

# openLISTEN: Freestyle Cross-Modal Instruction Compliance for Large Speech-Language Models with Limited Resources

Anonymous ACL submission

## Abstract

Recent advancements in Large Language Models (LLMs) have revolutionized text comprehension, yet bridging the gap to speech-native understanding remains a challenge due to the loss of paralinguistic features in cascaded systems and the high computational costs of end-to-end vocabulary expansion. To address these limitations, we propose *openLISTEN*, a resource-efficient Large Speech-Language Model (LSLM) trained entirely on consumer-grade GPUs. *openLISTEN* integrates Gated Cross-Attention (GCA) with Open-Domain Cross-Modal Instruction Tuning to learn robust audio-text alignment from only 500+ hours of paired data. Extensive evaluations on URO-Bench demonstrate that *openLISTEN* achieves competitive performance against state-of-the-art (SOTA) baselines with significantly fewer parameters, while controlled ablations under resource-efficient training consistently favor GCA over alternative fusion designs. Furthermore, empirical results on freestyle instruction compliance benchmarks indicate that our approach effectively mitigates the rigid response patterns and modality bias typically exacerbated by limited training data, thereby significantly enhancing instruction adherence and generalization in freestyle scenarios. The code will be available at <https://anonymous.4open.science/r/openLISTEN-8D11>

## 1 Introduction

Recent advancements in Large Language Models (LLMs) (Liu et al., 2024; OpenAI, 2023) have revolutionized text comprehension. However, real-world communication is inherently multimodal, relying heavily on speech rich in prosody and emotion. To bridge this gap, researchers are advancing *Large Speech-Language Models* (LSLMs). Early efforts adopted a *cascaded* paradigm (Huang et al., 2024; Liu et al., 2025), utilizing LLMs as controllers for Automatic Speech Recognition (ASR)

and Text-to-Speech (TTS) tools. While effective, this modular design converts speech to intermediate text, stripping away essential paralinguistic features and degrading interaction naturalness.

To overcome the limitations, recent work has shifted toward *end-to-end* paradigms using neural audio codecs (Zeghidour et al., 2022; Du et al., 2024a; Ji et al., 2025) to discretize speech into tokens, expanding LLM vocabularies to jointly model speech and text (Zhang et al., 2023; Rubenstein et al., 2023). However, this typically necessitates full LLM retraining and introduces significant challenges, including modality distribution mismatches, codebook inefficiencies, and the high computational cost of vocabulary expansion. These constraints hinder deployment in resource-limited settings. Alternatively, the direct projection methods like (Tang et al., 2024; Zeng et al., 2024; Chu et al., 2023, 2024) employ a lightweight speech adapter. By aligning speech features with a frozen LLM’s text embedding space, they circumvent vocabulary expansion, preserving multimodal capabilities while mitigating computational overhead.

Despite its advantages, this direct projection method presents substantial challenges in achieving effective cross-modal alignment. Fundamentally, the representation gap between continuous speech and discrete text, compounded by statistical distribution mismatches, hinders convergence. Most critically, the speech adapter must simultaneously perform two demanding tasks: modeling semantic content while decoupling acoustic features – a dual objective that substantially increases training complexity. Current solutions rely heavily on massive-scale, diverse multimodal datasets to improve alignment. For example, recent approaches rely on extensive scaling, often employing upwards of 500K or even 1,000K hours of pretraining data. However, as noted in (Yang et al., 2025), with limited training data and few cross-modal tasks, modality alignment remains suboptimal. This leads to a

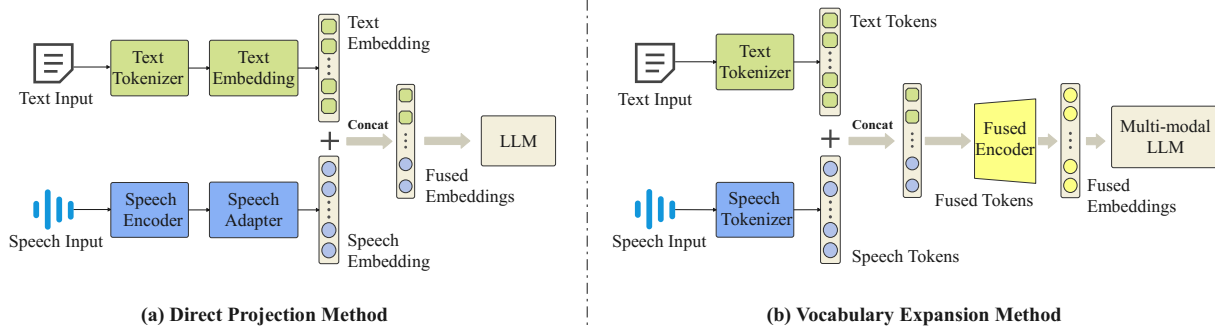


Figure 1: Schematic diagram of the two mapping paradigms: (a) The Direct Projection Method aligns the speech embeddings with the text embedding space through a speech adapter. (b) The Vocabulary Expansion Method expands the speech codebook into the original vocabulary space of the LLM using a speech tokenizer;

critical limitation: models develop rigid response patterns biased toward dominant modalities, often failing to properly respond to user instructions or remaining silent. Such behavior severely undermines the generalization and flexibility of LSLMs in freestyle instruction compliance scenarios.

To address these limitations, we propose *openLISTEN*, an LSLM designed for the seamless integration of speech and text processing. For speech-text pair inputs, *openLISTEN* employs *Gated Cross-Attention (GCA) mechanisms*, which significantly enhance speech-text representation alignment through the use of self-attention and gated selection mechanisms. Combined with open-domain instruction tuning, this architecture achieves robust cross-modal alignment and generalization capabilities using only over 500 hours text-audio pairs. Notably, compared to existing methods, *openLISTEN* requires only a few hundred hours of speech-text pairs for training, and can be trained on just four GeForce RTX 4090D GPUs, while maintaining excellent modal alignment and achieving efficient freestyle instruction compliance capabilities.

Our contributions are as follows:

- We propose *openLISTEN*, a resource-efficient LSLM that enables training entirely on consumer-grade GPUs, significantly lowering the barrier for LSLMs research.
- We introduce a GCA mechanism combined with an open-domain instruction tuning strategy, effectively addressing the instruction compliance issue in multimodal interaction.
- On URO-Bench (Yan et al., 2025) and freestyle instruction compliance benchmarks, *openLISTEN* achieves strong results, with controlled studies showing consistent gains

from GCA over alternative fusion designs.

## 2 Related work

According to (Yang et al., 2025), implementation generally falls into two categories: direct projection and vocabulary expansion methods. The direct projection method (see Fig. 1 (a)) uses a trainable speech adapter to map continuous speech representations into the LLM’s text space, preserving linguistic capabilities via minimal updates. LLaMA-Omni (Fang et al., 2025), SLAM-Omni (Chen et al., 2025), and Mini-Omni (Xie and Wu, 2024a,b) employ frozen Whisper encoders with LoRA (Hu et al., 2022). SALMONN (Tang et al., 2024) utilizes a dual-encoder system (Whisper and BEATs) fused via a Q-Former (Li et al., 2023b). Conversely, Qwen-Audio (Chu et al., 2023, 2024) co-optimizes the encoder and speech adapter, while Freeze-Omni (Wang et al., 2025a) uses a flow-based encoder pretrained on ASR data with trainable prompt embeddings. Xue et al. (Xue et al., 2025) focus on the performance of LSLMs on non-core languages. Despite parameter efficiency, these methods often suffer from poor modality alignment, causing errors in cross-modal instruction interpretation.

The vocabulary expansion method (see Fig. 1 (b)) incorporates discretized audio codebooks directly into the LLM vocabulary. SpeechGPT employs HuBERT (Hsu et al., 2021) with  $k$ -means discretization. Recognizing that discretization strategy is critical to performance, AudioPaLM systematically evaluated frameworks like w2v-BERT (Chung et al., 2021) and USM (Resler et al., 2017). However, these approaches require extensive cross-modal alignment datasets and full-parameter fine-tuning, increasing training complexity and restricting academic adoption.

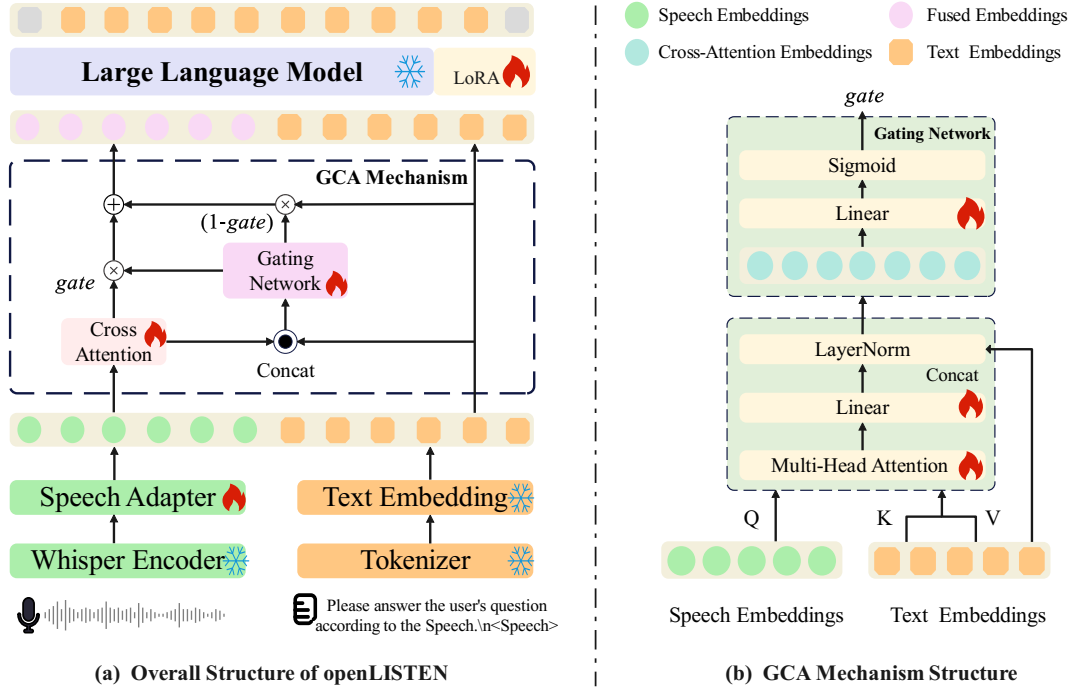


Figure 2: (a) The overall structure of the openLISTEN. Speech and text embeddings are fused via the GCA mechanism and then fed into the LLM. (b) GCA Mechanism Structure. Incorporating Multi-Head Cross-Attention and a Gating Network, this module dynamically integrates cross-modal features through a gated residual connection.

### 3 openLISTEN

This section outlines the architecture of openLISTEN, followed by details on the GCA mechanism, model training and optimization strategies.

#### 3.1 Overview

The overall architecture of openLISTEN is illustrated in Fig. 2(a). For multimodal alignment, we extract and compress speech embeddings using a frozen Whisper encoder, which was trained for speech recognition and translation but retains rich speech information capable of waveform reconstruction (Gong et al., 2022; Chu et al., 2023). These embeddings are then compressed and aligned via a speech adapter consisting of two fully connected layers. For text processing, we tokenize input text using the Llama-3.2-3B-Instruct tokenizer and fuse the resulting discrete tokens with the speech embeddings through our proposed GCA module (Fig. 2(b)). This alignment produces a unified representation, which serves as the final input to the LLM.

#### 3.2 Multimodal Alignment Through Gated Cross-Attention Mechanisms

For speech input, a Whisper encoder is employed to extract audio features from the user’s speech

at a frame rate of 50 Hz as  $A = [a_1, a_2, \dots, a_N]$ . The Whisper encoder automatically pads the input audio to a 30-second duration, leading to speech embeddings that are overly long and contain a significant amount of redundant information. This increases the computational load. To alleviate this, the extracted speech embeddings are downsampled along the temporal dimension to produce intermediate representations:  $A^I = [a_1^I, a_2^I, \dots, a_{N'}^I]$ . These intermediate representations  $A^I$  are then projected via a linear projector to  $A^P$  to ensure alignment with the text embedding dimensions of the LLM. Formally, this is defined as  $A^P = \text{Linear}(A^I)$ . Simultaneously, the instruction text is processed by a tokenizer and an embedding layer to obtain text embeddings  $T^I \in \mathbb{R}^{M \times d}$ , where  $M$  denotes the sequence length of the instruction.

After processing through the speech adapter, simple concatenation of speech and text embeddings often leads to suboptimal alignment, where the LLM may overlook specific instructions due to the dominance of long acoustic sequences. To address this, openLISTEN introduces a modality alignment module (see Fig. 2 (b)) that integrates fine-grained alignment with global semantic context.

To inject textual semantics into the acoustic timeline while ensuring effective gradient flow, we employ a multi-head cross-attention mechanism

with residual connections and layer normalization. Specifically, the speech embeddings  $A^P$  serve as queries ( $Q$ ), while the text embeddings  $T^I$  serve as both keys ( $K$ ) and values ( $V$ ). This process is defined as:

$$Attn = \text{MultiHead}(A^P, T^I, T^I), \quad (1)$$

$$A_{\text{cross}} = \text{LayerNorm}(Attn + A^P), \quad (2)$$

where  $A_{\text{cross}} \in \mathbb{R}^{N' \times d}$  represents the acoustic features enriched with local textual information.

While cross-attention captures local token-level dependencies, critical instruction semantics can be diluted over long audio sequences. To address this, we introduce a gating mechanism to incorporate global textual context. First, we compute a global text context vector  $v_{\text{ctx}}$  by mean-pooling across the text embedding sequence:

$$v_{\text{ctx}} = \frac{1}{M} \sum_{j=1}^M T_j^I \in \mathbb{R}^{1 \times d}. \quad (3)$$

To align with the temporal dimension of the acoustic features,  $v_{\text{ctx}}$  is broadcasted  $N'$  times along the temporal axis to form the context matrix  $T_{\text{ctx}} \in \mathbb{R}^{N' \times d}$ . Subsequently, a learnable linear layer followed by a Sigmoid activation computes the gate coefficient:

$$gate = \sigma(\text{Linear}([A_{\text{cross}}; T_{\text{ctx}}])), \quad (4)$$

where  $[\cdot]$  denotes concatenation along the feature dimension and  $\sigma$  is the Sigmoid function. The gate dynamically determines, for each time step and feature dimension, whether the model should rely more on the fine-grained cross-modal features or the global textual instruction. The final fused representation  $A_{\text{fused}}$  is computed as:

$$A_{\text{fused}} = gate \odot A_{\text{cross}} + (1 - gate) \odot T_{\text{ctx}}, \quad (5)$$

where  $\odot$  denotes element-wise multiplication (Hadamard product). This mechanism ensures that the model preserves the temporal structure of speech while remaining consistently guided by the global textual intent.

The modeling loss is defined using the cross-entropy loss. Specifically, the LLM autoregressively predicts the target text sequence of length  $L$  conditioned on the fused multimodal inputs. The loss function (Mao et al., 2023) is defined as:

$$\mathcal{L}_{\text{LLM}} = - \sum_{i=1}^L \log P(y_i^T | \mathcal{P}(S), y_{<i}^T), \quad (6)$$

$$\mathcal{P}(S) = \text{Concat}([T^I, A_{\text{fused}}]), \quad (7)$$

where  $y_{<i}^T$  denotes the previously generated target tokens,  $\mathcal{P}(S)$  denotes the fused multimodal input sequence provided to the LLM as conditioning context, and  $\text{Concat}(\cdot)$  concatenates embeddings along the sequence dimension to form the final input sequence of the LLM.

### 3.3 Open-Domain Cross-Modal Instruction Tuning

openLISTEN employs an *Open-Domain Cross-Modal Instruction Tuning (OCIT)* strategy designed to mitigate the *modality dominance* problem—where models tend to over-prioritize acoustic features while neglecting textual prompts. To address this, we propose the “One-Speech-Multi-Instruction” data construction strategy to curate a high-quality, diverse, and conflict-aware dataset.

**Diverse Synthesis and Quality Control.** We begin by collecting a large-scale textual Question-Answering (QA) corpus (Taori et al., 2023; Xu et al., 2024; Wang et al., 2023). To convert this text into speech, we leverage an *ensemble of SOTA zero-shot TTS models* (Wang et al., 2025b; Du et al., 2024b; Ju et al., 2024). This multi-model approach prevents the model from overfitting to the acoustic artifacts of a single synthesizer. To maximize ecological validity, we utilize a reference library of *authentic human speech collected from real-world scenarios* as acoustic prompts. This ensures the generated dataset covers a broad spectrum of realistic speaker identities and timbres. Post-synthesis, we apply a strict filtering protocol: samples with high Word Error Rates are discarded to guarantee semantic intelligibility.

**One-Speech-Multi-Instruction Strategy.** To dismantle the spurious correlation between acoustic distributions and task types, we construct multiple instruction-response pairs for a single speech segment, covering distinct tasks: (1) *Speech QA*; (2) *Speech Transcription*; and (3) *Summarization*. To further prevent overfitting to specific prompt templates, we utilize an LLM for instruction augmentation, rewriting textual instructions into synonymous sentences with varying lengths and syntactic structures. This comprehensive data pipeline ensures that the Gated Cross-Attention mechanism learns to dynamically modulate acoustic features based

on specific textual intents, rather than relying on memorized acoustic patterns.

## 4 Experiments and Results

This section introduces the evaluation tasks, experimental results and analysis, and illustrative use cases. Details of the training configuration are provided in Appendix B.

### 4.1 Evaluation Tasks

We structure the evaluation into two distinct streams to comprehensively validate openLISTEN’s performance. All experiments primarily use English test sets, except for the Multilingual subset in the Pro track, focusing on benchmarking holistic proficiency via standardized frameworks and verifying the GCA mechanism’s robustness through targeted adversarial scenarios.

To rigorously control confounding factors and isolate the contribution of the fusion design, in addition to comparisons with mainstream SOTA LSLMs, we construct a suite of controlled baselines and conduct experiments on two evaluation benchmarks. Specifically, we compare five variants: (i) the full model with GCA enabled, (ii) a GCA with the gating network removed, (iii) a model that removes GCA entirely, (iv) a FiLM-based fusion baseline (Perez et al., 2018), and (v) a Q-Former based fusion baseline (Li et al., 2023a). Across all variants, we keep the backbone architecture, training data, and optimization protocol fixed, and vary only the fusion module to ensure a fair comparison. For variants (iv) and (v), we follow the official open-source implementations of FiLM and Q-Former, respectively, and use the default setting of 32 query tokens for Q-Former. Implementation details for each variant are provided in our open-source codebase.

#### 4.1.1 Fixed Instruction Compliance

To benchmark openLISTEN against leading SOTA LSLMs across a holistic spectrum of capabilities, we employ the URO-Bench framework (Yan et al., 2025). Table 5 (see Appendix C) summarizes statistics for the Basic and Pro tracks.

#### 4.1.2 Freestyle Instruction Compliance

Existing benchmarks overlook LSLMs behavior under multimodal conflict. To validate openLISTEN’s suppression of *modality bias* via GCA, we introduce a two-part evaluation: *Adversarial Ro-*

*bustness Evaluation* and *Complex Constraint Evaluation*.

**Adversarial Robustness Evaluation.** To quantify the model’s anti-interference capability when the *intrinsic semantic intent of the speech modality* (e.g., spoken questions or instructions) conflicts with the *explicit constraints of the text modality*, we constructed an Adversarial Cross-Modal Evaluation Set containing approximately 330 samples. Within this set, we synthesized semantic adversarial conditions across three representative scenarios:

*Style Conflict:* The speech input carries strong acoustic emotional features, while the text instruction, enhanced by an LLM, explicitly demands a response in a conflicting style. This tests the model’s ability to shield against acoustic interference.

*Task Conflict:* The speech input consists of an emotional question that typically triggers a Question-Answering mode, the text instruction mandates verbatim transcription, compelling the model to suppress its generative tendencies.

*Language Conflict:* The speech requests a response in a foreign language, while the text instruction restricts the output to a fixed language. This scenario tests the model’s adherence to text-defined boundaries amidst conflicting language modalities.

We define three metrics to quantify conflict-scenario performance. *The Text Instruction Followed Rate (TIFR)* measures the model’s capacity to suppress acoustic interference and strictly adhere to text instructions. *The Text Correct Response Rate (TCR)* evaluates the accuracy of responses conditional on successful text adherence. Finally, *the Non-Compliance Rate (NFR)* aggregates all failure modes, including following speech intent and hallucinations. All metrics are derived from statistical analysis of GPT-5 evaluations.

**Complex Constraint Evaluation.** To further evaluate the model’s capability to execute complex instructions and adhere to strict formatting requirements in non-adversarial scenarios, we extended the evaluation to three scenarios:

*Direct Speech Translation:* We selected 200 English speech samples from the LibriTTS test set and instructed the model to generate German translations, evaluated using the sacreBLEU metric to assess cross-lingual capabilities.

*Contextual Story Continuation:* Using 118 complete audio story clips, we evaluated the model’s ability to generate contextually consistent continuations. GPT-5 scored the outputs on a scale of 1

Model	Scale	Understanding-Basic			Understanding-Pro				Overall↑
		Repeat↑	Summary↑	Gaokao Eval↑	Under Emotion↑	Code Switching↑	Clotho Eval↑	Safety↑	
openLISTEN	3B	49.14	74.35	21.12	42.23	36.57	12.97	40.56	39.56
LLaMA-Omni	8B	45.62	<b>80.68</b>	16.06	36.35	25.52	<b>22.52</b>	43.89	38.66
Freeze-Omni	7B	<b>70.89</b>	78.87	<b>26.29</b>	<b>48.27</b>	<b>37.90</b>	1.51	58.06	<b>45.97</b>
Mini-Omni	0.5B	5.07	32.20	30	29.06	20.38	0.00	<b>58.89</b>	20.80
SLAM-Omni	0.5B	12.26	66.21	1.32	45.84	21.14	10.94	48.33	29.43
<b>Ablation Study</b>									
w/o GCA	3B	10.73	66.05	18.15	29.30	32.09	16.98	20.70	27.71
w/o Gate	3B	0.00	20.00	0.00	20.00	20.00	0.00	39.67	14.24
w/ FiLM	3B	0.00	20.28	0.00	22.92	20.00	0.00	25.56	12.68
w/ Q-Former	3B	0.00	20.16	0.00	20.15	20.00	0.00	22.77	11.87

Table 1: Evaluation scores for basic and pro track English tests in speech *understanding* tasks.

to 5 across three dimensions: narrative coherence, linguistic fluency, and creative content quality.

*Forced instruction following:* To test strict adherence to predefined output templates, we curated a dataset of 66 speech samples covering five emotions. The model was required to identify the emotion and provide a rational explanation following a rigid format. GPT-5 evaluated the outputs on a scale of 1 to 5 based on instruction compliance, explanation coherence, and formatting accuracy.

## 4.2 Results and Analysis of Fixed Instruction Compliance

We compare openLISTEN with representative SOTA LSLMs on URO-Bench. As shown in Table 1, openLISTEN achieves competitive performance in the *Understanding* dimension with an overall score of 39.56, ranking second among all baselines. Despite utilizing a smaller 3B backbone, it outperforms the 8B LLaMA-Omni, particularly in instruction-sensitive sub-tasks such as *Repeat* and *Code Switching*. This indicates that the proposed alignment mechanism effectively preserves the instruction-following capabilities of the LLM even under acoustic conditioning, whereas larger models may struggle to balance acoustic constraints with textual instructions. In the *Reasoning* dimension, openLISTEN demonstrates a significant advantage, attaining the highest overall score of 43.24 as detailed in Table 2. The model significantly outperforming other comparative baselines, and maintains a leading position in complex reasoning tasks. This performance gap suggests that our fusion strategy better mitigates the degradation of logical reasoning often observed in projection-based multimodal models, allowing the model to perform structured multi-step inference

reliably even when processing speech inputs. In the *Oral Conversation* dimension, reported in Table 3, openLISTEN secures the top rank with an overall score of 60.27. While maintaining competitive performance on general chat benchmarks, the model demonstrates distinct advantages in style controllability and multilingual robustness. These results highlight that openLISTEN delivers top-tier conversational quality and maintains stable behavior under open-ended and style-conditioned interactions.

### 4.2.1 Ablation Study

To validate the effectiveness of the proposed architecture, we conducted controlled ablation studies analyzing the contributions of the GCA module and gating mechanisms. The results are presented in the bottom sections of Table 1, Table 2, and Table 3. The removal of the GCA module (*w/o GCA*) results in a consistent performance drop across all dimensions. For instance, the overall reasoning score decreases from 43.24 to 36.16. This decline confirms that naïve fusion is insufficient for high-fidelity speech modeling and that the granular cross-attention provided by GCA is necessary for capturing fine-grained acoustic details. However, the model retains basic interactive functionality, suggesting that the backbone architecture remains fundamentally sound even without this specific module. In sharp contrast, the variants without dynamic gating (*w/o Gate*) or those using static adapters like *w/ FiLM* and *w/ Q-Former* suffer catastrophic failure. Across open-ended tasks in the *Understanding* and *Conversation* dimensions, the scores for these baselines cluster around  $\sim 20$ . This specific value reflects the URO-Bench normalization lower bound for rubric-based grading, indicating

Model	Scale	Reasoning-Basic				Reasoning-Pro			Overall↑
		Storal Eval↑	Truthful Eval↑	Gsm8k Eval↑	MLC↑	MLCpro↑	MtBench Eval↑	Speaker Aware↑	
openLISTEN	3B	54.01	<b>49.83</b>	<b>4.64</b>	<b>52.03</b>	<b>55.67</b>	<b>50.39</b>	36.12	<b>43.24</b>
LLaMA-Omni	8B	50.65	45.13	3.89	44.44	47.62	–	–	27.39
Freeze-Omni	7B	<b>57.74</b>	46.95	2.81	42.56	5.49	–	–	22.22
Mini-Omni	0.5B	23.02	25.06	0.00	2.82	0.00	–	–	7.27
SLAM-Omni	0.5B	36.95	34.65	0.00	21.85	10.26	32.88	8.42	20.72
<b>Ablation Study</b>									
w/o GCA	3B	34.73	44.04	0.34	40.49	48.72	46.49	<b>38.30</b>	36.16
w/o Gate	3B	20.10	20.05	0.00	0.00	0.00	20.00	20.00	11.45
w/ FiLM	3B	20.03	20.23	0.00	0.00	0.00	20.18	20.24	11.53
w/ Q-Former	3B	22.13	20.05	0.00	0.00	0.00	20.00	20.00	11.74

Table 2: Evaluation scores for basic and pro track English tests in speech *reasoning* tasks. “–” entries arise because the official implementation is configured for multi-turn dialogue, and the corresponding evaluation outputs are unavailable, resulting in missing values.

Model	Scale	Oral Conversation-Basic			Oral Conversation-Pro		Overall↑
		AlpacaEval↑	CommonEval↑	WildchatEval↑	GenStyle↑	Multilingual↑	
openLISTEN	3B	<b>67.74</b>	58.37	69.28	<b>83.94</b>	<b>22.04</b>	<b>60.27</b>
LLaMA-Omni	8B	64.36	<b>58.40</b>	<b>72.19</b>	83.03	21.10	59.82
Freeze-Omni	7B	52.23	48.70	55.80	66.36	20.42	48.70
Mini-Omni	0.5B	30.99	29.80	31.42	40.30	20.83	30.67
SLAM-Omni	0.5B	48.98	41.03	52.61	64.24	20.54	45.48
<b>Ablation Study</b>							
w/o GCA	3B	67.54	58.16	62.86	72.27	21.18	56.40
w/o Gate	3B	20.10	20.13	20.06	20.00	20.00	20.06
w/ FiLM	3B	20.00	20.97	20.25	20.15	20.00	20.27
w/ Q-Former	3B	20.10	20.23	20.36	20.15	20.00	20.17

Table 3: Evaluation scores for basic and pro track tests in *oral conversation* tasks.

470 that these models revert to generic, template-like re- 490  
471 fusals or irrelevant responses rather than generating 491  
472 valid answers. Similarly, in closed-form Reasoning 492  
473 tasks, scores drop to near zero. These results col- 493  
474 lectively validate that static or un-gated injection of 494  
475 acoustic features tends to overwhelm the semantic 495  
476 space of the language model. The dynamic regula- 496  
477 tion provided by our gating mechanism is therefore 497  
478 critical for balancing acoustic evidence with text- 498  
479 tual constraints, ensuring that the LLM can process 499  
480 speech without losing its reasoning and generation 500  
481 capabilities. 501

### 482 4.3 Results and Analysis of Freestyle 502 483 Instruction Compliance 503

484 URO-Bench primarily evaluates multimodal input 505  
485 capabilities under fixed instruction settings, with 506  
486 an emphasis on the speech modality, but it does 507  
487 not cover freestyle multimodal input scenarios. To 508  
488 fill this gap, we introduce the freestyle instruction 509  
489 compliance tasks that jointly use text prompts and 510

speech input. Since many existing LSLM archi-  
tectures do not support freestyle instruction input,  
we conduct this set of experiments only on open-  
LISTEN and its controlled variants to ensure a fair  
comparison.

**Adversarial Robustness Evaluation.** Table 4  
shows that openLISTEN substantially strengthens  
*correct* conditional execution. It improves TCR  
from 26.35% to 53.45%, while also raising TIFR  
from 48.77% to 57.41% and reducing NFR from  
51.23% to 42.59%. Importantly, the gap between  
TIFR and TCR is highly diagnostic: w/o GCA vari-  
ant exhibits a large discrepancy, indicating that it  
often produces outputs that appear instruction com-  
pliance yet fail to resolve the conflict correctly. In  
contrast, openLISTEN keeps TCR close to TIFR,  
suggesting that the followed instructions are not  
merely superficial but are executed in a task-correct  
manner. Removing the gating network, as well as  
replacing GCA with FiLM or Q-Former, yields  
the same failure mode: TCR collapses to 0 while

Models	Adversarial Robustness Evaluation			Complex Constraint Evaluation		
	TCR $\uparrow$	TIFR $\uparrow$	NFR $\downarrow$	BLEU $\uparrow$	Story $\uparrow$	Format $\uparrow$
<b>openLISTEN</b>	<b>53.45 %</b>	<b>57.41 %</b>	<b>42.59 %</b>	<b>10.97</b>	<b>3.32</b>	<b>4.16</b>
w/o GCA	26.35 %	48.77 %	51.23 %	2.24	3.27	2.82
w/o Gate	0	35.19 %	64.81 %	0.19	1.49	1.12
w/ FiLM	0	28.70 %	71.30 %	0.70	2.30	1.37
w/ Q-Former	0	17.59 %	82.41 %	0.13	2.24	1.00

Table 4: Performance on freestyle instruction Compliance across fusion variants.

NFR increases substantially. This pattern indicates that, without a learned mechanism to dynamically regulate fusion, the model cannot reliably maintain textual constraint dominance under adversarial acoustic interference.

**Complex Constraints Evaluation.** The complex constraints results further clarify what GCA improves. The w/o GCA variant preserves story plausibility, with a Story score of 3.27, but fails on hard constraints, with BLEU 2.24 and Format 2.82. openLISTEN sharply increases BLEU to 10.97 and Format to 4.16, while maintaining story coherence at 3.32. This pattern indicates that GCA primarily enhances controllable execution under coupled constraints, rather than simply improving fluency. Removing the gate yields near-total degradation across all constraint-sensitive metrics, with BLEU 0.19, Story 1.49, and Format 1.12, consistent with unstable fusion that disrupts both content fidelity and structural adherence. All alternative fusion designs perform poorly on the Complex Constraints Evaluation and frequently fail to complete the task as specified, indicating that they do not provide sufficient controllability under coupled constraints.

In summary, robust freestyle instruction compliance hinges on two factors: token-level alignment that binds speech evidence to textual constraints, and a learned control signal that regulates acoustic injection into the language stream. Our computational efficiency analysis shows that GCA achieves substantial gains with minimal overhead (182MB parameters, about 6% of the 3B backbone, and 6.9G FLOPs); see Appendix E for details. Fig. 3 visualizes the input representations to openLISTEN. The results demonstrate that GCA effectively aligns the fused features with the textual space while preserving essential speech dynamics. Refer to Appendix D for further analysis.

#### 4.4 Case Studies

The Appendix F shows that openLISTEN continues the story in a speech-grounded and instruction-consistent manner, whereas w/o GCA exhibits no-

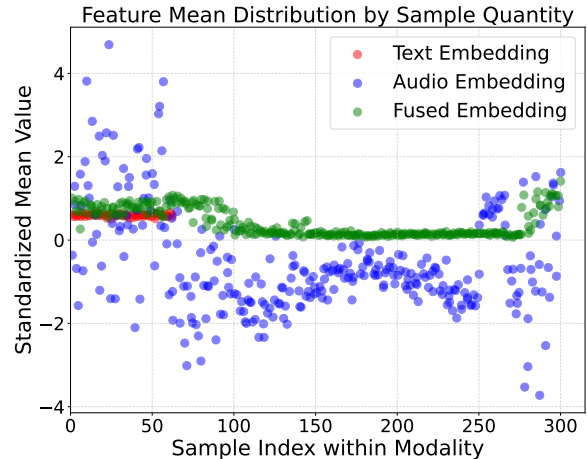


Figure 3: The standardized mean distribution of the speech/text embeddings and the fused embedding after the GCA module.

ticeable context drift. More critically, w/o Gate, FiLM, and Q-Former fail to follow the continuation instruction and generate off-task outputs, highlighting the necessity of GCA with gating for robust cross-modal compliance.

## 5 Conclusion

We present openLISTEN, a resource-efficient LSLM that enables speech–language modeling on consumer-grade hardware. By combining GCA with Open-Domain Cross-Modal Instruction Tuning, openLISTEN improves cross-modal alignment and reduces instruction non-compliance in data-constrained training regimes. Across URO-Bench and targeted freestyle instruction compliance benchmarks, controlled studies under identical training and optimization settings show consistent gains from GCA over alternative fusion designs, yielding more robust behavior under modality conflicts while preserving acoustic expressiveness. In future work, we will study how this architecture scales with larger backbones (e.g., 7B, 70B) and whether it unlocks emergent cross-modal reasoning capabilities.

576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
  
592  
593  
594  
595  
596  
597  
598  
  
599  
600  
601  
  
602  
603  
604  
  
605  
606  
607  
608  
  
609  
610  
611  
612  
613  
614  
  
615  
616  
617  
  
618  
619  
620  
621  
622  
623  
  
624  
625  
626  
627

## Limitations

Despite promising results, this study is subject to limitations primarily driven by computational constraints, restricting current validation to a 3B parameter backbone and leaving the scalability of GCA on larger architectures to be explored. Furthermore, the reliance on TTS-synthesized data may introduce prosodic biases and reduce ecological validity relative to natural human speech. Critically, the enhanced cross-modal instruction adherence presents potential dual-use risks, where adversarial speech prompts could be exploited to circumvent textual safety guardrails. Investigating how the GCA mechanism interacts with these risks and reinforcing safety protocols remains a priority for future work.

## References

Ge Bai, Jie Liu, Xingyuan Bu, and et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7421–7454, Bangkok, Thailand.

Wenxi Chen, Ziyang Ma, Ruiqi Yan, and et al. 2025. Slam-omni: Timbre-controllable voice interaction system with single-stage training. pages 2262–2282.

Yunfei Chu, Jin Xu, Qian Yang, and et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv: abs/2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, and et al. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv: abs/2311.07919*.

Yu-An Chung, Yu Zhang, Wei Han, and et al. 2021. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250, Cartagena, Colombia.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, and et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv: abs/2110.14168*.

Hui-Peng Du, Yang Ai, Rui-Chen Zheng, and et al. 2024a. Apcodec+: A spectrum-coding-based high-fidelity and high-compression-rate neural audio codec with staged training paradigm. In *Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 676–680, Beijing, China.

Zhihao Du, Yuxuan Wang, Qian Chen, and et al. 2024b. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv: abs/2412.10117*.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *Proc. Conference on Language Modeling (COLM)*, pages 1–11, Pennsylvania, USA. 628  
629  
630  
631  
632

Qingkai Fang, Shoutao Guo, Yan Zhou, and et al. 2025. Llama-omni: Seamless speech interaction with large language models. 633  
634  
635

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155, Singapore. 636  
637  
638  
639  
640

Jian Guan, Ziqi Liu, and Minlie Huang. 2022. A corpus for understanding and generating moral stories. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5069–5087, Seattle, USA. 641  
642  
643  
644  
645

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, and et al. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29(8):3451–3460. 646  
647  
648  
649  
650

Edward J. Hu, Yelong Shen, Phillip Wallis, and et al. 2022. Lora: Low-rank adaptation of large language models. In *Proc. International Conference on Learning Representations (ICLR)*, pages 1–4, Virtual Event. 651  
652  
653  
654  
655

Rongjie Huang, Mingze Li, Dongchao Yang, and et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proc. Annual AAAI Conference on Artificial Intelligence*, pages 23802–23804, Vancouver, Canada. 656  
657  
658  
659  
660

Shengpeng Ji, Ziyue Jiang, Wen Wang, and et al. 2025. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. pages 1–8. 661  
662  
663

Zeqian Ju, Yuancheng Wang, Kai Shen, and et al. 2024. Naturalspeech 3: zero-shot speech synthesis with factorized codec and diffusion models. In *Proc. International Conference on Machine Learning (ICML)*, pages 11–19, Vienna, Austria. 664  
665  
666  
667  
668

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. International Conference on Machine Learning ICML*, volume 202, pages 19730–19742, Honolulu, Hawaii, USA. 669  
670  
671  
672  
673  
674

Junnan Li, Silvio Savarese, and Steven C. H. Hoi. 2023b. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. International Conference on Machine Learning (ICML)*, pages 19730–19742, Honolulu, Hawaii, USA. 675  
676  
677  
678  
679  
680

681	Stephanie Lin, Jacob Hilton, and Owain Evans. 2022.	Yuancheng Wang, Haoyue Zhan, Liwei Liu, and et al.	734
682	TruthfulQA: Measuring how models mimic human	2025b. MaskGCT: Zero-shot text-to-speech with	735
683	falsehoods. In <i>Proc. Annual Meeting of the Asso-</i>	masked generative codec transformer. In <i>Proc. In-</i>	736
684	<i>ciation for Computational Linguistics (ACL)</i> , pages	<i>ternational Conference on Learning Representations</i>	737
685	3214–3252, Seattle, USA.	( <i>ICLR</i> ), pages 1–24, Singapore.	738
686	Aixin Liu, Bei Feng, Bing Xue, and et al. 2024.	Zhifei Xie and Changqiao Wu. 2024a. Mini-omni: Lan-	739
687	Deepseek-v3 technical report. <i>arXiv preprint arXiv:</i> ,	guage models can hear, talk while thinking in stream-	740
688	abs/2412.19437.	ing. <i>arXiv preprint arXiv</i> , abs/2408.16725.	741
689	Xubo Liu, Zhongkai Zhu, Haohe Liu, and et al. 2025.	Zhifei Xie and Changqiao Wu. 2024b. Mini-omni2:	742
690	Wavjourney: Compositional audio creation with large	Towards open-source gpt-4o with vision, speech	743
691	language models. <i>IEEE Transactions on Audio,</i>	and duplex capabilities. <i>arXiv preprint arXiv</i> ,	744
692	<i>Speech and Language Processing</i> , 33:2830–2844.	abs/2410.11190.	745
693	Anqi Mao, Mehryar Mohri, and Yutao Zhong. 2023.	Can Xu, Qingfeng Sun, Kai Zheng, and et al. 2024.	746
694	Cross-entropy loss functions: Theoretical analysis	Wizardlm: Empowering large pre-trained language	747
695	and applications. 202:23803–23828.	models to follow complex instructions. In <i>Proc. In-</i>	748
696	OpenAI. 2023. GPT-4 technical report. <i>arXiv preprint</i>	<i>ternational Conference on Learning Representations</i>	749
697	<i>arXiv:</i> , abs/2303.08774.	( <i>ICLR</i> ), pages 1–9, Vienna, Austria.	750
698	Ethan Perez, Florian Strub, Harm de Vries, Vincent	Hongfei Xue, Yufeng Tang, Hexin Liu, and et al. 2025.	751
699	Dumoulin, and Aaron C. Courville. 2018. Film: Vi-	Enhancing non-core language instruction-following	752
700	visual reasoning with a general conditioning layer. In	in speech llms via semi-implicit cross-lingual cot	753
701	<i>Proc. AAAI Conference on Artificial Intelligence,</i>	reasoning. In <i>Proc. ACM International Conference</i>	754
702	( <i>AAAI</i> ), pages 3942–3951, New Orleans, Louisiana,	<i>on Multimedia (MM)</i> , pages 10984–10993, Dublin,	755
703	USA.	Ireland.	756
704	Alec Radford, Jong Wook Kim, Tao Xu, and et al. 2023.	Ruiqi Yan, Xiquan Li, Wenxi Chen, and et al. 2025.	757
705	Robust speech recognition via large-scale weak su-	URO-bench: Towards comprehensive evaluation for	758
706	perception. In <i>Proc. International Conference on</i>	end-to-end spoken dialogue models. pages 17211–	759
707	<i>Machine Learning ICML</i> , pages 28492–28518, Hon-	17242.	760
708	olulu, Hawaii, USA.	Zhengdong Yang, Shuichiro Shimizu, Yahan Yu, and	761
709	Jaroslav Resler, Pavel Krč, and et al. 2017. Palm-usm	Chenhui Chu. 2025. When large language models	762
710	v1. 0: A new urban surface model integrated into	meet speech: A survey on integration approaches.	763
711	the palm large-eddy simulation model. <i>Geoscientific</i>	<i>arXiv preprint arXiv:</i> , abs/2502.19548.	764
712	<i>Model Development</i> , 10(10):3635–3659.	Neil Zeghidour, Alejandro Luebs, Ahmed Omran,	765
713	Paul K. Rubenstein, Chulayuth Asawaroengchai,	Jan Skoglund, and Marco Tagliasacchi. 2022.	766
714	Duc Dung Nguyen, and et al. 2023. Audiopalm:	Soundstream: An end-to-end neural audio codec.	767
715	A large language model that can speak and listen.	<i>IEEE/ACM Transactions on Audio, Speech, and Lan-</i>	768
716	<i>arXiv preprint arXiv:</i> , abs/2306.12925.	<i>guage Processing</i> , 30(10):495–507.	769
717	Changli Tang, Wenyi Yu, Guangzhi Sun, and et al. 2024.	Aohan Zeng, Zhengxiao Du, Mingdao Liu, and et al.	770
718	SALMONN: towards generic hearing abilities for	2024. Glm-4-voice: Towards intelligent and human-	771
719	large language models. In <i>Proc. International Con-</i>	like end-to-end spoken chatbot. <i>arXiv preprint</i>	772
720	<i>ference on Learning Representations (ICLR)</i> , pages	<i>arXiv:</i> , abs/2412.02612.	773
721	1–23, Vienna, Austria.	Dong Zhang, Shimin Li, Xin Zhang, and et al. 2023.	774
722	Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, and et al.	Speechgpt: Empowering large language models with	775
723	2023. Stanford alpaca: An instruction-following	intrinsic cross-modal conversational abilities. In	776
724	llama model.	<i>Proc. Findings of the Association for Computational</i>	777
725	Xiong Wang, Yangze Li, Chaoyou Fu, and et al. 2025a.	<i>Linguistics (EMNLP)</i> , pages 15757–15773, Singa-	778
726	Freeze-omni: A smart and low latency speech-to-	pore.	779
727	speech dialogue model with frozen LLM.		
728	Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, and		
729	et al. 2023. Self-instruct: Aligning language mod-		
730	els with self-generated instructions. In <i>Proc. An-</i>		
731	<i>nual Meeting of the Association for Computational</i>		
732	<i>Linguistics (ACL)</i> , pages 13484–13508, Toronto,		
733	Canada.		

## A Prompt Specifications for Automated Evaluation

This appendix summarizes the GPT-5 prompt setups for the freestyle instruction compliance benchmark, covering Complex Constraint and Adversarial Robustness evaluations. Figures 4 and 5 show the scoring prompts for Contextual Story Continuation and Forced Instruction Following, respectively, while Figure 6 presents the Adversarial Robustness prompt under text–audio conflicts.

**GPT-5 Scoring of Contextual Story Continuation**

You are an expert literary critic and editor evaluating creative writing. Your task is to assess how well a model continues a story based on a provided audio transcript context.

[ Story Context ]: {Context}  
[ Model Continuation ]: {Prediction}

Please evaluate the continuation on a scale of 1 to 5 based on the following detailed criteria:

- Narrative Coherence (1-5):**  
- Does the continuation maintain consistency with the characters' voices, the setting, and the plot established in the context?  
**5 points:** Seamless transition; perfectly mimics the tone and logic of the original context.  
**1 points:** Contradicts the context, changes character personalities abruptly, or introduces illogical plot points.
- Language Fluency (1-5):**  
- Is the writing smooth, grammatically correct, and natural?  
**5 points:** Native-level fluency, sophisticated vocabulary, and varied sentence structure.  
**1 points:** Broken grammar, repetitive phrasing, or unnatural wording.
- Creative Content Quality (1-5):**  
- Is the continuation engaging and original?  
**5 points:** Captivating storytelling that introduces interesting developments while avoiding clichés.  
**1 points:** Dull, generic, repetitive, or simply paraphrases the context without moving the story forward.

Return your evaluation strictly in the following format:  
[ Coherence ]: {1-5 points.}  
[ Fluency ]: {1-5 points.}  
[ Creativity ]: {1-5 points.}  
[ Reason ]: {Briefly state the rationale for your score.}

Figure 4: The GPT-5 scoring prompt configuration for the Contextual Story Continuation task.

## B Implementation Details

All experiments, including training and inference, are conducted on four NVIDIA GeForce RTX 4090D GPUs. openLISTEN integrates the Whisper-large-v3 (Radford et al., 2023) encoder as the speech encoder and utilizes Llama-3.2-3B-Instruct as the LLM backbone. The encoder representations are downsampled by a factor of 5. We set the micro-batch size to 2 per GPU and employ gradient accumulation to achieve an effective global batch size of 32. LoRA layers are configured with a rank

## GPT-5 Scoring of Forced Instruction Following

You are a strict evaluator focusing on mandatory instruction adherence and rigid formatting compliance. The model was required to identify an emotion and provide a rationale using a specific, rigid template: "The speaker's emotion is {{emotion}} and The Reasons is {{Reasons}}". The template contains deliberate grammatical quirks (e.g., "The Reasons is").

The model MUST reproduce these exactly.

[ Original Instruction ]: { Instruction }  
[ Model Response ]: { Prediction }

Please evaluate the response on a scale of 1 to 5 based on the following detailed criteria:

- Instruction Compliance (1-5):**  
**5 points:** The model explicitly identified a specific emotion and provided a rationale.  
**1 points:** The model failed to identify an emotion or failed to provide a reason.
- Explanation Coherence (1-5):**  
**5 points:** The provided reason is logically sound, directly referencing the speech context, and strongly supports the identified emotion.  
**1 points:** The reason is nonsensical, hallucinates facts not implied by the emotion, or contradicts the identified emotion.
- Formatting Accuracy (1-5):**  
**5 points:** Perfect character-level adherence to the template "The speaker's emotion is ... and The Reasons is ...". No extra words, introductions, or markdown.  
**3 points:** Follows the general structure but corrects the grammar (e.g., changes "The Reasons is" to "The reason is") or adds minor conversational filler.  
**1 points:** Completely ignores the required format.

Return your evaluation strictly in the following format:  
[ Compliance ]: {1-5 points.}  
[ Coherence ]: {1-5 points.}  
[ Formatting ]: {1-5 points.}  
[ Reason ]: {Briefly state the rationale for your score.}

Figure 5: The GPT-5 scoring prompt configuration for the Forced Instruction Following task.

of  $r = 8$  and an alpha of  $\alpha = 2$ . The training process utilizes a cosine annealing scheduler with a peak learning rate of  $1 \times 10^{-4}$ . To mitigate overfitting given the limited dataset size, the training duration is restricted to 3 epochs.

## C Details on Fixed Instruction Compliance

This evaluation rigorously assesses multifaceted competencies through both *Basic* (Cobbe et al., 2021; Lin et al., 2022; Guan et al., 2022) and *Pro* (Bai et al., 2024; Dubois et al., 2024) tracks, spanning three fundamental dimensions: *Understanding*, *Reasoning*, and *Oral Conversation*. Table 5 reports summary statistics for the Basic and Pro tracks. Built upon URO-Bench (Yan et al., 2025), these splits support systematic, *quantitative evaluation of core LSLM capabilities, with all metric scores computed under the official protocol*.

## D Cross-modal Alignment Visualization

Figure 7 visualizes the standardized mean distributions of text, audio, and fused embeddings across

### GPT-5 Scoring of Adversarial Robustness Evaluation

You are an expert impartial evaluator specializing in cross-modal linguistic alignment. Your task is to resolve conflicts between spoken intent and written directives with high precision. You possess a keen ability to distinguish between modality bias and strict instruction adherence, ensuring that model responses are judged solely on their compliance with explicit textual boundaries.

```
[ Original Speech Content ]: {item.get('source_text', 'N/A')}
[ Speech Implied Instruction ]: {item['speech_instruction_content']}
[ Text Explicit Instruction ]: {item['text_instruction_content']}
[ Model Response ]: {item['model_response']}
```

Please evaluate the model's response based on two criteria:

**1. Instruction Following: Which instruction did the model follow?**

- Return **"Text"** if it followed the text instruction.
- Return **"Neither"** if it followed neither or it's followed the speech instruction.

**2. Text Instruction Correctness: IF the model followed the TEXT instruction, is the response correct/appropriate?**

- Return **"Correct"** if the response correctly answers/follows the text instruction.
- Return **"Incorrect"** if the response is wrong or hallucinates, even if it attempted the text instruction.

- Return **"Not Applicable"** if the model did not follow the text instruction (i.e., followed speech or neither).

Return your evaluation strictly in the following format:

```
[ Instruction Followed ]: {Text | "Neither"}
[ Text Instruction Correctness ]: {"Correct" | "Incorrect" | "Not Applicable"}
[ Reason ]: {Short explanation}
```

Figure 6: The GPT-5 scoring prompt configuration for the Adversarial Robustness Evaluation task.

Types	Datasets	Lan.	#Samples	#Avg Words	Avg Dur. (s)
<b>Basic Track</b>					
Understanding	Repeat	en	252	21.76	8.04
	Summary	en	118	58.93	20.38
	GaokaoEval	en	303	47.38	20.44
Reasoning	StoralEval	en	201	66.46	20.52
	TruthfulEval	en	470	10.87	3.40
	Gsm8kEval	en	582	25.73	8.84
	MLC	en	177	22.43	7.56
Oral Conversation	AlpacaEval	en	199	16.37	5.67
	CommonEval	en	200	8.16	4.83
	WildchatEval	en	349	14.68	4.75
<b>Pro Track</b>					
Understanding	UnderEmotion	en	137	13.55	6.01
	CodeSwitching	en	70	10.37	3.24
	Safety	en	24	13.25	3.95
	ClothoEval	en	265	16.55	25.43
Reasoning	MLCpro	en	91	11.40	3.88
	MtBenchEval	en	190	7.48	2.51
	SpeakerAware	en	55	9.37	3.33
Oral Conversation	GenStyle	en	44	15.36	5.92
	Multilingual	multi	1,108	17.11	6.38

Table 5: Statistical information of the test datasets for both *Basic* and *Pro* tracks across understanding, reasoning, and oral conversation tasks.

Table 6: Model size and inference efficiency across fusion variants.

Method	Params (MB)	FLOPs (G)	Latency (ms)
w/ GCA	182.0	6.9	74.0
w/o Gate	146.5	6.3	67.3
w/ FiLM	110.4	0.2	73.6
w/ Q-Former	101.1	1.4	63.7

sample indices. As shown in subfigure (a), openLISTEN maintains a dynamic distribution in its fused embeddings (green), effectively capturing the variance inherent in the audio inputs (blue). This suggests that the model successfully integrates speech modality features into the latent space, allowing for context-aware generation.

In contrast, w/o Gate, FiLM, and Q-Former show a flat-line pattern: the fused embeddings become nearly constant across inputs, indicating weak input-conditioned fusion. This aligns with the off-task outputs and context drift noted in the appendix, and supports GCA as necessary for stable, speech-grounded instruction following.

## E Computational Performance Analysis

Table 6 shows that enabling GCA introduces only a small efficiency overhead compared with the w/o Gate variant: parameters increase from 146.5MB to 182.0MB, FLOPs from 6.3G to 6.9G, and latency from 67.3 ms to 74.0 ms. Despite this limited increase in model size and inference cost, GCA yields substantially larger gains in the main experiments. By contrast, FiLM and Q-Former are lighter and, but their performance in the primary evaluations is very poor and they often fail on instruction-sensitive settings. Overall, the results suggest that the improvements stem from GCA's gated, token-level cross-modal coupling rather than from using a cheaper fusion module.

## F Use Cases

Table 7 presents a qualitative case study under an open-ended story continuation instruction, where successful completion requires (i) grounding in the speech content, (ii) preservation of narrative entities and setting, and (iii) adherence to the continuation format.

openLISTEN generates a fluent and coherent continuation that faithfully reflects the core se-

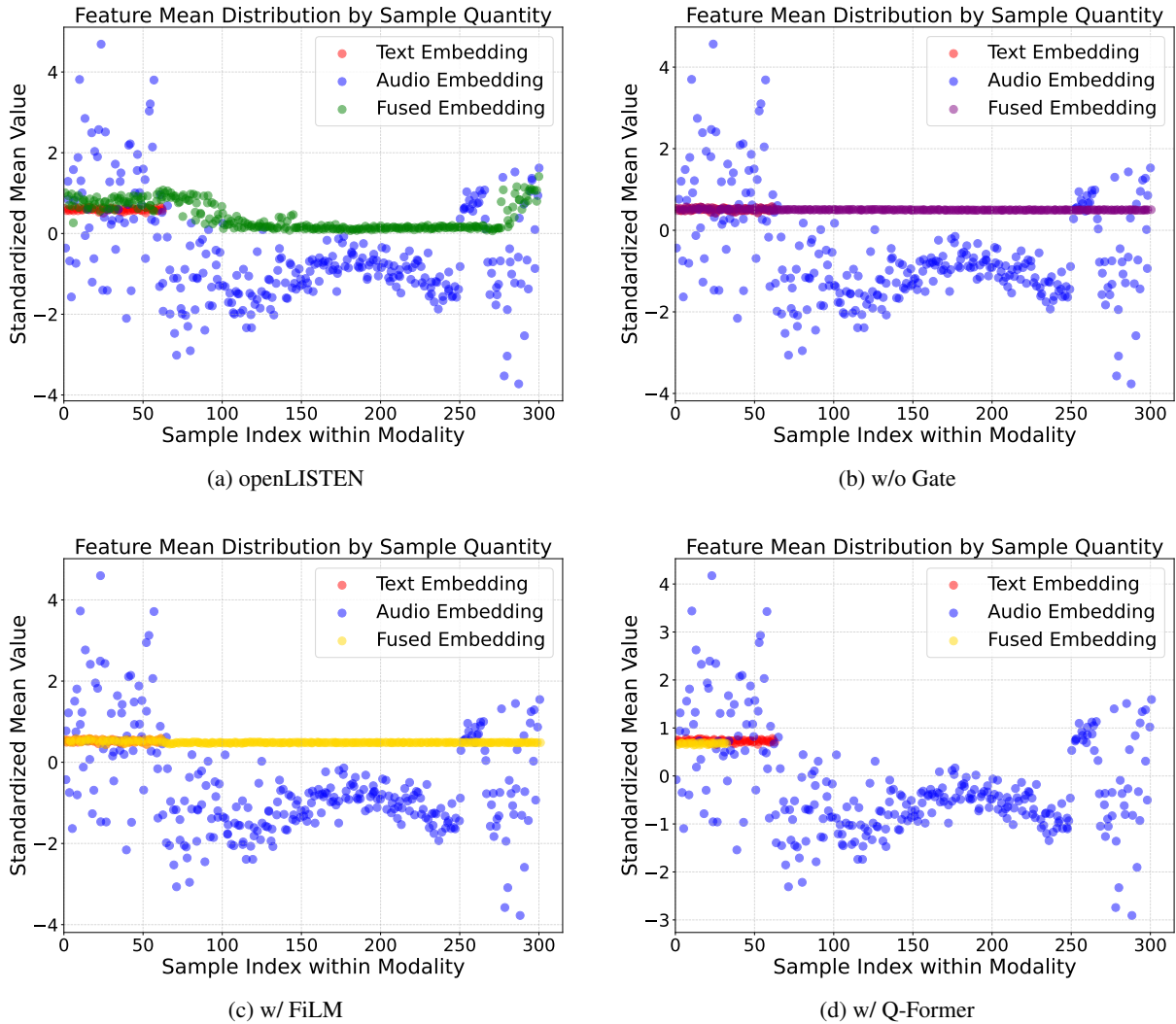


Figure 7: Comparison of different modules. The standardized mean distribution of the speech/text embeddings and the fused embedding after (a) openLISTEN, (b) w/o Gate, (c) w/ FiLM, and (d) w/ Q-Former modules.

860 mantic revelation in the speech (the duckling is  
 861 a swan), maintains narrative consistency, and ex-  
 862 tends the story without introducing contradictory  
 863 elements. This behavior indicates effective cross-  
 864 modal grounding and stable instruction execution  
 865 under open-ended conditions. The w/o GCA vari-  
 866 ant partially satisfies the continuation format but  
 867 introduces off-context content, such as new char-  
 868 acters and altered background details that are un-  
 869 supported by the input speech. This suggests that  
 870 removing GCA weakens fine-grained alignment  
 871 between speech evidence and text generation, lead-  
 872 ing to semantic drift even when surface-level task  
 873 structure is preserved. In contrast, w/o Gate, FiLM,  
 874 and Q-Former fail to comply with the instruction  
 875 entirely, producing outputs unrelated to both the  
 876 speech content and the requested task. These re-  
 877 sponses indicate a breakdown in cross-modal in-  
 878 struction conditioning, where acoustic or latent rep-

879 resentations no longer effectively constrain text  
 880 generation. Such failures are consistent with modal-  
 881 ity bias and unstable fusion behavior when dyn-  
 882 amic, token-level regulation is absent. Overall,  
 883 this use case highlights that robust freestyle in-  
 884 struction compliance in open-ended settings re-  
 885 quires not only cross-attention, but also a gated,  
 886 dynamically regulated fusion mechanism, as insti-  
 887 tiated by GCA in openLISTEN.

Table 7: Illustrative examples demonstrating freestyle instruction compliance in openLISTEN and its ablation variants. Correct responses are highlighted in green, while incorrect or irrelevant responses are highlighted in blue.

**Input Speech:** The ugly duckling was born different from his siblings, who all made fun of him. He felt lonely and rejected. As he grew older, the duckling discovered that he was not ugly at all, but instead, he was a beautiful swan. He was accepted by other swans and finally found happiness.

**Instruction:** Continue the story based on the content of the speech, ensuring that it fits the original story setting and characters. <speech>

Model	Output
openLISTEN	The duckling finally understood that he had never been ugly, he was a swan. Welcomed by the other swans, he began a new life with them. He glided across the pond under the warm sunshine, surrounded by his new friends, and for the first time felt truly at peace with himself.
w/o GCA	The duckling felt really sad and alone, you know? He was the only one who didn't have siblings. But then, he found a group of ducks who welcomed him with open wings. They showed him how to swim and dive, and he quickly became part of their family. The duckling was overjoyed to finally belong and was grateful for the love and support he received.
w/o Gate	how do you keep your eyes open while driving at night?
w/ FiLM	So, I'm wondering, can you tell me what the author of The Great Gatsby is trying to say about the American Dream in the story?
w/ Q-Former	How can I use the new features in the latest version of the software to boost my productivity and efficiency? What are the key benefits of these features, and how can I apply them to my work? What steps should I take to get the most out of the software? What are the potential risks or challenges I might face, and how can I address them? What are the long-term benefits of using this software, and how can I ensure its continued success?