MoST: Mixing Speech and Text with Modality-Aware Mixture of Experts

Yuxuan Lou [*1] **Kai Yang** [*2] **Yang You** [1]

Abstract

We present MoST (Mixture of Speech and Text), a novel multimodal large language model that seamlessly integrates speech and text processing through our proposed Modality-Aware Mixture of Experts (MAMoE) architecture. While current multimodal models typically process diverse modality representations with identical parameters—disregarding their inherent representational differences, we introduce specialized routing pathways that direct tokens to modality-appropriate experts based on input type. MAMoE simultaneously enhances modality-specific learning and cross-modal understanding through two complementary components: modality-specific expert groups that capture domain-specific patterns and shared experts that facilitate information transfer between modalities. Building on this architecture, we develop an efficient transformation pipeline that adapts the pretrained MoE model through strategic post-training on ASR and TTS datasets, followed by fine-tuning with a carefully curated speech-text instruction dataset. A key feature of this pipeline is that it relies exclusively on fully accessible, open-source datasets to achieve strong performance and data efficiency. Comprehensive evaluations across ASR, TTS, audio language modeling, and spoken question answering benchmarks show that MoST consistently outperforms existing models of comparable parameter counts. Our ablation studies confirm the effectiveness of modality-specific routing mechanism and shared experts design across all tested domains. To our knowledge, MoST represents the first fully open-source speech-text LLM built on a Mixture of Experts architecture. [1]

[*]Equal contribution [1]School of Computing, National University of Singapore, Singapore [2]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. Correspondence to: Yang You <youy@comp.nus.edu.sg>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

[1]We release MoST model, training code, inference code, and training data at https://github.com/NUS-HPC-AI-Lab/MoST

1. Introduction

The seamless integration of speech and text modalities represents a cornerstone objective in the pursuit of versatile and natural human-computer interaction. Foundation models, particularly Large Language Models (LLMs), have demonstrated remarkable capabilities in text understanding and generation (OpenAI et al., 2024; DeepSeek-AI et al., 2024), motivating efforts to extend their prowess to encompass spoken language. However, unifying inherently different modalities like continuous, high-dimensional speech waveforms and discrete, symbolic text sequences within a single model presents significant architectural and training challenges.

A prevalent approach in multimodal learning involves mapping different modalities into a shared representational space, often processed by a common set of transformer parameters (Nguyen et al., 2024; Défossez et al., 2024). While conceptually simple, forcing diverse modality representations through identical parameters disregards their inherent representational differences. Specifically for speech and text, this uniform treatment risks representational interference, where the distinct statistical properties and structural nuances of each modality are diluted. Furthermore, it can hinder the development of highly specialized processing pathways necessary to capture the fine-grained details unique to each domain—such as phonetic variations in speech or complex semantic relations in text.

Mixture of Experts (MoE) architectures have emerged as a powerful technique for scaling model capacity while maintaining computational efficiency (Shazeer et al., 2017; Lepikhin et al., 2020; Fedus et al., 2022). By sparsely activating only a subset of parameters (experts) for each input token, MoE models can significantly increase their parameter count without a proportional increase in inference cost. While promising for handling complex data, standard MoE routing mechanisms are typically modality-agnostic; tokens are routed based on learned gating functions over their representations, without explicit consideration of their originating modality. This leaves untapped the potential to dedicate specialized expert capacity based on known input characteristics.

In this work, we propose MoST (Mixture of Speech and Text), a novel speech-text LLM built upon a Modality-Aware Mixture of Experts (MAMoE) architecture. MAMoE directly addresses the limitations of uniform parameter processing by incorporating two principal components: **modality-specific expert groups** (e.g., text-only, audio-only) governed by a modality-aware routing mechanism, and **shared experts** accessible to all modalities, which facilitate cross-modal interaction. The modality-aware routing explicitly leverages token modality information to direct inputs towards appropriate modality-specific experts. Text tokens are primarily routed to text and shared experts, while audio tokens target audio and shared experts. This design allows for: (1) Specialized Learning: Modality-specific experts can capture the intricate patterns unique to either speech or text without interference. (2) Efficient Cross-Modal Interaction: Shared experts act as a bridge, facilitating knowledge transfer and cross-modal alignment crucial for tasks like Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) synthesis.

To effectively realize MoST, we develop an efficient LLM to Speech-Text Transformation Pipeline. Starting from a pretrained MoE LLM, we first conduct targeted post-training on large ASR and TTS datasets. This stage primes the modality-specific experts and adapts the shared experts for cross-modal tasks. Subsequently, we perform instruction fine-tuning using a carefully curated dataset comprising diverse speech and text instructions, enhancing the model's versatility and controllability. A key aspect of our approach is the exclusive use of fully accessible, open-source training data to achieve strong performance. This strategic choice distinguishes MoST from some prominent speech-text models, such as SpiritLM (Nguyen et al., 2024) and Moshi (Défossez et al., 2024), whose development incorporated datasets that are not publicly accessible.

Comprehensive evaluations across a range of speech-text benchmarks, including Automatic Speech Recognition (ASR), Text-to-Speech (TTS) synthesis, audio language modeling, and spoken question answering, demonstrate that MoST achieves state-of-the-art or competitive performance compared to existing models with similar parameter counts. Our ablation studies rigorously validate the effectiveness of the MAMoE architecture, confirming that the modality-aware routing mechanism is a key contributor to performance gains across both modality-specific and cross-modal tasks. To the best of our knowledge, MoST is the first fully open-source speech-text LLM leveraging a Mixture of Experts architecture with modality-specific routing.

Our main contributions are:

- **Modality-Aware Mixture of Experts (MAMoE)**: A novel multimodal MoE architecture that directs tokens to modality-specific expert groups and shared experts, enhancing specialized learning and cross-modal interaction.
- **Efficient Text-Speech Transformation Pipeline**: A data-efficient pipeline for adapting a pretrained LLM into a speech-text LLM via targeted post-training and instruction tuning.
- **Open Release of MoST Model**: We release the MoST model weights, training and inference code, and training dataset to facilitate future research and development in multimodal AI.

2. Related Work

End-to-end Speech-Text Modeling. Existing end-to-end speech-text models like Qwen2-Audio (Chu et al., 2024) (text output only) and AudioLM (Borsos et al., 2023) (no direct text integration) have specific limitations. Others such as SpiritLM (Nguyen et al., 2024) and Moshi (Défossez et al., 2024) employ dense transformers, which process all tokens uniformly, unlike MoST's Modality-Aware Mixture of Experts (MAMoE), which uses specialized and shared experts for modality-specific processing and cross-modal understanding.

Mixture of Experts. Standard Mixture of Experts (MoE) architectures (Lepikhin et al., 2020; Fedus et al., 2022; Du et al., 2022) scale models efficiently but typically feature modality-agnostic routing, limiting their multimodal effectiveness. While some research adapts MoE for general multimodal tasks (e.g., image-text) by grouping experts (Lin et al., 2024; Shen et al., 2024), MoST is the first to introduce MAMoE specifically for speech-text. Another key difference is that MoST's MAMoE employs shared experts for cross-modal interaction, which is novel.

Multimodal Adaptation from LLMs. LLMs are often adapted for multimodality via fine-tuning (Zhang et al., 2024), instruction tuning (Liu et al., 2023), or speech-specific methods (Fathullah et al., 2023). MoST presents a unique transformation pipeline: starting from a pretrained MoE LLM, it undergoes targeted ASR/TTS post-training, followed by fine-tuning with a curated speech-text instruction dataset. This approach offers significant data efficiency over conventional adaptation strategies.

3. MoST: Mixture of Speech and Text

3.1. Model Overview

MoST (Mixture of Speech and Text) is a unified foundation model built on a Mixture of Experts (MoE) architecture for seamless speech and text processing. Unlike discrete tokenization approaches (Borsos et al., 2023; Nguyen et al., 2024), MoST directly processes continu-

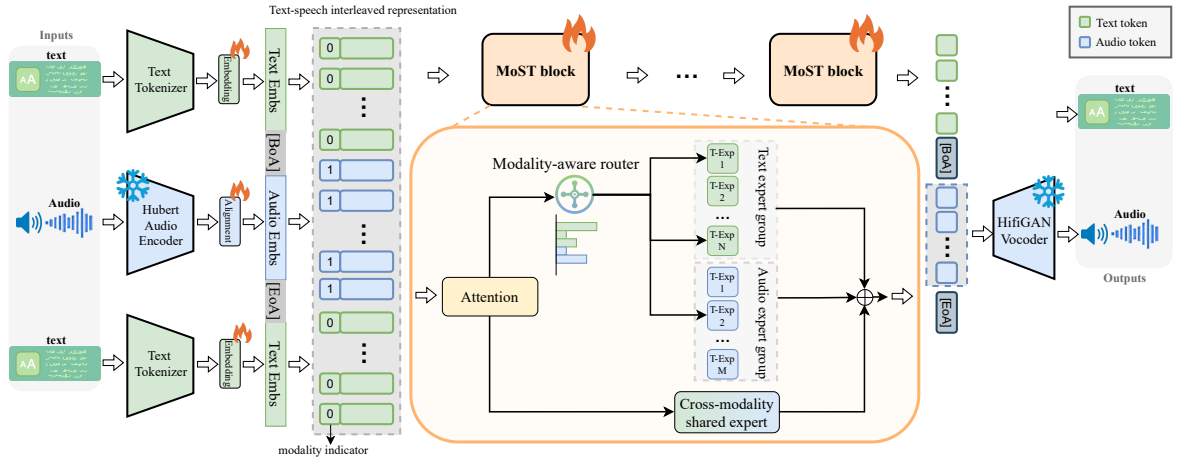

Figure 1. MoST overall architecture. MoST supports interleaved text and audio as input and can generate both text and audio. MoST blocks are transformer blocks featured with Modality-Aware Mixture of Experts (MAMoE). MAMoE comprises three critical components: modality-specific expert groups, cross-modal shared expert, and the modality-aware router. Modality-aware router utilizes the modality indicator information for modality-specific routing.

ous audio waveforms via audio encoder E_{audio}: $\mathbf{H}_{audio} = E_{audio}(\mathbf{A}) \in \mathbb{R}^{L_{audio} \times d_{model}}$. Text representations are obtained through standard tokenization and embedding: $\mathbf{H}_{text} = E_{text}(\mathbf{T}) \in \mathbb{R}^{L_{text} \times d_{model}}$. These modalities are combined into unified input $\mathbf{H}_{input}$.

The core transformer decoder D_{MoST}, adapted from a pre-trained MoE LLM, incorporates our Modality-Aware Mixture of Experts (MAMoE) in designated layers. MAMoE routes tokens to modality-specific experts and shared experts for cross-modal interaction: $\mathbf{H}_{output} = D_{MoST}(\mathbf{H}_{input})$.

MoST generates both text via language model head P_{text} and audio via vocoder G_{audio} conditioned on $\mathbf{H}_{output}$ and speaker embeddings, enabling diverse end-to-end speech-text tasks. Figure 1 illustrates the architecture.

3.2. Audio Encoder and Decoder

MoST processes continuous audio waveforms directly, preserving richer acoustic information than discrete tokenization methods (Borsos et al., 2023; Nguyen et al., 2024).

Audio Encoder Input waveforms $\mathbf{A}$ are encoded using frozen HuBERT (Hsu et al., 2021), yielding features $\mathbf{F}_{hubert} \in \mathbb{R}^{L_{audio} \times d_{hubert}}$, then projected to model dimension:

$$\mathbf{H}_{audio} = \mathbf{F}_{hubert}\mathbf{W}_{proj} \in \mathbb{R}^{L_{audio} \times d_{model}}. \quad (1)$$

These continuous embeddings replace placeholder tokens in the input sequence.

Audio Decoder For synthesis, MoST employs HiFi-GAN vocoder G_{audio} (Kong et al., 2020), generating waveforms $\hat{\mathbf{A}}$ conditioned on predicted HuBERT tokens $\mathbf{C}_{audio} = P_{hubert_tokens}(\mathbf{H}_{output})$ and speaker embeddings $\mathbf{s}$:

$$\hat{\mathbf{A}} = G_{audio}(\mathbf{C}_{audio}, \mathbf{s}). \quad (2)$$

Algorithm 1 Modality-Aware Expert Routing

Require: Token representation $\mathbf{h}_t$, modality indicator $\mathbf{m}_t$, gating weights $\mathbf{W}_g$, expert groups $\mathcal{E}_{text}$, $\mathcal{E}_{audio}$, number of experts to select $\mathbf{K}$
Ensure: Routed output $\mathbf{y}_{routed,t}$
Step 1: Compute Raw Gating Scores
$\mathbf{s}_t \leftarrow \text{softmax}(\mathbf{h}_t W_g) \in \mathbb{R}^N$ *// Raw scores over all N experts*
Step 2: Create Modality-Specific Mask
for $i = 1$ to N **do**
  if $m_t = 0$ (text) and $E_i \in \mathcal{E}_{text}$ **then**
    $M_{m_t,i} \leftarrow 1$
  else if $m_t = 1$ (audio) and $E_i \in \mathcal{E}_{audio}$ **then**
    $M_{m_t,i} \leftarrow 1$
  else
    $M_{m_t,i} \leftarrow 0$
  end if
end for
Step 3: Apply Modality Mask
$\mathbf{s}'_t \leftarrow \mathbf{s}_t \odot \mathbf{M}_{m_t}$ *// Element-wise multiplication*
Step 4: Select Top-K Experts
$\mathcal{I}_t \leftarrow \text{TopK}(\mathbf{s}'_t, K)$ *// Indices of top-K experts*
$\mathbf{w}'_t \leftarrow \{\mathbf{s}'_t[i] : i \in \mathcal{I}_t\}$ *// Corresponding weights*
Step 5: Compute Routed Output
$\mathbf{y}_{routed,t} \leftarrow \sum_{k=1}^{K} w'_{t,k} \cdot E_{i_{t,k}}(\mathbf{h}_t)$ *// Weighted expert outputs*
return $\mathbf{y}_{routed,t}$

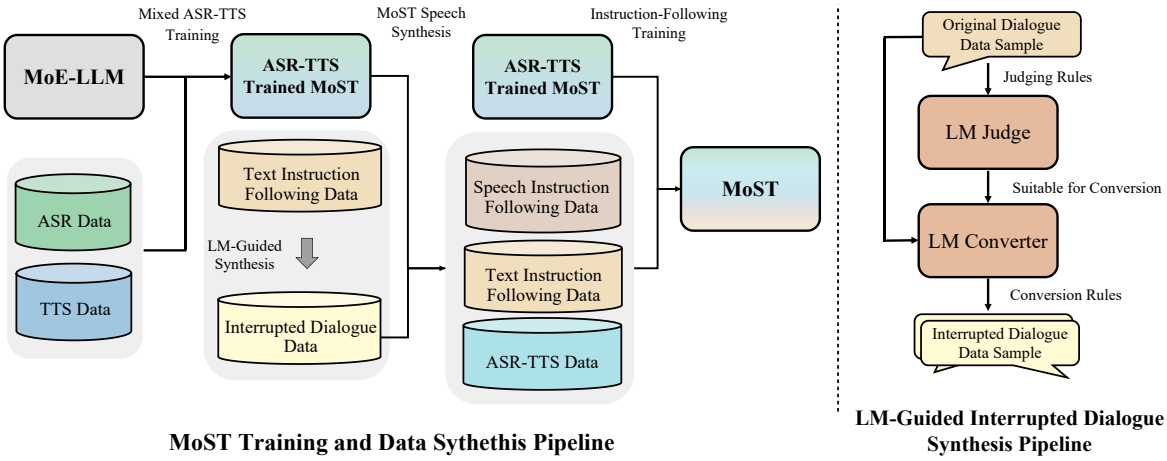

Figure 2. MoST training recipe overview. The left panel illustrates the systematic transformation pipeline from MoE LLM to MoST, encompassing data preparation and two-stage training protocol. The right panel details the interrupted dialogue synthesis process for instruction dataset construction.

3.3. Modality-Aware Mixture of Experts (MAMoE)

3.3.1. Modality-Specific Expert Groups

In the MAMoE layers, the set of N available routed experts, denoted $\mathcal{E} = \{E_1, E_2, \ldots, E_N\}$, is partitioned into disjoint modality-specific groups. Based on the primary modalities in MoST (speech and text), we define: **Text Expert Group** ($\mathcal{E}_{text}$), a subset of experts designated to primarily process text-derived token representations, and **Audio Expert Group** ($\mathcal{E}_{audio}$), a subset of experts designated to primarily process audio-derived token representations (specifically, features derived from the HuBERT encoder).

These sets are disjoint, $\mathcal{E}_{text} \cap \mathcal{E}_{audio} = \emptyset$, and their union constitutes the full set of routed experts, $\mathcal{E}_{text} \cup \mathcal{E}_{audio} = \mathcal{E}$. Each expert E_i is an MLP, consisting of gate, up, and down projection layers. This partitioning allows experts to specialize in capturing the distinct statistical patterns and representations characteristic of either speech or text.

3.3.2. Cross-Modal Shared Experts

While modality-specific experts promote specialized processing, effective multimodal understanding requires mechanisms for information integration across modalities. In MoST, this is facilitated by incorporating shared experts, implemented as a separate, parallel MLP block (E_{shared}). Unlike the routed experts $\mathcal{E}$, the shared expert block processes *all* tokens passing through the MAMoE layer.

Given the input hidden state $\mathbf{h}$ to the MAMoE layer, the output $\mathbf{y}_{routed}$ from the modality-specific routed experts (detailed in Sec 3.3.3) is combined with the output of the shared expert block:

$$\mathbf{y}_{mamoe} = \mathbf{y}_{routed} + E_{shared}(\mathbf{h}). \tag{3}$$

This parallel shared expert acts as a common pathway, allowing for the exchange and integration of information learned from different modalities, complementing the specialized processing occurring within the modality-specific expert groups.

3.3.3. Modality-Aware Router

MAMoE's core routing mechanism directs each token $\mathbf{h}_t$ to modality-specific experts $E_i \in \mathcal{E}$ based on its content and modality indicator m_t (e.g., $m_t = 0$ for text, $m_t = 1$ for audio). Algorithm 1 presents the step-by-step procedure for computing modality-aware routing scores.

During training, an auxiliary load balancing loss, based on $\mathbf{s}'_t$, promotes balanced expert utilization within each modality group. This modality-aware routing directs tokens to specialized pathways, enhancing expert capacity use and modality-specific representation learning.

4. Efficient Training Recipe of MoST

We develop an efficient training recipe to transform a MoE LLM into MoST. Our approach consists of carefully designed data preparation procedures followed by a two-stage training protocol. Comprehensive details of all procedures are provided in Appendix A.

4.1. Data Preparation

ASR and TTS Data Curation. We curate large-scale ASR and TTS datasets from open-source resources, primarily utilizing Common Voice (Ardila et al., 2020b) and LibriHeavy (Kang et al., 2024). For ASR data, we use the standard format with speech input and text transcription output. For TTS data, we reverse the modalities to create

Table 1. Comprehensive evaluation on speech processing and audio understanding tasks. **Speech Processing**: ASR and TTS performance (WER%↓) on LibriSpeech clean/other, VoxPopuli-V1.0-en (VP), and Common Voice 15-en (CV15). **Audio Understanding**: Audio language modeling accuracy (↑) on selected benchmarks; full results in Appendix F. Best results are **bold**; MoST row highlighted with (↑X.X) indicating improvement over previous best where MoST achieves state-of-the-art.

			Speech Processing (WER%↓)								**Audio Understanding** (Acc.↑)		
			LS-Clean		LS-Other		VP		CV15			sTopic-	
Model	**Type**	**S→S**	ASR	TTS	ASR	TTS	ASR	TTS	ASR	TTS	sWUGGY	StoryCloze	Avg.
Qwen2-Audio	Dense	✗	**1.8**	–	3.6	–	7.1	–	8.6	–	–	–	–
Phi-4 Multimodal	Dense	✗	2.1	–	**3.5**	–	6.3	–	9.2	–	–	–	–
SpeechGPT	Dense	✓	11.0	14.1	16.7	15.3	18.2	21.3	19.4	23.2	51.82	60.13	53.71
AudioLM	Dense	✓	9.5	9.2	12.0	12.1	15.0	22.1	17.6	25.1	71.50	–	–
SpiritLM	Dense	✓	6.0	6.7	11.0	9.5	14.3	19.4	15.4	22.4	40.14	83.32	57.67
Moshi	Dense	✓	5.5	7.0	12.0	7.2	8.8	10.6	9.4	14.2	51.14	46.34	48.99
SeamlessM4T-v2	Dense	✓	3.3	6.3	4.2	5.3	7.5	10.3	10.3	**10.8**	71.84	81.55	69.00
MinMo	Dense	✓	**1.8**	6.7	3.9	7.5	6.7	10.9	**8.0**	13.5	68.59	75.43	65.19
LLaMA-Omni2	Dense	✓	3.5	10.1	4.0	9.2	9.5	12.4	11.3	17.2	73.21	78.21	68.39
MoST (Ours)	MoE	✓	2.0	**6.0**(↓0.3)	3.7	**7.2**	**6.2**(↓0.1)	**10.1**(↓0.2)	8.4	11.5	**75.28**(↑2.1)	**83.64**(↑0.3)	**71.94**(↑2.9)

paired speech-text samples.

Speech-Text Instruction Dataset Construction. We construct a speech-text instruction-following dataset through two complementary strategies: (1) **Interrupted Dialogue Synthesis**: We augment existing dialogue datasets (Allal et al., 2025) with realistic interruptions using LLM-based synthesis to capture natural conversational dynamics and turn-taking patterns. (2) **Text-to-Speech Instruction Conversion**: We convert text-based instruction datasets into speech formats using our intermediate TTS-trained MoST, significantly expanding speech instruction data without manual collection.

4.2. Training Protocol

Stage 1: Cross-Modal Post-Training. We initialize MoST from DeepSeek-v2 Lite (DeepSeek-AI et al., 2024), a pre-trained MoE LLM. The initialized model undergoes targeted post-training on the curated ASR and TTS datasets. This stage serves dual purposes: priming modality-specific experts within MAMoE for speech processing and adapting shared experts for fundamental cross-modal alignment between speech and text representations.

Stage 2: Mixed Instruction Fine-Tuning. The model is fine-tuned using our constructed speech-text instruction dataset, strategically mixed with a substantial portion of ASR/TTS data from Stage 1. This mixed-task learning approach enhances complex instruction-following capabilities while preventing catastrophic forgetting of foundational speech processing abilities through continual exposure to core tasks.

5. Empirical Results

5.1. Evaluation Settings

To rigorously assess the performance of MoST, we establish a comprehensive evaluation framework encompassing standard benchmarks and strong baseline models.

Baselines. We compare MoST against several state-of-the-art open-source speech and text models, including **AudioLM** (Borsos et al., 2023), a prominent model for high-fidelity audio generation; **SpeechGPT** (Zhang et al., 2023), an LLM adapted for speech tasks; **SpiritLM** (Nguyen et al., 2024), a recent speech-language model with strong performance; **Moshi** (Défossez et al., 2024), another relevant model in speech processing; **Qwen2-Audio** (Chu et al., 2024), a powerful open-source model which has shown strong performance in various speech tasks; **Phi-4 Multimodal** (Microsoft et al., 2025), a compact yet powerful multimodal language model that achieves very strong ASR and AST performance; **SeamlessM4T-v2** (Communication et al., 2023), an advanced model for speech translation and other speech-related tasks; **MinMo** (Chen et al., 2025), a recent multimodal model for speech and text; and **LLaMA-Omni2** (Fang et al., 2025), a recent omnimodal language model with speech capabilities.

Benchmarks and Metrics. Evaluation spans ASR, TTS, audio language modeling (ALM), and spoken question answering (SQA). For ASR/TTS, we use LibriSpeech-clean and -other (Kang et al., 2024), VoxPopuli-V1.0-en (Wang et al., 2021), and Common Voice 15-en (WER for ASR, TTS; lower is better) (Ardila et al., 2020a). ALM is assessed on sWUGGY, sBLIMP (Both from Nguyen et al. (2020)); sTopic-StoryCloze and sStoryCloze (Both from Hassid et al. (2024)) (accuracy). SQA uses Llama Q (Nachmani et al.,

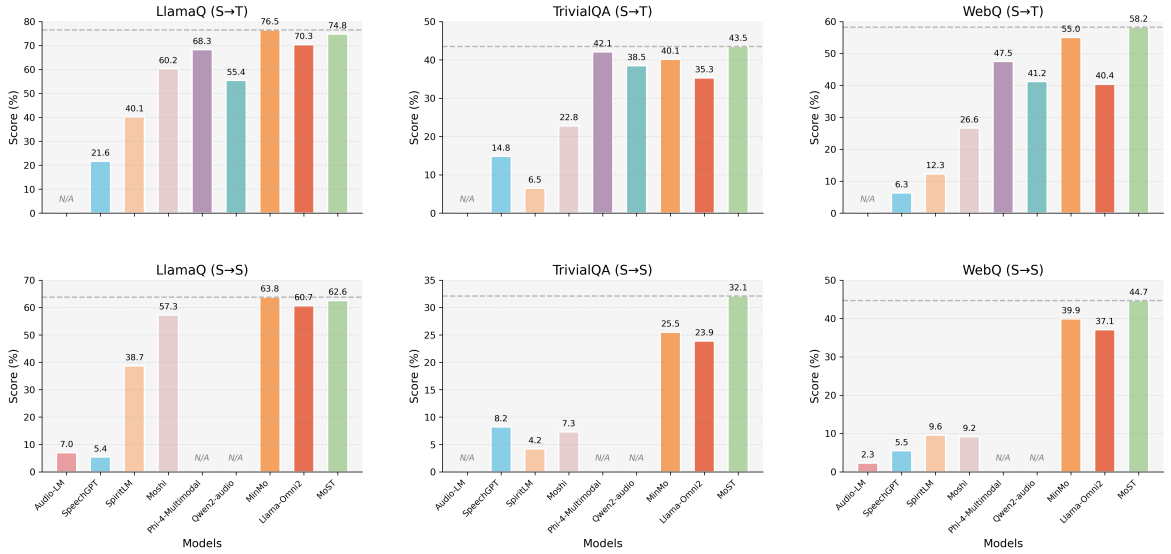

Figure 3. Performance comparison on Spoken Question Answering tasks. MoST achieves competitive or superior performance on all three datasets.

2024) (S→T, S→S), Trivial QA (Joshi et al., 2017) (S→T, S→S), and WebQ (Berant et al., 2013) (S→T, S→S) (task-specific scores/accuracy).

5.2. Evaluation on TTS and ASR Tasks

We evaluate MoST on ASR and TTS across multiple datasets in Table 1 (left panel). MoST demonstrates competitive ASR performance, achieving 2.0% WER on LS-Clean and 3.7% WER on LS-Other, while excelling in challenging cross-dataset scenarios with state-of-the-art 6.2% WER on VoxPopuli-V1.0-en. In TTS synthesis, MoST achieves state-of-the-art performance on VoxPopuli with 10.1% WER, consistently demonstrating strong cross-domain generalization. These results highlight MAMoE's effectiveness in cross-modal tasks and our pipeline's role in creating a versatile speech-text model with strong generalization capabilities.

5.3. Evaluation on Audio Language Modeling Tasks

MoST and baseline models are evaluated on four audio language modeling benchmarks—sWUGGY, sBLIMP (Lakhotia et al., 2021), sTopic-StoryCloze, and sStoryCloze (Hassid et al., 2023)—designed to probe audio language modeling capabilities such as grammatical acceptability and discourse coherence. We report accuracies derived from negative log-likelihood (NLL) scores normalized by sequence length, ensuring a fair comparison across variable-length inputs and tasks.

As shown in Table 1 (right panel), MoST achieves competitive performance with an average score of 71.94, outperforming Phi-4 Multimodal (69.00), MinMo (65.19), and LLaMA-Omni2 (68.39). MoST achieves state-of-the-art performance on sWUGGY (75.28) and sTopic-StoryCloze (83.64), demonstrating strong performance across multiple benchmarks. Full results on all ALM benchmarks are provided in Appendix F. This strong performance underscores MAMoE's efficacy in learning nuanced audio-specific linguistic patterns for enhanced spoken language understanding.

5.4. Evaluation on Spoken Question Answering Tasks

MoST's Spoken Question Answering (SQA) capabilities are evaluated on Llama Q (Nachmani et al., 2024) (S→T and S→S), Trivial QA (Joshi et al., 2017) (S→T and S→S), and WebQ (Berant et al., 2013) (S→T and S→S). As shown in Figure 3, MoST demonstrates strong performance across all SQA tasks, achieving the best or competitive results in most settings. On Llama Q, MoST achieves 74.8 (S→T) and 62.6 (S→S), demonstrating robust speech-to-text understanding while maintaining competitive speech-to-speech performance. On Trivial QA, MoST scores 43.5 (S→T) and a notably high 32.1 (S→S), substantially exceeding most competitors. On WebQ, MoST achieves the best performance with 58.2 (S→T) and 44.7 (S→S), significantly outperforming all baselines including MinMo and LLaMA-Omni2. This strong performance, particularly in challenging S→S settings, highlights MoST's robust end-to-end speech understanding and generation, attributed to the MAMoE architecture.

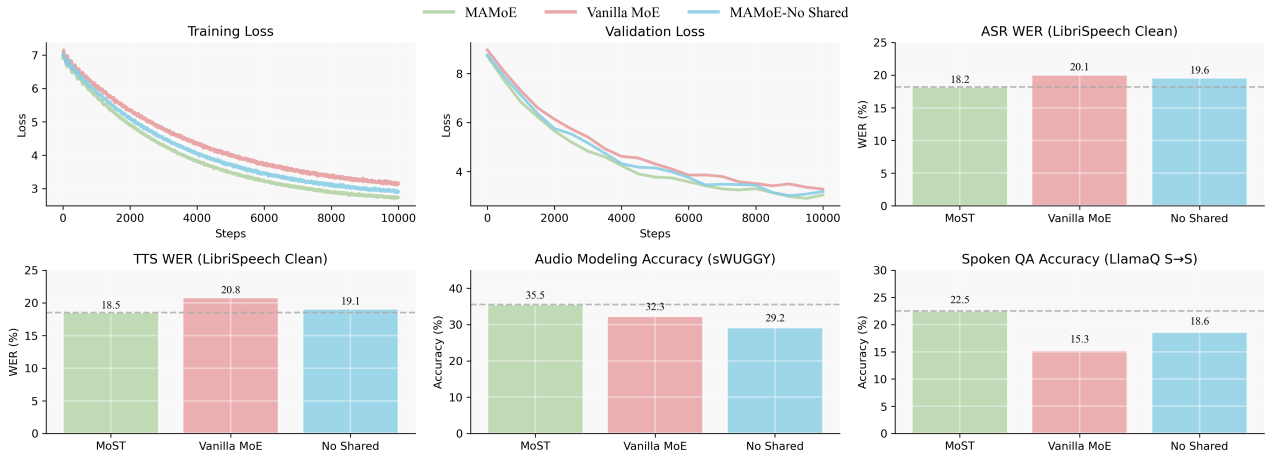

Figure 4. MAMoE design variants analysis. The 2×2 grid displays: ASR, TTS performance (WER) on LibriSpeech dev-clean; Audio language modeling performance (Accuracy) on sWUGGY; Spoken QA performance (Accuracy) on LlamaQ. All metrics are reported at 10,000 training steps. Full MAMoE consistently outperforms ablated versions.

6. Ablation Study: Architecture Benefits with Controlled Comparisons

6.1. Controlled Initialization Comparison

To rigorously isolate the architectural contributions of MAMoE from LLM initialization quality, we conduct a controlled comparison using identical Llama3.2 3B (Grattafiori et al., 2024) initialization across three architectural variants: **Dense Baseline** (standard transformer without MoE), **Traditional MoE Upcycling** (Komatsuzaki et al., 2023) (modality-agnostic routing where all tokens compete for all experts), and **MoST-Style Upcycling** (the proposed MAMoE with modality-specific expert groups and shared experts). This controlled experimental design ensures that performance differences stem solely from architectural choices rather than varying initialization quality.

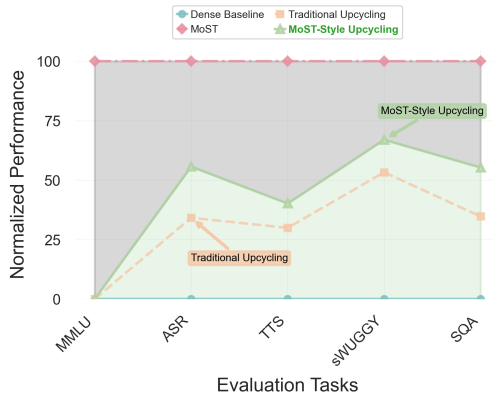

Figure 5. Controlled comparison with identical Llama3.2 3B initialization. We report (100%-WER) for ASR and TTS tasks and accuracy for other tasks.

Figure 5 demonstrates the clear architectural progression benefits. Both MoE variants significantly outperform the dense baseline when starting from identical initialization, confirming MoE's fundamental effectiveness for multimodal tasks. Most importantly, our MoST-style upcycling substantially surpasses traditional upcycling across all evaluation metrics, providing compelling evidence for MAMoE's architectural advantages independent of initialization quality. The consistent performance gains—ranging from 7.3% improvement in ASR to 21.8% in SQA—validate that our proposed modality-aware routing mechanism delivers robust benefits regardless of the starting model. While stronger initialization (DeepSeek-V2 Lite) yields superior absolute performance for the full MoST model, this controlled comparison establishes that our architectural innovations provide consistent improvements across diverse speech-text tasks.

6.2. MAMoE Design Variants Analysis

To validate the individual contributions of MAMoE's core components—modality-aware routing and shared experts—we conduct systematic ablations. We compare MoST against two variants: **MoST-VanillaMoE** (standard MoE routing where all tokens compete for all experts) and **MoST-NoShared** (MAMoE routing but without shared experts, making all experts modality-specific). Both variants maintain the same total number of routed experts as full MoST for fair comparison, precisely isolating the contributions of modality-specific routing and shared expert capacity.

All three models were trained on the combined speech-text instruction-following dataset for 10,000 steps (batch size 256). As illustrated in Figure 4, the full MoST model consistently outperforms both ablated variants across the comprehensive task suite. The comparison with MoST-VanillaMoE

highlights the significant benefits of MAMoE's modality-specific routing mechanism, while MoST's advantage over MoST-NoShared underscores the crucial role of shared experts in enhancing model generalization and facilitating cross-modal understanding. The training and validation loss curves (Figure 4, top row) further corroborate these findings, showing more favorable learning dynamics for the full MAMoE architecture.

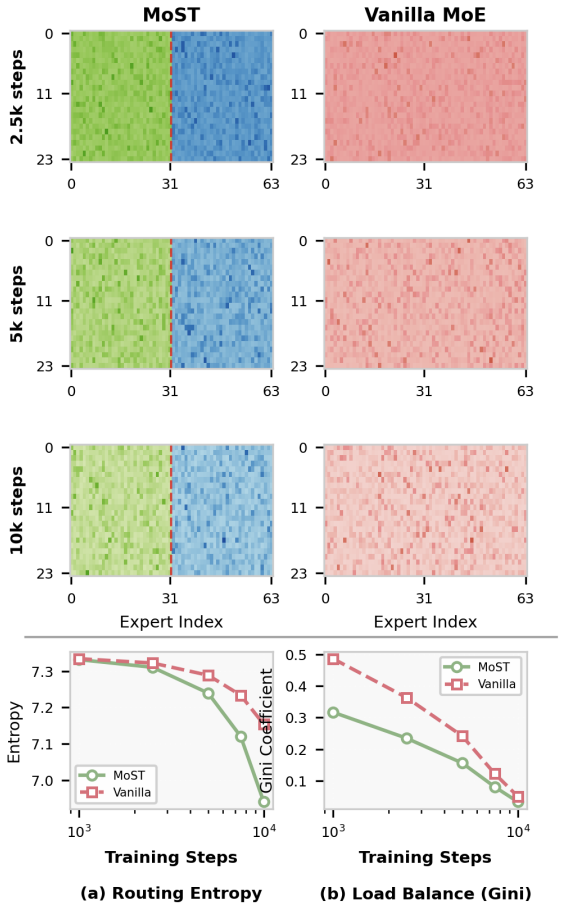

Figure 6. Expert routing analysis for MAMoE versus vanilla MoE. **Top:** Routing frequency heatmaps showing MAMoE's modality-specific routing (left) versus vanilla MoE's diffuse patterns (right) at 2,500, 5,000, and 10,000 training steps. **Bottom:** Quantitative metrics showing MAMoE's superior specialization (lower entropy) and load balancing (lower Gini coefficient) throughout training.

6.3. Expert Routing Analysis

To provide deeper insights into MAMoE's routing mechanism, we analyze expert utilization patterns and quantitative routing metrics. Figure 6 presents a comprehensive visualization combining routing frequency heatmaps (left) with quantitative routing metrics (right) for MoST with MAMoE versus vanilla MoE.

The routing heatmaps reveal that MAMoE exhibits distinct modality-specific routing patterns that become increasingly pronounced during training. Text and audio tokens are systematically directed to their respective expert groups, fostering specialized processing pathways. The quantitative analysis (right panel) confirms that MAMoE achieves progressively lower routing entropy during training, indicating enhanced expert specialization. Besides, MAMoE maintains superior load balancing with consistently lower Gini coefficients compared to vanilla MoE, demonstrating the effectiveness of our routing design.

7. Conclusion

We introduced MoST, a novel multimodal LLM integrating speech and text with a Modality-Aware Mixture of Experts architecture. MAMoE's modality-specific routing and shared experts enhance specialized processing and cross-modal interaction. Our efficient pipeline transforms a pre-trained MoE LLM into a data-efficient speech-text model.

Empirical evaluations confirm MoST achieves state-of-the-art or competitive results across ASR, TTS, audio language modeling, and SQA, outperforming strong baselines.

We release MoST's model checkpoints, codebase, and curated data to encourage further research and development in efficient and effective multimodal models.

While our index-based 50% expert partition proves effective, it represents a simple initialization strategy. More sophisticated approaches—such as clustering-based or knowledge-preserving partitioning—offer promising directions for future modality-aware MoE research (detailed in Appendix D).

Future work presents several other exciting avenues. A primary direction is extending the MAMoE concept to incorporate additional modalities, particularly vision, to create a truly comprehensive audio-visual-language model. Further research could also explore scaling MoST to larger parameter counts, investigating different types of expert specializations, and evaluating its capabilities on a broader range of downstream speech and multimodal tasks.

Acknowledgement

Yang You's research group is being sponsored by NUS startup grant (Presidential Young Professorship), Singapore MOE Tier-1 grant, ByteDance grant, ARCTIC grant, SMI grant and Alibaba grant. This work is also supported by NSF 2048280, 2325121, 2244760, 2331966 and ONR N0001423-1-2300:P00001. Additionally, We thank the open-source community for releasing code and datasets that made this research possible.

Impact Statement

The paper presents work whose goal is to advance the field of multimodal machine learning by introducing Modality-Aware Mixture of Experts (MAMoE) and offers a scalable and efficient alternative to dense transformer paradigms for future foundation model design. Furthermore, by achieving competitive performance exclusively via open-source datasets and releasing the full training pipeline, this work lowers barriers to entry and fosters rigorous reproducibility in large-scale speech-text research, encouraging further exploration into efficient, sparse multimodal architectures. As for potential social consequences, while our approach enables benefits such as enhanced accessibility tools, we also acknowledge the importance of mitigating potential misuse of generative speech capabilities, such as voice spoofing.

References

Allal, L. B., Lozhkov, A., Bakouch, E., Blázquez, G. M., Penedo, G., Tunstall, L., Marafioti, A., Kydlíček, H., Lajarín, A. P., Srivastav, V., Lochner, J., Fahlgren, C., Nguyen, X.-S., Fourrier, C., Burtenshaw, B., Larcher, H., Zhao, H., Zakka, C., Morlon, M., Raffel, C., von Werra, L., and Wolf, T. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. Common voice: A massively-multilingual speech corpus, 2020a. URL https://arxiv.org/abs/1912.06670.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. Common voice: A massively-multilingual speech corpus, 2020b. URL https://arxiv.org/abs/1912.06670.

Berant, J., Chou, A., Frostig, R., and Liang, P. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1160.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., and Zeghidour, N. Audiolm: a language modeling approach to audio generation, 2023. URL https://arxiv.org/abs/2209.03143.

Chen, Q., Chen, Y., Chen, Y., Chen, M., Chen, Y., Deng, C., Du, Z., Gao, R., Gao, C., Gao, Z., Li, Y., Lv, X., Liu, J., Luo, H., Ma, B., Ni, C., Shi, X., Tang, J., Wang, H., Wang, H., Wang, W., Wang, Y., Xu, Y., Yu, F., Yan, Z., Yang, Y., Yang, B., Yang, X., Yang, G., Zhao, T., Zhang, Q., Zhang, S., Zhao, N., Zhang, P., Zhang, C., and Zhou, J. Minmo: A multimodal large language model for seamless voice interaction, 2025. URL https://arxiv.org/abs/2501.06282.

Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., and Zhou, J. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Communication, S., Barrault, L., Chung, Y.-A., Meglioli, M. C., Dale, D., Dong, N., Duppenthaler, M., Duquenne, P.-A., Ellis, B., Elsahar, H., Haaheim, J., Hoffman, J., Hwang, M.-J., Inaguma, H., Klaiber, C., Kulikov, I., Li, P., Licht, D., Maillard, J., Mavlyutov, R., Rakotoarison, A., Sadagopan, K. R., Ramakrishnan, A., Tran, T., Wenzek, G., Yang, Y., Ye, E., Evtimov, I., Fernandez, P., Gao, C., Hansanti, P., Kalbassi, E., Kallet, A., Kozhevnikov, A., Gonzalez, G. M., Roman, R. S., Touret, C., Wong, C., Wood, C., Yu, B., Andrews, P., Balioglu, C., Chen, P.-J., Costa-jussà, M. R., Elbayad, M., Gong, H., Guzmán, F., Heffernan, K., Jain, S., Kao, J., Lee, A., Ma, X., Mourachko, A., Peloquin, B., Pino, J., Popuri, S., Ropers, C., Saleem, S., Schwenk, H., Sun, A., Tomasello, P., Wang, C., Wang, J., Wang, S., and Williamson, M. Seamless: Multilingual expressive and streaming speech translation, 2023. URL https://arxiv.org/abs/2312.05187.

DeepSeek-AI, Liu, A., Feng, B., Wang, B., Wang, B., Liu, B., Zhao, C., Dengr, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Xu, H., Yang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Chen, J., Yuan, J., Qiu, J., Song, J., Dong, K., Gao, K., Guan, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Pan, R., Xu, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Zheng, S., Wang, T., Pei, T., Yuan, T., Sun, T., Xiao, W. L., Zeng, W., An, W., Liu, W., Liang, W., Gao, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Chen, X., Nie, X., Sun, X., Wang, X., Liu, X., Xie, X., Yu, X., Song, X., Zhou, X., Yang, X., Lu, X., Su, X., Wu, Y., Li, Y. K., Wei, Y. X., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Zheng, Y., Zhang, Y., Xiong, Y., Zhao, Y., He, Y., Tang, Y., Piao, Y., Dong, Y., Tan, Y., Liu, Y., Wang, Y., Guo, Y., Zhu, Y., Wang, Y., Zou, Y., Zha, Y., Ma, Y.,

Yan, Y., You, Y., Liu, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Huang, Z., Zhang, Z., Xie, Z., Hao, Z., Shao, Z., Wen, Z., Xu, Z., Zhang, Z., Li, Z., Wang, Z., Gu, Z., Li, Z., and Xie, Z. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL https://arxiv.org/abs/2405.04434.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., Zoph, B., Fedus, L., Bosma, M., Zhou, Z., Wang, T., Wang, Y. E., Webster, K., Pellat, M., Robinson, K., Meier-Hellstern, K., Duke, T., Dixon, L., Zhang, K., Le, Q. V., Wu, Y., Chen, Z., and Cui, C. Glam: Efficient scaling of language models with mixture-of-experts, 2022. URL https://arxiv.org/abs/2112.06905.

Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., and Zeghidour, N. Moshi: a speech-text foundation model for real-time dialogue, 2024. URL https://arxiv.org/abs/2410.00037.

Fang, Q., Zhou, Y., Guo, S., Zhang, S., and Feng, Y. Llama-omni2: Llm-based real-time spoken chatbot with autoregressive streaming speech synthesis, 2025. URL https://arxiv.org/abs/2505.02625.

Fathullah, Y., Wu, C., Lakomkin, E., Jia, J., Shangguan, Y., Li, K., Guo, J., Xiong, W., Mahadeokar, J., Kalinli, O., Fuegen, C., and Seltzer, M. Prompting large language models with speech recognition abilities, 2023. URL https://arxiv.org/abs/2307.11795.

Fedus, W., Zoph, B., and Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity, 2022. URL https://arxiv.org/abs/2101.03961.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H.,

Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Hassid, M., Remez, T., Nguyen, T. A., Gat, I., Conneau, A., Kreuk, F., Copet, J., Defossez, A., Synnaeve, G., Dupoux, E., et al. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36: 63483–63501, 2023.

Hassid, M., Remez, T., Nguyen, T. A., Gat, I., Conneau, A., Kreuk, F., Copet, J., Defossez, A., Synnaeve, G., Dupoux, E., Schwartz, R., and Adi, Y. Textually pretrained speech language models, 2024. URL https://arxiv.org/abs/2305.13009.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units, 2021. URL https://arxiv.org/abs/2106.07447.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Barzilay, R. and Kan, M.-Y. (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Kang, W., Yang, X., Yao, Z., Kuang, F., Yang, Y., Guo, L., Lin, L., and Povey, D. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context, 2024. URL https://arxiv.org/abs/2309.08105.

Komatsuzaki, A., Puigcerver, J., Lee-Thorp, J., Ruiz, C. R., Mustafa, B., Ainslie, J., Tay, Y., Dehghani, M., and Houlsby, N. Sparse upcycling: Training mixture-of-experts from dense checkpoints, 2023. URL https://arxiv.org/abs/2212.05055.

Kong, J., Kim, J., and Bae, J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, 2020. URL https://arxiv.org/abs/2010.05646.

Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., et al. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021.

Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., and Chen, Z. Gshard: Scaling giant models with conditional computation and automatic sharding, 2020. URL https://arxiv.org/abs/2006.16668.

Lin, X. V., Shrivastava, A., Luo, L., Iyer, S., Lewis, M., Ghosh, G., Zettlemoyer, L., and Aghajanyan, A. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts, 2024. URL https://arxiv.org/abs/2407.21770.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Microsoft, :, Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., Chen, D., Chen, D., Chen, J., Chen, W., Chen, Y.-C., ling Chen, Y., Dai, Q., Dai, X., Fan, R., Gao, M., Gao, M., Garg, A., Goswami, A., Hao, J., Hendy, A., Hu, Y., Jin, X., Khademi, M., Kim, D., Kim, Y. J., Lee, G., Li, J., Li, Y., Liang, C., Lin, X., Lin, Z., Liu, M., Liu, Y., Lopez, G., Luo, C., Madan, P., Mazalov, V., Mitra, A., Mousavi, A., Nguyen, A., Pan, J., Perez-Becker, D., Platin, J., Portet, T., Qiu, K., Ren, B., Ren, L., Roy, S., Shang, N., Shen, Y., Singhal, S., Som, S., Song, X., Sych, T., Vaddamanu, P., Wang, S., Wang, Y., Wang, Z., Wu, H., Xu, H., Xu, W., Yang, Y., Yang, Z., Yu, D., Zabir, I., Zhang, J., Zhang, L. L., Zhang, Y., and Zhou, X. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025. URL https://arxiv.org/abs/2503.01743.

Nachmani, E., Levkovitch, A., Hirsch, R., Salazar, J., Asawaroengchai, C., Mariooryad, S., Rivlin, E., Skerry-Ryan, R., and Ramanovich, M. T. Spoken question answering and speech continuation using spectrogram-powered llm, 2024. URL https://arxiv.org/abs/2305.15255.

Nguyen, T. A., de Seyssel, M., Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., Dunbar, E., and Dupoux, E. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling, 2020. URL https://arxiv.org/abs/2011.11588.

Nguyen, T. A., Muller, B., Yu, B., Costa-jussa, M. R., Elbayad, M., Popuri, S., Ropers, C., Duquenne, P.-A., Algayres, R., Mavlyutov, R., Gat, I., Williamson, M., Synnaeve, G., Pino, J., Sagot, B., and Dupoux, E. Spirit lm: Interleaved spoken and written language model, 2024. URL https://arxiv.org/abs/2402.05755.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., and Dean, J. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer, 2017. URL https://arxiv.org/abs/1701.06538.

Shen, L., Chen, G., Shao, R., Guan, W., and Nie, L. Mome:

Mixture of multimodal experts for generalist multimodal large language models, 2024. URL https://arxiv.org/abs/2407.12709.

Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., and Dupoux, E. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation, 2021. URL https://arxiv.org/abs/2101.00390.

Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., and Qiu, X. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities, 2023. URL https://arxiv.org/abs/2305.11000.

Zhang, D., Yu, Y., Dong, J., Li, C., Su, D., Chu, C., and Yu, D. Mm-llms: Recent advances in multimodal large language models, 2024. URL https://arxiv.org/abs/2401.13601.

A. Detailed Training Recipe

To effectively realize MoST, we develop an efficient LLM-to-Speech-Text Transformation Pipeline. Starting from a pretrained MoE LLM, we first conduct targeted post-training on large ASR and TTS datasets. This stage primes the modality-specific experts and adapts the shared experts for cross-modal tasks. Subsequently, we perform instruction fine-tuning using a carefully curated speech-text instruction dataset. Our training strategy involves mixed-task learning and incorporates techniques to prevent catastrophic forgetting of foundational abilities during fine-tuning. Critically, this pipeline demonstrates remarkable data efficiency, achieving strong performance with significantly fewer training tokens compared to prominent open-source speech-text models like SpiritLM (Nguyen et al., 2024) and Moshi (Défossez et al., 2024).

Table 2 provides a comprehensive overview of our two-stage training protocol, including key hyperparameters, task mixing ratios, and dataset composition details for each stage.

Table 2. Training Configuration Overview. Detailed breakdown of training stages, configurations, task mixing ratios, and dataset compositions.

Training Stage	Configuration	Task Mixing Ratios	Dataset Composition Details
Stage 1: Cross-Modal Post-Training	**Steps:** 500k **Batch Size:** 512	**ASR:** 0.4 **TTS:** 0.4 **Text LM:** 0.2	**ASR/TTS:** LibriHeavy (0.6), Common Voice (0.2), VoxPopuli (0.2) **LM:** RefinedWeb (to prevent forgetting)
Stage 2: Mixed Instruction Tuning	**Steps:** 10k **Batch Size:** 128	**Speech Instruct:** 0.4 **Text Instruct:** 0.4 **ASR:** 0.1 **TTS:** 0.1	**Speech Instruct:** Synthesized from SmolTalk (via MoST Stage 1) **Text Instruct:** Standard open-source instruction data

A.1. ASR and TTS Data Preparation (Stage 1)

The initial stage of our LLM-to-Speech-Text transformation pipeline involves post-training on large-scale Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) datasets. This step is crucial for priming the modality-specific experts within the MAMoE architecture and adapting the shared experts for fundamental cross-modal alignment tasks.

We utilize two primary open data sources for this stage: the Common Voice corpus (Ardila et al., 2020b) and the full LibriHeavy dataset (Kang et al., 2024), selected for their scale and diversity in speech patterns and linguistic content. The ASR data follows the standard format where the input is speech (represented as acoustic features or tokens) and the output is the corresponding text transcription, guided by an instruction like "Transcribe the following speech into text."

To generate the necessary TTS data, we systematically transform the ASR datasets. For each ASR sample {`instruction: "Transcribe...", input: speech, output: text`}, we create a corresponding TTS sample by reversing the input and output modalities and modifying the instruction. The resulting TTS sample format is {`instruction: "Convert the following text into speech.", input: text, output: speech`}. This conversion process, (further details are provided in this appendix), ensures that the model is trained on paired speech-text data for both recognition and synthesis tasks, effectively leveraging existing ASR resources to create aligned TTS training data. This phase prepares the model for more complex instruction-following tasks in the subsequent fine-tuning stage.

A.2. Speech-Text Instruction-Following Data Construction (Stage 2)

Following the initial post-training phase, MoST undergoes instruction fine-tuning to enhance its ability to follow complex commands and engage in nuanced interactions involving both speech and text. This requires a diverse dataset of speech-text instructions. We construct this dataset through two complementary strategies: synthesizing interrupted dialogues and converting text-based instructions into a speech format.

A.2.1. Interrupted Dialogue Synthesis Pipeline

Standard instruction datasets often lack the dynamic turn-taking and interruptions characteristic of natural human conversation. To address this, we develop a pipeline to augment existing dialogue datasets with realistic interruptions. We utilize the

open-source SmolTalk dataset (Allal et al., 2025) as our base, which contains multi-turn user-assistant conversations.

Our synthesis pipeline employs a large language model (LLM) in a two-step process. First, the LLM analyzes each dialogue from the text instruction-following dataset to determine its suitability for adding a user interruption. Suitability criteria include clear role separation (user/assistant), the presence of longer assistant turns amenable to interruption, and contextual appropriateness for an interruption. Second, for dialogues identified as suitable, the LLM is prompted to rewrite the conversation by inserting a natural user interruption during one of the assistant's longer responses. This interruption typically takes the form of a brief clarifying question related to the assistant's immediately preceding statement. The LLM is instructed to modify the assistant's response to briefly address the interruption before seamlessly transitioning back to its original point (e.g., "A host cell is... Now, as I was saying..."). The remainder of the dialogue remains unchanged. This process yields a dataset enriched with more complex conversational flows, better preparing MoST for interactive scenarios.

A.2.2. Text-to-Speech Instruction Dataset Conversion

To enable our model with both speech and text instruction-following abilities, we employ a strategy where we leverage the capabilities of our own MoST model, specifically after its initial post-training on ASR and TTS data (as described in Appendix A.1).

For a given text-based instruction sample (e.g., {`instruction: text_prompt, input: text_input, output: text_response`}), we use the ASR-TTS trained MoST model to synthesize the speech equivalent of the `text_response`. This converts the sample into a speech-output format (e.g., {`instruction: text_prompt, input: text_input, output: speech_response`}). Depending on the target task, we can also synthesize speech for the `text_input` or even the `instruction` itself, creating varied multimodal instruction formats. This approach allows us to efficiently repurpose vast quantities of high-quality text instructions for speech-based fine-tuning without the need for manual speech data collection, significantly broadening the range of tasks MoST is trained on.

A.3. Mixed Speech-Text Instruction Fine-tuning and Hyper-parameters (Stage 3)

Our training process for MoST (after Stage 1 post-training described in Appendix A.1 and Stage 2 data construction in Appendix A.2) proceeds to instruction fine-tuning.

Instruction Fine-tuning Strategy. This stage uses the curated dataset comprising synthesized interrupted dialogues and text-to-speech converted instructions. The goal is to teach the model complex reasoning, interaction, and command-following involving both modalities. Crucially, to mitigate catastrophic forgetting of the foundational ASR and TTS capabilities learned in Stage 1, the instruction fine-tuning data is mixed with a significant portion of the original ASR and TTS data. This ensures that MoST retains its core speech processing abilities while acquiring higher-level instruction-following competence. The mixing ratio between new instruction data and the ASR/TTS data is carefully managed throughout this stage.

B. Training Settings

This section details the hyper-parameters used for training MoST and the computing resources utilized for the experiments.

B.1. Training Hyper-parameters

Across all training stages (post-training and fine-tuning), we utilize the AdamW optimizer (Loshchilov & Hutter, 2019) with a cosine learning rate schedule, including a warm-up phase. Specific values for the peak learning rate, weight decay, batch size, and the exact number of training steps or epochs for each stage, along with the mixing ratios used during instruction fine-tuning are as follows:

Peak Learning Rate: 5e-5, Weight Decay: 0.01, Batch Size for ASR-TTS Post-training: 256, Batch Size for Mixed Instruction-Following Fine-tuning: 128, Fine-tuning Steps: 10000, Warmup Steps: 1000, Mixing Ratio (Instruction:ASR/TTS): 1:1. These parameters were chosen based on preliminary experiments to ensure stable and efficient convergence.

B.2. Experiment Computing Resources

The training experiments for MoST were conducted using a distributed setup comprising 48 NVIDIA A100 GPUs. Specifically, this involved 6 nodes, each equipped with 8 A100 GPUs.

C. Model Configurations

Table 3 details the model configurations used for MoST.

Table 3. Model configurations from `config_mostc.json`

Parameter	Value
architectures	`["MoSTCForCausalLM"]`
attention_bias	false
attention_dropout	0.0
audio_expert_indices	`[32, 33, ..., 62, 63]`
auto_map	
AutoConfig	`configuration_MoST.MoSTCConfig`
AutoModel	`modeling_most.MoSTCModel`
AutoModelForCausalLM	`modeling_most.MoSTCForCausalLM`
aux_loss_alpha	0.001
bos_token_id	100000
eos_token_id	100001
hidden_act	silu
hidden_size	2048
intermediate_size	10944
num_hidden_layers	27
rope_scaling	
beta_fast	32
beta_slow	1
factor	40
type	yarn
vocab_size	102400
use_cache	true
torch_dtype	bfloat16
hubert_model_path	`/.../hubert_MoSTC/mhubert_base_25hz.pt`
quantizer_model_path	`/.../hubert_MoSTC/L11_quantizer_500.pt`

D. Limitations, Social Impacts and Future Work

MoST presents both potential societal benefits (e.g., accessibility) and risks (e.g., voice spoofing).

D.1. Expert Partition Strategy

A key design decision in MAMoE is the 50% partition of the pretrained MoE LLM's experts into text-specific and audio-specific groups based on expert indices. We acknowledge that this hard partition, while effective, represents a relatively simple initialization strategy with potential limitations:

Rationale for Index-Based Partition. In standard sparse MoE models like DeepSeek-V2, experts are randomly initialized and do not exhibit semantic clustering (e.g., specialized "text experts" vs. "math experts") prior to training. Consequently, a random index-based split at initialization is statistically neutral—no particular subset of experts has inherent advantages for any specific modality. Our Stage 1 Cross-Modal Post-Training is explicitly designed to specialize these initially random experts into their assigned modalities through targeted ASR/TTS training.

Role of Shared Experts. Critically, the Shared Experts in MAMoE process every token regardless of modality, providing a

mechanism to preserve general capabilities and facilitate cross-modal knowledge transfer. This architectural choice helps mitigate the risk of losing valuable knowledge during the hard partition, as evidenced by MoST's strong performance on text-based benchmarks (Appendix E).

Future Directions. While our approach demonstrates strong empirical results, we recognize that the index-based partition may not represent the theoretical optimum. More sophisticated initialization strategies represent important directions for future research, including: (1) *Clustering-based initialization*: Analyzing pretrained expert activations on representative data to group experts by their learned specializations before partitioning; (2) *Knowledge-preserving partitioning*: Developing methods to identify and preserve critical expert subsets for each modality; (3) *Adaptive expert allocation*: Dynamically adjusting the ratio of modality-specific experts during training based on task requirements.

Future work presents several exciting avenues. A primary direction is extending the MAMoE concept to incorporate additional modalities, particularly vision, to create a truly comprehensive audio-visual-language model. Further research could also explore scaling MoST to larger parameter counts, investigating different types of expert specializations, and evaluating its capabilities on a broader range of downstream speech and multimodal tasks.

E. Text-Based Evaluations

To comprehensively assess MoST's capabilities beyond speech-specific tasks, we evaluate its performance on standard text-based benchmarks. Table 4 presents results across four diverse text benchmarks: MMLU (general knowledge), TriviaQA (factual question answering), GSM8K (mathematical reasoning), and HumanEval (code generation).

Table 4. Performance on Text-Based Evaluation Benchmarks. MoST demonstrates strong text understanding and generation capabilities, outperforming most speech-native models while maintaining competitive performance with text-focused systems. Higher scores are better. **MoST (Ours)** results are highlighted with green background.

Model	MMLU	TriviaQA	GSM8K	HumanEval
MinMo	**58.5**	65.8	58.1	42.3
Llama-Omni2	44.7	35.2	21.8	13.2
Moshi	49.8	48.5	40.3	25.4
Qwen2-Audio	45.2	34.1	24.3	11.7
SpiritLM	36.9	42.0	21.5	9.7
MoST (Ours)	55.4	**67.1**	**66.9**	**58.2**

MoST significantly outperforms other open-source speech-native models (SpiritLM, Moshi, Qwen2-Audio) and recent end-to-end models like Llama-Omni2 across text tasks. MoST also maintains strong general reasoning capabilities (GSM8K 66.9) and coding abilities (HumanEval 58.2), confirming that our shared-expert design successfully mitigates catastrophic forgetting. While MinMo achieves a higher MMLU score (58.5 vs 55.4), MoST demonstrates superior performance on TriviaQA, GSM8K, and HumanEval, highlighting the effectiveness of our MAMoE architecture in preserving and enhancing text-based capabilities during multimodal training.

F. Full Audio Language Modeling Results

Table 5 presents the complete evaluation results on all four audio language modeling benchmarks. MoST achieves state-of-the-art performance on sWUGGY (75.28), sTopic-StoryCloze (83.64), and the overall average (71.94), while remaining competitive on sBLIMP and sStoryCloze. AudioLM achieves the best performance on sBLIMP (64.70), and LLaMA-Omni2 leads on sStoryCloze (68.55). These results demonstrate MoST's strong overall audio language understanding capabilities across diverse linguistic phenomena.

G. Use of LLMs

As this work is centered on Large Language Models (LLMs), they are inherently part of our study. We clarify the roles of LLMs in four distinct but limited capacities:

- **Base model**. The core of our proposed MoST model is built upon an existing LLM as its backbone, consistent with the

Table 5. Full results on Audio Language Modeling Tasks. Accuracy (↑) is reported for all benchmarks. Best results are **bold**. MoST row highlighted with (↑X.X) indicating improvement over previous best where MoST achieves state-of-the-art.

Model	sWUGGY	sBLIMP	sTopic-StoryCloze	sStoryCloze	Average
AudioLM	71.50	**64.70**	–	–	–
SpeechGPT	51.82	49.75	60.13	53.13	53.71
SpiritLM	40.14	48.28	83.32	58.95	57.67
Moshi	51.14	53.31	46.34	45.16	48.99
Phi-4 Multimodal	71.84	60.21	81.55	62.39	69.00
MinMo	68.59	55.43	75.43	61.29	65.19
LLaMA-Omni2	73.21	53.59	78.21	**68.55**	68.39
MoST (Ours)	**75.28**(↑2.1)	63.42	**83.64**(↑0.3)	65.43	**71.94**(↑2.9)

research objective of studying and advancing LLM-based methods. And we include it as part of reproducibility as well.

- **LM-guided synthesis.** In data synthesis section, LLMs were used as judges and converters to evaluate candidate outputs and perform format conversion. These roles were auxiliary, and MoST's effectiveness does not depend on LLM-specific heuristics.
- **Baselines.** We adopted open-source LLMs as baselines to ensure fair and transparent comparisons, with results reproduced from official or widely accepted implementations.
- **Language correction.** LLMs were employed only for grammar correction and minor error detection in the manuscript, without altering the scientific content or analysis.

Overall, LLMs were employed only in auxiliary or benchmarking roles. Importantly, LLMs were never used for idea generation or the formulation of scientific arguments. All core algorithmic design, experimental findings, and analyses remain fully attributable to our proposed methodology.