BrainSpeech: Parallel Semantic–Acoustic Generation for SpeechLMs

Anonymous ACL submission

Abstract

Recent advances in Speech Language Models (SpeechLMs) have enabled end-to-end spoken dialogue systems. However, most existing SpeechLMs rely on a single-stream or sequential modeling paradigm, in which semantic reasoning and acoustic generation are either flattened into one stream or processed in a fixed Thinker–Talker order. Both paradigms constrain the interaction between semantic reasoning and acoustic generation, limiting the naturalness and expressiveness of spoken dialogue. To address this, we propose BrainSpeech, a novel SpeechLM framework that models semantic reasoning and acoustic generation as two parallel yet interacting streams within a unified large language model. To enable effective cross-stream interaction, we introduce a dedicated attention mechanism that explicitly exchanges semantic and acoustic information during generation. Building on this dual-stream formulation, we further develop a three-stage training strategy that preserves the reasoning capability of the underlying language model while reducing the reliance on large-scale spoken dialogue corpora. In addition, we design a streaming decoding strategy that supports real-time generation of continuous, high-fidelity speech. Experiments on multiple spoken dialogue benchmarks demonstrate that BrainSpeech produces more natural and expressive speech and achieves superior speech-to-speech performance compared to larger open-source SpeechLMs[1].

1 Introduction

Spoken dialogue systems (Long et al., 2025) play a crucial role in natural human–computer interaction by lowering interaction barriers and enabling more intuitive and seamless user experiences. Traditional speech interaction pipelines (Wang et al., 2024a) typically consist of three dedicated components: Automatic Speech Recognition (ASR) (Saini and Kau, 2013) for transcribing speech, a large language model (LLM) (Tsai et al., 2023) for generating textual responses, and Text-to-Speech (TTS) (Liu et al., 2025) for synthesizing speech outputs. However, such cascaded architectures suffer from accumulated interaction latency and limited modeling of emotional prosody. To overcome these limitations, recent studies have increasingly explored end-to-end Speech Language Model (SpeechLM)-based frameworks that directly map speech inputs to speech responses (Xie and Wu, 2024a; Chen et al., 2025b; Fang et al., 2024). In these systems, speech inputs are first encoded by an audio encoder and aligned to the LLM's textual representation space via an adaptor, after which speech responses are generated using either discrete audio tokens (Zeng et al., 2024) or continuous hidden states (Wang et al., 2024b) produced by the LLM.

Existing end-to-end SpeechLMs can be categorized into sequential-stream and single-stream-based systems depending on the manner in which understanding and generation are carried out. Sequential-stream-based systems (Yang et al., 2024; Chen et al., 2025a; Long et al., 2025; Wang et al., 2024b; Fang et al., 2025) perform text-based semantic reasoning before generating speech from text tokens or continuous semantic representations. These systems better preserve the comprehension capabilities of LLMs but tend to lose paralinguistic information such as emotional and prosodic cues. Single-stream-based systems (Zeng et al., 2024; Chen et al., 2025b; Xie and Wu, 2024b; KimiTeam et al., 2025; Wu et al., 2025) employ a unified LLM to simultaneously perform thinking and generation, generate text and audio tokens in a parallel or interleaved manner. Although these systems are advantageous for modeling paralinguistic information,

[1] Audio samples are available at https://brainspeech.github.io/

they face challenges in preserving the knowledge of the LLM.

In contrast, both single-stream and sequential-stream designs deviate from the dynamics of human spoken communication, where thinking and speaking are supported by distinct yet tightly coupled cognitive processes that unfold concurrently. Motivated by this observation, we propose BrainSpeech, a novel dual-stream SpeechLM architecture that performs semantic reasoning and acoustic generation in parallel within a unified large language model. To facilitate effective coordination between the two streams, we introduce a dedicated cross-stream attention mechanism that enables explicit information exchange between semantic and acoustic representations during generation. Combined with a three-stage training strategy, this design preserves the reasoning capability of the underlying language model while enhancing implicit emotional expressiveness, without relying on scarce spoken dialogue corpora. Notably, our training relies solely on open-source textual dialogue data and TTS datasets. Furthermore, to support seamless and natural streaming speech synthesis, we develop an overlapping-buffer-based decoding algorithm that smooths inter-segment transitions via waveform splicing, together with a forced-chunking strategy that stabilizes processing latency through iterative batch accumulation.

In conclusion, our contributions are as follows:

- We propose BrainSpeech, the first SpeechLM framework built upon a dual-stream architecture that simultaneously executes semantic and acoustic generation in separate streams within a unified LLM.
- A dedicated attention mechanism is designed to enable explicit interaction between semantic and acoustic information across the dual-stream architecture.
- Our model architecture, along with our three-stage training strategy, ensures system performance and eliminates the reliance on large-scale spoken dialogue corpora.
- For inference, we develop an overlapping-buffer-based audio decoding algorithm that smooths inter-segment transitions, alongside a forced-chunking method to ensure real-time generation of continuous, high-fidelity audio.
- Extensive experiments demonstrate that our model even matches the understanding capability of larger models and shows superior performance on acoustic quality and emotional expressiveness among all existing SpeechLMs.

2 Related Work

Sequential-stream-based SpeechLMs, including models such as MinMo (Chen et al., 2025a), Qwen2.5-Omni (Yang et al., 2024), and LLaMA-Omni 2 (Fang et al., 2025), retain the original LLM architecture to preserve textual capabilities, using its hidden states as input to a speech decoder for audio generation. Although effective in preserving LLM knowledge, these models struggle to capture rich paralinguistic acoustic attributes such as emotion and prosody, as only text tokens are used for autoregressive modeling.

Single-stream-based SpeechLMs employ a unified LLM to simultaneously perform thinking and generation, producing text and audio tokens in an interleaved or parallel manner. For the interleaved manner, as seen in GLM-4-Voice (Zeng et al., 2024), VITA-Audio (Long et al., 2025) and Step-Audio 2 (Wu et al., 2025), the model alternates between text and audio tokens during generation, typically requiring large-scale interleaved speech-text data and extensive pre-training. For the parallel manner, adopted by Mini-Omni (Xie and Wu, 2024a) and SLAM-Omni (Chen et al., 2025b), text and audio tokens are modeled autoregressively in parallel. However, these systems tend to compromise the retention of original LLM knowledge.

3 Methodology: BrainSpeech

Inspired by human cognitive processes, where *what to say* and *how to say* occur simultaneously via distinct yet interdependent mechanisms, we introduce BrainSpeech, a novel SpeechLM which employs a dual-stream architecture to decouple semantic and acoustic generation within a unified framework. This design preserves the extensive knowledge of LLMs and enables the acoustic generation module to be trained independently on standard speech corpora, eliminating the dependency on scarce spoken dialogue data.

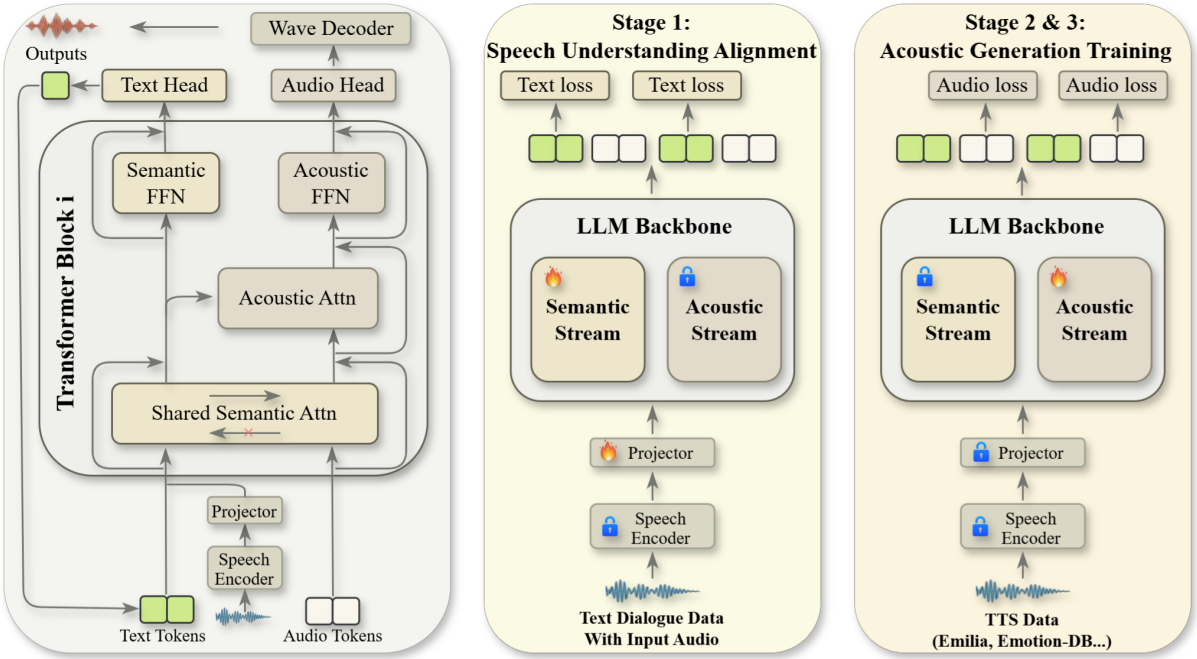

Figure 1: The overall framework and training paradigm of BrainSpeech. (a) Model Architecture. Text and audio tokens are processed by the Semantic Stream (Shared Semantic Attn + Semantic FFN) and Acoustic Stream (Shared Semantic Attn + Acoustic Attn + Acoustic FFN), respectively. (b) Stage 1: Speech Understanding Alignment. Trains the Semantic Stream and Projector to enable speech understanding. (c) Stages 2 & 3: Acoustic Generative Training. Trains the Acoustic Stream on TTS data to enable speech generation.

3.1 Model Architecture

As illustrated in Figure 1, the architecture of BrainSpeech is a dual-stream architecture LLM that maintains semantic and acoustic processing streams within each block. To enable speech understanding, speech queries are first processed by the pre-trained speech encoder (Radford et al., 2023) and mapped to the LLM's space through a trainable linear projector. Then, the hidden states of input speech and text tokens are processed sequentially through semantic attention and Feed-Forward Networks (FFN) inherited from LLMs, focusing exclusively on semantic generation. In contrast, the output audio tokens traverse a specialized path for acoustic generation. They first utilize the shared semantic attention to extract high-level semantic guidance, followed by a dedicated acoustic attention to model acoustic information (e.g. prosody and pitch) and finally acoustic FFN to integrate and refine the modeled acoustic information for coherent acoustic representation synthesis. This design ensures the effective integration of semantic and acoustic information. Finally, the generated discrete audio tokens are converted into continuous speech waveforms using the decoder of XCodec (Ye et al., 2025).

3.2 Sequence Format and Terminal Control

In natural human communication, textual responses are typically accompanied by corresponding speech signals to enhance interactive naturalness. Therefore, BrainSpeech generates responses using an interleaved sequence format:

$$\mathbf{T}_{s1}, \texttt{<b>}, \mathbf{A}_{s1}, \texttt{<e>}, \ldots, \mathbf{T}_{sl}, \texttt{<lb>}, \mathbf{A}_{sl}, \texttt{<e>} \quad (1)$$

To support efficient streaming generation, BrainSpeech imposes a fixed token budget (10:40) for text and audio tokens within intermediate segments. However, a critical challenge is to prevent additional audio token generation in intermediate segments, as this would disrupt the interleaved structure and cause a divergence from the training distribution. We address this by introducing a special token <lb> (last begin) to explicitly identify the final segment. This control mechanism strictly constrains generation to a fixed token count for intermediate segments, effectively deferring the generation of remaining audio tokens until the <lb> token is predicted, preserving structural integrity during inference.

3.3 Asymmetric Causal Masking

Given the interleaved format, BrainSpeech employs an asymmetric attention mask to disentangle the generation dependencies. Since *what to say* (text) should remain independent of *how to say* (audio), BrainSpeech masks text tokens from attending to audio tokens. This unidirectional constraint preserves the purity of semantic generation and prevents contamination from the update of acoustic features. Formally, let $\mathbf{H}^l = [\mathbf{h}_1^l, \ldots, \mathbf{h}_N^l]$ denote the hidden states of layer l. For the text token at index i, the update process is constrained to:

$$\begin{aligned} \tilde{\mathbf{h}}_i^l &= \text{SemanticAttn}\left(\mathbf{h}_i^l, \{\mathbf{h}_j^l : j \in \mathbf{T}, j \leq i\}\right) \\ \mathbf{h}_i^{l+1} &= \text{SemanticFFN}(\tilde{\mathbf{h}}_i^l) \end{aligned} \tag{2}$$

Audio tokens utilize a standard causal mask to attend to all preceding text and audio context, thereby ensuring the acoustic features are semantically aligned with text while maintaining prosodic coherence.

3.4 Training Paradigm

For the training paradigm, we employ a progressive three-stage training pipeline. This strategy eliminates the need for large-scale spoken dialogue data by decoupling the optimization of understanding and generation capabilities.

Stage 1: Speech Understanding Alignment. The primary goal is to equip LLMs with the capabilities of speech understanding. We denote the parameters of the projector as $\theta_{\mathcal{P}}$, the semantic stream as $\theta_{\mathcal{S}}$, and the acoustic stream as $\theta_{\mathcal{A}}$. In this stage, we freeze the speech encoder and maximize the likelihood of generating the correct text response $\mathbf{T}$ given the input speech query $\mathbf{S}$. The optimization objective is:

$$\mathcal{L}_{\text{Stage 1}} = -\sum_{i=1}^{|\mathbf{T}|} \log P(\mathbf{T}_i | \mathbf{T}_{<i}, \mathbf{S}; \theta_{\mathcal{P}}, \theta_{\mathcal{S}}) \tag{3}$$

where $\theta_{\mathcal{A}}$ is not involved, and BrainSpeech is optimized to bridge the modality gap in this stage.

Stage 2: Acoustic Generation Pre-training. This stage trains the acoustic stream to generate audio tokens conditioned on the text response $\mathbf{T}$. Specifically, we freeze $\theta_{\mathcal{S}}$ to preserve the capabilities of understanding and only update $\theta_{\mathcal{A}}$. The training objective is to maximize the conditional log-likelihood of the target audio sequence

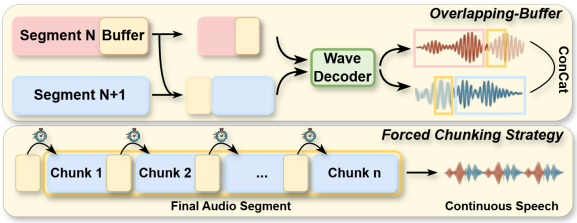

Figure 2: Inference of BrainSpeech.

$\mathbf{A}$ given $\mathbf{T}$. We minimize the following negative log-likelihood loss:

$$\mathcal{L}_{\text{Stage 2\&3}} = -\sum_{t=1}^{|\mathbf{A}|} \log P(\mathbf{A}_t | \mathbf{A}_{<t}, \mathbf{T}; \theta_{\mathcal{A}}) \tag{4}$$

This stage establishes a robust mapping from semantic content to acoustic features.

Stage 3: Expressive Acoustic Generation Fine-tuning. To enhance both acoustic fidelity and emotional expressiveness, we fine-tune $\theta_{\mathcal{A}}$ using a composite dataset. Similar to Stage 2, $\theta_{\mathcal{S}}$ remains frozen. This stage enables BrainSpeech to produce high-quality speech, effectively capturing context-aware prosody and emotion inferred from the text.

3.5 Inference of BrainSpeech

Generating long-form audio in interleaved segments would introduce acoustic artifacts. As shown in Figure 2, we develop specific decoding strategies to ensure high-fidelity speech while enabling continuous real-time streaming.

Overlapping-Buffer: To address boundary discontinuities caused by independent decoding in a segment, we introduce overlapping-buffer. Specifically, we concatenate the trailing tokens of the preceding segment to the current segment, ensuring continuity between segments. We achieve seamless transitions by splicing waveforms at the midpoint of the buffer, preserving the initial half from the prior segment while utilizing the second half from the current segment. Finally, the buffer is refreshed with the current segment's trailing tokens for the next iteration.

Forced-Chunking: The final audio segment frequently exhibits significant length variation. To ensure stable and real-time decoding, we employ forced-chunking. Tokens are accumulated and processed in chunks aligned with the intermediate segment size, iteratively applying overlapping-buffer to guarantee robust generation independent of total duration.

4 Data Construction

To train BrainSpeech effectively across understanding and generation tasks without relying on large-scale spoken dialogue data, we construct specialized datasets for each training stage. Detailed formatting examples for each stage are provided in Appendix A.

4.1 Data of Speech Understanding Alignment

To align the speech encoder with LLMs, we derive data from high-quality text-only instruction datasets (e.g., InstructS2S (Fang et al., 2024)). We employ commercial TTS models (Higgs (Boson AI, 2025) and Minimax-Speech-02) to synthesize questions into speech. Then, the original text query is replaced by the continuous speech embeddings extracted by the speech encoder. This effectively enables the model to comprehend speech by leveraging its established textual understanding capabilities.

4.2 Data of Acoustic Generation Pre-training

For acoustic generation pre-training, we utilize Emilia (He et al., 2024), a TTS dataset with 2w hours, as training data. To follow the dialogue format required by our architecture, we implement a shuffle-and-pair strategy. Specifically, we randomly sample two distinct audio-text pairs: one serves as the audio "query" and the other as the text and audio "response". Although the context is intentionally randomized, this design forces the model to focus purely on the mapping from the current response text to its corresponding audio tokens, ensuring the stability of audio generation regardless of text context.

4.3 Data of Expressive Acoustic Generation Fine-tuning

To achieve high-fidelity and expressive synthesis, we create a composite dataset, aiming to equip BrainSpeech with both superior acoustic quality and rich emotional expressiveness. We organize the data into two categories:

- *High Quality Dialogue Data:* We leverage the data utilized in stage 1. This process ensures BrainSpeech masters the generation of speech with superior acoustic quality.
- *Expressive Speech Corpora:* EmotionDB (Yang et al., 2025), a corpus of high-fidelity human speech characterized by rich emotional dynamics, is used. Therefore, $\theta_{\mathcal{A}}$ can capture intricate paralinguistic nuances, thereby enriching the generated responses with diverse and vivid emotional tones.

5 Experimental Setup

5.1 Model Configuration

LLM Backbone and Speech Encoder: We adopt Llama-3.2-3B-Instruct as our base LLM. For the speech encoder, we utilize Whisper-large-v3. The speech encoder are frozen throughout the training process to preserve the capabilities of robust acoustic feature extraction. To bridge the modality gap, we employ a trainable projector consisting of a two-layer MLP.

Initialization of Dual-Stream Architecture: The parameters of $\theta_{\mathcal{S}}$ are initialized from the base LLM. Conversely, $\theta_{\mathcal{A}}$ are randomly initialized.

Acoustic Tokenizer: We employ XCodec as our acoustic tokenizer which employs a residual vector quantization (RVQ) (Zeghidour et al., 2021) scheme with a codebook size of 1024 across 8 quantization layers. Additionally, we implement the delay pattern. Subsequently, we set the audio boundaries by enclosing this shifted sequence with two special tokens (`<|audio_stream_bos|>` and `<|audio_stream_eos|>`). Finally, the input representation at each position is obtained by summing the embeddings of the vertically aligned codes across all layers.

5.2 Training Details

In our training pipeline, all models are trained using AdamW with BF16. The details are listed as follows:

Stage 1: We train $\theta_{\mathcal{P}}$ and $\theta_{\mathcal{S}}$ on InstructS2S-200k for 3 epochs. We use a batch size of 16 and a peak learning rate of 5×10^{-5}. This stage is conducted on 8 NVIDIA H20 GPUs.

Stage 2: In this stage, we freeze $\theta_{\mathcal{S}}$ and train $\theta_{\mathcal{A}}$ for 2 epochs on Emilia. The batch size is 64, and the learning rate is set to 3×10^{-4} with a warmup of 4,000 steps. This stage is distributed across 32 NVIDIA H20 GPUs.

Stage 3: Finally, we train for 2 epochs with a batch size of 16 and a learning rate of 5×10^{-5}, using 8 NVIDIA H20 GPUs.

5.3 Evaluation on Spoken Question Answering

We evaluate the spoken question answering (SQA) capability of BrainSpeech on three widely used

Table 1: Results on Spoken Question Answering (SQA) benchmarks. We compare BrainSpeech with baselines across different scales. "S→T" denotes Speech-to-Text accuracy, and "S→S" denotes Speech-to-Speech accuracy.

Model	Llama Question		TriviaQA		Web Question	
	S→T	S→S	S→T	S→S	S→T	S→S
MinMo (7B)	78.9	64.1	48.3	37.5	55.0	39.9
GLM-4-Voice (9B)	64.7	50.7	39.1	26.5	32.2	15.9
VITA-Audio (7B)	76.3	64.6	43.6	39.5	44.2	40.0
Moshi (7B)	62.3	21.0	22.8	7.3	26.6	9.2
Llama-Omni (8B)	64.0	45.3	34.2	22.9	31.2	10.7
Llama-Omni2 (7B)	70.3	60.7	38.2	33.5	34.5	31.3
Llama-Omni2 (0.5B)	45.7	38.7	18.5	14.2	17.7	16.8
Llama-Omni2 (1.5B)	62.0	52.7	29.4	24.8	28.2	26.6
Llama-Omni2 (3B)	64.3	55.7	33.6	28.1	30.5	28.0
BrainSpeech (3B)	68.7	63.7	40.5	37.4	39.4	36.0

public benchmarks: Llama Question (Nachmani et al., 2023), TriviaQA (Joshi et al., 2017), and Web Question (Berant et al., 2013).

We benchmark BrainSpeech against a comprehensive suite of representative speechLMs, spanning a wide range of parameter scales (0.5B to 9B). Specifically, we include recent state-of-the-art models such as MinMo (7B) (Chen et al., 2025a), Moshi (7B) (Défossez et al., 2024), GLM-4-Voice (9B) (Zeng et al., 2024) and VITA-Audio (7B) (Long et al., 2025). Furthermore, we compare BrainSpeech with Llama-Omni (8B) (Fang et al., 2024) and Llama-Omni2 family (Fang et al., 2025) (ranging from 0.5B to 7B).

Following established protocols, we employ two evaluation metrics to assess the models:

- *Speech-to-Text (S→T):* The model receives speech and generates text responses directly. This metric evaluates the pure semantic generation and knowledge retrieval capabilities of the model.

- *Speech-to-Speech (S→S):* The model receives speech and generates audio responses. The generated speech is then transcribed into text using an ASR model. This metric assesses the capabilities of semantic generation and acoustic quality.

The results are summarized in Table 1. BrainSpeech demonstrates superior performance, achieving the highest accuracy among models of comparable parameter size. Notably, BrainSpeech remains highly competitive even when compared to larger models.

Specifically, these results validate the effectiveness of the architecture of BrainSpeech. By freezing θ_S during stage 2 and 3, BrainSpeech prevents the interference of acoustic gradients, safeguarding the reasoning capabilities of LLMs against catastrophic forgetting, which is the common issue in fine-tuning. Furthermore, the remarkably narrow gap between S→T and S→S metrics attests to the high fidelity of output speech. This confirms that BrainSpeech successfully decouples *what to say* from *how to say*, generating speech with high quality.

5.4 Evaluation on General Capabilities

To further assess the capabilities of BrainSpeech in open-ended scenarios and complex reasoning tasks, we conduct extensive evaluations on subsets of the VoiceBench benchmark (Chen et al., 2024). We confirm the performance across diverse domains:

- *Open-ended QA:* We utilize AlpacaEval and CommonEval to evaluate the ability to handle diverse queries. Following standard protocols, we employ GPT as an impartial judge to assign a score between 1 and 5. For these tasks, we evaluate both audio and text inputs to assess the capabilities of understanding.

- *Reasoning and Knowledge:* We employ MMSU and OpenBookQA to assess multi-choice reasoning capabilities using Accuracy (Acc). For these tasks, we employ Text as input. This design allows us to rigorously evaluate the knowledge retention of θ_S.

Table 2: Evaluation results on VoiceBench subsets. We report Scores (1-5) obtained by DeepSeek (Guo et al., 2025) for AlpacaEval and CommonEval, and Accuracy (%) for MMSU and OpenBookQA. "A" denotes Audio input, and "T" denotes Text input. "-" indicates the modality is unsupported. For MMSU and OpenBookQA, metrics are derived from text input, falling back to the result of audio input only when text input is unsupported.

Model	**AlpacaEval**		**CommonEval**		**MMSU**	**OpenBookQA**
	A (Score)	T (Score)	A (Score)	T (Score)	Acc	Acc
Moshi (7B)	2.01	-	1.60	-	24.04	26.15
LLaMA-Omni (8B)	3.58	3.70	3.26	3.52	59.01	79.34
VITA-Audio (7B)	2.21	2.73	1.93	2.49	59.41	66.96
GLM-4-Voice (9B)	3.87	-	3.29	-	62.69	44.84
Step-Audio2 (8B)	3.77	3.86	3.14	3.69	66.52	74.89
Mini-Omni (0.5B)	1.95	2.34	2.02	2.55	26.74	30.55
Mini-Omni2 (0.5B)	2.32	2.65	2.18	2.86	27.13	32.09
LLaMA-Omni2 (3B)	3.56	-	3.28	-	21.27	68.06
BrainSpeech (3B)	3.58	3.79	3.15	3.60	53.29	72.31

Baselines: We compare BrainSpeech against a wide array of baselines. We include models such as Mini-Omni/Omni2 (0.5B), LLaMA-Omni2 (3B), and larger-scale models including LLaMA-Omni (8B), Moshi (7B), VITA-Audio (7B), GLM-4-Voice (9B), and Step-Audio 2 (8B).

The quantitative results are presented in Table 2. BrainSpeech exhibits remarkable performance across all tasks. In AlpacaEval and CommonEval, BrainSpeech demonstrates robust instruction-following capabilities. Furthermore, in MMSU and OpenBookQA, BrainSpeech achieves high accuracy, significantly outperforming the comparable LLaMA-Omni2 (3B) and even surpassing larger 7B-scale baselines. This provides strong evidence that our dual-stream architecture, successfully safeguards the model's semantic generation, preventing the catastrophic knowledge degradation often observed in other speechLMs.

5.5 Evaluation on Speech Synthesis Quality

To evaluate the acoustic quality and expressiveness of BrainSpeech, we conduct assessments on speech generated by audio prompt. The test set aggregates the VoiceBench subsets mentioned above with GenEmotion-en from URO-Bench (Yan et al., 2025), which evaluates the capability to synthesize speech with target emotion based on audio instructions. This ensures a comprehensive evaluation covering both dialogue prosody and specific synthesis capabilities. We employ a combination of objective and subjective metrics:

- *Objective Metrics:* UTMOS (Saeki et al., 2022) is used to assess the generated audio. We also report Content Enjoyment (CE) and Production Quality (PQ) from AudioBox Metrics (Vyas et al., 2023) to assess the overall quality.

- *Subjective Metrics:* Naturalness Mean Opinion Score (NMOS) is used as the metric. We conduct a rigorous human evaluation with 10 listeners. Participants are presented with paired samples (questions and responses) and asked to rate the audio on a scale of 1 to 5 (increments of 0.5 allowed). The criteria strictly focus on prosody, rhythm, and the authenticity of human-like speech (5 = Excellent/Native-like, 1 = Poor/Robotic).

The results are summarized in Table 3. In terms of objective metrics, BrainSpeech achieves highly competitive results, particularly in CE and PQ, surpassing large-scale baselines like GLM-4-Voice. This indicates that $\theta_{\mathcal{A}}$ generates high-fidelity speech.

For the subjective evaluation, BrainSpeech outperforms baselines, validating the efficacy of BrainSpeech and the fine-tuning strategy. Our inference strategy effectively resolves boundary artifacts. In contrast, VITA-Audio suffers severely from such discontinuities. Trained on high-fidelity and expressive corpora, BrainSpeech successfully captures the subtle nuances of human articulation, exhibiting superior affective expressiveness and human-level naturalness.

Table 3: Evaluation of speech synthesis quality. We report objective metrics (UTMOS, AudioBox CE/PQ) and subjective human ratings for naturalness (NMOS 1-5). BrainSpeech achieves superior objective and subjective scores.

Model	Objective Metrics			Subjective
	UTMOS	CE	PQ	NMOS
Freeze-Omni (7B)	4.36	6.14	7.75	2.91
LLaMA-Omni (8B)	4.01	6.06	6.95	2.76
VITA-Audio (7B)	3.08	5.76	7.21	1.43
GLM-4-Voice (9B)	4.07	6.20	7.45	3.41
Step-Audio2 (8B)	**4.51**	6.09	7.66	3.56
Mini-Omni2 (0.5B)	4.44	6.29	7.73	3.08
LLaMA-Omni2 (3B)	4.22	**6.34**	7.74	3.79
Qwen2.5-Omni (3B)	4.12	6.31	7.71	3.68
BrainSpeech (3B)	4.41	6.31	**8.00**	**3.96**

5.6 Ablation Study

To validate the effectiveness of the architecture and training strategy, we examine three variants of models:

- *w/o Acoustic Stream:* A unified model where acoustic attention is removed, retaining only the acoustic FFN and the shared semantic attention. The model is fully fine-tuned on data used in stage 1 to learn understanding and generation simultaneously.
- *Serial Architecture:* Adopts Thinker-Talker where acoustic attention and FFN are appended after the base LLMs. Similar to the above, it is fully fine-tuned on data used in stage 1.
- *BrainSpeech (Stage 2 Only):* BrainSpeech trained only up to stage 2 (acoustic generation pre-training).

We evaluate these variants on semantic understanding (AlpacaEval/CommonEval Audio Scores) and speech synthesis quality (UTMOS, CE, PQ).

The results are shown in Figure 3. First, regarding architecture, "w/o Acoustic Stream" and "Serial Architecture" suffer significant degradation in semantic performance. This confirms that independent parameter spaces are crucial, and acoustic gradients during full fine-tuning harms the knowledge of base LLMs. Second, regarding the training stage, while stage 2 lays the foundation of audio generation, the full model's performance highlights the necessity of stage 3 for expressive synthesis. It significantly elevates acoustic quality metrics while maintaining robust semantic alignment established in earlier stages.

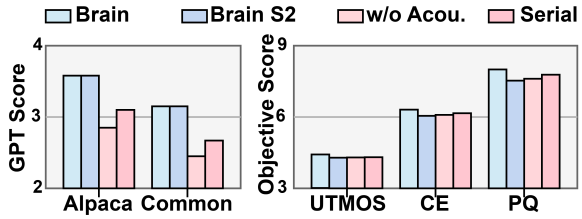

Figure 3: Ablation study results. BrainSpeech best preserves semantics while the fine-tuning stage significantly boosts audio quality.

6 Conclusion

In this paper, we introduced BrainSpeech, an end-to-end spoken dialogue system that successfully reconciles the conflict between semantic and acoustic generation in SpeechLMs. By employing a biologically inspired dual-stream architecture, BrainSpeech decouples the optimization objectives of *what to say* and *how to say*. This design allows us to safeguard the model's inherent reasoning capabilities and generate high-fidelity speech. Furthermore, our progressive three-stage training paradigm further eliminates the dependency on scarce spoken dialogue data. Extensive evaluations demonstrate that BrainSpeech achieves state-of-the-art performance among models of comparable scale, delivering responses with superior naturalness and emotional expressiveness. We believe BrainSpeech establishes a new baseline for future research into efficient, expressive, and intelligent assistants.

Limitations

The current implementation prioritizes efficiency with a 3B backbone. While competitive, the potential of scaling to larger foundation models remains underexplored. We hypothesize that scaling up will yield substantial improvements, as larger backbones possess richer semantic representations. This would facilitate a more comprehensive capture of the interplay between linguistic and paralinguistic features, thereby further elevating the quality of both textual and acoustic responses. Future work will investigate these scaling laws to broaden system applicability.

References

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

Boson AI. 2025. Higgs Audio V2: Redefining Expressiveness in Audio Generation. https://github.com/boson-ai/higgs-audio. GitHub repository. Release blog available at https://www.boson.ai/blog/higgs-audio-v2.

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

Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, and 1 others. 2025b. Slam-omni: Timbre-controllable voice interaction system with single-stage training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2262–2282.

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

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

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*.

Qingkai Fang, Yan Zhou, Shoutao Guo, Shaolei Zhang, and Yang Feng. 2025. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis. *arXiv preprint arXiv:2505.02625*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, and 1 others. 2024. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890. IEEE.

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

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

Qingyu Liu, Yushen Chen, Zhikang Niu, Chunhui Wang, Yunting Yang, Bowen Zhang, Jian Zhao, Pengcheng Zhu, Kai Yu, and Xie Chen. 2025. Cross-lingual f5-tts: Towards language-agnostic voice cloning and speech synthesis. *arXiv preprint arXiv:2509.14579*.

Zuwei Long, Yunhang Shen, Chaoyou Fu, Heting Gao, Lijiang Li, Peixian Chen, Mengdan Zhang, Hang Shao, Jian Li, Jinlong Peng, and 1 others. 2025. Vita-audio: Fast interleaved cross-modal token generation for efficient large speech-language model. *arXiv preprint arXiv:2505.03739*.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *arXiv preprint arXiv:2204.02152*.

Preeti Saini and Parneet Kau. 2013. Automatic speech recognition: A review. *Springer US*.

Yun Da Tsai, Yu Che Tsai, Bo Wei Huang, Chun Pai Yang, and Shou De Lin. 2023. Automl-gpt: Large language model for automl. *arXiv preprint arXiv:2309.01125*.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, and 1 others. 2023. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*.

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

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

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

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

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

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. Uro-bench: A comprehensive benchmark for end-to-end spoken dialogue models. *arXiv preprint arXiv:2502.17810*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Guanrou Yang, Chen Yang, Qian Chen, Ziyang Ma, Wenxi Chen, Wen Wang, Tianrui Wang, Yifan Yang, Zhikang Niu, Wenrui Liu, and 1 others. 2025. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. *arXiv preprint arXiv:2504.12867*.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, and 1 others. 2025. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25697–25705.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

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

A Formatting Examples of Data

This section illustrates the interleaved dialogue template employed by BrainSpeech. As demonstrated in the listing below, the assistant's response follows a strict alternating structure of text and audio segments. To ensure fine-grained modality alignment, the text response is segmented into segments of uniform length (text_sequence). Correspondingly, audio segments introduced by the <|audio_out_bos|> token contain a **fixed number** of acoustic tokens, maintaining a constant text-to-audio ratio. The special token <|audio_last_out_bos|> is reserved for the final audio segment, which encapsulates all **residual acoustic tokens** ensuring the completeness of the speech generation.

Speech QA Interleaved Data Format

```
{
 "messages": [
  {
   "role": "user",
   "content": "<|audio_bos|> <|
       AUDIO|> <|audio_eos|>"
  },
  {
   "role": "assistant",
   "content": "text_sequence_1 <|
       audio_out_bos|>
<|AUDIO_OUT|> <|audio_eos|>
    text_sequence_2
<|audio_out_bos|> <|AUDIO_OUT|> <|
    audio_eos|> text_sequence_3
<|audio_last_out_bos|> <|AUDIO_OUT
    |> <|audio_eos|>"
  }
 ]
}
```

B Dataset Statistics

Dataset	Samples	Avg.Words	Avg.AudioLen
Llama Question	300	8.46	3.00
TriviaQA	1000	12.02	4.72
Web Question	2032	6.76	2.33
AlpacaEval	199	16.32	5.67
CommonEval	200	8.06	4.83
OpenBookQA	455	44.28	18.89
MMSU	3074	53.16	23.61
GenEmotion-en	54	15.35	4.93

Table 4: Statistical information of the datasets used in experiments.

This appendix provides statistical details of the datasets utilized in the experiments. "Samples" indicates the number of samples, "Average Number of Words" (Avg.Words) denotes the average word count of the input of text modality, and "Average Audio Length" (Avg.AudioLen) refers to the average duration of the input of audio modality (unit: seconds).

C Audio Generation with Minimax-Speech-02: Parameter Inference via Qwen2.5

To generate high-quality and expressive audio using Minimax-Speech-02, we leverage its API for audio synthesis. A key step in this process involves inferring critical acoustic and paralinguistic parameters that shape the naturalness of the generated speech. Specifically, we employ Qwen2.5 to analyze input text and predict implicit speech attributes, including pitch, volume, speed, emotion, and gender. Then, these parameters are fed into Minimax-Speech-02 to guide audio generation.

C.1 Parameter Inference Protocol with Qwen2.5

We design a structured instruction for Qwen2.5 to ensure consistent and standardized parameter prediction. The instruction specifies the requirements for each parameter, including valid value ranges, emotion categories, and gender options, as follows:

- **Pitch**: Classified on a scale of [-3, 3], where negative values indicate lower pitch and positive values indicate higher pitch relative to a neutral baseline.
- **Volume**: Rated on a scale of [0, 4], with 0 representing the softest volume and 4 representing the loudest.
- **Speed**: Normalized on a scale of [0.5, 2.0], where 0.5 denotes half the average speech speed and 2.0 denotes double the average speed.
- **Emotion**: Restricted to seven categorical labels: happy, sad, angry, fearful, disgusted, surprised, or calm.
- **Gender**: Specified as male, female or neutral to align with typical speech synthesis gender configurations.

Qwen2.5 processes the input text alongside the instruction and returns the predicted parameters

in a machine-readable format, which is then integrated into the Minimax-Speech-02 API call.

C.2 Example of Parameter Inference and API Input Format

The following illustrates the instruction, input text, and corresponding API-compatible format for parameterized audio generation:

Example of Parameter Inference and API Input Format

{
"instruction": "Analyze the following text and infer the likely implicit pitch, volume, speed, and emotion that would be used if the text were spoken aloud. Classify pitch on scale [-3, 3], volume on scale [0, 4] and speed on scale [0.5, 2.0]. Determine emotion as one of the following: happy, sad, angry, fearful, disgusted, surprised, calm. Give the gender like male or female. Return the result in the format: {'pitch': , 'volume': , 'speed': , 'emotion': , 'gender': }",
"input": "How can cross training benefit groups like runners, swimmers, or weightlifters?",
"output": ques_0"
}

This approach ensures that the generated audio not only accurately conveys the semantic content of the input text but also exhibits natural prosody and emotional alignment, enhancing the overall expressiveness and usability of the synthesized speech.

D GPT Evaluation Prompt for VoiceBench

This appendix presents the full prompt template used to instruct GPT in automatically evaluating the quality of text responses of models. The prompt standardizes the evaluation criteria for speech interaction scenarios, focusing on relevance, accuracy, and conciseness of responses on a 1–5 scale (1 = poorest, 5 = optimal). The prompt (with placeholders `{prompt}` and `{response}` replaced by transcribed user instructions and model responses) is:

GPT Evaluation Prompt

I need your help to evaluate the performance of several models in the speech interaction scenario. The models will receive a speech input from the user, which they need to understand and respond to with a speech output.
Your task is to rate the model's responses based on the provided user input transcription [Instruction] and the model's output transcription [Response].
Please evaluate the response on a scale of 1 to 5:
1 point: The response is largely irrelevant, incorrect, or fails to address the user's query. It may be off-topic or provide incorrect information.
2 points: The response is somewhat relevant but lacks accuracy or completeness. It may only partially answer the user's question or include extraneous information.
3 points: The response is relevant and mostly accurate, but it may lack conciseness or include unnecessary details that don't contribute to the main point.
4 points: The response is relevant, accurate, and concise, providing a clear answer to the user's question without unnecessary elaboration.
5 points: The response is exceptionally relevant, accurate, and to the point. It directly addresses the user's query in a highly effective and efficient manner, providing exactly the information needed.
Below are the transcription of user's instruction and models' response:
[Instruction]: prompt
[Response]: response
After evaluating, please output the score only without anything else. You don't need to provide any explanations.

The prompt was applied uniformly across AlpacaEval and CommonEval subsets to ensure consistent evaluation of model performance in these scenarios.

E Additional Experiments about Latency

This appendix presents an analysis of latency performance between BrainSpeech and state-of-the-art speechLMs, focusing on two critical metrics for real-time interaction: the latency to generate the first text token (from user input to model's first text output) and the latency to fully generate the first audio segment (from input to the first speech segment). We compare BrainSpeech against four baselines: GLM-4-Voice (9B), LLaMA-Omni (8B), LLaMA-Omni2 (3B), and VITA-Audio (7B).

Model	FTT (ms)	FAS (ms)
LLaMA-Omni (8B)	289.46	**346.73**
LLaMA-Omni2 (3B)	204.79	567.84
VITA-Audio (7B)	414.69	1215.89
GLM-4-Voice (9B)	422.77	1562.81
BrainSpeech (3B)	**203.75**	1050.21

Table 5: Latency comparison of first token/segment generation (unit: milliseconds). "FTT" refers to "Latency of First Text Token", while "FAS" refers to "Latency of First Audio Segment".

As summarized in Table 5, BrainSpeech outperforms GLM-4-Voice (9B) and VITA-Audio (7B) in the two latency metrics. While BrainSpeech exhibits higher latency than LLaMA-Omni (8B) and LLaMA-Omni2 (3B), it achieves significantly superior speech quality compared to these two models (as validated in Table 3). Notably, the higher "FAS" of BrainSpeech relative to LLaMA-Omni2 (3B) stems from the per-segment audio token count: BrainSpeech generates 40 tokens per audio segment, whereas LLaMA-Omni2 (3B) uses only 10 tokens per segment. In practice, we can also reduce the number of audio tokens per segment for BrainSpeech to obtain lower "FAS" latency if needed. Additionally, the latency of BrainSpeech is sufficiently low to support real-time audio interaction already: users only experience a minimal wait for the first audio segment, with no additional delays for subsequent content generation.

F More Discussion About BrainSpeech

In this section, we further elaborate on the broader implications of BrainSpeech beyond its empirical performance presented in the main paper.

As highlighted earlier, BrainSpeech achieves high-quality text responses and more natural, higher-fidelity audio outputs without relying on scarce multimodal dialogue data. Instead, it leverages only easily accessible resources: pure text dialogue data, synthesized TTS data, and TTS datasets. However, the value of BrainSpeech extends beyond these.

Notably, BrainSpeech aligns more closely with the working mechanism of the brain of human: different brain regions collaborate to perform various tasks, each specializing in its domain, rather than using the whole region (which risks interference between distinct types of information). We can analogize the parameters responsible for text processing in BrainSpeech to the "logical center" of the brain of human: text is a representation constructed by human, rather not naturally occurring in the physical world, and semantic processing based on text enables us to understand the world systematically.

By separating this text (logical) module from the audio module, BrainSpeech mitigates the cross-modal interference that often arises when a single parameter space is forced to handle multiple modalities simultaneously. This modular design also shows scalability to more modalities (e.g., images, video): instead of mixing modal information in a shared parameter space, each modality can be processed by a specialized sub-module that collaborates with the core text logical center. This preserves the clarity of each modality's processing while enabling coherent multi-modal interaction, extending the strengths of BrainSpeech beyond text-audio dialogue to broader multi-modal scenarios.