DIFFA: Large Language Diffusion Models Can Listen and Understand

Jiaming Zhou[1,2*], Hongjie Chen[2], Shiwan Zhao[1], Jian Kang[2], Jie Li[2], Enzhi Wang[1], Yujie Guo[1], Haoqin Sun[1], Hui Wang[1], Aobo Kong[1], Yong Qin[1†], Xuelong Li[2†],

[1]College of Computer Science, Nankai University
[2]Institute of Artificial Intelligence (TeleAI), China Telecom, China
zhoujiaming@mail.nankai.edu.cn, qinyong@nankai.edu.cn

Abstract

Recent advances in large language models (LLMs) have shown remarkable capabilities across textual and multimodal domains. In parallel, large language diffusion models have emerged as a promising alternative to the autoregressive paradigm, offering improved controllability, bidirectional context modeling, and robust generation. However, their application to the audio modality remains underexplored. In this work, we introduce **DIFFA**, the first diffusion-based large audio-language model designed to perform spoken language understanding. DIFFA integrates a frozen diffusion language model with a lightweight dual-adapter architecture that bridges speech understanding and natural language reasoning. We employ a two-stage training pipeline: first, aligning semantic representations via an ASR objective; then, learning instruction-following abilities through synthetic audio-caption pairs automatically generated by prompting LLMs. Despite being trained on only 960 hours of ASR and 127 hours of synthetic instruction data, DIFFA demonstrates competitive performance on major benchmarks, including MMSU, MMAU, and VoiceBench, outperforming several autoregressive open-source baselines. Our results reveal the potential of large language diffusion models for efficient and scalable audio understanding, opening a new direction for speech-driven AI.

Code — https://github.com/NKU-HLT/DIFFA
Extended version — https://arxiv.org/abs/2507.18452

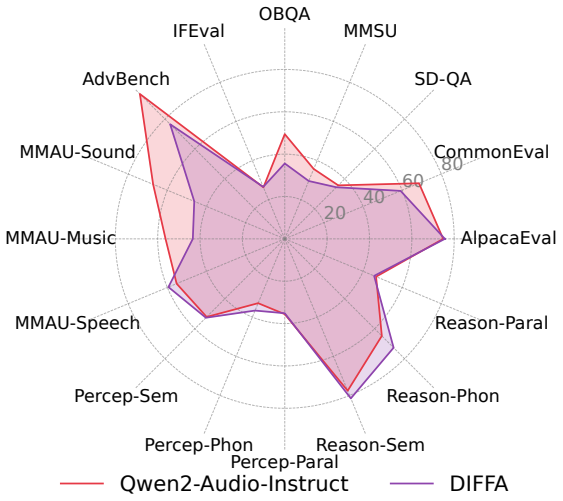


Figure 1: DIFFA vs. Qwen2-Audio-Instruct. The abbreviations correspond to MMSU benchmark's capabilities: Perception-Semantics (Percep-Sem), Perception-Phonology (Percep-Phon), Perception-Paralinguistics (Percep-Paral), Reasoning-Semantics (Reason-Sem), Reasoning-Phonology (Reason-Phon), and Reasoning-Paralinguistics (Reason-Paral).

1 Introduction

Large language models (LLMs) have catalyzed a paradigm shift in artificial intelligence, pushing the frontiers of natural language understanding, computer vision, and multimodal reasoning. In the domain of speech and audio processing, large audio-language models (LALMs) have similarly benefited from these advances in LLMs. By bridging continuous acoustic signals with discrete linguistic representations, LALMs enable end-to-end modeling of spoken interaction. This capability not only advances fundamental research in speech understanding and generation, but also opens up practical opportunities for building more natural, robust, and versatile human-computer communication systems.

Existing LALMs typically follow two design paradigms. The first couples a speech encoder with an LLM, often through lightweight adapters that project continuous acoustic representations into the input space of the language model (e.g., Qwen2-Audio (Chu et al. 2024), Audio-Flamingo (Kong et al. 2024a)). The second discretizes audio into speech tokens via speech tokenizers and subsequently trains directly on these tokens under the LLM training paradigm (e.g., SpeechGPT (Zhang et al. 2023), Moshi (Défossez et al. 2024)). Despite their strong results, both paradigms predominantly rely on autoregressive (AR) decoding, which suffers from well-known drawbacks such as exposure bias, slow generation, and limited flexibility for bidirectional or partially conditioned inference.

*This work was done during an internship at TeleAI.
†Yong Qin and Xuelong Li are corresponding authors.

To address these limitations, diffusion large language models (dLLMs) [1] (Austin et al. 2021; Shi et al. 2024) have emerged as a promising alternative. By framing generation as an iterative denoising process, dLLMs support non-autoregressive decoding, parallel prediction, and improved controllability(Shi et al. 2024). Recent advances such as LLaDA (Nie et al. 2025) demonstrate that dLLMs can rival autoregressive counterparts like LLaMA-3 (Dubey et al. 2024), while exhibiting stronger robustness and training efficiency. Furthermore, LLaDA-V (You et al. 2025) extends this paradigm to vision–language tasks, confirming the competitiveness and generality of diffusion modeling in multimodal learning.

However, the audio modality remains notably underexplored in the context of dLLMs. While such models have demonstrated promising results in text domains, their applicability to audio-language understanding has not been systematically investigated. The unique characteristics of audio—such as acoustic variability, complex temporal structures, and rich paralinguistic information—motivate an exploration into whether dLLMs can be effectively extended to this domain with their flexible decoding mechanisms and bidirectional context modeling.

To bridge this gap, we explore the potential of adapting large language diffusion models for audio understanding. Specifically, we investigate whether such models can effectively process audio inputs and perform on par with, or surpass, strong autoregressive LALMs. Toward this goal, we introduce **DIFFA**, a **DIFF**usion-based large **A**udio-language framework. DIFFA adopts a modular and efficient design: a pretrained speech encoder (Whisper (Radford et al. 2023)), two lightweight adapters (semantic and acoustic), and a frozen diffusion-based LLM. To avoid catastrophic forgetting, we train only the adapters and keep both the language model and speech encoder frozen. The training procedure is divided into two stages: first, we align the semantic adapter under an ASR objective using LibriSpeech (Panayotov et al. 2015); then, we fine-tune both adapters on synthetic instruction data using the "What can you hear from the audio?" prompting scheme inspired by the DESTA-2 (Lu et al. 2025). Figure 1 visually compares the performance of DIFFA and Qwen2-Audio-Instruct across multiple benchmarks. Notably, DIFFA attains competitive outcomes using merely 960 hours of ASR data and 127 hours of synthetic data, whereas Qwen2-Audio-Instruct depends on a far larger dataset of 510,000 hours.

Our contributions are summarized as follows:

- We propose the first diffusion-based LALM, DIFFA, enabling large-scale audio-text understanding without relying on autoregressive modeling.
- We introduce a dual-adapter training framework and a two-stage training strategy that aligns speech representations to a frozen diffusion LLM, supporting audio understanding and instruction following without explicit supervised fine-tuning data.
- Despite using only 960 hours of ASR data, 127 hours of synthetic instruction data and 72 A800 GPU hours, DIFFA achieves competitive performance across multiple benchmarks, including MMSU, MMAU, and VoiceBench.
- We will release the training pipeline, inference code, and data generation scripts to promote research on diffusion-based LALMs with minimal compute and data requirements.

[1] In this paper, the terms large language diffusion model and diffusion large language model (dLLM) are used interchangeably to describe the same concept, consistent with common usage in prior work.

2 Related Work

Large Audio-Language Models (LALMs). Existing LALMs generally adopt one of two design paradigms. The first integrates a speech encoder with an LLM through lightweight adapters that map continuous acoustic features into the language model's input space (e.g., Qwen2-Audio (Chu et al. 2024), Qwen2.5-Omni (Xu et al. 2025), SALMONN (Tang et al. 2024a), Audio-Flamingo (Kong et al. 2024a)). The second discretizes speech into tokens using quantizers or self-supervised encoders, and feeds these tokens to the LLM as an additional input stream (e.g., SpeechGPT (Zhang et al. 2023), Moshi (Défossez et al. 2024)). While effective, both paradigms are predominantly built upon autoregressive (AR) decoding.

Large Language Diffusion Models. Diffusion-LM (Shi et al. 2024; Sahoo et al. 2024) generate text by iterative denoising. LLaDA (Nie et al. 2025) scales diffusion models to LLMs with likelihood-based training, and LLaDA-V (You et al. 2025) extends it to vision. Large language diffusion models offer bidirectional, parallel decoding—yet remain unexplored in the audio domain.

Supervision-Efficient Modality Alignment. Efforts such as DESTA (Zhang et al. 2023; Lu et al. 2025) and BLSP (Wang et al. 2023) align speech and text using synthetic instructions (e.g., "What can you hear from the audio?") without human annotation. While prior models use cascaded ASR+LLM pipelines, our model performs end-to-end instruction following via dual adapters and a frozen diffusion LLM.

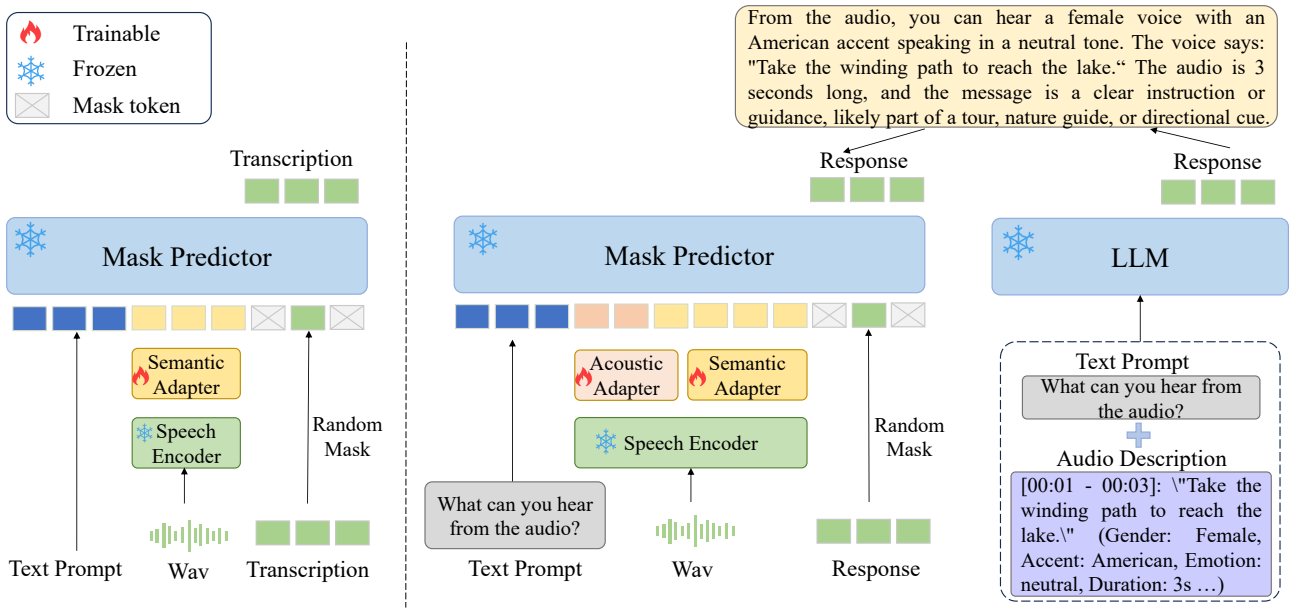
3 Methods

In this section, we present the overall framework of our proposed **DIFFA**, including its formulation, data construction, training strategy, and inference procedure.

3.1 Preliminaries

LLaDA (Nie et al. 2025) is a non-autoregressive language modeling paradigm that introduces a discrete random masking process and learns a *mask predictor* to approximate its reverse. Unlike traditional autoregressive models, LLaDA allows for bidirectional dependency modeling and efficient likelihood-based training.

LLaDA defines a forward masking process to sample a corrupted sequence x_t, where each token is independently replaced with a special mask token M with probability $t \in$



(a) Stage 1 : ASR (b) Stage 2 : Q&A

Figure 2: Training process of our DIFFA framework. Stage 1 performs semantic alignment via an ASR objective, aligning the speech encoder with the diffusion language model. Stage 2 enables modality alignment by prompting the model to describe what it hears from the audio, following an audio caption instruction paradigm.

$(0, 1]$. The mask predictor $p_\theta(x_0|x_t)$ is parameterized by a standard Transformer decoder and trained to reconstruct masked tokens:

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

where L denote the length of target sequence. This objective yields a tractable upper bound of the negative log-likelihood (Shi et al. 2024; Ou et al. 2025), enabling parallel token prediction.

Supervised fine-tuning (SFT) under LLaDA follows a similar approach. Given a prompt p_0 and response r_0, the response tokens are masked independently to obtain r_t. The loss is computed as:

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

where L' is the response length.

During inference, LLaDA decodes iteratively from a fully masked sequence. At each denoising step, the model predicts masked tokens and re-applies masks to low-confidence positions, gradually refining predictions over T steps.

3.2 Data Construction

Inspired by the DESTA series, we construct a dataset by prompting LLaDA-based or instruction-tuned language models (e.g., LLaMA3, Qwen3) with audio transcriptions: "*[00:01 - 00:03]: "Take the winding path to reach the lake."* *(Gender: Female, Accent: American, Emotion: neutral, Duration: 3s . . .)* " and acoustic attributes using the prompt: *"What can you hear from the audio?"*. The generated response serves as the supervision signal, paired with the corresponding audio. This enables extbfmodality alignment without any explicit supervised fine-tuning data.

Furthermore, motivated by self-distillation techniques (Yang et al. 2024), we introduce a rewriting step to mitigate the domain shift arising from different model pre-training distributions. Specifically, we first employ Qwen3-8B to generate an initial set of captions. Subsequently, our LLaDA model rewrites these captions to align the textural style with its own internal data distribution. We denote this variant as LLaDA-rewrite-Qwen3.

3.3 Model Architecture and Training Strategy

Let (a_0, p_0, r_0) denote the audio input, textual prompt, and target response, respectively. We employ a frozen Whisper-small encoder to extract frame-level acoustic features from a_0, and integrate them into the LLaDA-8B-Instruct backbone via two lightweight adapters.:

Semantic Adapter. A 2-layer convolutional network with a subsampling rate of 4, followed by a 2-layer linear projection. It compresses the encoder's 50 Hz output to 12.5 Hz.

Acoustic Adapter. A 2-layer Q-former (Li et al. 2023) blocks with 64 trainable query vectors. It extracts speech-specific features from intermediate encoder states.

Two-Stage Training. The whole training process is shown in Figure 2. In stage 1, the semantic adapter is trained on 960 hours Librispeech using an ASR-style objective to align the speech encoder with the language model. In stage

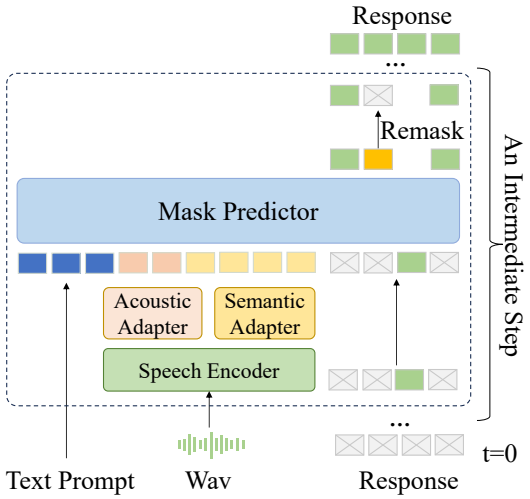


Figure 3: Inference procedure of DIFFA.

2, both adapters are fine-tuned on our 127-hour synthetic dataset under the audio captioning objective. The final audio representation is the concatenation of outputs from both adapters and prepended as prefix tokens to the LLM input. In both stages, audio and prompt tokens remain unmasked during training. We use <endoftext> as a padding and end-of-sequence token during training, which must also be predicted. The LLaDA model and Whisper encoder remain frozen throughout. At each training step, the tokens of r_0 are independently replaced with a special mask token M with probability $t \in (0, 1]$. And then a forward masking process to sample a corrupted sequence r_t. We optimize the model using a diffusion-style masked prediction objective:

$$L_a = -\mathbb{E}_{t,a_0,p_0,r_0,r_t} \left[\frac{1}{t} \sum_{i=1}^{L'} \mathbf{1}[r_t^i = \mathrm{M}] \log p_\theta(r_0^i \mid a_0, p_0, r_t) \right], \quad (3)$$

where r_t is the masked response and L' is its length.

3.4 Inference Procedure

At inference time, we first pad the prompt and audio input, then initialize the response r_T as a fully masked sequence of desired length. The model iteratively refines r_t over T denoising steps.

At step $t \to s$, the model predicts masked tokens:

$$\hat{r}_t = \arg\max p_\theta(r_0 | a_0, p_0, r_t), \quad (4)$$

then re-masks $\lceil s/t \rceil$ proportion of tokens with the lowest confidence to form r_s.

We follow a semi-autoregressive strategy (Nie et al. 2025), generating the sequence block-wise from left to right. Within each block, tokens are predicted in parallel and partially remasked.

This iterative inference scheme balances generation quality and efficiency, while maintaining the benefits of parallel decoding and bidirectional context modeling inherent in diffusion-based LLMs.

Dataset	Samples	Total Dur. (h)
VCTK-Corpus	20,000	19.91
Accentdb	16,874	19.28
IEMOCAP	20,000	24.82
dailytalk	20,000	18.17
VoxCeleb1	20,000	45.83
Total	96,874	127.01

Table 1: Statistics of Datasets

4 Experimental Setup

4.1 Datasets

In our experiments, we employ only open-source datasets Librispeech for ASR task in stage 1 and five dataset to construct dataset for Q&A in stage 2: VCTK-Corpus (Yamagishi, Veaux, and MacDonald 2019), Accentdb (Ahamad, Anand, and Bhargava 2020), IEMOCAP (Busso et al. 2008), dailytalk (Lee, Park, and Kim 2023), VoxCeleb (Nagrani, Chung, and Zisserman 2017). The details of datasets are presented in Table 1.

Compared to DESTA-2, our dataset follows a similar construction paradigm but excludes the *PromptTTS* and *Mixed Noise & Reverb* subsets due to lack of access. Our dataset includes 10 annotated attributes—*gender, age, accent, emotion, pitch, volume, speaking speed, duration, intent, and spoken text*—which is slightly fewer than the 12 attributes used in DESTA-2.

4.2 Model Configuration and Training Setup

For the speech encoder, we adopt the Whisper-Small encoder, which contains 88.2 million parameters. As the language backbone, we use LLaDA-8B-Instruct, a large language diffusion model trained with a masked denoising objective inspired by diffusion-based frameworks. It is built upon a Transformer decoder architecture with 32 layers, 32 attention heads, a hidden size of 4096, and approximately 8.1 billion parameters. The architecture follows LLaMA (Touvron et al. 2023; Dubey et al. 2024), with key modifications including RMSNorm (Zhang and Sennrich 2019) for normalization, SwiGLU (Shazeer 2020) for nonlinearity, and rotary position embeddings (RoPE) (Su et al. 2024) for positional encoding.

In our experiments, all parameters of LLaDA-8B-Instruct are frozen. We introduce lightweight trainable adapters to integrate audio features, following prior work on parameter-efficient multimodal learning. The semantic adapter contains 14.4 million parameters and the acoustic adapter 22.3 million. In Stage 1, we train the semantic adapter using the LibriSpeech dataset for 10 epochs. We adopt a learning rate of $1e^{-4}$ with 1000 warm-up steps and a global batch size of 128. In Stage 2, both the semantic and acoustic adapters are jointly trained on our generated dataset for 10 epochs. We use a learning rate of $5e^{-5}$ with 2000 warm-up steps and a global batch size of 64. All experiments are optimized using the Adam optimizer and conducted on 4 NVIDIA A800 GPUs with 80GB memory each.

Models	Perception				Reasoning				Overall
	Semantics	Phonology	Paralinguistics	Avg	Semantics	Phonology	Paralinguistics	Avg	Avg
Human	87.10	94.32	92.88	91.24	82.16	87.60	89.12	86.77	89.72
Gemini-1.5-Pro (Team et al. 2024)	57.06	**53.60**	31.23	**46.10**	79.47	**83.46**	46.33	76.16	**60.68**
Qwen2.5-Omni (Xu et al. 2025)	55.12	37.33	**39.35**	42.50	**88.00**	81.37	48.36	**79.83**	60.57
Kimi-Audio (Ding et al. 2025)	57.64	42.30	35.74	43.52	81.77	76.65	**55.22**	76.03	59.28
MiniCPM (Team 2025)	56.56	34.05	36.48	40.54	80.71	74.72	46.71	73.57	56.53
GPT-4o-Audio (OpenAI et al. 2024)	**59.70**	41.56	21.44	39.67	80.83	78.74	26.25	71.96	56.38
MERaLiON (He et al. 2024)	54.49	33.69	25.84	35.74	80.32	77.18	41.49	73.68	54.10
Qwen2-Audio-Instruct (Chu et al. 2024)	52.14	32.87	35.56	39.02	77.62	64.81	46.67	68.90	53.27
Gemini-2.0-Flash	47.17	41.30	30.62	40.83	70.69	70.69	36.16	47.83	51.03
Megrez-3B-Omni (Li et al. 2025)	41.36	32.52	26.35	32.48	73.53	66.11	40.42	67.05	49.03
DIVA(Held et al. 2024)	44.36	33.72	27.45	33.95	62.32	74.24	40.00	65.04	48.31
Qwen-Audio-Chat (Chu et al. 2023)	57.21	38.52	24.70	35.69	58.61	59.78	25.60	55.93	46.92
Step-Audio (Huang et al. 2025)	31.56	29.39	24.01	28.72	49.10	50.09	45.27	47.27	37.42
BLSP (Wang et al. 2023)	31.35	20.96	23.75	28.36	47.91	42.31	42.08	44.97	35.96
GLM-4-Voice (Zeng et al. 2024)	27.80	24.52	27.34	26.18	46.10	48.16	44.35	46.76	35.51
Random	24.30	25.70	26.10	24.90	23.80	25.40	25.40	25.02	25.37
DIFFA	52.67	36.65	35.12	40.28	81.53	72.68	45.67	72.92	56.04

Table 2: Performance breakdown on the MMSU benchmark across perception and reasoning dimensions.

Model	Sound	Music	Speech	Average
Gemini 2.5 Pro (Comanici et al. 2025)	75.08	68.26	**71.47**	**71.60**
Qwen2.5-Omni (Xu et al. 2025)	**78.10**	65.90	70.60	71.53
Phi-4-multimodal (Abouelenin et al. 2025)	65.47	64.37	67.27	65.70
Audio Flamingo 2 Reasoning (Ghosh et al. 2025)	75.98	**74.25**	43.54	64.59
GPT-4o Audio (OpenAI et al. 2024)	64.56	56.29	66.67	62.51
Audio Flamingo 2 (Ghosh et al. 2025)	71.47	70.96	44.74	62.39
Qwen2-Audio-Instruct (Chu et al. 2024)	67.27	56.29	55.26	59.61
GPT-4o mini Audio (OpenAI et al. 2024)	50.75	39.22	69.07	53.01
Gemini Pro v1.5 (Team et al. 2024)	56.75	49.40	58.55	52.97
M2UGen (Liu et al. 2023)	43.24	37.13	33.33	37.90
MusiLingo (Deng et al. 2024)	43.24	40.12	31.23	38.20
SALMONN (Tang et al. 2024b)	41.14	37.13	26.43	34.90
MuLLaMa (Liu et al. 2024)	33.03	32.34	17.42	27.60
GAMA-IT (Ghosh et al. 2024)	30.93	26.74	10.81	22.83
GAMA (Ghosh et al. 2024)	31.83	17.71	12.91	20.82
LTU (Gong et al. 2024)	20.42	15.97	15.92	17.44
Audio Flamingo Chat (Kong et al. 2024b)	25.23	17.66	6.91	16.60
DIFFA	46.25	43.41	59.46	49.71

Table 3: Evaluation results on the MMAU benchmark. Each model is assessed across three core audio domains: sound, music, and speech.

4.3 Benchmarks

MMSU (Wang et al. 2025) is a large-scale benchmark aimed at evaluating the perception and reasoning capabilities of SpeechLLMs in authentic spoken language scenarios. It consists of 5,000 carefully curated audio-question-answer triplets across 47 diverse tasks, covering a wide spectrum of linguistic and paralinguistic phenomena—including phonetics, prosody, semantics, emotion, and speaker traits.

MMAU (Sakshi et al. 2025) is a benchmark designed to evaluate advanced audio understanding through human-annotated multiple-choice questions paired with audio clips. It covers three core domains—speech, music, and environmental sounds—and targets 27 distinct skills that require complex reasoning and expert-level knowledge. In our experiments, we use the Test-mini split of MMAU for evaluation.

VoiceBench (Chen et al. 2024b) is a comprehensive benchmark designed to evaluate the capabilities of LLM-based voice assistants. It primarily consists of audio queries synthesized via text-to-speech (TTS) from existing text-based benchmarks, simulating realistic user interactions in spoken form.

Model	AlpacaEval	CommonEval	SD-QA	MMSU*	OBQA	IFEval	AdvBench	Overall
GPT-4o-Audio (OpenAI et al. 2024)	**4.78**	**4.49**	**75.50**	**80.25**	**89.23**	**76.02**	98.65	**86.43**
Kimi-Audio (Ding et al. 2025)	4.46	3.97	63.12	62.17	83.52	61.10	**100.00**	76.93
Baichuan-Omni-1.5 (Li et al. 2024)	4.50	4.05	43.40	57.25	74.51	54.54	97.31	71.14
GLM-4-Voice (Zeng et al. 2024)	3.97	3.42	36.98	39.75	53.41	25.92	88.08	55.99
DiVA (Held et al. 2024)	3.67	3.54	57.06	25.76	25.49	39.16	98.27	55.70
Qwen2-Audio (Chu et al. 2024)	3.74	3.43	35.72	35.72	49.45	26.33	96.73	55.34
Step-Audio (Huang et al. 2025)	4.13	3.09	44.21	28.33	33.85	27.96	69.62	49.77
LLaMA-Omni (Fang et al. 2025)	3.70	3.46	39.69	25.93	27.47	14.87	11.35	37.50
VITA (Fu et al. 2024)	3.38	2.15	27.94	25.70	29.01	22.82	26.73	34.68
Slam-Omni (Chen et al. 2024a)	1.90	1.79	4.16	26.06	25.27	13.38	94.23	33.84
Mini-Omni2 (Xie and Wu 2024b)	2.32	2.18	9.31	24.27	26.59	11.56	57.50	31.32
Mini-Omni (Xie and Wu 2024a)	1.95	2.02	13.92	24.69	26.59	13.58	37.12	27.90
Moshi (Défossez et al. 2024)	2.01	1.60	15.64	24.04	25.93	10.12	44.23	27.45
DIFFA	3.78	2.96	34.45	29.57	35.60	26.56	76.54	48.22

Table 4: Evaluation results on VoiceBench. Metrics cover diverse QA and alignment tasks. Note that MMSU* in VoiceBench is derived from MMLU-Pro, which differs from the MMSU benchmark.

5 Experiments

5.1 Evaluation on MMSU

To assess the advanced reasoning abilities of our diffusion-based model, we evaluate DIFFA on the **MMSU** benchmark, which assesses fine-grained spoken language understanding across perception (semantics, phonology, paralinguistics) and reasoning dimensions. As shown in Table 2, our model achieves an average accuracy of 56.04%, highlighting its ability to handle a wide range of linguistically grounded tasks.

Although the overall performance trails top proprietary models like Gemini-1.5-Pro (60.68%), DIFFA outperforms many strong autoregressive baselines, such as Qwen2-Audio-Instruct (53.27%) and Gemini-2.0-Flash (51.03%). This competitive result, despite relying solely on synthetic supervision and lightweight adapters, supports the viability of diffusion-based approaches for nuanced speech understanding.

DIFFA performs particularly well on semantic reasoning tasks (81.53%), benefiting from its strong language modeling backbone and ASR-aligned speech encoder. However, like most models in the benchmark, it exhibits lower accuracy in phonological and paralinguistic tasks—areas that demand precise acoustic perception beyond textual semantics. These trends mirror the broader challenges outlined in the MMSU benchmark, where human-level performance (89.72%) remains a distant target.

This result provides the first empirical evidence that large language diffusion models can serve as viable backbones for large-scale audio-language understanding, even without autoregressive decoding. Despite using only 127 hours of synthetic training data—orders of magnitude less than the tens of thousands of supervised fine-tuning (SFT) hours used by many baselines—DIFFA demonstrates strong generalization and reasoning capabilities.

Overall, DIFFA demonstrates robust generalization across complex linguistic dimensions, establishing a strong diffusion-based baseline. Future work should focus on enhancing the model's sensitivity to prosodic and phonological cues to bridge the gap with human-level performance.

5.2 Evaluation on MMAU

We further evaluate DIFFA on the MMAU benchmark, which tests 27 skills across three audio domains: sound, music, and speech. As shown in Table 3, DIFFA achieves an average accuracy of 49.71%, outperforming several widely used autoregressive LALMs, such as SALMONN (34.90%), GAMA-IT (22.83%), and LTU (17.44%). It also approaches the performance of commercial models like GPT-4o mini Audio (53.01%) and Gemini Pro v1.5 (52.97%).

A domain-level breakdown shows that DIFFA achieves the highest performance on speech-related tasks (59.46%), likely benefiting from its speech-caption-focused training paradigm. In contrast, models such as Audio Flamingo 2 perform better on music and environmental sounds but underperform in speech understanding (e.g., 44.74% for speech), underscoring the advantage of targeted, speech-centric training.

Overall, these results position DIFFA as a promising diffusion-based alternative to autoregressive LALMs. With limited resources and a parameter-efficient adapter tuning scheme, it achieves competitive results across complex audio reasoning benchmarks, indicating strong potential for scaling with larger or higher-quality datasets.

5.3 Evaluation on VoiceBench

VoiceBench evaluates semantic understanding from audio-as-question prompts, covering knowledge, instructions, and safety—making it a rigorous benchmark for spoken query comprehension.

Despite being trained on only 960 hours of ASR data and 127 hours of synthetic instructions, DIFFA achieves 34.45% on SD-QA and 35.60% on OBQA, demonstrating promising capability in factual spoken QA. This is particularly notable when compared to models like Qwen2-Audio (35.72% SD-QA, 49.45% OBQA), which are trained on hundreds of

LLM Backbone	Data Source	Adapter	MMAU				MMSU		
			Sound	Music	Speech	Avg	Perception	Reasoning	Avg
LLaMA 3.1	LLaMA 3	Dual	22.82	26.65	35.74	28.40	32.36	44.83	38.40
LLaDA	LLaMA 3	Dual	**47.75**	**45.51**	**61.86**	**51.71**	**37.27**	**73.20**	**54.72**
LLaDA	Qwen3	Single	44.44	**44.31**	54.35	47.70	36.98	69.83	52.88
LLaDA	Qwen3	Dual	**46.25**	43.41	**59.46**	**49.71**	**40.28**	**72.92**	**56.04**

Table 5: Ablation study on model architecture and adapter design. All models use 8B Instruct version.

Data Source	MMAU	MMSU	Voicebench	Avg
LLaMA 3	**51.71**	54.72	37.17	47.86
LLaDA	51.31	56.18	43.52	50.34
Qwen3	49.71	56.04	**48.22**	**51.32**
rewrite-Qwen3	50.41	**56.43**	46.60	51.15

Table 6: Ablation study on the impact of different instruction data sources. All models use 8B Instruct version.

thousands of hours of proprietary data. The results suggest that diffusion-based models, even with limited training, can capture core semantic structures in speech.

On IFEval, DIFFA reaches 26.56%, slightly surpassing Qwen2-Audio (26.33%) and GLM-4-Voice (25.92%), indicating a basic capacity for instruction comprehension from audio inputs. However, the performance gap with top models such as Kimi-Audio (61.10%) underscores the challenge of aligning audio-conditioned instruction execution without large-scale supervised tuning.

In AdvBench, DIFFA attains 76.54%, outperforming many strong baselines and approaching GLM-4-Voice. This highlights the potential of lightweight, diffusion-based models to learn safety-aligned behavior with minimal data.

In summary, DIFFA establishes a competitive baseline on VoiceBench despite using orders of magnitude less training data than competing systems. These findings validate the feasibility of diffusion-based LLMs for semantic audio understanding and offer a data-efficient alternative to current autoregressive paradigms.

5.4 Ablation Study

We perform a comprehensive ablation study to assess the impact of three key factors on model performance: (1) the choice of language model backbone, (2) adapter design, and (3) instruction data source. Besides, effect of inference hyperparameters are provided in Appendix.

Impact of Diffusion-Based Language Modeling. As shown in Table 5, replacing the autoregressive LLaMA 3.1 backbone with the diffusion-based LLaDA architecture leads to a substantial improvement across all metrics. Specifically, with dual adapters, the LLaDA variant achieves 51.71 on MMAU and 54.72 on MMSU, significantly outperforming its LLaMA counterpart. This highlights the advantage of diffusion language model for audio understanding tasks.

Effect of Adapter Design. We compare single and dual adapter configurations using the same LLaDA backbone. The single adapter setup uses only a semantic adapter aligned with the speech encoder's output. The dual adapter adds an acoustic adapter that extracts low-level acoustic cues from intermediate encoder states. Results show that dual adapters yield a +2.01 gain on MMAU and +3.16 on MMSU, confirming the advantage of combining semantic and acoustic information for richer audio understanding.

Instruction Data Source. Table 6 examines the role of different instruction data sources on final performance. All variants show comparable results, with Qwen3-generated instruction data leading to the best overall accuracy. Rewriting Qwen3 data with LLaDA brings only marginal improvements, suggesting that enhancing data quality at generation time may be more effective than post-hoc refinement. Interestingly, models trained on LLaDA-generated data outperform those based on LLaMA-3 instructions, indicating that the inductive bias of diffusion-based generation may yield more aligned supervision.

6 Conclusion

In this work, we introduce **DIFFA**, a diffusion-based large audio-language model that combines a frozen diffusion language backbone with lightweight dual adapters for audio understanding and instruction following. Unlike prior LALMs, DIFFA replaces the autoregressive LLM backbone with a diffusion LLM — a fundamentally different modeling paradigm which we extend to the audio domain for the first time. Our dual-adapter design explicitly disentangles semantic and acoustic signals, which is crucial for stable diffusion decoding. Importantly, DIFFA achieves performance comparable to Qwen2-Audio while using less than 0.22% of its data (960h + 127h vs. 510k h), and it surpasses an LLaMA3.1-based baseline under identical configurations, demonstrating both conceptual novelty and empirical efficiency. As an initial exploration of diffusion-based modeling in the audio domain, our results suggest that such models offer a promising alternative to autoregressive LALMs. While our experiments are conducted on relatively small-scale open-source corpora, we plan to scale to broader and more diverse data in future work. We hope this study encourages further research into efficient, flexible, and controllable speech-driven AI systems.

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

Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Ahamad, A.; Anand, A.; and Bhargava, P. 2020. AccentDB: A Database of Non-Native English Accents to Assist Neural Speech Recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 5351–5358. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Austin, J.; Johnson, D.; Ho, J.; Tarlow, D.; and Berg, R. 2021. Structured Denoising Diffusion Models in Discrete State-Spaces. *arXiv: Learning,arXiv: Learning*.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 335–359.

Chen, W.; Ma, Z.; Yan, R.; Liang, Y.; and et al. 2024a. Slam-omni: Timbre-controllable voice interaction system with single-stage training. *arXiv preprint arXiv:2412.15649*.

Chen, Y.; Yue, X.; Zhang, C.; Gao, X.; Tan, R. T.; and Li, H. 2024b. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; Zhou, C.; and Zhou, J. 2024. Qwen2-Audio Technical Report. *arXiv preprint arXiv:2407.10759*.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-Audio: Advancing Universal Audio Understanding via Unified Large-Scale Audio-Language Models. *arXiv preprint arXiv:2311.07919*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Défossez, A.; Mazaré, L.; Orsini, M.; Royer, A.; Pérez, P.; Jégou, H.; Grave, E.; and Zeghidour, N. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Deng, Z.; Ma, Y.; Liu, Y.; Guo, R.; Zhang, G.; Chen, W.; Huang, W.; and Benetos, E. 2024. MusiLingo: Bridging Music and Text with Pre-trained Language Models for Music Captioning and Query Response. In *NAACL-HLT (Findings)*, 3643–3655.

Ding, D.; Ju, Z.; Leng, Y.; Liu, S.; Liu, T.; Shang, Z.; Shen, K.; Song, W.; Tan, X.; Tang, H.; et al. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Fang, Q.; Guo, S.; Zhou, Y.; Ma, Z.; Zhang, S.; and Feng, Y. 2025. LLaMA-Omni: Seamless Speech Interaction with Large Language Models. In *The Thirteenth International Conference on Learning Representations*.

Fu, C.; Lin, H.; Long, Z.; Shen, Y.; Dai, Y.; Zhao, M.; Zhang, Y.-F.; Dong, S.; Li, Y.; Wang, X.; et al. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Ghosh, S.; Kong, Z.; Kumar, S.; Sakshi, S.; Kim, J.; Ping, W.; Valle, R.; Manocha, D.; and Catanzaro, B. 2025. Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities. In *Forty-second International Conference on Machine Learning*.

Ghosh, S.; Kumar, S.; Seth, A.; Evuru, C. K. R.; Tyagi, U.; Sakshi, S.; Nieto, O.; Duraiswami, R.; and Manocha, D. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. In *EMNLP*, 6288–6313.

Gong, Y.; Luo, H.; Liu, A. H.; Karlinsky, L.; and Glass, J. R. 2024. Listen, Think, and Understand. In *The Twelfth International Conference on Learning Representations*.

He, Y.; Liu, Z.; Sun, S.; Wang, B.; Zhang, W.; Zou, X.; Chen, N. F.; and Aw, A. T. 2024. Meralion-audiollm: Bridging audio and language with large language models. *arXiv preprint arXiv:2412.09818*.

Held, W.; Li, E.; Ryan, M.; Shi, W.; Zhang, Y.; and Yang, D. 2024. Distilling an end-to-end voice assistant without instruction training data. *arXiv preprint arXiv:2410.02678*.

Huang, A.; Wu, B.; Wang, B.; Yan, C.; Hu, C.; Feng, C.; Tian, F.; Shen, F.; Li, J.; Chen, M.; et al. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Kong, Z.; Goel, A.; Badlani, R.; Ping, W.; Valle, R.; and Catanzaro, B. 2024a. Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities. *arXiv preprint arXiv:2402.01831*.

Kong, Z.; Goel, A.; Badlani, R.; Ping, W.; Valle, R.; and Catanzaro, B. 2024b. Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities. In *International Conference on Machine Learning*, 25125–25148. PMLR.

Lee, K.; Park, K.; and Kim, D. 2023. DailyTalk: Spoken Dialogue Dataset for Conversational Text-to-Speech. In *ICASSP 2023*, 1–5.

Li, B.; Li, Y.; Li, Z.; Liu, C.; Liu, W.; Niu, G.; Tan, Z.; Xu, H.; Yao, Z.; Yuan, T.; et al. 2025. Megrez-omni technical report. *arXiv preprint arXiv:2502.15803*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Y.; Sun, H.; Lin, M.; Li, T.; Dong, G.; Zhang, T.; Ding, B.; Song, W.; Cheng, Z.; Huo, Y.; et al. 2024. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*.

Liu, S.; Hussain, A. S.; Sun, C.; and Shan, Y. 2023. M2UGen: Multi-modal Music Understanding and Generation with the Power of Large Language Models. *arXiv preprint arXiv:2311.11255*.

Liu, S.; Hussain, A. S.; Sun, C.; and Shan, Y. 2024. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024*, 286–290. IEEE.

Lu, K.-H.; Chen, Z.; Fu, S.-W.; Yang, C.-H. H.; Balam, J.; Ginsburg, B.; Wang, Y.-C. F.; and Lee, H.-y. 2025. Developing instruction-following speech language model without speech instruction-tuning data. In *ICASSP 2025*, 1–5. IEEE.

Nagrani, A.; Chung, J. S.; and Zisserman, A. 2017. VoxCeleb: a large-scale speaker identification dataset. In *Interspeech 2017*.

Nie, S.; Zhu, F.; You, Z.; Zhang, X.; Ou, J.; Hu, J.; Zhou, J.; Lin, Y.; Wen, J.-R.; and Li, C. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

OpenAI; :; Hurst, A.; Lerer, A.; Goucher, A. P.; and et al., A. P. 2024. GPT-4o System Card. arXiv:2410.21276.

Ou, J.; Nie, S.; Xue, K.; Zhu, F.; Sun, J.; Li, Z.; and Li, C. 2025. Your Absorbing Discrete Diffusion Secretly Models the Conditional Distributions of Clean Data. In *The Thirteenth International Conference on Learning Representations*.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *ICASSP 2015*, 5206–5210. IEEE.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Sahoo, S. S.; Arriola, M.; Schiff, Y.; Gokaslan, A.; Marroquin, E.; Chiu, J. T.; Rush, A.; and Kuleshov, V. 2024. Simple and Effective Masked Diffusion Language Models. *ArXiv*, abs/2406.07524.

Sakshi, S.; Tyagi, U.; Kumar, S.; Seth, A.; Selvakumar, R.; Nieto, O.; Duraiswami, R.; Ghosh, S.; and Manocha, D. 2025. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. In *The Thirteenth International Conference on Learning Representations*.

Shazeer, N. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Shi, J.; Han, K.; Wang, Z.; Doucet, A.; and Titsias, M. 2024. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; MA, Z.; and Zhang, C. 2024a. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; and Zhang, C. 2024b. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *ICLR*.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

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

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wang, C.; Liao, M.; Huang, Z.; Lu, J.; Wu, J.; Liu, Y.; Zong, C.; and Zhang, J. 2023. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*.

Wang, D.; Wu, J.; Li, J.; Yang, D.; Chen, X.; Zhang, T.; and Meng, H. 2025. MMSU: A Massive Multi-task Spoken Language Understanding and Reasoning Benchmark. *arXiv preprint arXiv:2506.04779*.

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

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

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yamagishi, J.; Veaux, C.; and MacDonald, K. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92).

Yang, Z.; Pang, T.; Feng, H.; Wang, H.; Chen, W.; Zhu, M.; and Liu, Q. 2024. Self-Distillation Bridges Distribution Gap in Language Model Fine-Tuning. In *ACL 2024*, 1028–1043. Bangkok, Thailand.

You, Z.; Nie, S.; Zhang, X.; Hu, J.; Zhou, J.; Lu, Z.; Wen, J.-R.; and Li, C. 2025. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*.

Zeng, A.; Du, Z.; Liu, M.; Wang, K.; Jiang, S.; Zhao, L.; Dong, Y.; and Tang, J. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Zhang, B.; and Sennrich, R. 2019. Root mean square layer normalization. *Advances in neural information processing systems*, 32.

Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.