made significant strides, but generating spoken dialogue remains challenging due to features like paralinguistic cues, interruptions, and unclear turn boundaries in speech (Abdullin et al., 2023; Kim et al., 2023; Lin et al., 2024). Approaches such as dGSLM (Nguyen et al., 2023) have tackled overlapping speech generation using dual-tower transformers without text inputs, while CHATS (Mitsui et al., 2023) builds on this by generating natural spoken dialogues from text transcriptions. However, these methods lack mechanisms to condition overall conversational behavior throughout a dialogue.

Conditional speech generation. Recent models like PromptTTS (Guo et al., 2023), Audiobox (Vyas et al., 2023), and CosyVoice (Du et al., 2024) have advanced text-to-speech (TTS) by incorporating natural language instructions to vary prosodic features like accent, emotion, pitch, and speed. However, these models focus on isolated utterances and lack the ability to maintain consistency across full conversations, including conversational behaviors such as turn-taking and

backchanneling. As a result, they struggle to replicate natural dialogue flow.

Spoken dialogue behaviors. Research has shown that conversational behaviors like backchannels and interruptions play a crucial role in communication, signaling listener engagement and influencing dialogue flow (Reece et al., 2023; Bavelas et al., 2000). These behaviors vary based on context and individual traits (Blomsma et al., 2024). Traditional dialogue systems have attempted to account for such nuances by manually inserting fillers or backchannels, but they lack the flexibility to adapt dynamically to the conversation (De Sevin et al., 2010). Recent efforts focus on equipping LLMs to handle interruptions more naturally, enhancing interaction fluidity (Ma et al., 2024; Zhang et al., 2024; Wang et al., 2024), highlighting the importance of integrating these behaviors for more natural dialogue.

Spoken language models. Spoken language models (SLMs) have evolved from textless approaches like GSLM (Lakhotia et al., 2021), which convert speech into discrete units and training similarly to text-based language models. Recent advancements enhance SLMs by directly integrating speech representations into pre-trained LLMs. For instance, SpeechGPT (Zhang et al., 2023) incorporates speech tokens into LLMs through instruction tuning, enabling more natural and context-aware audio interactions. AudioGPT (Huang et al., 2024), LLaMA-Omni (Fang et al., 2024), and Qwen-Audio (Chu et al., 2023) integrate multimodal inputs, including text and various audio types, to support multimodal interactions without relying on speech transcription.

3 Behavior-driven Data Generation

Our framework for dataset generation operates in four stages: (1) generating dialogues with behavior-specific traits, (2) inserting backchannels at contextually appropriate moments, (3) captioning the speech style of each utterance for TTS integration,

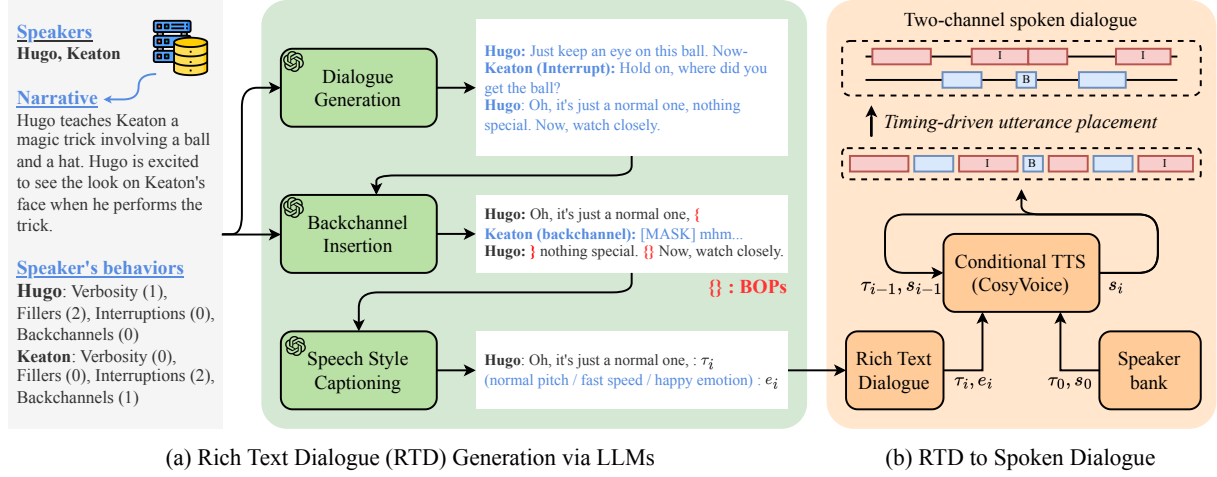


Figure 1: Overall data generation pipeline. (a) Narrative and conversational behaviors are sampled. The conversation is then simulated using LLMs, incorporating backchannel opportunity points (BOPs) detection, backchannel insertion, and speech style captioning to model diverse full-duplex interactions (§3.1). (b) The generated text dialogues are converted into spoken form by CosyVoice and timing-driven utterance placement (§3.2).

and (4) converting text dialogues into spoken form. Figure 1 overviews the framework. All detailed prompts for LLMs can be found in Appendix A.

3.1 Rich Text Dialogue Generation via LLMs

Sample narrative and conversational behaviors.

To generate diverse social spoken dialogues, we sample two key elements: narratives and conversational behaviors. We sample narratives from the SODA dataset (Kim et al., 2023), since it covers a broad spectrum of social interactions. Conversational behaviors include four traits, each categorized into three levels (0, 1, 2), as shown in Table 1. For each conversation, two speakers are sampled with separate behavioral traits to ensure a variety of interaction styles. We further categorize interruptions into seven types: agreement, disagreement, floor taking, tangentialization, clarification, assistance, and topic change, as follows Goldberg (1990). Based on the sampled traits of speakers, specific categories of interruptions are applied, ensuring realistic conversational dynamics.

Dialogue generation via LLMs. With narrative and speaker-specific conversational behaviors, we guide GPT-4o to generate complete multi-turn spoken conversations of 8-12 turns. To enhance the naturalness of the dialogue, we enable the LLM to depict *laughter*, using notations such as [laughter] or <laughter>yeah<laughter/> (indicating a laugh while saying "yeah") to represent it. If an interruption occurs during the conversation, the corresponding utterance is marked as

Interrupt for the speech synthesizing process.

Backchannel insertion. Backchannels typically occur at natural pause points in dialogue, often referred to as backchannel opportunity points (BOPs). Previous approaches (Kawahara, 2019) manually insert backchannels from a predefined set of responses (e.g., "hmm," "yeah") with no consideration of conversational context. In contrast, we use LLMs to automatically detect BOPs and generate diverse contextually appropriate backchannels based on predefined conversational behaviors, as illustrated in Figure 1.

Specifically, we first instruct LLMs to segment utterances at natural pause points to identify BOPs. Next, we select a subset of these BOPs based on the speaker’s backchannel behavior level. At these BOPs, we insert "speaker (backchannel): [MASK]", allowing LLMs to generate contextually appropriate backchannel responses by filling in the [MASK]. We use GPT-4o-mini for BOP detection and GPT-4o for generating backchannel responses.

Speech style captioning for TTS. We use GPT-4o-mini to generate speaking styles, including pitch, speaking rate, and emotion for each sentence. The generated styles are then input into the conditional TTS model, as described in §3.2.

3.2 Spoken Dialogue Generation from Text

Conditional TTS model. To convert rich text dialogues into spoken form, we utilize a CosyVoice-Instruct (Du et al., 2024), conditional text-to-speech (TTS) model that incorporates style at-

Dataset	Full-Duplex	Behavior Label	Public Accessibility	Category	# Dialogues	Audio (hrs)	Naturalness (\uparrow)	Emotion (\uparrow)	Sound Quality (\uparrow)
Fisher	✓			recorded	5,850	984	-	-	-
Switchboard	✓			recorded	2,400	260	-	-	-
CANDOR	✓		✓	recorded	1,650	850	3.73 ± 0.08	3.58 ± 0.07	3.54 ± 0.07
MELD	✓		✓	recorded	1,400	12	3.79 ± 0.07	3.68 ± 0.07	3.61 ± 0.08
DailyTalk			✓	recorded	2,541	20	3.89 ± 0.07	3.77 ± 0.07	3.82 ± 0.07
StyleTalk			✓	TTS-converted	2,364	7	3.88 ± 0.07	3.77 ± 0.07	3.78 ± 0.07
Behavior-SD	✓	✓	✓	TTS-converted	108,174	2,164	3.94 ± 0.07	3.78 ± 0.07	3.87 ± 0.06

Table 2: Statistics and human evaluation results of various spoken dialogue datasets. Citations are provided in the main text due to space constraints (§4).

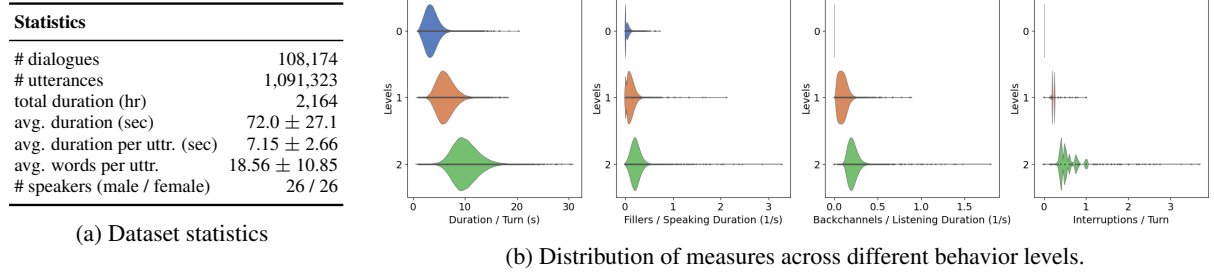


Figure 2: Overall statistics and distribution analysis of the dataset.

tributes specific to each sentence. We randomly sample 26 female and 26 male speakers. These speakers are then stored in a speaker bank, from which we randomly sample to construct the dataset. Further details about the speaker bank are provided in Appendix B.1. The final output is rendered as two-channel audio, where each channel corresponds to a single speaker in the dialogue.

Consistent utterance continuation. Each utterance is synthesized independently, so maintaining speaker consistency and smooth transitions is critical. To achieve this, we provide the previous utterance of the current speaker as additional input to the TTS model. CosyVoice-Base supports speech prompting, where generation starts from a provided speech token. To ensure continuity, we augment the input with the current utterance’s text τ_i , style instruction e_i , the speaker’s initial text and speech (τ_0, s_0) , and the previous utterance’s text and speech (τ_{i-1}, s_{i-1}) , forming the sequence: $[e_i, \tau_0, \tau_{i-1}, \tau_i, s_0, s_{i-1}, \text{<EOP>}]$. This structure enables temporal coherence and speaker consistency. Importantly, the input does not include the previous utterance of the other speaker. CosyVoice-Instruct, however, uses only style instructions without prompting. Our approach combines both, improving consistency and coherence across utterances.

Timing-driven utterance placement. The previously generated utterances are categorized as Backchannel, Interrupt, or None. When merging them, their timing is adjusted to mimic the natural flow of conversation. These placements are based on Gaussian distributions derived from statistics in a real-world spoken dialogue corpus (Reece et al., 2023). Specifically, none-type utterances are appended with an inter-utterance gap sampled from $\mathcal{N}(0.4s, 0.2s)$, backchannels follow after a brief delay sampled from $\mathcal{N}(0.2s, 0.02s)$, and interrupts overlap the preceding speech by a duration drawn from $\mathcal{N}(0.45s, 0.05s)$.

4 The Behavior-SD Dataset

Using the pipeline in the previous section, we contribute Behavior-SD (Behavior-driven Spoken Dialogues) as a large-scale, full-duplex dataset that captures a diverse range of social interactions and conversational behaviors. A sample from our dataset is shown in Figure 3.

Table 2 presents the statistics and human evaluation results of various spoken dialogue datasets. We compare our Behavior-SD with several existing datasets: human-recorded conversations from CANDOR (Reece et al., 2023), Switchboard (Godfrey et al., 1992), Fisher (Cieri et al., 2004), and MELD (Poria et al., 2019), human-read text conversations from DailyTalk (Lee et al., 2023), and text dialogues converted to speech using TTS from




 Narrative
Abigail is observant. He recognizes Caelyn’s scent of lavender and vanilla. He knows she uses the same shampoo and body wash.
 Behaviors
Abigail: V(0) F(1) B(1) I(0) Caelyn: V(0) F(0) B(0) I(2) → Interruption scenarios: Topic change, Assistance
 Dialogue
Abigail: Is that lavender and vanilla? Caelyn: Yes, my shampoo and body wash scent. Abigail: It’s... um, nice. Very relaxing. Caelyn: Thank you. {Yeah} [laughter] {[laughter]} I chose it for that reason. Abigail: I thought so. It fits the calming vibe... You know. Caelyn (interrupt): Speaking of calming, did you try the meditation app? Abigail: No, not yet. But you think it’s— Caelyn (interrupt): Helpful for stress, {oh, really?} definitely. Abigail: I’ve been meaning to, but it slipped my mind. Caelyn: Well, make sure to give it a try. Abigail: Yeah, sure. Sounds good.

Figure 3: A rich text dialogue sample of Behavior-SD.

StyleTalk (Lin et al., 2024). We run human evaluation on Amazon Mechanical Turk (MTurk), and ask participants to rate each 30-second audio samples on three aspects: dialogue naturalness, emotion appropriateness, and sound quality, using a 1-5 scale. Dialogue naturalness measures whether the conversation flow feels like a real-life interaction. Emotion appropriateness assesses whether the emotions in the dialogue align with the context. Sound quality evaluates how clear the audio is and the extent to which it is free of noise.

As shown in Table 2, Behavior-SD is the only publicly available dataset that includes speaker behavior labels, setting it apart from other datasets. Moreover, Behavior-SD contains over 100K dialogues and 2K hours of audio, far surpassing the sizes of the other datasets. The human evaluation demonstrates that Behavior-SD is preferred to other datasets in all three aspects. These findings indicate that Behavior-SD, with its larger scale and the inclusion of behavior labels, offers a distinct advantage over existing datasets. Moreover, the higher human evaluation scores of Behavior-SD in naturalness, emotion appropriateness, and sound quality further demonstrate its suitability for the research of behavior-rich spoken dialogues.

Figure 2 presents the detailed statistics of our dataset, showing the distribution of key conversa-

tional features across different behavior levels. It illustrates how varying behavior categories impact the length and structure of conversations. For example, the duration of utterances differs notably between behavior levels, indicating different conversational patterns. Additionally, features such as filler words, backchannels, and interruptions show distinct distribution trends across behavior levels.

5 BeDLM: Behavior-conditioned Dialogue Language Models

This section introduces BeDLM, a spoken dialogue language model designed to generate simultaneous two-channel speech conditioned on conversational behaviors and a narrative. Our BeDLM addresses several key challenges: (1) enabling LLMs to comprehend spoken dialogue forms beyond mere textual representations and (2) equipping LLMs to interpret conversational behaviors and effectively incorporate them into dialogue generation.

To leverage the rich capabilities of LLMs, we utilize a pre-trained LLM instead of training a transformer-based language model from scratch. This allows us to benefit that LLMs already has extensive knowledge of language patterns and dialogue structures, thereby allowing BeDLM to generate more nuanced and contextually appropriate outputs. In our work, we use Llama3.2-1B (Dubey et al., 2024; Meta, 2024) as a base model.

5.1 Speech Representations

Following previous approaches on SLMs (Zhang et al., 2023; Hassid et al., 2024; Nguyen et al., 2023), we utilize HuBERT (Hsu et al., 2021) to discretize speech signals into 500 unique units, which serve as a compact representation of speech. These units are incorporated into the LLM as additional tokens in its vocabulary, labeled $\langle S0 \rangle$, $\langle S1 \rangle$, ..., $\langle S499 \rangle$. To reconstruct audio from these discrete speech representations, we employ a two-stage approach: (1) HuBERT2Mel, which converts HuBERT representations into mel spectrograms, and (2) HiFi-GAN (Kong et al., 2020a), which synthesizes raw waveforms from the generated mel spectrograms.

5.2 Streamlined Spoken Dialogue Representations

To model full-duplex spoken dialogue, handling gaps and overlapping speech (*e.g.*, interruptions, backchannels) is crucial. Although dGSLM employs a dual-tower model to predict two channels

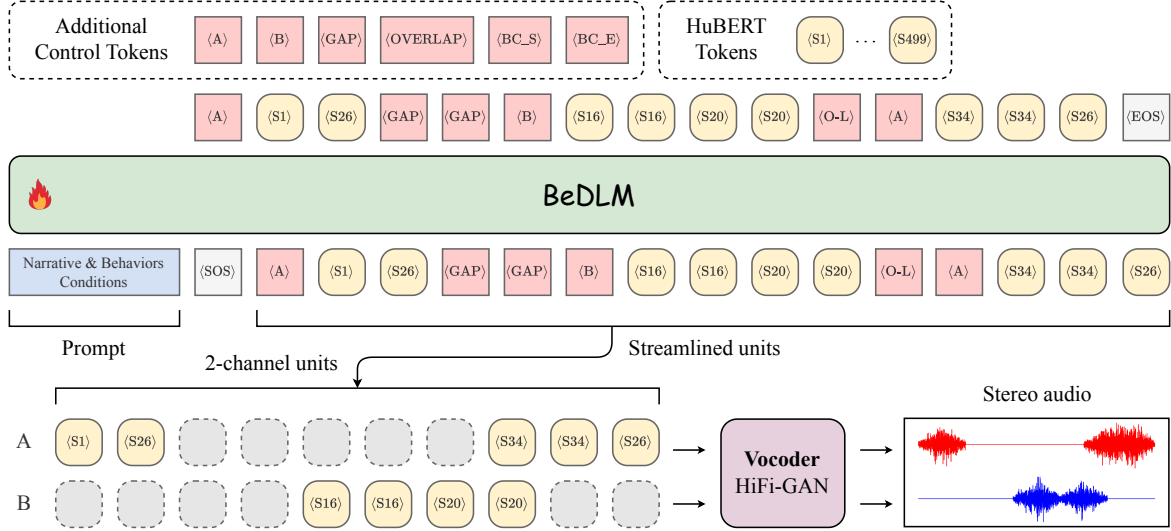


Figure 4: Overview of the proposed conditional spoken-dialogue generation pipeline. Narrative and behavior conditions serve as prompts for BeDLM to generate streamlined units, which are then converted into a two-channel unit sequence using additional control tokens (*e.g.*, ⟨A⟩, ⟨B⟩, ⟨GAP⟩). Each channel’s unit sequence is individually synthesized into audio by HiFi-GAN, and the resulting audio signals are combined to form stereo audio.

simultaneously, this approach inherently restricts the model’s ability to fully leverage the power of LLMs. Furthermore, alternating predictions between channels (*e.g.*, [A][B][A][B][A][B]) is inefficient, as it doubles the number of tokens required to represent the same duration of speech.

We introduce a novel duplex dialogue representation that uses control tokens to mark overlapping speech and backchannels. The tokens ⟨A⟩ and ⟨B⟩ indicate the beginning of each speaker’s turn. ⟨GAP⟩ and ⟨OVERLAP⟩ represent silences and overlaps, with each token covering 40ms. For instance, the sequence ⟨GAP⟩⟨GAP⟩⟨GAP⟩⟨A⟩ denotes a 120ms silence before Speaker A begins, whereas ⟨OVERLAP⟩⟨OVERLAP⟩⟨B⟩ indicates an 80ms period of overlapping speech during Speaker B’s turn. Backchannels are delineated by enclosing them between the ⟨BC_S⟩ and ⟨BC_E⟩ tokens, enabling precise control over conversational dynamics. More details are in Figure 6.

This method efficiently models two-channel dialogue into single-channel stream while preserving LLM performance, converting the tokenized sequences into two-channel audio via a vocoder for natural speech generation.

5.3 Conditional Spoken Dialogue Generation

To enable LLMs to generate spoken dialogues that align with a given narrative and various conversational behaviors, we train them using our Behavior-SD dataset. However, generating spoken dialogues

directly from a narrative is challenging, so we first conduct pre-training by generating text-based dialogues from the conditions. This process is done through supervised fine-tuning using the prompt specified in Figure 7.

Next, as illustrated in Figure 4, we fine-tune the LLM to directly generate streamlined spoken dialogue representations when provided with a prompt that includes the narrative and behavior conditions.

6 Experiments

In this section, we provide the details of model training and baselines for spoken dialogue generation.

6.1 Model Training

Finetuning HuBERT and HiFi-GAN. To enhance the expressiveness of BeDLM, we fine-tune the HuBERT (Hsu et al., 2021), HuBERT2Mel, and HiFi-GAN (Kong et al., 2020a) models using the Behavior-SD corpus. Specifically, the k-means quantizer of the HuBERT model is adapted to capture key acoustic features. The HuBERT2Mel module is then trained to convert the HuBERT representations into mel spectrograms. Finally, HiFi-GAN is fine-tuned to synthesize raw waveforms from the generated mel spectrograms. Training details for each module are presented in Appendix C.

BeDLM training. We first create a supervised-finetuning dataset from our Behavior-SD dataset

Models	LM settings		Mean Opinion Scores (\uparrow)		
	SLM (or LM)	Output Sequence	Naturalness	Meaningfulness	Sound Quality
Ground-Truth	-	-	4.14 \pm 0.08	4.02 \pm 0.08	4.12 \pm 0.08
Resynthesized	-	-	4.09 \pm 0.08	4.02 \pm 0.08	4.04 \pm 0.08
Cascaded	GPT-4o	Text dialogue	4.12 \pm 0.08	3.97 \pm 0.08	4.09 \pm 0.08
Cascaded	Llama3-70B	Text dialogue	4.08 \pm 0.09	3.97 \pm 0.09	4.10 \pm 0.09
dGSLM	DLM	2-channel units	3.96 \pm 0.09	3.88 \pm 0.09	3.92 \pm 0.09
BeDLM	Llama3.2-1B	Alternating units	3.90 \pm 0.09	3.86 \pm 0.09	3.85 \pm 0.10
BeDLM (<i>w/o pretrain</i>)	Llama3.2-1B	Streamlined units	4.00 \pm 0.08	3.89 \pm 0.08	3.99 \pm 0.09
BeDLM (ours)	Llama3.2-1B	Streamlined units	4.09 \pm 0.09	4.04 \pm 0.08	4.05 \pm 0.09

Table 3: Comparison of Naturalness, Meaningfulness, and Sound Quality across various models.

Models	LM settings		Adherence Score				
	SLM (or LM)	Output Sequence	N(\uparrow)	V(\downarrow)	F(\downarrow)	B(\downarrow)	I(\downarrow)
Ground-Truth	-	-	4.48	0.00	0.00	0.00	0.00
Resynthesized	-	-	4.48	0.12	0.20	0.12	0.10
Cascaded	GPT-4o	Text dialogue	4.58	0.40	0.18	1.73	0.91
Cascaded	Llama3-70B	Text dialogue	4.09	0.71	0.47	1.18	0.68
dGSLM	DLM	2-channel units	-	0.90	0.64	3.25	2.62
BeDLM	Llama3.2-1B	Alternating units	1.23	1.44	1.09	2.22	1.21
BeDLM (<i>w/o pretrain</i>)	Llama3.2-1B	Streamlined units	2.80	0.26	0.10	0.87	0.64
BeDLM (ours)	Llama3.2-1B	Streamlined units	3.11	0.25	0.15	0.87	0.58

Table 4: Comparison of adherence score across various models. Narrative adherence scores (N) ranging from 1-5, as rated by GPT-4o. Behavioral adherence scores are provided for various conversational behaviors: verbosity (V), filler word usage (F), backchannel usage (B), and interruption usage (I).

for the pre-training, we fine-tune the pre-trained Llama3.2-1B model over 15k iterations, utilizing 4 A40 GPUs. Training was performed with a total batch size of 32×2048 (65.5k tokens) and a learning rate of 5×10^{-5} .

In the spoken dialogue generation training stage, we use audio files from the behavior-SD dataset. Each two-channel audio is converted into streamlined units by incorporating control tokens based on backchannels, overlaps, and gaps. These streamlined units are paired with the corresponding narrative and behavior conditions to build a supervised fine-tuning dataset. In our experiments, we trained the model on 94K dialogues (1.87K hours) using 8 A40 GPUs with a total batch size of 32×4096 (130K tokens). Training spanned 18k iterations with a learning rate of 5×10^{-5} .

6.2 Baselines

To evaluate BeDLM’s ability to adhere to behavior conditions and generate natural spoken dialogues, we compared it against several baseline models.

GT and Resynthesized audios. We randomly select 500 narrative and behavioral instances from

the Behavior-SD test set, using their corresponding audio as the ground truth. Additionally, we resynthesize audio using HuBERT units derived from these ground-truth audios as a baseline for comparison.

dGSLM. We train dGSLM (Nguyen et al., 2023) on the Behavior-SD dataset. As dGSLM is an unconditional model for generating spoken dialogues, we add each speaker’s behavior condition tokens at the start of the speech units to guide generation. However, narrative conditions are not included in dGSLM’s inputs.

Cascaded LLM + TTS. We implement a baseline using Llama3-70B (Dubey et al., 2024) and GPT-4o (2024-08-06) to generate text dialogues based on narrative and behavior conditions, which are then converted to speech using CosyVoice-SFT (Du et al., 2024). Cascaded models often struggle to produce parsable text for spoken dialogue generation, complicating direct comparisons with BeDLM. To ensure fairness, we exclude conditions that cascaded models could not handle across all baselines and proposed models, focusing only

on conditions where all models could generate viable outputs. This approach ensures a balanced and fair comparison. The text generation prompt for cascaded models is detailed in Appendix D.

7 Results

7.1 Human Opinion Scores

For each spoken dialogue, three human evaluators assess the following aspects: dialogue naturalness, meaningfulness, and sound quality. Each sample is rated on a 5-point Likert scale, ranging from 1 to 5. Detailed instructions provided to the evaluators can be found in Appendix E.

Table 3 shows human evaluation scores, with BeDLM using streamlined units and pre-training achieving the best results, closely matching ground truth in naturalness and meaningfulness. Its sound quality is comparable to that of HuBERT and HiFi-GAN resynthesis. Cascaded models perform well in naturalness and sound quality but struggle with speaker confusion and misplaced backchannels, lowering meaningfulness. dGSLM consistently scores the lowest. BeDLM variants without pre-training or using alternating units show weaker performance, highlighting the importance of pre-training for natural, meaningful dialogue generation.

7.2 Adherence Scores

To evaluate the model’s adherence to narrative conditions, we use GPT-4o to score the alignment between the provided narrative and WhisperX (Bain et al., 2023) transcriptions on a 5-point scale (1=Bad, 5=Excellent), based on 100 samples. The evaluation prompts are shown in Figure 13.

For behavioral adherence, we measure key features (verbosity, filler words, backchannels, interruptions) for model-generated and ground-truth dialogues across 500 samples. We calculate the Wasserstein distance between the distributions of each behavior level and average the normalized distances to arrive at the final adherence score.

Table 4 compares the dialogue generation models based on narrative and behavioral alignment. Our model, BeDLM, excels in behavioral coherence, outperforming cascaded models and dGSLM in managing verbosity, backchannels, and interruptions. Although BeDLM’s narrative adherence (N) score is lower than cascaded models using larger LLMs like GPT-4o and Llama3-70B, it maintains a smoother conversational flow. BeDLM without

pre-training shows lower narrative adherence but similar behavioral alignment, while the alternating unit variant underperforms, highlighting the importance of streamlined units and pre-training.

7.3 Speaker Consistency.

To evaluate the effectiveness of our speaker conditioning method in maintaining speaker identity throughout a dialogue, we conducted experiments on 1,000 sampled utterances using randomly selected prompt speech. Table 5 presents the mean cosine similarity computed using the WavLM-Base+ model (Chen et al., 2021). The cosine similarity between (s_0, s_i) reflects the consistency of the speaker’s identity across the dialogue, while the cosine similarity between (s_{i-1}, s_i) indicates the smoothness of transitions between consecutive utterances. Including the first utterance s_0 as prompt speech is essential for both speaker consistency and smooth transitions, whereas incorporating the current speaker’s previous utterance, s_{i-1} does not have a significant impact. However, relying solely on the initial prompt s_0 can lead to perceptible discontinuities, often due to breathing sounds or moments of silence. To address this, our proposed method, which concatenates the initial and previous utterances (s_0, s_{i-1}) , yields the best performance in maintaining speaker consistency and smooth transition throughout the dialogue.

TTS Prompt	Cosine Similarity (\uparrow)	
	(s_0, s_i)	(s_{i-1}, s_i)
-	0.680 ± 0.09	0.687 ± 0.10
s_{i-1}	0.779 ± 0.15	0.850 ± 0.12
s_0	0.889 ± 0.09	0.867 ± 0.10
s_0, s_{i-1}	0.885 ± 0.09	0.872 ± 0.10

Table 5: Cosine similarity for different prompts

7.4 Qualitative results

Figure 5 shows two example waveforms from full-duplex dialogues generated by our BeDLM. Both follow the same narrative but differ in behavior conditions. In panel (a), the first speaker avoids using backchannels, whereas in panel (b), backchannels occur frequently. This clear contrast in waveforms demonstrates our model’s ability to capture varying backchannel conditions in conversation. Furthermore, our model effectively reflects the second speaker’s behavior conditions, such as interruptions and verbosity.

Narrative: Keilyn was critical of the holes he found in Danna's research. He felt that Danna had not been thorough in his work and that it reflected poorly on Keilyn.

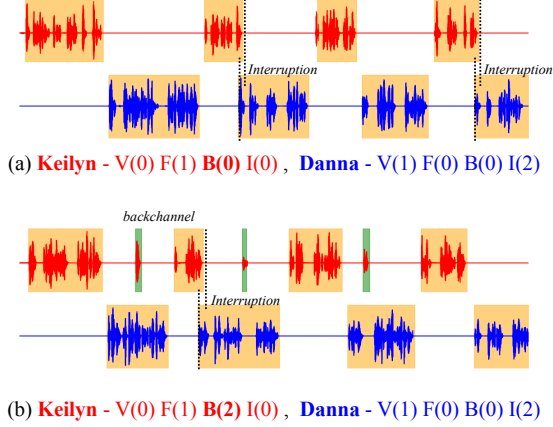


Figure 5: Waveforms of spoken dialogues generated by BeDLM with the same narrative but different behavior conditions. (a) shows no backchannels, while (b) includes frequent backchannels, reflecting the conditioned behaviors in each case.

8 Conclusion

We present a scalable framework for generating spoken dialogues with a diverse combination of conversational behaviors using LLMs and TTS models. Our contribution includes the Behavior-SD dataset, comprising over 100K spoken dialogues with behavior annotations, and the BeDLM, generating spoken dialogues that outperform baseline models in naturalness, diversity, and behavioral coherence. Future work could explore expanding multi-speaker dialogues, improving behavior control granularity, and integrating sophisticated real-time interaction capabilities to enhance practical applications in various spoken dialogue systems.

9 Limitations and Potential Risks

Limitation. One limitation of our system is the occasional mispronunciation, limited control over complex emotions, and limited non-lexical vocalizations other than laughter. To assess the pronunciation accuracy of Behavior-SD, we measure word error rates using the automatic speech recognition model, as detailed in Appendix B.3. Additionally, while the scale of our model and dataset are smaller than LLMs, they still effectively capture the semantics for generating coherent dialogues. However, this may impact the depth of nuance.

Potential Risks. The dataset is synthesized using LLMs and TTS models, which are known to inherit biases from their training data. As a result, Behavior-SD may not fully capture the nuances of diverse cultural, social, or linguistic variations in the real world, potentially leading to biased or unnatural dialogues. To address this issue, we designed our dataset generation process to incorporate diverse narratives from knowledge graphs and various conversational styles, ensuring broad representation. Additionally, we applied a Max-Min sampling algorithm to select a diverse and inclusive range of voices, as detailed in Appendix B.1. However, we acknowledge that real-world conversational complexity is vast, and continuous efforts are required to improve inclusivity and fairness.

Furthermore, while BeDLM generates human-like speech, it raises ethical concerns, particularly regarding deepfake audio, misinformation, and voice impersonation. To address this, we emphasize that our work is intended for research purposes and support the development of watermarking or detection mechanisms to safeguard against misuse in deceptive applications.

Acknowledgements

We want to thank the reviewers for their valuable feedback. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220156, Fundamental research on continual meta-learning for quality enhancement of casual videos and their 3D metaverse transformation), the IITP grant funded by the Korea government (MSIT) (No. RS-2019-II191082, SW StarLab), the SNU-Global Excellence Research Center establishment project, the IITP (Institute of Information & Communications Technology Planning & Evaluation)-ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2025-RS-2024-00437633), and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (RS-2023-00274280).

References

Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2023. Synthetic dialogue dataset generation using llm

- agents. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, page 181.
- Max Bain, Jaesung Huh, Tengda Han, and Andrew Senior. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.
- Janet B Bavelas, Linda Coates, and Trudy Johnson. 2000. Listeners as co-narrators. *Journal of personality and social psychology*.
- Elisabetta Bevacqua, Etienne de Sevin, Sylwia Julia Hyniewska, and Catherine Pelachaud. 2012. [A listener model: introducing personality traits](#). *Journal on Multimodal User Interfaces*, page 12.
- Peter Blomsma, Julija Vaitonytė, Gabriel Skantze, and Marc Swerts. 2024. Backchannel behavior is idiosyncratic. *Language and Cognition*.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, and Furu Wei. 2021. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv preprint arXiv:2110.13900*.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.
- Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: A resource for the next generations of speech-to-text. In *LREC*.
- Etienne De Sevin, Sylwia Julia Hyniewska, and Catherine Pelachaud. 2010. Influence of personality traits on backchannel selection. In *Intelligent Virtual Agents: 10th International Conference, IVA 2010, Philadelphia, PA, USA, September 20-22, 2010. Proceedings 10*, pages 187–193. Springer.
- Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. 2024. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.
- John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*.
- Julia A. Goldberg. 1990. Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics*, 14(6):883–903.
- Google. 2024. Notebook LM. <https://blog.google/technology/ai/notebooklm-audio-overviews/>.
- Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 2023. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP*, pages 1–5. IEEE.
- Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, et al. 2024. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. [Hubert: Self-supervised speech representation learning by masked prediction of hidden units](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.
- Rongjie Huang, Mingze Li, Dongchao Yang, Jia-tong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23802–23804.
- Tatsuya Kawahara. 2019. Spoken dialogue system for a human-like conversational robot erica. In *9th International Workshop on Spoken Dialogue System Technology*.
- Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *EMNLP*.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020a. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *NeurIPS*, 33:17022–17033.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020b. Hifi-gan: High-fidelity generative adversarial networks for text-to-speech synthesis. <https://github.com/jik876/hifi-gan>. Accessed: 2025-02-07.
- Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman

- Mohamed, et al. 2021. On generative spoken language modeling from raw audio. In *ACL*, volume 9, pages 1336–1354.
- Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Yinghao Aaron Li, Xilin Jiang, Jordan Darefsky, Ge Zhu, and Nima Mesgarani. 2024. Styletalker: Finetuning audio language model and style-based text-to-speech model for fast spoken dialogue generation. In *First Conference on Language Modeling*.
- Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. 2024. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. In *ACL*.
- Ziyang Ma, Yakun Song, Chenpeng Du, Jian Cong, Zhuo Chen, Yuping Wang, Yuxuan Wang, and Xie Chen. 2024. Language model can listen while speaking. *arXiv preprint arXiv:2408.02622*.
- Meta. 2024. Llama 3.2. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/.
- Kentaro Mitsui, Yukiya Hono, and Kei Sawada. 2023. Towards human-like spoken dialogue generation between ai agents from written dialogue. *arXiv preprint arXiv:2310.01088*.
- Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2023. Generative spoken dialogue language modeling. In *ACL*.
- OpenAI. 2024. GPT-4o Voice Mode. <https://openai.com/index/hello-gpt-4o/>.
- Mohamed Osman. 2022. Emo-tts:parallel transformer-based text-to-speech model with emotional awareness. In *2022 5th International Conference on Computing and Informatics (ICCI)*, pages 169–174.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL*, pages 527–536, Florence, Italy.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. 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*, pages 28492–28518. PMLR.
- Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*.
- Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. Spokenwoz: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *Advances in Neural Information Processing Systems*, volume 36, pages 39088–39118. Curran Associates, Inc.
- Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. 2023. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*.
- Peng Wang, Songshuo Lu, Yaohua Tang, Sijie Yan, Yuanjun Xiong, and Wei Xia. 2024. A full-duplex speech dialogue scheme based on large language models. *arXiv preprint arXiv:2405.19487*.
- Kenta Yamamoto, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara. 2018. Dialogue behavior control model for expressing a character of humanoid robots. In *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1732–1737.
- Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773.
- Xinrong Zhang, Yingfa Chen, Shengding Hu, Xu Han, Zihang Xu, Yuanwei Xu, Weilin Zhao, Maosong Sun, and Zhiyuan Liu. 2024. Beyond the turn-based game: Enabling real-time conversations with duplex models. *arXiv preprint arXiv:2406.15718*.

A Rich Text Dialogue Generation Pipeline

In this section, we offer further details outlined in Section 3.1 on generating rich text dialogues, including the prompts used for generation and the filtering processes involved.

A.1 Prompts for Rich Text Dialogue Generation

Dialogue generation. We use Figure 8 as the prompt for GPT-4o. At this stage, speaker names and narratives sampled from SODA are incorporated. From here, the process varies depending on the behavioral traits assigned to each speaker. Specifically, if both speakers are conditioned not to interrupt, we manually exclude any constraints related to interruption from the prompt. Additionally, each speaker’s behavior is mapped to text based on its level, as shown in Table 6, and incorporated into the prompt. The dialogue generated by GPT-4o based on this prompt is then parsed and proceeds to the next stage.

Backchannel insertion. As described in section 3.1, to insert backchannels, we first identify backchannel opportunity points (BOPs). To do this, we use the prompt in Figure 9 with GPT-4o-mini to find the natural pause points within each utterance. Then, based on the backchannel insertion level, we subsample the BOPs by selecting either 0–30% (Level 1) or 30–60% (Level 2) of the total.

To facilitate this, we insert placeholders such as "speaker1 (backchannel): [MASK1]" or "speaker2 (backchannel): [MASK2]" at the subsampled BOPs to guide the LLM in generating contextually appropriate backchannels at the specified positions. After adding masked utterances for each BOP into the original dialogue, we use the prompt from Figure 10 for GPT-4o.

Speech style captioning. To achieve more natural and emotionally rich speech synthesis, it is crucial to capture the speech style. Since elements like pitch, speed, and emotion can vary even within a single utterance, we do not obtain the speech style description at the utterance level. Instead, we utilize NLTK’s sentence tokenizer¹ to obtain speaking styles on a sentence-wise basis. For this, we use the prompt in Figure 11 to caption the pitch, speaking style, and emotion for each sentence.

¹https://www.nltk.org/api/nltk.tokenize.html#nltk.tokenize.sent_tokenize

A.2 Interruption Scenario Banks

To ensure that the LLM covers diverse interruption scenarios and that Behavior-SD is behaviorally rich, we provide examples of interruption scenarios in the prompt during dialogue generation. Specifically, we randomly select 1-3 samples from an interruption scenario bank. The scenario bank is composed of examples from seven categories of interruption (Goldberg, 1990): disagreement, floor taking, topic change, tangentialization, agreement, assistance, and clarification. Figure 14 shows interruption scenarios examples.

B TTS Pipeline

This section expands on the conversion of rich text dialogues into spoken dialogues in Section 3.2. We present methodologies for enhancing speaker consistency, improving the pronunciation of vocalization backchannels, and evaluating the pronunciation accuracy of synthesized speech.

B.1 Speaker Bank

When a speaking style is provided as an instruction, TTS models generate random voices corresponding to different speakers. However, this method does not ensure consistent speaker identity throughout a dialogue. To address this limitation, we propose selecting the prompt speech for the TTS model from a pre-sampled "speaker bank" to ensure speaker consistency across multiple utterances. For each synthesized utterance, our method ensures that subsequent utterances are generated using the same speaker’s voice.

To ensure a diverse and representative selection of speaker embedding, we applied a Max-Min sampling algorithm, which maximizes the minimum pairwise distance between embeddings to enhance speaker distinction. We applied a Max-Min sampling algorithm on a speaker bank of 30k samples to ensure diversity in speaker selection, ultimately choosing 26 male and 26 female speakers for the final set.

B.2 Enhancing Vocalization Backchannels through Voice Cloning

To mitigate issues related to suboptimal pronunciation of vocalization backchannels, such as "hmm" and "mhm," in the CosyVoice TTS system, we utilized ElevenLabs’ voice cloning² and text-to-

²<https://elevenlabs.io/voice-cloning>

Behavior type	Level	Text
Verbosity	0	{speaker} speak very briefly, sharing only essential content without any extra details
	1	{speaker} speak at a moderate length, providing a balanced amount of content—enough to convey your point clearly but without overexplaining
	2	{speaker} speak in detail, often providing lengthy explanations or descriptions whenever possible
Filler words	0	{speaker} never use filler words
	1	{speaker} moderately use filler words
	2	{speaker} frequently use filler words
Interruptions	0	{speaker} never interrupt
	1	{speaker} moderately interrupt
	2	{speaker} frequently interrupt

Table 6: Mapping of speaker behaviors to text based on levels.

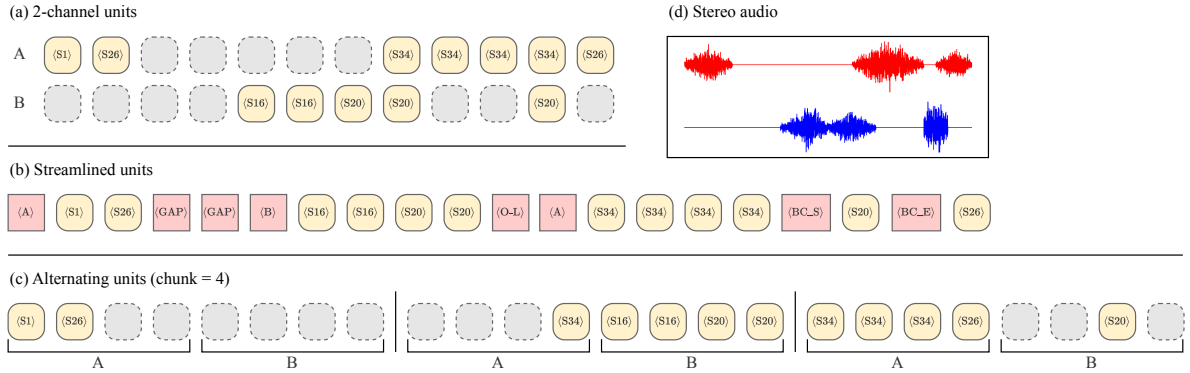


Figure 6: Examples of spoken dialogue representations: (a) two-channel units, (b) streamlined units, and (c) alternating units with a chunk size of 4. (d) These representations can be easily converted to stereo audio using a vocoder.

speech³ technologies. Voice cloning was applied to the 52 speakers in our dataset, and the backchannels were ranked by frequency. We selected the 100 most common backchannels for synthesis, with each one synthesized up to 100 times. These synthesized backchannels were then integrated into the dialogue synthesis process to enhance the naturalness of the system’s output.

B.3 Word Error Rate Analysis of TTS Output

We evaluate the pronunciation accuracy of Behavior-SD using Whisper-Large V3 (Radford et al., 2023) for automatic speech recognition (ASR). Specifically, we compute the word error rate (WER) by comparing the ASR-transcribed text with the ground-truth reference. On the test split of Behavior-SD, the model achieves a WER of 3.55%, whereas the original text-to-speech system, CosyVoice (Du et al., 2024), attained a lower

WER of 2.89% on the LibriSpeech test split. The relatively higher WER of Behavior-SD can be primarily attributed to distributional differences in the text, including the insertion of filler words and the misrecognition of proper names by the ASR model. These results indicate that the synthesized speech maintains high pronunciation accuracy, with errors largely stemming from conversational artifacts and occasional TTS mispronunciations. Further improvements in the pronunciation accuracy of Behavior-SD could be achieved by adopting a TTS model with better pronunciation accuracy in place of CosyVoice.

C BeDLM Details

C.1 Streamlined Spoken Dialogue Representations

Figure 6 presents various representations of full-duplex spoken dialogues. Our streamlined units effectively capture diverse aspects of full-duplex

³<https://elevenlabs.io/text-to-speech>

interactions, including turn-taking, overlapping speech, gaps, and backchannels.

C.2 Inference

During the inference stage of BeDLM, we utilize the same prompt detailed in Figure 7. We then autoregressively generate the next streamlined unit, using a temperature of 0.8, top- k sampling with k set to 60, and top- p sampling with p set to 0.8.

C.3 HuBERT

Due to computational constraints, we train only the k-means clustering component of HuBERT while adopting the dense model parameters from TWIST (Hassid et al., 2024). This process enhances HuBERT’s ability to encode speaker attributes in the Behavior-SD corpus, particularly in capturing laughter. To achieve robust clustering, the k-means model is trained for 25,600 iterations.

C.4 HuBERT2Mel

HuBERT2Mel converts HuBERT token sequences, in combination with a speaker identity embedding, into mel spectrograms that correspond to the target speaker’s voice. These spectrograms are then used to fine-tune HiFi-GAN, improving both the quality of waveform generation and training efficiency compared to unit-based HiFi-GAN approaches.

The model predicts mel spectrograms derived from 22,050 Hz speech, utilizing a hop size of 256, a window size of 1024, and an FFT size of 1024, with frequency limits of 0 Hz (f_{min}) and 8,000 Hz (f_{max}). HuBERT2Mel processes token and speaker embeddings through a combination of transformer and convolutional layers incorporating gated linear units (GLU). The model then applies upsampling to align its output length with the mel spectrogram’s temporal resolution. Finally, mel spectrogram predictions are generated via fully connected layers with residual connections.

Training is conducted for 60,000 steps using an L1 loss function, optimizing the model toward ground-truth mel spectrograms.

C.5 HiFi-GAN

We fine-tune a pre-trained HiFi-GAN model (Kong et al., 2020a) using pairs of original speech waveforms and their corresponding mel spectrograms, which are generated by first converting the speech into HuBERT tokens and then processing them through HuBERT2Mel. For initialization, we adopt

the Universal-V1 checkpoints from the official implementation (Kong et al., 2020b). Fine-tuning is performed on a 1/100 subset of the Behavior-SD training split for 100,000 steps, refining the model to enhance the perceptual quality of BeDLM-generated speech.

During inference, HiFi-GAN processes mel spectrograms generated by HuBERT2Mel and synthesizes raw audio at a sampling rate of 22,050 Hz, ensuring high-quality speech reconstruction.

D Cascaded Model Description

The text dialogue generation prompts used for the cascaded models are illustrated in Figure 12.

E Human Evaluation Guidelines

The evaluators are instructed as follows:

- **Dialogue Naturalness:** Are backchannels and laughter appropriately included to create a human-like interaction? Is there a seamless transition between the speaker and listener at the right moments? Does the conversation flow smoothly and not awkward?
- **Meaningfulness:** Does the dialogue have meaningful content, and is it possible to understand what is being said?
- **Sound Quality:** Is the sound clear and easy to hear, free from noise or other distractions?

These instructions are adapted from (Mitsui et al., 2023).

Variables
narrative, speaker1, speaker2, speaker1_behaviors, speaker2_behaviors, text_dialogue (or streamlined_units)
Prompts for LLM
<p>Generate a dialogue between two speakers based on the given narrative. Follow the specific behaviors for each speaker.</p> <p>Narrative:</p> <ul style="list-style-type: none"> - {narrative} <p>{speaker1} behaviors:</p> <p>{speaker1_behaviors}</p> <p>{speaker2} behaviors:</p> <p>{speaker2_behaviors}</p> <p>Ensure that the dialogue reflects both the narrative and the specified behaviors for each speaker</p> <p>{% < SOT >{text_dialogue}< EOS > if pretraining phase else < SOS >{streamlined_units}< EOS > %}</p>

Figure 7: A prompt designs narrative and conversational behaviors for conditional generation. During the pretraining phase, the corresponding text dialogues are used as the target for the given condition.

Variables
speaker1, speaker2, narrative, speaker1_behaviors, speaker2_behaviors, interrupt_scenarios
Prompts for GPT-4o (2024-08-06)
<p># Instructions</p> <p>Your task is to write a natural and conversational dialogue transcription between two persons.</p> <p>Ensure that the generated dialogue should follow the constraints and output format outlined below.</p> <p># Constraints</p> <ul style="list-style-type: none"> - The conversation should be between two persons ({speaker1} and {speaker2}). - The conversation should follow the narrative provided. - The speakers use non-lexical vocalizations like [laughter], <laughter>yeah</laughter>. - The conversation should not contain any non-verbal gestures, only verbal responses. - The speakers should behave according to the behaviors described below. - The conversation should include interruption scenarios (cut-off and take a turn). - Interruption utterances are marked by appending "(interrupt)" after the speaker's name, and the other speaker's previous cut-off utterance should be marked with [interrupted] to indicate where the interruption occurred. - The conversation should include the below interruption scenarios: <p>{interrupt_scenarios}</p> <ul style="list-style-type: none"> - The conversation should contain 8-12 utterances. <p># Narrative</p> <p>{narrative}</p> <p># Speakers behaviors</p> <p>{speaker1_behaviors}</p> <p>{speaker2_behaviors}</p> <p># Output</p> <p>The conversation should be in the following format:</p> <p><Format></p> <p>Here is a generated dialogue (N turns):</p> <p>1) {speaker1}: utterance</p> <p>2) {speaker2}: utterance</p> <p>3) {speaker1}: utterance</p> <p>...</p> <p>N) {speaker2}: utterance</p> <p></Format></p>

Figure 8: A prompt for dialogue generation.

Variables
utterance
Prompts for GPT-4o-mini (2024-07-18)
Instructions Your task is to reformat the given utterance according to the constraints below:
Constraints - Split the utterance into smaller parts at natural pause points, such as after commas, conjunctions, or at the end of phrases. - Each smaller part becomes its own line.
Example ## Original utterance Jenifer: They do, which feels really rewarding. I've built up a good reputation because of my love and attention to [interrupted]
Reformatted Dialogue Jenifer: They do, Jenifer: which feels really rewarding. Jenifer: I've built up a good reputation Jenifer: because of my love and attention to [interrupted]
Task ## Original utterance {utterance}

Figure 9: A prompt for backchannel opportunity points detection.

Variables
speaker1, speaker2, dialogue
Prompts for GPT-4o (2024-08-06)
<p># Instructions</p> <p>Your task is to insert appropriate backchannels into the provided dialogue at the locations marked "{speaker1} (backchannel): [MASK1]" or "{speaker2} (backchannel): [MASK2]".</p> <ul style="list-style-type: none"> - Only place backchannels in these marked locations. - A backchannel is a vocalization or short word/phrase that shows the speaker is engaged and listening. - Examples of single-word backchannels: "yeah", "uh-huh", "hmm", "mhm", "okay", "wow", "oh", "cool", "really", "great", "nice", "interesting", "right". - Examples of multi-word backchannels (no more than 3 words): "yeah, yeah", "okay, okay", "oh, really?", "that's great". - Examples of vocalization backchannels: "[laughter]", "<laughter>yeah</laughter>". - If a backchannel is needed, replace [MASK1] or [MASK2] with an appropriate backchannel from the examples or a similar expression. - Ensure the number of inserted backchannels follows the specified constraints. - The backchannels should be appropriate for the context and the speaker's style. <p># Example 1</p> <p>## Given dialogue</p> <p>A: I just found out that</p> <p>A: there's a new art exhibit downtown.</p> <p>B (backchannel): [MASK2]</p> <p>A: It's all about abstract sculptures.</p> <p>B (backchannel): [MASK2]</p> <p>A: I think it might be really interesting to check out.</p> <p>B: That sounds amazing. When are you planning to go?</p> <p>## Response</p> <p>A: I just found out that</p> <p>A: there's a new art exhibit downtown.</p> <p>B (backchannel): Really?</p> <p>A: It's all about abstract sculptures.</p> <p>B (backchannel): Oh, cool.</p> <p>A: I think it might be really interesting to check out.</p> <p>B: That sounds amazing. When are you planning to go?</p> <p># Example 2</p> <p>## Given dialogue</p> <p>A: I just found out that there's a new art exhibit downtown. It's all about abstract sculptures. I think it might be really interesting to check out.</p> <p>B: That sounds amazing.</p> <p>A (backchannel): [MASK1]</p> <p>B: When are you planning to go?</p> <p>## Response</p> <p>A: I just found out that there's a new art exhibit downtown. It's all about abstract sculptures. I think it might be really interesting to check out.</p> <p>B: That sounds amazing.</p> <p>A (backchannel): Yeah!</p> <p>B: When are you planning to go?</p> <p># Task</p> <p>## Given dialogue</p> <p>{dialogue}</p>

Figure 10: A prompt for backchannel insertion.

Variables
dialogue
Prompts for GPT-4o-mini (2024-07-18)
<p># Instructions</p> <p>Your task is to generate captions for each utterance in the given dialogue. Follow the format and use only the specified values for pitch, speed, and emotion.</p> <p># Constraints</p> <ul style="list-style-type: none"> - Pitch: low, normal, high - Speed: slow, normal, fast - Emotion: neutral, happy, sad, angry, fearful <p># Output</p> <ul style="list-style-type: none"> - The captions are appended to the end of each utterance in the following format: - ('Pitch' pitch, 'Speed' speed, 'Emotion' emotion) <p># Example</p> <p>## Given dialogue:</p> <p>Rylea: Hey, man, what's going on?</p> <p>Shavon (backchannel): Uh-huh.</p> <p>Rylea: You seem really down.</p> <p>Rylea: Is everything okay?</p> <p>Shavon: Honestly, no... I'm just having a hard time right now.</p> <p>## Response:</p> <p>Rylea: Hey, man, what's going on? (normal pitch, normal speed, neutral emotion)</p> <p>Shavon (backchannel): Uh-huh. (normal pitch, fast speed, neutral emotion)</p> <p>Rylea: You seem really down. (normal pitch, normal speed, sad emotion)</p> <p>Rylea: Is everything okay? (normal pitch, normal speed, neutral emotion)</p> <p>Shavon: Honestly, no... I'm just having a hard time right now. (normal pitch, normal speed, sad emotion)</p> <p># Task</p> <p>## Given dialogue:</p> <p>{dialogue}</p>

Figure 11: A prompt for speech style captioning.

Variables
speaker1, speaker2, narrative, speaker1_behaviors, speaker2_behaviors
Prompts for LLM
Instructions Your task is to write a natural and conversational dialogue transcription between two persons. Ensure that the generated dialogue should follow the constraints and output format outlined below. # Constraints <ul style="list-style-type: none"> - The conversation should be between two persons ({speaker1} and {speaker2}). - The conversation should follow the narrative provided. - The conversation should not contain any non-verbal gestures, only verbal responses. - The listener's backchannel should be included within braces {{ }} to indicate their position. For example: You know, {{Hmm?}} I always carry a book with me. {{Oh, really?}} There's just something about having a story... - The conversation should include interruption scenarios (cut-off and take a turn). - Interruption utterances are marked by appending "(interrupt)" after the speaker's name, and the other speaker's previous cut-off utterance should be marked with [interrupted] to indicate where the interruption occurred. - The speakers should behave according to the behaviors described below. - The conversation should contain 8-12 utterances. # Narrative {narrative} # Speakers behaviors {speaker1_behaviors} {speaker2_behaviors} # Output The conversation should be in the following format: <Format> Here is a generated dialogue (N turns): 1) {speaker1}: [utterances] {{backchannels}} [utterances] {{backchannels}} ... 2) {speaker2}: [utterances] {{backchannels}} [utterances] {{backchannels}} ... 3) {speaker1}: [utterances] {{backchannels}} [utterances] {{backchannels}} N) {speaker2} : utterance </Format>

Figure 12: A prompt for cascaded model.

Variables
narrative, transcription
Prompts for GPT-4o (2024-08-06)
Evaluate the spoken dialogue between two speakers based on how well it aligns with the given narrative. The speaker names might be slightly wrong due to transcription, so focus on the overall content. Use a scale of 1 to 5, where: <ul style="list-style-type: none"> - 1 (Bad): The dialogue is entirely irrelevant to the narrative at any point. - 2 (Poor): The dialogue is loosely related to the narrative in a few minor points. - 3 (Fair): The dialogue is related to the narrative in some main points. - 4 (Good): The dialogue follows the narrative in the main points but sometimes deviates from the focus. - 5 (Excellent): The dialogue fully aligns with the narrative and does not deviate from it. Narrative: {narrative} Dialogue:: {transcription} The response should be a only single digit from 1 to 5.

Figure 13: A prompt for GPT-4o to evaluate the narrative adherence score.

<p>Disagreement</p> <p>When the listener interrupts to take the control of the interaction, and disrupts the flow of dialogue, which can be seen as a conflict.</p> <p>Speaker A: I think our team should focus on the marketing campaign next month [interrupted]</p> <p>Speaker B (interrupt): No, that's not the priority right now. We need to fix our product issues first [interrupted]</p> <p>Speaker A (interrupt): But if we don't market, we won't have customers to use the product.</p>	<p>Tangentialization</p> <p>The listener grabs the turn and sums up the information received from the current speaker to prevent listening to more unwanted information.</p> <p>Speaker A: The quarterly report indicates that we have a 15% increase in sales, which is due to the new marketing strategy we implemented. Additionally, we have [interrupted]</p> <p>Speaker B (interrupt): So, sales are up 15% thanks to marketing. Got it. What's the next step?</p> <p>Speaker A: I was going to explain the other factors, but we can move on.</p>
<p>Clarification</p> <p>The listener expects the current speaker to clarify or explain the information about which the listener is not clear.</p> <p>Speaker A: Our latest strategy involves leveraging social media influencers to [interrupted]</p> <p>Speaker B (interrupt): So, sales are up 15% thanks to marketing. Got it. What's the next step?</p> <p>Speaker A: Yes, exactly. That will help us meet their expectations better.</p>	<p>Assistance</p> <p>The listener interrupts to provide the current speaker with a word, a phrase, or an idea to help complete the utterance.</p> <p>Speaker A: We need to focus on the customer feedback to improve our [interrupted]</p> <p>Speaker B (interrupt): Products and services.</p> <p>Speaker A: Yes, exactly. That will help us meet their expectations better.</p>
<p>Floor Taking</p> <p>The listener grabs the floor and expands on the current speaker's topic.</p> <p>Speaker A: Our project timeline needs to be adjusted because [interrupted]</p> <p>Speaker B (interrupt): Absolutely, we need at least two more weeks to ensure quality control and testing.</p> <p>Speaker A: And we should also consider adding more team members to speed up the process.</p>	<p>Topic Change</p> <p>The listener grabs the turn and changes the current topic of conversation.</p> <p>Speaker A: We need to discuss the budget allocation for the upcoming quarter [interrupted]</p> <p>Speaker B (interrupt): Speaking of budgets, did you hear about the new software upgrade we're getting next month?</p> <p>Speaker A: Oh, that sounds crucial. Let's dive into the details of the software upgrade then.</p>
<p>Agreement</p> <p>The listener shows understanding or support to the speaker.</p> <p>Speaker A: The new policy will help improve our workflow efficiency [interrupted]</p> <p>Speaker B (interrupt): Exactly, it will streamline our processes significantly.</p> <p>Speaker A: And reduce the time spent on manual tasks.</p>	

Figure 14: Samples from the interruption scenario bank.