DIFFA-2: A Practical Diffusion Large Language Model for General Audio Understanding

Jiaming Zhou[1,2*], **Xuxin Cheng**[2], **Shiwan Zhao**[1], **Yuhang Jia**[1,2*], **Cao Liu**[2], **Ke Zeng**[2], **Xunliang Cai**[2†], **Yong Qin**[1†]
[1]College of Computer Science, Nankai University, [2]Meituan LongCat Interaction Team
Correspondence: zhoujiaming@mail.nankai.edu.cn, qinyong@nankai.edu.cn

Abstract

Autoregressive (AR) large audio language models (LALMs) such as Qwen-2.5-Omni have achieved strong performance on audio understanding and interaction, but scaling them remains costly in data and computation, and strictly sequential decoding limits inference efficiency. Diffusion large language models (dLLMs) have recently been shown to make effective use of limited training data, and prior work on DIFFA indicates that replacing an AR backbone with a diffusion counterpart can substantially improve audio understanding under matched settings, albeit at a proof-of-concept scale without large-scale instruction tuning, preference alignment, or practical decoding schemes. We introduce **DIFFA-2**, a practical diffusion-based LALM for general audio understanding. DIFFA-2 upgrades the speech encoder, employs dual semantic and acoustic adapters, and is trained with a four-stage curriculum that combines semantic and acoustic alignment, large-scale supervised fine-tuning, and variance-reduced preference optimization, using only fully open-source corpora. Experiments on MMSU, MMAU, and MMAR show that DIFFA-2 consistently improves over DIFFA and is competitive to strong AR LALMs under practical training budgets, supporting diffusion-based modeling is a viable backbone for large-scale audio understanding. Our code is available at https://github.com/NKU-HLT/DIFFA.git.

1 Introduction

Diffusion large language models (dLLMs) (Nie et al., 2025; Zhu et al., 2025) have recently emerged as a promising alternative to conventional autoregressive (AR) decoders. Instead of generating tokens strictly left-to-right, dLLMs perform iterative denoising over partially masked sequences, enabling any-order token modeling and naturally supporting parallel token updates during decoding. Recent study (Ni et al., 2025) in the text domain further shows that dLLMs can act as strong *data learners*: when the amount of unique training data is constrained, they continue to improve and can even surpass AR models by leveraging super-dense compute and implicit Monte Carlo-style data augmentation. These properties are particularly appealing for audio understanding, where high-quality audio–text supervision across speech, sound, and music is far more expensive to collect than text-only data, and where the latency of strictly sequential AR decoding becomes a bottleneck for long-form and interactive applications.

Despite these advantages, state-of-the-art large audio language models (LALMs) are still predominantly AR-based: systems such as Qwen-3-Omni (Xu et al., 2025b), Qwen-2.5-Omni (Xu et al., 2025a), and Kimi-Audio (Ding et al., 2025) couple powerful speech encoders with AR LLMs and achieve strong results on a wide range of audio understanding and dialogue benchmarks. This raises a natural question: *can dLLMs be turned into competitive and practical audio backbones that match these AR LALMs under realistic data and latency budgets?* Initial evidence is encouraging. A recent work on diffusion-based large audio-language model (DIFFA) (Zhou et al., 2025) compares an 8B AR backbone with its diffusion counterpart under matched data, adapter design, and training recipe, and finds substantial gains on audio understanding benchmarks such as MMAU (Sakshi et al., 2025) and MMSU (Wang et al., 2025b) after simply replacing the AR backbone. This suggests that the generative paradigm itself can strongly influence audio performance and that dLLMs have significant potential as audio backbones. However, that study remains largely a proof of concept: the model is trained mainly on speech-centric supervision with a relatively small Whisper encoder (Radford

[*]This work was done during an internship at Meituan.
[†]Xunliang Cai and Yong Qin are corresponding authors. Correspondence to: qinyong@nankai.edu.cn

Findings of the Association for Computational Linguistics: ACL 2026, pages 4789–4805
July 2-7, 2026

et al., 2023), keeps the diffusion backbone frozen, and does not exploit large-scale instruction data, preference-based objectives, or practical inference acceleration. It therefore does not answer whether dLLMs can be scaled into *strong and practical* LALMs that reliably compete with state-of-the-art AR models.

In this paper, we present **DIFFA-2**, a substantially strengthened diffusion-based large audio language framework that addresses this question. DIFFA-2 aims to turn dLLMs from a proof-of-concept backbone into a competitive and practical audio model through comprehensive semantic–acoustic alignment and scalable training. Concretely, it adopts a four-stage training strategy that progressively aligns audio representations, incorporates large-scale supervised fine-tuning, and applies variance-reduced preference optimization (VRPO) (Zhu et al., 2025), and combines this with factor-based parallel decoding (Wu et al., 2025) at inference time to enable practical audio-based interaction without relying on AR decoding. Our contributions are as follows:

- We present **DIFFA-2**, a strengthened diffusion-based large audio language model with improved acoustic modeling and semantic–acoustic alignment for unified understanding of speech, sound, and music.
- We introduce a progressive four-stage training curriculum that combines semantic and acoustic alignment, large-scale supervised fine-tuning, and preference-based reinforcement learning using fully open-source data, together with factor-based parallel decoding for practical diffusion inference.
- With only 11,000 hours of automatic speech recognition (ASR) data and 3,767 hours of supervised fine-tuning data, DIFFA-2 updates only about 1.1% of parameters and achieves strong performance on audio understanding benchmarks including MMSU, MMAU, and MMAR (Ma et al., 2025), remaining highly competitive with strong AR-based LALMs on these benchmarks.
- We will open-source both the training and inference pipeline to facilitate future research on dLLM-based audio models.

2 Preliminaries

LLaDA (Nie et al., 2025) is a non-autoregressive language modeling framework based on a discrete random masking process. Instead of factorizing the sequence likelihood in a left-to-right manner, LLaDA introduces a stochastic corruption mechanism and trains a mask predictor to approximate the reverse denoising process. This design allows the model to leverage bidirectional context and enables parallel token prediction during inference.

Formally, given a clean target sequence $x_0 = (x_0^1, \ldots, x_0^L)$, LLaDA defines a forward masking process that independently replaces each token with a special mask symbol M with probability $t \in (0, 1]$, resulting in a corrupted sequence x_t. The mask predictor $p_\theta(x_0 \mid x_t)$, parameterized by a standard Transformer decoder, is trained to reconstruct the original tokens at masked positions. The pre-training objective is given by

$$\mathcal{L}(\theta) \triangleq -\mathbb{E}_{t,x_0,x_t} \left[\frac{1}{t} \sum_{i=1}^{L} \mathbb{I}[x_t^i = \mathrm{M}] \log p_\theta(x_0^i \mid x_t) \right], \tag{1}$$

where L denotes the sequence length. This objective yields a tractable upper bound on the negative log-likelihood (Shi et al., 2024; Ou et al., 2025), while avoiding autoregressive factorization.

Supervised fine-tuning (SFT) under the LLaDA framework follows the same corruption–reconstruction principle. Given a prompt–response pair (p_0, r_0), only the response sequence is subject to random masking, producing a corrupted response r_t, while the prompt remains fully observed. The SFT objective is defined as

$$-\mathbb{E}_{t,p_0,r_0,r_t} \left[\frac{1}{t} \sum_{i=1}^{L'} \mathbb{I}[r_t^i = \mathrm{M}] \log p_\theta(r_0^i \mid p_0, r_t) \right], \tag{2}$$

where L' denotes the response length.

At inference time, LLaDA performs generation through an iterative decoding procedure. Starting from a fully masked response sequence, the model predicts token values at masked positions and selectively re-applies masks to low-confidence tokens. By repeating this denoising process for a fixed number of steps, the model gradually refines its predictions and produces the final output sequence.

3 Methods

3.1 Model Architecture

DIFFA-2 follows the overall framework of DIFFA, but substantially strengthens acoustic representation and cross-modal alignment. The model con-

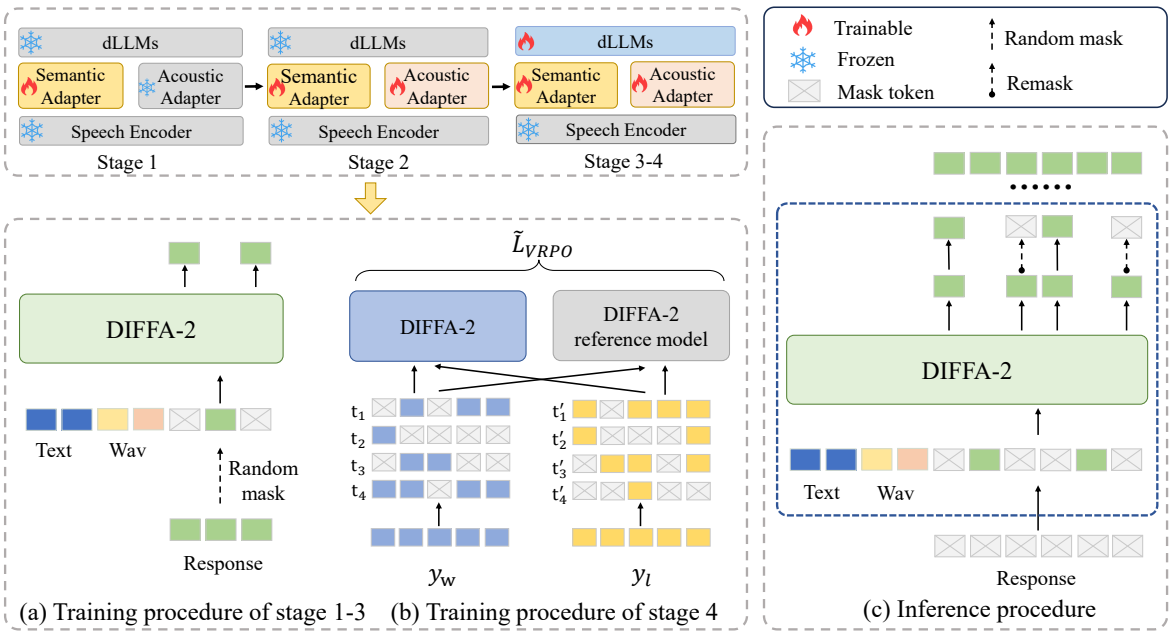

Figure 1: Overview of DIFFA-2, including the dual-adapter architecture, multi-stage training pipeline (Stages 1–4), and iterative diffusion-based inference for general audio understanding.

sists of three main components: a frozen Whisper-Large-V3 encoder, a dual-adapter audio interface, and a dLLM backbone. The *semantic adapter* consists of a two-layer convolution subsampling module followed by a two-layer linear projection, reducing the temporal resolution from 50 Hz to 12.5 Hz and aligning temporally aggregated audio features with textual semantics. The *acoustic adapter* is implemented as a two-layer Q-former (Li et al., 2023) with 64 trainable query vectors attending to intermediate encoder states, enabling effective capture of paralinguistic and non-linguistic acoustic cues, such as prosody, emotion-related patterns, environmental sounds, and music. The key design idea is to expose the backbone to two complementary views of the audio signal for speech, sound, and music: a content-oriented stream that is temporally aligned with textual semantics, and a compact acoustic summary stream that highlights prosodic, stylistic, and non-linguistic cues, while keeping the diffusion backbone lightweight to train.

3.2 Training Data Overview

Our training pipeline follows a progressive curriculum corresponding to the four stages detailed in Sec. 3.3. Broadly, we utilize large-scale transcription data for semantic alignment, diverse audio-centric instruction data for acoustic enrichment, and curated preference pairs for reinforcement-style alignment. Detailed data construction and statistics are provided in Appendix A.

Transcription Data for Semantic Alignment. For Stage 1, we leverage major ASR corpora, including LibriSpeech (Panayotov et al., 2015) and GigaSpeech (Chen et al., 2021), by framing speech recognition as an instruction-following task. We utilize Qwen3-32B (Yang et al., 2025) to generate 25 distinct instruction templates, which are applied to original transcriptions.

SFT Data. To move beyond pure speech recognition and enrich the model's acoustic and paralinguistic understanding (Stages 2 and 3), we construct four complementary types of audio question answering (AQA) data: (i) Caption-grounded AQA: Using multi-domain audio and speech corpora with existing captions or paralinguistic annotations, we prompt a strong LLM to synthesize diverse and high-quality grounded answers. This ensures broad yet controlled coverage of speech, environmental sounds, and music. (ii) Direct audio QA via text-to-speech (TTS), where text QA pairs from general text datasets are converted into speech via TTS, covering simple and complex queries. (iii) Multiple-choice AQA: This subset, derived from existing benchmarks, emphasizes fine-grained discrimination and objective evaluation of specific audio attributes. (iv) ASR Subset: a subset from ASR data.

Preference Alignment Data. In Stage 4 (preference optimization), we construct high-quality preference triplets consisting of an audio input, a question, and a reference answer. We prompt an LLM to generate "rejected" responses that are fluent and superficially plausible but contain subtle audio-related factual errors (e.g., incorrect gender, rhythm, or sound events). Only pairs where the reference is unambiguously superior are retained. These chosen–rejected pairs are utilized for VRPO to sharpen the model's sensitivity to nuanced paralinguistic cues and complex audio reasoning.

3.3 Progressive Four-Stage Training

To fully exploit the dual-adapter architecture and the heterogeneous training data, we adopt a progressive four-stage training curriculum. Throughout all stages, the Whisper-large-v3 encoder is kept frozen, and supervised objectives operate only on the adapters and the diffusion backbone.

Stage 1: Semantic alignment on ASR. In the first stage, we freeze the dLLM backbone and train only the semantic adapter under an ASR-style objective. The goal is to align it with the semantic space, so that its outputs can be seamlessly consumed by the diffusion backbone.

Stage 2: Joint semantic–acoustic alignment. In stage 2, we still keep the diffusion backbone frozen, but jointly align both semantic and acoustic adapters with synthesized SFT data that explicitly incorporates paralinguistic and acoustic cues, thereby enhancing both semantic and acoustic understanding in speech, audio, and music.

Stage 3: Unfreezing the diffusion backbone with LoRA. In this stage, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) to fine-tune the diffusion backbone, thereby strengthening the model's audio understanding capabilities. LoRA strikes a balance between adaptation capacity and training efficiency while effectively mitigating catastrophic forgetting. The formulation of the SFT loss applied across the first three stages is detailed in Section 3.4.

Stage 4: Preference optimization with VRPO. Finally, to refine instruction following and acoustic sensitivity, we apply VRPO (Zhu et al., 2025) in Stage 4. Unlike standard direct preference optimization (DPO) (Rafailov et al., 2023), which can suffer from high variance in the evidence lower bound (ELBO) estimates for dLLMs, VRPO employs variance reduction to stabilize training (details in Sec. 3.5). Using chosen–rejected AQA pairs, DIFFA-2 learns to prefer responses that are more faithful to subtle audio cues (e.g., prosody, emotion, background sound) from curated preference data, further enhancing its audio understanding capabilities.

3.4 Supervised Fine-Tuning of DIFFA-2

Given a textual prompt x, an audio input a, and a target response y, DIFFA-2 models $p_\theta(y \mid x, a)$ through a diffusion denoising process over text tokens, with audio embeddings inserted into the prompt. Audio embeddings (from the adapters) and text prompt tokens remain fully visible and are never masked; only the response tokens are corrupted and denoised.

Let r_0 denote the full response sequence, with length L'. During training, the special token `<endoftext>` is used both as padding and as the end-of-sequence marker, and the model is required to predict it. At each training iteration, tokens in r_0 are independently masked by a special mask token M with probability $t \in (0, 1]$, yielding a noised response r_t. The model is then optimized using a diffusion-style masked prediction objective:

$$\mathcal{L}_{sft-a} = -\,\mathbb{E}_{t,a,p,r_0,r_t} \left[\frac{1}{t} \sum_{i=1}^{L'} \mathbf{1}[r_t^i = \mathrm{M}] \log p_\theta(r_0^i \mid a, p, r_t) \right], \tag{3}$$

where r_t represents the masked response corresponding to a masking ratio of t. This loss trains DIFFA-2 to reconstruct masked tokens in r_t given both audio and textual context, enabling the backbone to exploit bidirectional context and multimodal cues.

3.5 Variance-Reduced Preference Optimization

For preference alignment, we adopt VRPO proposed in (Zhu et al., 2025). This method leverages a DPO-style objective function with Monte Carlo estimates of ELBO, and incorporates optimal budget allocation as well as antithetic sampling strategies to explicitly reduce the variance of these estimates. As illustrated in Figure 1, we perform N (=4) independent sampling steps, and crucially, share identical masking patterns between

the policy model and the reference model when calculating $\widehat{\log p_\theta}$ and $\widehat{\log p_{\text{ref}}}$. This design effectively implements antithetic sampling over diffusion trajectories, thereby stabilizing the preference learning process even when handling long, acoustically rich input sequences.

We first estimate log-likelihoods $\widehat{\log p_\theta}(y \mid x, a)$ using Monte Carlo ELBO estimation:

$$\widehat{\log p_\theta}(y \mid x, a) = \frac{1}{K} \sum_{k=1}^{K} \text{ELBO}_\theta^{(k)}(y \mid x, a), \tag{4}$$

where K is the sample budget. We analogously obtain $\widehat{\log p_{\text{ref}}}(y \mid x, a)$ using a frozen reference model. VRPO then applies a DPO-style objective to the estimated log-ratios:

$$s_\theta(y) = \widehat{\log p_\theta}(y \mid x, a) - \widehat{\log p_{\text{ref}}}(y \mid x, a), \tag{5}$$

$$\mathcal{L}_{\text{VRPO}} = -\log \sigma\Big(\beta\big[s_\theta(y^+) - s_\theta(y^-)\big]\Big), \tag{6}$$

where y^+ and y^- denote the preferred and rejected responses, respectively, and β controls the strength of preference enforcement.

3.6 Inference Procedure

At inference time, we pad the prompt and audio input to the target length and initialize the response sequence r_T as fully masked. DIFFA-2 then performs iterative denoising over T steps, gradually refining r_t from coarse to fine. At each transition $t \to s$, the model predicts the masked tokens conditioned on the audio input a, prompt p, and current corrupted sequence r_t:

$$\hat{r}_t = \arg\max p_\theta(r_0 \mid a, p, r_t), \tag{7}$$

and re-masks the lowest-confidence fraction of tokens to form r_s, enabling iterative refinement with bidirectional context. Following LLaDA (Nie et al., 2025), we adopt a semi-autoregressive strategy that decodes the response in left-to-right blocks, while predicting tokens within each block in parallel and selectively re-masking them across diffusion steps.

To further accelerate decoding, we use the factor-based parallel decoding strategy from fast-dLLMs (Wu et al., 2025), which adaptively chooses how many tokens to update in parallel based on model confidence rather than a fixed threshold. Intuitively, the algorithm allows more aggressive parallel decoding when the model is confident and reduces parallelism in uncertain regions. The formal decision rule and implementation details are provided in Appendix B.4.

4 Experimental Setup

4.1 Training and Inference Setup

DIFFA-2 is trained with the four-stage curriculum in Section 3 using only fully open-source corpora. In total, we use about 11,000 hours of ASR data in Stage 1 and 3,767 hours of curated supervised fine-tuning data in Stages 2–3, together with roughly 3,000 preference pairs for Stage 4; a detailed breakdown of datasets, sample counts, and prompts is given in Appendix A. We adopt LLaDA-8B-Instruct as the dLLM backbone and update only lightweight components (semantic adapter, acoustic adapter, and LoRA parameters), resulting in roughly **1.1%** trainable parameters overall (see Table B.3). Inference configurations for each benchmark are summarized in Appendix B.4.

4.2 Baselines

We compare DIFFA-2 with both proprietary and open-source audio LLMs. As proprietary reference points, we include GPT-4o-Audio (OpenAI et al., 2024) and Gemini 2.0 Flash (Team et al., 2025). Among open-source models, we focus on strong omni/audio baselines such as Qwen3-Omni (Xu et al., 2025b), Qwen2.5-Omni (Xu et al., 2025a), Kimi-Audio (Ding et al., 2025), and the first-generation DIFFA (Zhou et al., 2025). The full list of baselines is given in Appendix B.1.

4.3 Benchmarks

We evaluate DIFFA-2 on four representative benchmarks. Among them, **MMSU** (Wang et al., 2025b), **MMAU** (Sakshi et al., 2025), and **MMAR** (Ma et al., 2025) are audio understanding benchmarks and constitute the primary focus of our study, while **VoiceBench** (Chen et al., 2024b) is included only as an auxiliary evaluation of semantic dialogue ability. Further details on benchmarks are provided in Appendix B.2.

5 Experiments

5.1 Results on Benchmarks

MMSU. Table 1 reports a fine-grained breakdown on MMSU. Among open LALMs of comparable size, DIFFA-2 attains the best overall accuracy (60.45), slightly outperforming Kimi-Audio (59.28), Qwen2.5-Omni (59.09), and other 7–11B

Models	Size	Perception				Reasoning				Overall
		Seman.	Phono.	Para.	Avg	Seman.	Phono.	Para.	Avg	
Qwen3-Omni	30B-A3B	72.13	55.83	38.85	53.20	86.64	82.19	43.58	78.88	65.63
GPT-4o-Audio	-	59.70	41.56	21.44	39.67	80.83	78.74	26.25	71.96	56.38
Gemini 2.0 Flash	-	47.17	41.30	30.62	40.83	70.69	70.69	36.16	47.83	51.03
DIFFA-2	8B	60.63	39.04	**41.92**	**45.58**	**85.29**	**77.58**	43.58	**76.40**	**60.45**
DIFFA-2 (w/ FPD)	8B	59.69	39.57	40.93	45.06	85.02	77.28	42.09	76.00	60.10
Kimi-Audio	7B	57.64	42.30	35.74	43.52	81.77	76.65	**55.22**	76.03	59.28
Qwen2.5-Omni	7B	**61.11**	**43.96**	33.20	43.97	82.40	76.77	46.87	75.21	59.09
MiniCPM-O	8B	56.56	34.05	36.48	40.54	80.71	74.72	46.71	73.57	56.53
DIFFA	8B	52.67	36.65	35.12	40.28	81.53	72.68	45.67	72.92	56.04
Qwen2-Audio	8B	52.14	32.87	35.56	39.02	77.62	64.81	46.67	68.90	53.27
Qwen-Audio-Chat	8B	57.21	38.52	24.70	35.69	58.61	59.78	25.60	55.93	46.92
Phi-4-multimodal	8B	38.72	34.86	29.56	33.41	57.81	65.94	42.09	57.59	44.96
Baichuan-Audio	11B	39.63	31.26	27.09	31.48	57.96	63.92	34.35	55.70	43.09
GLM-4-Voice	9B	27.80	24.52	27.34	26.18	46.10	48.16	44.35	46.76	35.51
Salmonn	7B	31.55	29.08	28.71	29.83	36.43	26.22	25.26	30.04	30.01
LTU	7B	21.34	22.46	18.73	20.81	22.65	25.53	24.74	24.37	22.61

Table 1: Performance breakdown on the MMSU benchmark across perception and reasoning dimensions in Semantics (Seman.), Phonology (Phono.), and Paralinguistics (Para.) domains. "w/ FPD" denotes factor-based parallel decoding.

Model	Size	Sound		Music		Speech		Avg	
		Test-mini	Test	Test-mini	Test	Test-mini	Test	Test-mini	Test
Qwen3-Omni	30B-A3B	78.68	73.70	69.46	72.22	69.37	66.46	72.50	70.77
Gemini 2.0 Flash	-	71.17	68.93	65.27	59.30	75.08	72.87	70.50	67.03
GPT-4o-Audio	-	64.56	63.20	56.29	49.93	66.67	69.33	62.50	60.82
DIFFA-2	8B	**76.28**	**70.83**	63.47	60.10	**69.06**	**70.18**	**69.60**	**67.00**
Qwen2.5-Omni	7B	72.97	69.53	61.68	62.50	60.96	67.93	65.20	66.64
DIFFA-2 (w/ FPD)	8B	75.38	69.43	61.08	59.57	68.47	70.15	68.30	66.34
Kimi-Audio	7B	75.68	70.70	**66.77**	**65.93**	62.16	56.57	68.20	64.40
MiniCPM-O	8B	71.47	-	65.57	-	63.06	-	66.70	-
Phi-4-multimodal	8B	65.47	62.67	64.37	61.97	67.27	63.80	65.70	62.81
Qwen2-Audio	8B	67.27	61.17	56.29	55.67	55.26	55.37	59.60	57.40
DIFFA	8B	46.25	-	43.41	-	59.46	-	49.71	-
Baichuan-Audio	11B	-	59.46	-	49.10	-	42.47	-	50.34
Qwen-Audio-Chat	8B	55.25	56.73	44.00	40.90	30.03	27.95	43.10	41.86
Salmonn	7B	41.14	42.10	37.13	37.83	26.43	28.77	34.90	36.23
GLM-4-Voice	9B	-	27.63	-	27.84	-	35.44	-	30.30
LTU	7B	20.42	20.67	15.97	15.68	15.92	15.33	17.44	17.23

Table 2: Performance breakdown on the MMAU benchmark. "w/ FPD" denotes factor-based parallel decoding.

baselines, while remaining within roughly 5 points of the larger proprietary Qwen3-Omni. DIFFA-2 also achieves the highest perception average (45.58), with clear gains on paralinguistic perception (41.92) and solid performance on semantic and phonological perception. On reasoning, DIFFA-2 reaches the best average (76.40) among open models, with noticeable improvements in semantic and phonological reasoning compared with Kimi-Audio and Qwen2.5-Omni. Relative to the first-generation DIFFA, DIFFA-2 improves both perception and reasoning (overall +4.41 points), indicating that the upgraded acoustic front-end and four-stage training pipeline translate into stronger speech understanding on MMSU.

MMAU. On MMAU (Table 2), DIFFA-2 again performs competitively with strong AR LALMs. It achieves the best average accuracy among open models on both *Test-mini* and *Test* splits (69.60 and 67.00) surpassing Qwen2.5-Omni and Kimi-Audio, and approaching larger or proprietary systems such as Qwen3-Omni and Gemini. DIFFA-2 is particularly strong on sound and speech, where it attains the highest scores among open models, while its music performance is slightly behind Kimi-Audio and MiniCPM-O but remains competitive without any music-specialized design. Compared with DIFFA, DIFFA-2 gains nearly 20 points in average

Models	Size	Single Modality (%)			Mixed Modalities (%)				Avg (%)
		Sound	Music	Speech	Sound-Music	Sound-Speech	Music-Speech	All	
Qwen3-Omni	30B-A3B	59.39	54.37	70.41	90.91	74.77	63.41	70.83	65.90
Gemini 2.0 Flash	-	61.21	50.97	72.11	81.82	72.48	65.85	70.83	65.60
GPT-4o-Audio	-	53.94	50.97	70.41	63.64	72.48	62.20	75.00	63.50
Qwen2.5-Omni	7B	**55.76**	**41.75**	**54.42**	**45.45**	55.96	**57.32**	**54.17**	**51.40**
DIFFA-2	8B	54.55	**41.75**	53.40	**45.45**	**58.26**	54.88	37.50	50.80
DIFFA-2 (w/ FPD)	8B	53.94	36.41	52.72	36.36	56.88	53.66	45.83	50.20
MiniCPM-O	8B	49.70	36.41	50.00	36.36	53.67	45.12	**54.17**	48.60
Baichuan-Omni-1.5	7B	41.21	33.01	40.48	36.36	48.62	39.02	41.67	40.70
DIFFA	8B	37.58	31.07	39.46	36.36	43.12	45.12	25.00	37.20
Salmonn	13B	30.30	29.61	34.69	9.09	34.86	35.37	41.67	33.20
Salmonn	7B	30.91	25.73	34.35	9.09	37.61	28.05	37.50	32.80
Qwen2-Audio	8B	33.33	24.27	32.31	9.09	31.19	30.49	25.00	30.00
OpenOmni	8B	20.61	22.33	35.37	18.18	27.06	23.17	25.00	27.00
Qwen-Audio-Chat	8B	27.88	20.39	22.11	9.09	25.23	25.61	20.83	23.50
LTU	7B	19.39	19.90	13.95	18.18	24.77	21.95	16.67	19.20

Table 3: Performance breakdown on the MMAR benchmark. The 'All' category denotes the comprehensive evaluation across mixed sound, music, and speech modalities. "w/ FPD" denotes factor-based parallel decoding.

Model	WER ↓		RTF ↓	
	clean	other	clean	other
LLaMA-Audio (S1)	**2.43**	**5.09**	0.1402	0.1418
DIFFA-2 (S1)	2.72	5.34	0.6792	0.7489
DIFFA-2 (S1 w/ FPD)	3.05	5.68	**0.0820**	**0.0867**

Table 4: Word error rate (WER%) and real-time factor (RTF) of Stage-1 with dLLMs and AR backbone on Librispeech-clean and Librispeech-other testing set. "w/ FPD" denotes factor-based parallel decoding.

accuracy on *Test-mini* (49.71 → 69.60), suggesting that the enhanced acoustic modeling and multi-stage training generalize well to high-level audio understanding across sound, music, and speech.

MMAR. The MMAR benchmark (Table 3) evaluates single and mixed audio modalities with more compositional queries. DIFFA-2 achieves an average accuracy of 50.80%, substantially improving over DIFFA (37.20%, +13.6 points) and outperforming other 8B open baselines such as MiniCPM-O (48.60%) and Qwen2-Audio (30.00%). On single-modality tasks (sound, music, speech), DIFFA-2 consistently improves over DIFFA and narrows the gap to Qwen2.5-Omni. For the most challenging Sound–Music–Speech mixtures, DIFFA-2 underperforms relative to Qwen2.5-Omni, likely reflecting the lack of mixed-modality supervision in the training data. Overall, MMAR indicates that DIFFA-2 effectively extends diffusion-based modeling to multi-source audio composition, while complex three-way mixtures remain a challenging regime compared with the strongest AR-based LALMs.

VoiceBench. We additionally evaluate on VoiceBench to assess dialogue-style spoken interaction. As shown in Appendix C, DIFFA-2 lags behind heavily instruction-tuned omni models such as GPT-4o-Audio and Qwen3-Omni, but still improves over DIFFA and several open-source baselines, which is consistent with our design focus on audio understanding rather than extensive conversational tuning.

Summary. Across MMSU, MMAU, and MMAR, DIFFA-2 consistently outperforms the first-generation DIFFA and often matches or surpasses strong open AR LALMs of similar size, while approaching larger proprietary systems on several metrics. Gains are particularly clear in semantic and phonological reasoning and in sound and speech understanding, whereas performance on VoiceBench remains behind omni models that are heavily tuned for dialogue and alignment. These results indicate that, under realistic data constraints, diffusion-based backbones can serve as competitive audio understanding models, and that additional dialogue-centric supervision could further close the gap on interactive voice assistant benchmarks.

5.2 Ablation Study

We first compare Stage-1 ASR performance between diffusion and autoregressive backbones (Table 4). Under the same ASR-style training on LibriSpeech, the AR baseline LLaMA-Audio (S1) attains slightly lower word error rate (WER) than DIFFA-2 (S1), which is consistent with the advantage of strictly left-to-right decoding for monotonic transcription. When we enable factor-based par-

Model	MMAU			Overall	MMSU (Perception)				MMSU (Reasoning)				Overall
	Sound	Music	Speech		Sem.	Phon.	Para.	Avg	Sem.	Phon.	Para.	Avg	
LLaMA-Audio (S2)	65.47	54.79	62.16	60.80	39.21	30.91	30.23	32.69	51.35	63.97	44.18	55.45	43.71
LLaMA-Audio (S3)	73.87	65.87	62.46	67.40	50.39	**39.89**	36.67	41.22	75.45	73.18	45.07	70.33	55.31
DIFFA-2 (S2)	72.07	52.69	66.97	63.90	52.91	36.58	31.32	38.54	84.03	75.74	**46.57**	75.50	56.43
DIFFA-2 (S3)	74.77	62.57	67.27	68.20	59.53	37.54	40.63	44.16	84.66	76.15	44.48	75.70	59.41
DIFFA-2 (S4)	**76.28**	**63.47**	**69.07**	**69.60**	**60.63**	39.04	**41.92**	**45.58**	**85.29**	**77.58**	43.58	**76.40**	**60.45**

Table 5: Ablation of DIFFA-2 and LLaMA-Audio's multi-stage training on MMAU and MMSU. LLaMA-Audio is based on LLaMA 3.1 backbone and then trained with the same data and settings.

allel decoding for DIFFA-2 (S1, w/ FPD), WER increases moderately, but the real-time factor (RTF) drops substantially and becomes lower than that of the AR baseline. This indicates that a diffusion backbone does not inherently improve low-level recognition accuracy, but its inference latency can be made competitive with, or better than, an AR backbone by using an appropriate parallel decoding scheme in ASR task. On audio understanding benchmarks, DIFFA-2 with factor-based parallel decoding attains accuracy that is close to the standard setting, suggesting that it provides a practical knob to trade off accuracy and latency for diffusion-based audio models.

Table 5 reports the multi-stage training ablation on MMAU and MMSU, comparing DIFFA-2 with LLaMA-Audio under matched data and training curriculum. With only Stage 2 (adapter-only alignment), DIFFA-2 already achieves higher overall scores than LLaMA-Audio (S2) on both benchmarks, suggesting that the diffusion backbone benefits more from the same semantic–acoustic alignment. Advancing from Stage 2 to Stage 3 improves both models, and DIFFA-2 gains more on reasoning-oriented metrics. Adding Stage 4 (VRPO) further improves DIFFA-2, yielding the best overall performance and more balanced gains across perception and reasoning.

Overall, the ablation highlights a difference between transcription and audio understanding. For token-level ASR, the AR backbone retains a small advantage in WER, in line with its sequential decoding nature. For holistic audio QA, DIFFA-2 with a diffusion backbone achieves stronger performance under the same data and multi-stage training, which may be attributed to the corruption–reconstruction training objective of dLLMs, which has been shown in text domains to make more effective use of limited data, making it better aligned with audio understanding benchmarks, although we do not completely disentangle backbone pre-training effects.

6 Related Work

6.1 Large Audio Language Models

Recent LALMs primarily adopt autoregressive backbones. A common design couples a speech encoder to an LLM via lightweight bridging modules (e.g., Qwen2-Audio (Chu et al., 2024), SALMONN (Tang et al., 2024a), Audio-Flamingo2 (Ghosh et al., 2025), with omni models further supporting streaming and multi-modality (Xu et al., 2025a,b). Another line tokenizes audio into discrete sequences (Zhang et al., 2023; Défossez et al., 2024), while Kimi-Audio (Ding et al., 2025) fuses discrete and continuous representations. These models largely rely on AR decoding.

6.2 Diffusion Large Language Models

Diffusion large language models generate sequences by iteratively denoising corrupted tokens, offering bidirectional context modeling and parallel token updates. Early work on diffusion for discrete text (Austin et al., 2021; Shi et al., 2024; Sahoo et al., 2024) established the feasibility of this paradigm, and LLaDA (Nie et al., 2025; Zhu et al., 2025) scaled it to large language models with strong performance on understanding and reasoning tasks. Recent efforts improve inference efficiency via training-free acceleration, including KV-cache-like reuse with confidence-aware parallel decoding and adaptive length prediction (Li et al., 2025a; Wei et al., 2025). Our work is closely related to diffusion LMs such as LLaDA and fast-dLLMs (Wu et al., 2025). LLaDA and its variants focus on text generation and do not consider audio encoders or audio-specific training curricula.

7 Conclusions

This paper presents **DIFFA-2**, an enhanced dLLMs-based LALM for audio understanding. Despite

having only 1.1% (99M) trainable parameters and utilizing a modest 14.8k hours of open-source data, DIFFA-2 achieves substantial performance gains over its predecessor. Evaluations on MMSU, MMAU, and MMAR benchmarks demonstrate that DIFFA-2 is competitive with leading autoregressive models. These results establish dLLMs-based modeling as a highly competitive alternative for universal audio understanding tasks.

Limitations

Although DIFFA-2 achieves strong results on audio understanding benchmarks, several limitations remain. First, our training objectives and data curation are geared toward fine-grained audio understanding rather than open-domain spoken dialogue. Consequently, DIFFA-2 is exposed to only limited conversational and alignment-style supervision, which is reflected in its mid-range performance on VoiceBench compared with heavily instruction-tuned AR Omni-models. Designing a more balanced training recipe that jointly targets audio understanding and spoken dialogue is an important direction for future work. Second, we focus exclusively on text-based audio understanding and do not consider speech generation or streaming, full-duplex interaction. DIFFA-2 is evaluated in an offline speech-in/text-out setting; integrating it into end-to-end speech-in/speech-out systems and assessing user-centric metrics such as latency and interaction quality are important next steps. Third, we apply a simple training-free factor-based parallel decoding scheme to reduce diffusion steps and observe clear latency gains with negligible accuracy loss, but DIFFA-2 is not yet uniformly faster than strong AR audio LLMs under all settings. We view this as a systems-level design choice rather than a fundamental limitation of dLLMs backbones; adapting more advanced training-free acceleration methods from text dLLMs to audio is a promising but orthogonal direction for future work.

Acknowledgments

This work has been supported by the National Key R&D Program of China (Grant No.2022ZD0116307) and NSF China (Grant No.62271270).

References

Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, and 1 others. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Afroz Ahamad, Ankit Anand, and Pranesh Bhargava. 2020. Accentdb: A database of non-native english accents to assist neural speech recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5353–5360. European Language Resources Association.

Jacob Austin, DanielD. Johnson, Jonathan Ho, Daniel Tarlow, and Riannevanden Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *arXiv: Learning,arXiv: Learning*.

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

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, and et al. 2024a. Slam-omni: Timbre-controllable voice interaction system with single-stage training. *arXiv preprint arXiv:2412.15649*.

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

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. 2016. Fma: A dataset for music analysis. *Preprint*, arXiv:1612.01840.

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

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. 2025. Scaling rich style-prompted text-to-speech datasets. *Preprint*, arXiv:2503.04713.

Seungheon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023. Lp-musiccaps: Llm-based pseudo music captioning. In *Ismir 2023 Hybrid Conference*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2019. Clotho: An audio captioning dataset. *Preprint*, arXiv:1910.09387.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, and 1 others. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. 2017. Neural audio synthesis of musical notes with wavenet autoencoders. *Preprint*, arXiv:1704.01279.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. LLaMA-omni: Seamless speech interaction with large language models. In *The Thirteenth International Conference on Learning Representations*.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, and 1 others. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Forty-second International Conference on Machine Learning*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*.

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. *Preprint*, arXiv:2205.03433.

Haolin He, Xingjian Du, Renhe Sun, Zheqi Dai, Yujia Xiao, Mingru Yang, Jiayi Zhou, Xiquan Li, Zhengxi Liu, Zining Liang, and 1 others. 2025. Measuring audio's impact on correctness: Audio-contribution-aware post-training of large audio language models. *arXiv preprint arXiv:2509.21060*.

William Held, Ella Li, Michael Ryan, Weiyan Shi, Yanzhe Zhang, and Diyi Yang. 2024. Distilling an end-to-end voice assistant without instruction training data. *arXiv preprint arXiv:2410.02678*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, and 1 others. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. 2024. Speechcraft: A fine-grained expressive speech dataset with natural language description. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1255–1264.

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

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 119–132. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jiaqi Wang, and Dahua Lin. 2025a. Beyond fixed: Training-free variable-length denoising for diffusion large language models. *arXiv preprint arXiv:2508.00819*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, and 1 others. 2025b. Baichuan-audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, and 1 others. 2024. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, and Chao-Han Huck Yang et al. 2025. Desta2.5-audio: Toward general-purpose large audio language model with self-generated cross-modal alignment. *Preprint*, arXiv:2507.02768.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, Kai Li, Keliang Li, Siyou Li, Xinfeng Li, Xiquan Li, Zheng Lian, Yuzhe Liang, Minghao Liu, Zhikang Niu, and 15 others. 2025. MMAR: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *Preprint*, arXiv:2303.17395.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. In *Interspeech 2017*.

Jinjie Ni, Qian Liu, Longxu Dou, Chao Du, Zili Wang, Hang Yan, Tianyu Pang, and Michael Qizhe Shieh. 2025. Diffusion language models are super data learners. *arXiv preprint arXiv:2511.03276*.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, and Adam Perelman et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. 2025. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *The Thirteenth International Conference on Learning Representations*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Karol J. Piczak. 2015. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536. Association for Computational Linguistics.

Paul Primus, Florian Schmid, and Gerhard Widmer. 2025. Tacos: Temporally-aligned audio captions for language-audio pretraining. *arXiv preprint arXiv:2505.07609*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *ArXiv*, abs/2406.07524.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. 2024. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024a. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024b. Salmonn: Towards generic hearing abilities for large language models. In *ICLR*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, and et al. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

OpenBMB MiniCPM-o Team. 2025. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L. Seltzer. 2024. Towards measuring fairness in speech recognition: Fair-Speech dataset. In *Interspeech 2024*, pages 1385–1389.

Chen Wang, Tianyu Peng, Wen Yang, Yinan Bai, Guangfu Wang, Jun Lin, Lanpeng Jia, Lingxiang Wu, Jinqiao Wang, Chengqing Zong, and 1 others. 2025a. Opens2s: Advancing fully open-source end-to-end empathetic large speech language model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 906–917.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. 2025b. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark. *arXiv preprint arXiv:2506.04779*.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long MA. 2025c. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *Forty-second International Conference on Machine Learning*.

Linye Wei, Wenjue Chen, Pingzhi Tang, Xiaotian Guo, Le Ye, Runsheng Wang, and Meng Li. 2025. Orchestrating dual-boundaries: An arithmetic intensity inspired acceleration framework for diffusion language models. *Preprint*, arXiv:2511.21759.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. 2025. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *Preprint*, arXiv:2505.22618.

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

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. Dataset from University of Edinburgh.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

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

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Jiaming Zhou, Hongjie Chen, Shiwan Zhao, Jian Kang, Jie Li, Enzhi Wang, Yujie Guo, Haoqin Sun, Hui Wang, Aobo Kong, and 1 others. 2025. Diffa: Large language diffusion models can listen and understand. *arXiv preprint arXiv:2507.18452*.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *Preprint*, arXiv:2505.19223.

A Data Details and Prompt Templates

A.1 ASR Data

We use LibriSpeech (Panayotov et al., 2015) and GigaSpeech (Chen et al., 2021) as ASR corpora. For each transcript, we construct instruction-style ASR samples by applying 25 instruction templates generated by Qwen-32B (e.g., "Please transcribe the following audio..." or "What is the exact content of this recording?"). The full list of instruction templates is shown in Table A.1.

Examples of ASR Prompts in Stage 1

```
1. Please transcribe the audio to text.
2. Convert this speech to text.
3. What is being said in this audio?
4. Transcribe the following audio clip.
5. Please write down what you hear in the
audio.
6. Convert the spoken words to written text.
7. What words are spoken in this recording?
8. Please provide a transcription of this
audio.
9. Turn this speech into text format.
10. Write out what is said in the audio
file.
```

Figure A.1: Examples of ASR Prompts in Stage 1

Dataset	Samples	Duration (h)
Sound & General Audio		
AudioCaps	181,453	383.47
Clotho	75,090	18.04
ESC50	9,000	1.39
TACOS	33,320	57.78
VocalSound	20,208	23.51
WavCaps_AudioSetSL	108,308	296.95
WavCaps_Freesound30s	155,287	432.80
Subtotal	***582,666***	***1,213.94***
Music		
FMA_medium	16,896	140.74
LP-MusicCaps-MTT	15,560	126.54
MusicCaps	2,568	7.14
Nsynth	296,382	329.31
Subtotal	***331,406***	***603.73***
Speech		
AccentDB_extended	50,622	19.28
CompA-R	197,218	170.60
EmoV_DB	68,930	9.49
IEMOCAP	50,304	5.23
MELD	6,450	0.93
SpeechCraft	228,008	483.82
VCTK-Corpus	176,968	44.04
AlpacaTrain	19,307	25.17
MetaFairASR	46,342	48.66
NaturalQuestions	9,022	10.25
Opens2s	49,368	67.47
Paraspeechcaps	184,552	497.07
TrivalQA	79,565	104.29
VoxCeleb1	297,284	340.39
WebQuestions	2,576	2.49
Subtotal	***1,466,516***	***1,829.18***
Total	**2,380,588**	**3,646.85**

Table A.1: Dataset Statistics categorized by domain (Sound, Music, and Speech). Duration represents unique audio hours.

Prompts of Audio QA Data Creation in Stage 2

```
[System]
You are a helpful voice assistant.
Imagine you can hear the audio clips.
Focus on the audios and respond directly to
the prompts.

[User]
This is the audio: {Audio Description}.
{Text Prompt}

[Assistant]

- ...
```

Figure A.2: Prompts of Audio QA Data Creation in Stage 2

A.2 SFT Data

Part 1: Audio-caption-based AQA. We collect a diverse set of audio and speech datasets covering speech, environmental sounds, and music, including ParaSpeechCaps (Diwan et al., 2025), AudioCaps (Kim et al., 2019), WavCaps (Mei et al., 2023), VocalSound (Gong et al., 2022), NSynth (Engel et al., 2017), FMA-medium (Defferrard et al., 2016), ESC-50 (Piczak, 2015), Clotho (Drossos et al., 2019), AccentDB (Ahamad et al., 2020), EmoV-DB (Adigwe et al., 2018), IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), VCTK (Yamagishi et al., 2019), Meta FAIR ASR (Veliche et al., 2024), and VoxCeleb1 (Nagrani et al., 2017). From the Desta 2.5 (Lu et al., 2025) dataset, we select a subset of question types and paralinguistic annotations, and use Qwen3-32B to generate an answer for each audio–question pair. The exact prompts are listed in Figure A.2.

Part 2: Direct Audio QA. We construct three categories of direct audio QA data: simple QA, complex QA, and empathetic QA. We first collect text-only QA pairs from Alpaca (Taori et al., 2023), Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and WebQuestions (Berant et al., 2013), and then synthesize speech with CosyVoice2 (Du et al., 2024), using speaker prompts randomly sampled from the LibriSpeech training set. We categorize samples into simple or complex based on answer length and apply different instruction templates for each type. We

Examples of Prompts in Stage 2

```
[Simple QA Prompts]
1. Listen to the audio question and reply
with a concise and factual answer.
2. Respond to the spoken query with a
brief, accurate answer based on the audio.
3. Give a straightforward response to the
question in the audio.
4. Respond directly to the audio question,
stating only what is necessary.

[Complex QA Prompts]
1. Listen to the audio question and provide
a comprehensive, detailed answer covering
all aspects.
2. Respond to the spoken message with an
in-depth explanation addressing every part
of the query.
3. Offer a full, structured answer to the
voice question, including background and
supporting details.

[Sympathetic QA Prompts]
1. Listen to the voice message and respond
in a natural, conversational manner,
showing empathy and genuine understanding.
2. Reply to the spoken message as if
engaged in a friendly, real-life
conversation, maintaining warmth and
authenticity.
3. Respond in a smooth, natural tone that
conveys a human touch and kindness.
```

Figure A.3: Examples of prompts in Stage 2

further augment this with the English subset of OpenS2S (Wang et al., 2025a). Detailed statistics and prompt templates are shown in Tables A.1 and Figure A.3.

Part 3: ASR Data. We randomly sample 5% of the ASR data used in Stage 1 and add it to the Stage 2 training set to preserve ASR-related acoustic grounding.

Part 4: Multi-choice AQA Data We employ the AudioMCQ (He et al., 2025) corpus, which integrates multi-choice AQA instances derived from AudioCaps, Clotho, CompA-R (Ghosh et al., 2024), MusicCaps (Agostinelli et al., 2023), LP-MusicCaps (Doh et al., 2023), SpeechCraft (Jin et al., 2024), and TACOS (Primus et al., 2025). We follow the official data construction recipe of AudioMCQ without chain-of-thought, and then combine it with the Stage 2 data used for adapter training. Detailed dataset statistics are provided in Table A.1.

A.3 Preference Data and Prompts

To construct preference data for VRPO, we start from high-quality audio QA instances sampled from SFT data. Given an audio input, a question, and a reference answer, we prompt a language model to generate a fluent but partially incorrect answer that introduces subtle audio-related errors. The full prompt templates and examples for preference data construction are listed in Figure A.4.

B Additional Experimental Details

All models and datasets utilized in this study are released under open-source licenses, in compliance with the terms of their respective original distributions.

B.1 Baseline Models

For completeness, we list here all baseline models used in our experiments. In the main text, we highlight a subset of representative systems (e.g., Qwen2.5-Omni, Kimi-Audio, Qwen2-Audio, MiniCPM-O, and DIFFA); the full set of models covered in our evaluation is summarized in Table B.1.

Model	Reference
GPT-4o-Audio	(OpenAI et al., 2024)
Gemini 2.0 Flash	(Team et al., 2025)
Qwen3-Omni	(Xu et al., 2025b)
Qwen2.5-Omni	(Xu et al., 2025a)
Kimi-Audio	(Ding et al., 2025)
MiniCPM-O	(Team, 2025)
Qwen2-Audio	(Chu et al., 2024)
Baichuan-Omni-1.5	(Li et al., 2024)
Baichuan-Audio	(Li et al., 2025b)
GLM-4-Voice	(Zeng et al., 2024)
Step-Audio	(Huang et al., 2025)
LLaMA-Omni	(Fang et al., 2025)
Slam-Omni	(Chen et al., 2024a)
Freeze-Omni	(Wang et al., 2025c)
Mini-Omni / Mini-Omni2	(Xie and Wu, 2024a,b)
Moshi	(Défossez et al., 2024)
DiVA	(Held et al., 2024)
VITA	(Fu et al., 2024)
LTU	(Gong et al., 2024)
Salmonn	(Tang et al., 2024b)
Qwen-Audio-Chat	(Chu et al., 2023)
DIFFA	(Zhou et al., 2025)

Table B.1: Baseline models used in our experiments.

B.2 Benchmark Details

MMSU (Wang et al., 2025b) is a large-scale benchmark for assessing perception and reasoning in

Preference Data Generation Prompt

```
[System]
You are creating training data for an audio
question answering model (Audio-QA).
You will NOT see the raw audio, only a
textual description of it.

You are given a QUESTION and a
REFERENCE_ANSWER which should be treated as
the GOOD answer.
Your task is to generate ONE BAD_ANSWER to
the same question.

Requirements for BAD_ANSWER:
- It must be fluent and look superficially
reasonable.
- BUT it must be partially incorrect with
respect to the audio description (e.g.,
wrong attribute, wrong event, missing key
detail).
- The error should be about the
AUDIO-related content (e.g., rhythm, sound
type, emotion, gender), rather than just
style or verbosity.
- DO NOT make the bad answer obviously
nonsensical or completely unrelated.
- It must still directly answer the
question (not "I don't know").

At the end, you must also CHECK whether the
REFERENCE_ANSWER is clearly better than the
BAD_ANSWER.
If they are too similar or you are not
confident, mark the pair as unusable.

Output your result in STRICT JSON with the
following fields only:
{
  "bad_answer": "...",
  "explanation": "...",
  "usable": true or false
}
Do NOT add any extra text outside the JSON
object.

[User]
Audio Description:
{audio_desc}

Question:
{question}

Reference Answer (GOOD):
{gold_answer}

[Assistant]:
{bad_answer}
```

Figure A.4: Preference Data Generation Prompt

realistic spoken-language scenarios. It contains 5,000 audio–question–answer triplets across 47 tasks, covering both linguistic and paralinguistic phenomena such as phonetics, prosody, semantics, emotion, and speaker traits. Tasks are divided into perception- and reasoning-oriented categories, enabling a comprehensive evaluation of fine-grained audio understanding.

Stage	LR	Batch	Warmup	Epochs
Stage 1	1×10^{-4}	1280	1000	12
Stage 2	5×10^{-5}	196	1000	10
Stage 3	5×10^{-5}	196	1000	10
Stage 4	5×10^{-6}	4	200	1

Table B.2: Stage-wise training configuration of DIFFA-2.

MMAU (Sakshi et al., 2025) evaluates advanced audio understanding via human-annotated multiple-choice questions over speech, music, and environmental sounds. It emphasizes high-level reasoning and expert knowledge rather than low-level perception. We report results on the *Test-mini* split.

MMAR (Ma et al., 2025) is a multi-task audio reasoning benchmark designed to assess reasoning consistency and generalization across heterogeneous audio modalities. It focuses on joint perceptual grounding and reasoning over diverse audio inputs, providing a complementary evaluation of audio reasoning robustness.

VoiceBench (Chen et al., 2024b) targets semantic dialogue ability in spoken interaction. It is constructed by converting text-based benchmarks into audio queries using TTS and evaluates general knowledge, instruction following, and safety. Since VoiceBench primarily measures semantic dialogue performance rather than audio understanding, we include it only as a supplementary evaluation.

B.3 Training Details

DIFFA-2 is trained using a four-stage schedule with progressively reduced learning rates. LoRA is applied to the dLLMs backbone in Stage 3 with a rank of 8 and a scaling factor α of 16, while preference optimization is performed in Stage 4. We employ 64 NVIDIA A100 GPUs for the first three stages and 4 A100 GPUs for the final stage. The entire training pipeline takes approximately 5 days to complete.

In the DIFFA-2 vs. LLaMA-Audio comparison we use the same datasets, stage-wise curriculum, LoRA configuration, and optimization hyperparameters.

Module	#Params	Trainable
Whisper-Large-V3 Encoder	637M	0
Semantic Adapter	36.4M	36.4M
Acoustic Adapter	47.9M	47.9M
dLLMs Backbone	8.03B	0
LoRA Modules	14.7M	14.7M
Total	**8.77B**	**99.0M**

Table B.3: Statistics of parameters

B.4 Inference Details

At inference time, we pad the prompt and audio input to the target length and initialize the response sequence r_T as fully masked. DIFFA-2 then performs iterative denoising over T steps, progressively refining the response from coarse to fine (Figure 1).

At each denoising transition $s_1 \to s_2$, the model predicts masked tokens conditioned on the audio input a, prompt p, and the current corrupted sequence r_t:

$$\hat{r}_{s_1} = \arg\max p_\theta(r_0 \mid a, p, r_{s_1}). \tag{8}$$

Token-level confidence scores are used to re-mask a fraction of low-confidence tokens—proportional to $\lceil s_2/s_1 \rceil$—to form the next intermediate sequence r_{s_2}, enabling iterative refinement with full bidirectional context.

Following LLaDA (Nie et al., 2025), we also adopt a semi-autoregressive decoding strategy that generates the response in a block-wise left-to-right manner. Within each block, tokens are decoded in parallel and selectively re-masked across diffusion steps, balancing generation quality and efficiency.

Factor-based parallel decoding. To further accelerate decoding, we adopt the factor-based parallel decoding strategy proposed in fast-dLLMs (Wu et al., 2025). This strategy extends threshold-based decoding by adaptively determining how many tokens to decode in parallel based on model confidence, rather than relying on a fixed confidence threshold.

Given the marginal confidence estimates of candidate tokens within a decoding block, we first sort confidences in descending order and select the largest number of tokens n such that

$$(n+1)\big(1 - c^{(n)}\big) < f, \tag{9}$$

where $c^{(n)}$ denotes the n-th highest confidence and f is a decoding factor hyperparameter. This criterion allows more aggressive parallel decoding when the model is confident, while conservatively reducing parallelism in uncertain regions.

By leveraging this factor-based strategy on top of the diffusion denoising process, DIFFA-2 achieves improved inference efficiency while preserving generation quality.

Benchmark	Answer length	Block length	Steps
Librispeech	128	128	128
MMSU	16	16	16
MMAU	16	16	16
MMAR	16	16	16
AlpacaEval	128	32	128
CommonEval	128	32	128
SD-QA	128	32	128
MMSU*	16	16	16
OBQA	16	16	16
IFEval	256	32	256
AdvBench	128	32	128

Table B.4: Inference hyperparameters used for each benchmark

Inference hyperparameters used for each benchmark is presented in Table B.4. Key parameters include the maximum answer length, the block length for incremental decoding, and the total denoising steps. We synchronize the denoising budget with the sequence length to achieve peak generation quality. For factor-based parallel decoding, we set f to 1.0. Note that the steps are invalid when the factor parallel decoding is applied. All inference experiments are conducted on a single NVIDIA A100 GPU. The prompt for audio understanding tasks is provided in Figure B.1.

C Additional Experiments

VoiceBench (Table C.1) focuses on spoken dialogue, instruction following, and safety rather than pure audio understanding. DIFFA-2 is trained with very limited dialogue-style audio data, and its overall score (59.63) is therefore clearly lower than heavily instruction-tuned AR omnimodels such as GPT-4o-Audio, Kimi-Audio, and Qwen2.5-Omni. Nevertheless, DIFFA-2 remains competitive with or better than several open-source baselines (e.g., GLM-4-Voice, DiVA, Qwen2-Audio, Freeze-Omni), and improves markedly over DIFFA (48.22) across most VoiceBench metrics. This gap between strong performance on MMSU/MMAU/MMAR and mid-range performance on VoiceBench highlights the design focus of DIFFA-2: it is optimized primarily for fine-grained audio understanding rather than large-scale conversational tuning.

Model	AlpacaEval	CommonEval	SD-QA	MMSU*	OBQA	IFEval	AdvBench	Overall
GPT-4o-Audio	**4.78**	**4.49**	**75.50**	**80.25**	**89.23**	**76.02**	98.65	**86.43**
Kimi-Audio	4.46	3.97	63.12	62.17	83.52	61.10	**100.00**	76.93
Qwen-2.5-Omni	4.50	3.84	56.40	61.70	80.90	53.50	99.20	74.04
Phi-4-multimodal	3.81	3.82	39.78	42.19	65.93	45.35	100.00	63.69
DIFFA-2	3.86	3.67	40.78	38.13	61.53	40.76	85.58	59.63
GLM-4-Voice	3.97	3.42	36.98	39.75	53.41	25.92	88.08	55.99
DiVA	3.67	3.54	57.06	25.76	25.49	39.16	98.27	55.70
Qwen2-Audio	3.74	3.43	35.72	35.72	49.45	26.33	96.73	55.34
Freeze-Omni	4.03	3.46	53.45	28.14	30.98	23.40	97.30	54.72
Step-Audio	4.13	3.09	44.21	28.33	33.85	27.96	69.62	49.77
DIFFA	3.78	2.96	34.45	29.57	35.60	26.56	76.54	48.22
LLaMA-Omni	3.70	3.46	39.69	25.93	27.47	14.87	11.35	37.50
VITA	3.38	2.15	27.94	25.70	29.01	22.82	26.73	34.68
Slam-Omni	1.90	1.79	4.16	26.06	25.27	13.38	94.23	33.84
Mini-Omni2	2.32	2.18	9.31	24.27	26.59	11.56	57.50	31.32
Moshi	2.01	1.60	15.64	24.04	25.93	10.12	44.23	27.45

Table C.1: Performance breakdown on VoiceBench. Metrics cover diverse direct QA and alignment tasks. Note that MMSU* in VoiceBench is derived from MMLU-Pro, which differs from the MMSU benchmark.

Prompt for Audio Understanding Tasks

```
[System]
You are a helpful voice assistant.

[User]
This is the audio: {Audio Infomation}.

Choose the most suitable answer from
options A, B, C, and D to respond the
question in next line. Do not provide any
additional explanations or content.

Question:{Text Question}

Options:{Text Options}

[Assistant]

- ...
```

Figure B.1: Prompt for audio understanding tasks

D AI Usage

In this work, AI assistants are employed for language refinement and manuscript polishing, including enhancing clarity, coherence, and grammatical accuracy. We assume full responsibility for the content, ensuring compliance with academic standards and the absence of misconduct or plagiarism.