NEXUS: AN OMNI-PERCEPTIVE AND -INTERACTIVE MODEL FOR LANGUAGE, AUDIO, AND VISION

Che Liu[1,2*], Yingji Zhang[3*], Dong Zhang[1*], Weijie Zhang[1*], Chenggong Gong[1*], Haohan Li[1*],

Yu Lu[1], Shilin Zhou[1,6], Yue Lu[1], Ziliang Gan[1], Ziao Wang[7], Junwei Liao[8], Haipang Wu[1], Ji Liu[1],

André Freitas[3,10], Qifan Wang[9], Zenglin Xu[5], Rongjuncheng Zhang[1,4,♠], Yong Dai[1,5,♠,†]

[1]HiThink Research, China
[2]Imperial College London, UK
[3]University of Manchester, UK
[4]Zhejiang University, China
[5]Fudan University, China
[6]Soochow University, China
[7]Baptist University, HK
[8]Microsoft, USA
[9]Meta AI, USA
[10]Idiap Research Institute, Switzerland

ABSTRACT

Human beings perceive the real world through a spectrum of sensory modalities, encompassing auditory, visual, and linguistic faculties. This work proposes an industry-level omni-modal large language model (LLM) pipeline that integrates auditory, visual, and linguistic modalities to overcome challenges such as limited tri-modal datasets, high computational costs, and complex feature alignments. Our pipeline consists of three main components: First, a modular, end-to-end framework enabling flexible configuration of various encoder–LLM–decoder architectures. Second, a lightweight training strategy that pre-trains audio-language alignment on the state-of-the-art vision-language model Qwen2.5-VL, thus avoiding the costly pre-training of vision-specific modalities. Third, an audio synthesis pipeline that generates high-quality audio-text data from diverse real-world scenarios, supporting applications such as Automatic Speech Recognition and Speech-to-Speech chat. To this end, we introduce an industry-level omni-modal LLM, **Nexus**. Extensive experiments validate the efficacy of our pipeline, yielding the following key findings: (1) In the visual understanding task, Nexus exhibits superior performance compared with its backbone model - Qwen2.5-VL-7B, validating the efficiency of our training strategy. (2) Within the English Spoken Question-Answering task, the model achieves better accuracy than the same-period competitor (i.e, MiniCPM-o2.6-7B) in the LLaMA Q. benchmark. (3) In our real-world ASR testset, Nexus achieves outstanding performance, indicating its robustness in real scenarios. (4) In the Speech-to-Text Translation task, our model outperforms Qwen2-Audio-Instruct-7B. (5) In the Text-to-Speech task, based on pretrained vocoder (e.g., Fishspeech1.4 or CosyVoice2.0), Nexus is comparable to its backbone vocoder on Seed-TTS benchmark. (6) An in-depth analysis of tri-modal alignment reveals that incorporating the audio modality enhances representational alignment between vision and language.

[*] denotes equal contribution. This paper was completed during Che Liu and Shilin Zhou's internships at Hithink Research.

♠ denotes Project Leader. † denotes corresponding author, daiyongya@outlook.com.

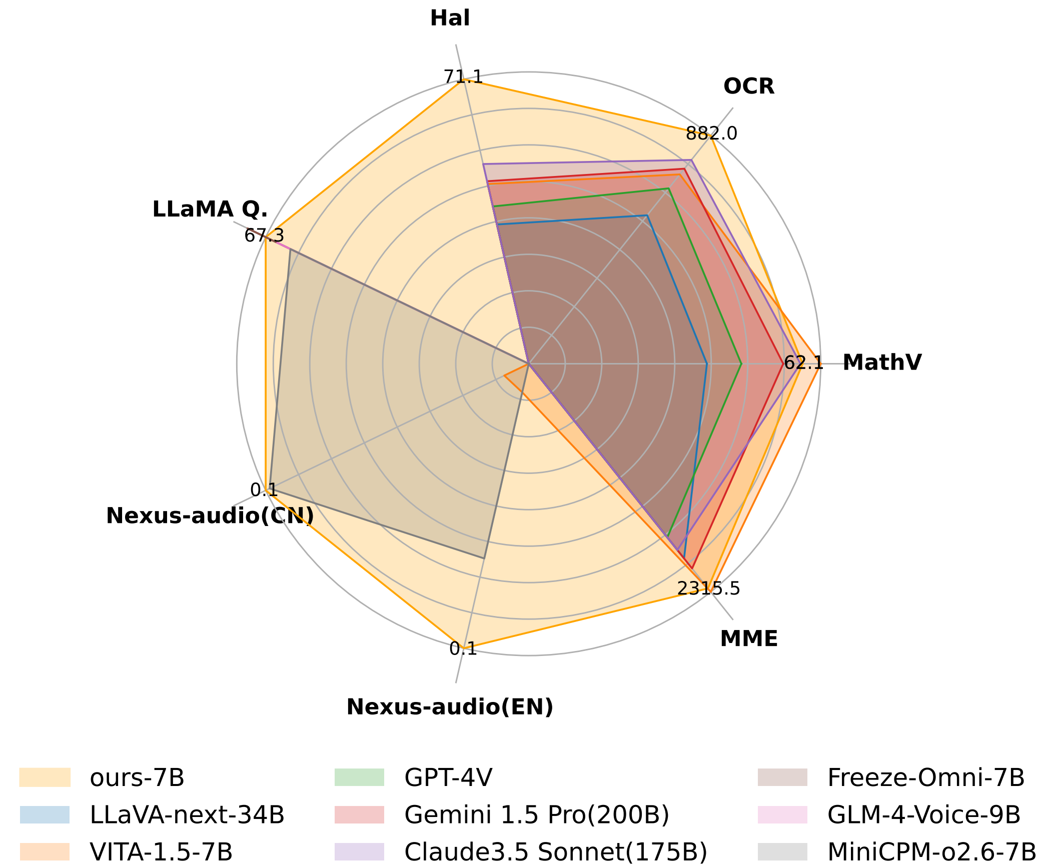

Figure 1: An overview of the model's performance. Nexus is capable of handling a wide range of tasks, demonstrating strong performance across various benchmarks. For Nexus-audio (EN) and (CH), the scores are expressed as reciprocal (1/scores).

1 Introduction

Since the mid-20th century, significant efforts have been devoted to designing systems of the capability to simulate human cognition, with the ultimate goal of developing Artificial General Intelligence (AGI). Notable algorithms include Rule-based systems, Neuro-symbolic AI systems, Neural networks systems (Hinton, 2006), etc. Among them, neural networks systems have emerged as a cornerstone thanks to their remarkable generalization capabilities. In this context, various neural architectures have been proposed, such as CNN (LeCun et al., 1995) and Transformer (Vaswani, 2017). The discovery of neural scaling laws (Kaplan et al., 2020) has further advanced the field, paving the way for Transformer-based Large Language Models (LLMs), such as ChatGPT, which centre on language as the medium for perceiving and interacting with both human society and the physical world. By aligning multiple modalities around the language, Multimodal Large Language Models (MLLMs) have gained widespread recognition as a promising pathway toward AGI.

MLLMs, involving language, vision, and auditory modalities, can generally be divided into three categories: vision-language model (Wang et al., 2024), audio-language model (Chu et al., 2024b), and vision-language-audio model (Fu et al., 2025). While alignment among three modalities offers a more comprehensive simulation of human cognitive processes than two-modality alignment, it presents significantly greater challenges. These challenges include limited accessibility to tri-modal datasets, higher computational resource requirements, and the complexity of aligning features across three modalities. These limitations hinder the advancement of omni-perceptive and interactive AGI systems. To address these challenges, we propose *a comprehensive, industry-level omni-modal LLM* *pipeline that comprises a modularised end-to-end framework, a lightweight training strategy, and an audio data synthesis pipeline*:

First, in the modularised framework, each modality (audio, image/video, and text) is processed using its own dedicated pretrained encoder. This design ensures that the inherent characteristics and fine-grained details unique to each modality are effectively preserved within their respective latent representations. Subsequently, these modality-specific representations are integrated and aligned within a LLM, and the resulting unified representations are subsequently supplied to decoupled decoders designed for each modality. This modular formalism design facilitates efficient and flexible configuration of various encoder–LLM–decoder architectures, thereby potentially contributing to the advancement of AGI by integrating components such as language models, world models, and embodied intelligence.

Second, most omni-modal models, such as Qwen2.5-omni (Xu et al., 2025), VITA1.5 (Fu et al., 2024c), etc., necessitate a preliminary pre-training stage for vision alignment, understanding, and reasoning. This stage typically demands substantial computational resources due to the processing of long sequences of image tokens, which is consequently less accessible to smaller research institutions and companies. To mitigate this limitation, we only pre-train audio-language alignment over the current state-of-the-art vision-language model, Qwen2.5-VL (Bai et al., 2025)[1]. Due to the inherent alignment between audio and text modalities, this training strategy effectively integrates audio understanding and generation capabilities while preserving the model's established vision comprehension and reasoning functions.

Third, high-quality audio training data are often less accessible in both academy and industry. To address this limitation, we propose an audio synthesis pipeline that generates high-quality audio-text data spanning a broad spectrum of real-world scenarios, such as live streaming and dialogue, thereby supporting various downstream tasks such as Automatic Speech Recognition (ASR) and Speech-to-Speech chat. By incorporating real-world audio data at the pretraining stage, the model can be fast deployed for real-world applications. Based on our audio synthesis pipeline, we introduce a real-life ASR testset, named as Nexus-audio.

To this end, we introduce **Nexus**[2], an industry-level omni-modal LLM. Extensive experiments validate the efficacy of our pipeline, yielding the following key findings: (1) In the visual understanding task, Nexus exhibits superior performance compared with its backbone model - Qwen2.5-VL-7B, validating the efficiency of our training strategy. (2) Within the English Spoken Question-Answering task, the model achieves better accuracy than the same-period competitor (i.e, MiniCPM-o2.6-7B) in the LLaMA Q. benchmark (Nachmani et al., 2024). (3) In our real-world ASR testset, Nexus achieves outstanding performance, indicating its robustness in real scenarios. (4) In the Speech-to-Text Translation task, our model outperforms Qwen2-Audio-Instruct-7B. (5) In the Text-to-Speech task, based on pretrained vocoder (e.g., Fishspeech1.4 or CosyVoice2.0) Nexus is comparable to its backbone vocoder on Seed-TTS benchmark Anastassiou et al. (2024).

To further investigate the interplay among three modalities, we conducted an in-depth analysis of tri-modal alignment. Our findings indicate that (6) the inherent correspondence between audio and language can serve to facilitate and reinforce vision-language alignment, indicating the efficiency of our training strategy. The main contributions of this paper are summarised as follows:

- We propose a comprehensive, industry-level omni-modal LLM pipeline that comprises a modularised end-to-end framework, a lightweight training strategy, and an audio data synthesis pipeline.
- Based on the proposed pipeline, we introduce Nexus, an industry-level omni-modal LLM. It is designed to support any combination of audio, image/video, and text inputs, and it is capable of generating outputs in either audio or language modalities.
- We systematically design the speech data synthesis pipeline to obtain high-quality, real-life speech datasets, covering various real-world scenarios, such as corporate meetings, live broadcasting scenarios, etc., to facilitate the rapid deployment of the model in real-world scenarios. Moreover, we introduce the Nexus-audio testset to evaluate its robustness in real life ASR task.

[1]In this report, Qwen2.5-VL refers to the instruction version.

[2]"Nexus": connect the strongest models to form an even more powerful integrated model.

- We comprehensively evaluate the performance of Nexus on different benchmarks and conduct an in-depth analysis of vision-language alignment by incorporating audio modality within the representation space. Experimental results indicate the merits of our proposed pipeline.

To the best of our knowledge, we are the first to propose the omni-modal LLM from the modularised design perspective that *"connect the strongest models to form an even more powerful integrated model"*. We intend to release the pretrained checkpoint, experimental pipeline, and test set, thereby establishing a new standard to foster innovation in addressing complex, real-world applications.

2 Related Work

In this section, we present the existing works of Multimodal Large Language Models (MLLMs) in Section 2.1, and analyse existing audio modality evaluation approaches in Section 2.2.

2.1 MLLMs

Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding and reasoning over textual knowledge (Yang et al., 2024a). Building on these advancements, a bunch of studies (Dai et al., 2022) have extended the understanding and alignment capabilities of LLMs to visual knowledge, addressing challenges such as vision-language alignment and instruction-following, by introducing Multimodal Large Language Models (MLLMs). Notable models include Qwen-VL (Wang et al., 2024), InternVL (Chen et al., 2024), etc. However, in complex visual reasoning tasks, such as video analysis, MLLMs typically struggle to perform well due to the heterogeneity among diverse modalities, where different modalities have distinct feature granularities, representations, and geometries in a shared latent space (Zhang et al., 2024). To further improve their performance in visual understanding tasks, recent works have expanded MLLMs by incorporating audio capabilities, such as LLaMA-Omni (Fang et al., 2024a), VITA-1.0 (Fu et al., 2024c), Mini-Omni2 (Xie & Wu, 2024b), Baichuan-Omni (Li et al., 2024a), and MinMo (Chen et al., 2025a). However, these models exhibit certain limitations: LLaMA-Omni and MinMo lacks visual capabilities, Baichuan-Omni does not have end-to-end Text To Speech (TTS) capabilities, and VITA-1.0 is constrained by the limited capability of its backbone model.

Recent models, such as VITA-1.5 (Fu et al., 2025), MiniCPM-o2.6 [3], Baichuan-omni-1.5[4], and OpenOmni (Luo et al., 2025), have been proposed and have achieved state-of-the-art performance. However, these models usually require a visual pretraining stage. While pretraining from scratch can enhance multimodal alignment, this strategy significantly increases training time and resource demands due to long sequence of image or video tokens and large pretraining vision data, which is less friendly with institutions with limited computation resources. To improve training efficiency, we start by pre-training over MLLM, specifically Qwen2.5-VL-7B. Owing to the natural alignment between audio and language modalities, the integration of the audio modality has minimal impact on the pretrained geometric and manifold structures underlying the vision-language alignment. As a result, the model's established capabilities in vision comprehension and reasoning are effectively preserved.

2.2 Audio Modality Evaluation

With the rapid advancement of MLLMs, a wide range of multimodal benchmarks has been proposed and widely adopted for evaluation, ranging over vision (Fu et al., 2024b), audio (Fu et al., 2024a), omni (Li et al., 2024b), etc. In the speech domain, existing benchmarks, e.g., Fleurs (Conneau et al., 2023), Aishell2 (Du et al., 2018a), LibriSpeech (Pratap et al., 2020b), and Common Voice (Ardila et al., 2019), provide linguistically diverse datasets under varying acoustic conditions, facilitating Automatic Speech Recognition (ASR) evaluation. Additionally, AIR-Bench (Yang et al., 2024b) introduces a pioneering generative evaluation framework encompassing speech, natural sounds, and music, employing novel audio mixing strategies and GPT-4 for unified, objective, and reproducible assessments.

[3] https://github.com/OpenBMB/MiniCPM-o

[4] `https://github.com/baichuan-inc/Baichuan-Omni-1.5`

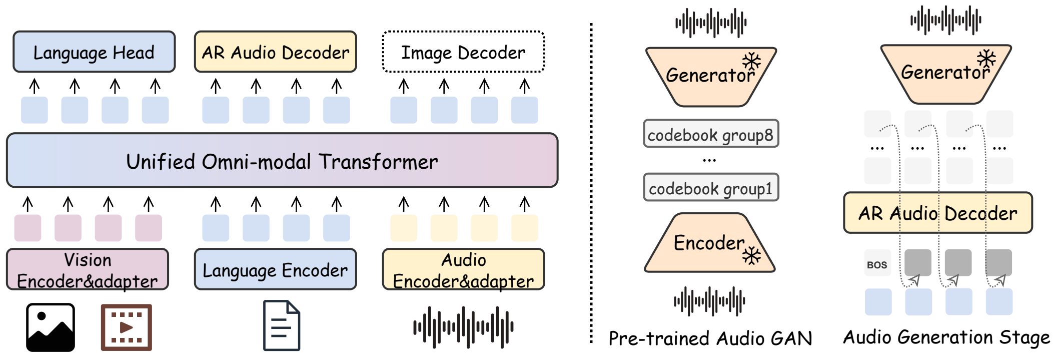

Figure 2: Modularised Architecture, which is designed to accept any combination of input modalities and generates output in either the language or audio modality, where the Auto-Regressive (AR) audio decoder takes a special start token embedding and the last language embedding as input to generate the hierarchical discrete audio codes in an auto-regressive manner. These codes are subsequently fed into a pretrained audio generator to produce the final audio output.

However, the assessment of speech models should be closely aligned with real-world environments to ensure that the evaluation results faithfully represent the demands and requirements of practical and industrial applications. We observe that these existing speech benchmarks present three limitations. First, the small data scale of these benchmarks typically results in significant performance variance. Second, the benchmarks primarily focus on controlled scenarios, making it difficult to evaluate models in dynamic and complex real-world settings, e.g., meetings or live broadcast environments. Third, quantitative evaluation scores from the existing ASR benchmarks generally fail to align closely with the actual user interaction experience, thereby limiting their practical utility. To address these limitations, we propose Nexus-audio testset, comprising 8.1k and 2.8k of Chinese and English ASR samples, spanning various application domains, such as corporate meetings and live broadcasting scenarios.

3 NEXUS

In this section, we introduce the architecture of Nexus in Section 3.1, training data composition and audio data synthesis pipeline in Section 3.2, pre-training strategies in Section 3.3, respectively.

3.1 MODULARISED ARCHITECTURE

Visual encoder. Following Qwen2.5-VL (Bai et al., 2025), the vision encoder employs a redesigned Vision Transformer (ViT) architecture by incorporating M-RoPE and window attention to support native input resolutions while accelerating the computation of the entire visual encoder. During both training and inference, the height and width of the input images are resized to multiples of 28 before being fed into the ViT. The vision encoder processes images by splitting them into patches with a stride of 14, generating a set of image features.

Audio encoder/decoder. To enable audio capabilities in the vision-language MLLM, we incorporate an audio encoder-decoder architecture. In this setup, the encoder is responsible for mapping the speech features into the semantic space of the MLLM, while the decoder transforms the semantic code back into speech, as shown in Figure 2. Specifically, the audio encoder comprises a pre-trained Whisper-large-v3 (Radford et al., 2023), and a two-layer MLP adapter. Specifically, the first layer of the MLP concatenates five consecutive speech feature vectors into a single vector, which is subsequently processed by the second MLP layer.

The auto-regressive audio decoder is a 6-layer decoder-only transformer designed to generate discrete speech codes, which are then passed to a pre-trained generator to produce the final waveforms. Specifically, we use the audio codebook from the FireFly GAN in Fish-Speech (Liao et al., 2024) as the audio codebook. Following MusicGen (Copet et al., 2023), the audio decoder predicts sequences

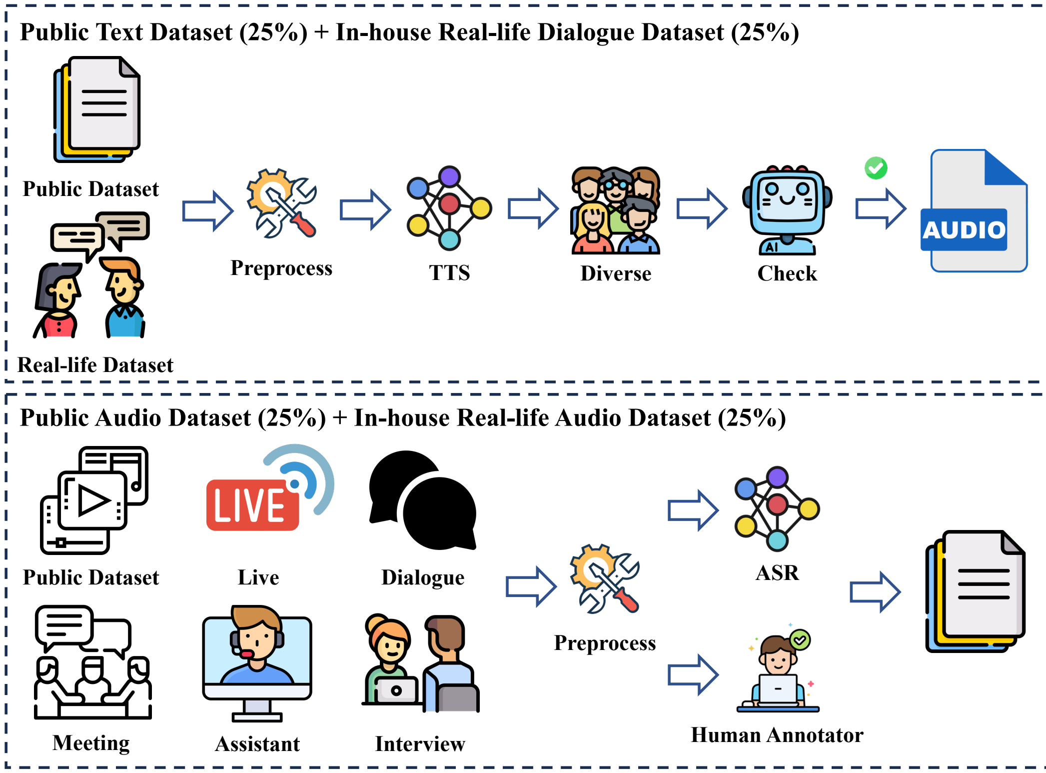

Figure 3: Audio dataset synthesis pipeline, the upper component represents the text-to-audio branch, while the lower component corresponds to the audio-to-text branch. Both components incorporate an equal proportion of in house, real-world samples.

of eight speech codes in a delayed manner. At each time step, the decoder averages the embeddings of the eight speech codes generated in the previous time step with the hidden state of the LLM at the current time step. The resulting averaged embedding is then fed into the audio transformer to predict the next eight speech tokens. Finally, the predicted speech tokens are input into the generator of the FireFly GAN to synthesise the final waveforms.

Language model. Nexus utilises the language model component of Qwen2.5-VL-7B (Bai et al., 2025) as the foundational language model.

3.2 Training Data

Datasets composition. As detailed in Table 1, the training data for Nexus encompasses a wide range of domains within speech processing. These domains include (i) recognition tasks such as ASR and Speech Emotion Recognition (SER), (ii) reasoning tasks include Automatic Audio Caption and Speech Emotion Recognition, (iii) translation tasks like Speech-to-Text Translation (S2TT), generation tasks such as Speech Synthesis, and (iv) interactive tasks involving Speech-to-Speech Chat and Speech-to-Text Chat. The pretrained datasets include (1) public corpora (50%): Aishell1 (Bu et al., 2017), Wenetspeech (Zhang et al., 2022), Librispeech (Panayotov et al., 2015a), Kensho (O'Neill et al., 2021), PeopleSpeech (Galvez et al., 2021), MLS (Pratap et al., 2020a), and Gigast (Ye et al., 2023) and (2) real-life in-house data (50%).

Table 1 presents the statistical information for the audio training dataset. Unlike most Omni-MLLMs, which typically report duration in hours Chen et al. (2025a), our summary is based on the number of audio samples. The total duration of the dataset is approximately 30k hours, this is considerably smaller than the durations employed in other works, such as MinMo (pretrained over 1.4 million hours) Chen et al. (2025a) and Qwen2-Audio (pretrained over 520k hours) Chu et al. (2024a).

Table 1: Datasets description. The target model for each stage is described in Figure 4. The template for each task is provided in Appendix A.

Stage	Task Description	Num
1	Automatic Speech Recognition	5.6M
	Speech-to-Text Translation	1.5M
	Speech-to-Text Chat	0.2M
	Automatic Audio Caption	0.14M
	Speech Emotion Recognition	0.75M
2	Speech-to-Text Chat	0.23M
	Automatic Speech Recognition	20k
	Speech-to-Text Translation	20k
3	Speech Synthesis	40M
	Speech-to-Speech chat	0.22M

Audio dataset synthesis pipeline. We integrate both English and Chinese speech and textual datasets to generate paired speech-text data for ASR, speech synthesis (text-to-speech), Speech-to-Text Chat, and Speech-to-Speech Chat tasks. For the speech corpus (depicted at the bottom of Figure 3), we employ Whisper (Radford et al., 2022) to transcribe speech into text. For noisy samples in our in-house dataset, human annotators are engaged to ensure accurate transcription.

For the textural corpus (top in Figure 3), we employ the zero-shot in-context generation technique proposed by CosyVoice (Du et al., 2024b) to convert the text into natural-sounding speech. However, many conversational and instruction textural datasets are not ideal for speech synthesis due to the nature of the tasks, which do not facilitate effective speech interaction. To address this issue, we enhance the quality of our synthesised speech data by implementing the following pre-processing rules:

1. Length Filtering: texts exceeding a length threshold are excluded from the dataset. Specifically, texts longer than 200 Chinese characters or 200 English words are removed.
2. Non-text Element Filtering: texts containing an excessively high proportion of non-text elements are excluded. Specifically, texts with a non-text character ratio exceeding 0.2 are removed. Text characters in this context include punctuation marks (e.g., !, ?), alphabetic characters ([a-z]), and Chinese characters.
3. Pattern Matching Filtering: texts that included URLs, file paths, or common LaTeX formulas were identified and removed using regular expressions.

For Speech-to-Text Chat and Speech-to-Speech Chat tasks, some texts are unsuitable for a conversational context. To address this limitation, we implement the following rules:

1. Excluding queries that are not appropriate for the Chat task. For example, questions involving metathetical questions are filtered out.
2. Rewriting the question in a conversational tone and generating a conversational response via Qwen2-72B (Chu et al., 2024b).

To enhance the diversity of voices, we randomly select a speaker from a pool of two thousand individuals for each corresponding language to synthesise the speech, ensuring that the voice aligns with the language of the text. Following the text-to-speech (TTS) process, an automated verification procedure is executed using the pretrained Whisper model, which checks whether the synthesised sample can be approximately reconverted to the original source sample, thereby ensuring the quality of the data synthesis.

Finally, leveraging the synthesised audio dataset, we introduce the **Nexus-audio testset**, which comprises 8.1k Chinese and 2.8k English ASR samples. This testset spans a range of application domains, such as corporate meetings, live broadcasting scenarios, etc.

3.3 Training Strategies

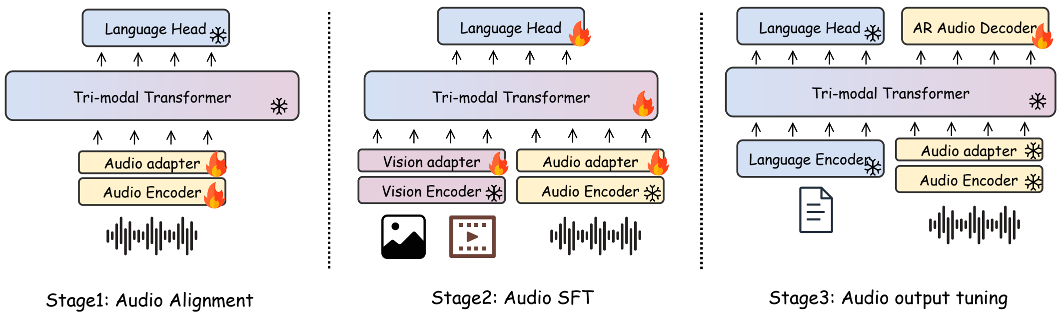

Figure 4: Overview of the training stage in Nexus. The first stage aims to map the speech features into semantic space, the second stage aims to enable the audio instruction-following capability, and the last stage aims to enable speech generation capability.

Stage 1: audio alignment. The first stage aims to facilitate the capacity of LLMs to comprehend and process input speech features. To accomplish this, we utilise 8 million bilingual speech-text pairs for alignment training, covering five training objectives, as illustrated in Table 1. Consistent with the methodology employed in Qwen2.5-VL-7B, our training data is structured in the ChatML format. An example of this format is illustrated as below:

Input Format

<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|>*[audio.wav]*<|audio_eos|>Please help me transcribe my current speech into text.
<|im_end|>
<|im_start|>assistant
[text] <|im_end|>

Stage 2: audio SFT The second stage aims to enable the model to follow speech instructions by incorporating speech-based questions and text-based answers. In this stage, we freeze the visual and audio encoders while unfreezing the visual and audio adapters, and backbone LLM. This strategy is designed to enhance the capability of Nexus to respond accurately to multimodal instructions, thereby improving its performance in understanding and processing both visual and auditory information.

By keeping the core encoder parameters fixed, we ensure that the foundational representations remain stable while allowing the adapters to fine-tune and specialize in the nuances of specific tasks or instructions. This strategic adjustment aims to refine the sensitivity of Nexus to multimodal cues without disrupting the robustness of its underlying architecture.

Stage 3: audio output. Different from the previous large speech model connected to the TTS module (Fu et al., 2024c), we exploit a decoder to generate speech end-to-end. First, we train Nexus on the Text-to-Speech dataset to make the model learn text-to-speech alignment. Then, we utilize the speech-to-speech dataset to train the model to learn speech understanding and speech synthesis capabilities. At this stage, only the parameters of the audio decoder are trainable.

4 Evaluation

We systematically evaluate our model performance from two perspectives: (i) downstream tasks alignment analysis in Section 4.1 and (ii) representation space alignment analysis in Section 4.2.

4.1 Multimodal Tasks Alignment Analysis

In this section, we evaluate the performance of Nexus on the following five tasks: vision understanding, Spoken English QA tasks, ASR, Speech-to-Text translation, and Text-to-Speech tasks. These tasks cover a wide range of modalities and capabilities, providing a comprehensive assessment of the versatility and effectiveness of MLLMs across various domains.

Vision-Language Evaluation We conduct a comprehensive quantitative evaluation of Nexuson a range of vision understanding tasks using eight distinct benchmarks: (1) HallusionBench (Hal) (Guan et al., 2023): Evaluates the model's susceptibility to hallucination and its handling of visual illusions. (2) MathVista (MathV) (Lu et al., 2023): Assesses mathematical reasoning within visual contexts. (3) OCRBench (OCR) (Fu et al., 2024d): Measures optical character recognition (OCR) capabilities on text-intensive images. (4) Video-MME (Fu et al., 2024b): Examines video question-answering reasoning abilities across six domains over diverse temporal ranges. (5) MMMU (Yue et al., 2024): Challenges the model with multi-disciplinary tasks that demand college-level subject knowledge and deliberate reasoning. (6) AI2D: A dataset comprising over 5,000 grade school science diagrams. (7) MMVet (Yu et al., 2024): Evaluates integrated capabilities including recognition, OCR, domain knowledge, language generation, spatial awareness, and mathematical reasoning. (8) MME (Fu et al., 2024a): Measures both perceptual and cognitive abilities across 14 subtasks.

As illustrated in Table 2, we observe that Nexus achieves strong performance to current popular vision-language MLLMs on the MathV, OCR, and Hal benchmarks. Furthermore, when compared to Qwen2.5-VL, Nexus demonstrates superior performance, highlighting its competitive capabilities in vision understanding and reasoning tasks, indicating our model and training strategy can effectively maintain vision-language alignment capability by incorporating audio modality **(Finding1)**.

Table 2: **Evaluation on Vision Understanding Benchmarks.** The best two values are shown in **bold** and <u>underlined</u>. The blue row refers to the main competitors. In MMMU benchmark, * indicates that the measurement is developed over the validation set. We can observe that Nexus shows performance comparable to the leading open-source models and advanced closed-source counterparts.

Model	LLM-size	Video-MME	MMMU	MathV	Hal	AI2D	OCR	MMVet	MME
Vision-Language Models									
VILA-1.5	Vicuna-v1.5-13B	44.2	41.1	42.5	39.3	69.9	460.0	45.0	1718.2
LLaVA-Next	Yi-34B	51.6	48.8	40.4	34.8	78.9	574.0	50.7	2006.5
CogVLM2	Llama3-Instruct-8B	50.5	42.6	38.6	41.3	73.4	757.0	57.8	1869.5
InternLM-Xcomposer2	InternLM2-7B	56.2	41.4	59.5	41.0	81.2	532.0	46.7	2220.4
Cambrian	NousHermes2-Yi-34B	54.2	50.4	50.3	41.6	79.5	591.0	53.2	2049.9
InternVL-Chat-1.5	InternLM2-20B	57.1	46.8	54.7	47.4	80.6	720.0	55.4	2189.6
Ovis1.5	Gemma2-It-9B	58.1	49.7	65.6	48.2	**84.5**	752.0	53.8	2125.2
InternVL2	InternLM2.5-7B	<u>61.5</u>	51.2	58.3	45.0	<u>83.6</u>	794.0	54.3	2215.1
MiniCPM-V 2.6	Qwen2-7B	57.5	49.8	60.6	48.1	82.1	852.0	60.0	2268.7
Qwen2.5-VL	Qwen2.5-7B	56.0	51.8*	61.1	**71.7**	80.7	<u>877.0</u>	-	2299.1
Omni-modal Models									
VITA-1.5-Audio	Qwen2-7B	-	52.1	**66.2**	44.9	79.3	732.0	49.6	**2352.0**
EMova-8B	LLaMA-3.1-8B	-	-	61.1	-	82.8	824.0	55.8	2205.0
Baichuan-Omni-1.5	-	58.2	47.3	51.9	47.8	-	-	65.4	2186.9
Megrez-3B-Omni	Megrez-3B	-	51.8	62.0	50.1	82.0	-	-	2315.0
Proprietary									
GPT-4V	-	50.4	59.3	48.2	39.3	71.4	678.0	49.0	1790.3
GPT-4o mini	-	54.8	60.0	52.4	46.1	77.8	785.0	**66.9**	2003.4
Gemini 1.5 Pro	200B	59.1	60.6	57.7	45.6	79.1	754.0	64.0	2110.6
GPT-4o	-	61.6	<u>62.8</u>	56.5	51.7	77.4	663.0	<u>66.5</u>	<u>2328.7</u>
Claude3.5 Sonnet	175B	**62.2**	**65.9**	61.6	49.9	80.2	788.0	66.0	1920.0
Our Model									
Nexus	Qwen2.5-VL-7B	57.0	53.2*	<u>62.1</u>	<u>71.1</u>	81.2	**882.0**	-	2315.5

Audio-Language Evaluation Next, we evaluate the audio-language alignment on Spoken Question-Answering (SQA), Automatic Speech Recognition (ASR), Speech-to-Text translation (S2TT), and Text-to-Speech (TTS) tasks.

SQA task We evaluate the performance on the English spoken QA benchmark: LLaMA Q. (Nachmani et al., 2024). As presented in Table 3, our model achieves top performance, higher than same-period competitor MiniCPM-o2.6-7B (highlighted in blue), demonstrating its competitive capabilities in this task.

Table 3: **Evaluation on Audio English QA Benchmarks.** The accuracy (%) of different models in English question answering on three sets. The parameter size is derived from the backbone LLMs. Our model achieves top accuracy in the LLaMA Q. benchmark, outperforming the same-period competitor MiniCPM-o2.6-7B.

Model	Modality	LLaMA Q.↑
SpeechGPT-7B (Zhang et al., 2023)	Audio&Text	21.60
Spectron-1B (Nachmani et al., 2023)	Audio&Text	22.90
Moshi-7B (Défossez et al., 2024)	Audio&Text	62.30
GLM-4-Voice-9B (Zeng et al., 2024)	Audio&Text	64.70
MiniCPM-o2.6-7B	Audio&Text	61.00
Mini-Omni-0.5B (Xie & Wu, 2024a)	Audio&Text	22.00
Llama-Omni-8B (Fang et al., 2024b)	Audio&Text	45.30
Our Models		
Nexus	Audio&Text	**67.33**

ASR task In the ASR task, we focus on both Mandarin (CH) and English (EN), evaluating performance on the following benchmarks: AIShell-2 (Du et al., 2018b), Librispeech (Panayotov et al., 2015b), and our real-scenario benchmark. As demonstrated in Table 4, our model achieves the best performance on the real scenario ASR testset, Nexus-audio, indicating the robustness of our model in real life **(Finding2)**.

Table 4: **Evaluation on ASR Benchmarks.** Nexus has demonstrated strong performance in both real-life Mandarin and English ASR tasks. It outperforms specialized speech models, achieving better results in both languages. AIShell-2 is measured over all categories (Mac/iOS/Android), TC: test-clean, TO: test-other. We can observe that Nexus outperforms other models in real scenarios.

Model	CN (CER↓)		Eng (WER↓)		
	AIShell-2	Nexus-audio	Librispeech TC	Librispeech TO	Nexus-audio
Speech LLMs					
Qwen2-Audio-Instruct-7B	3.00/3.00/2.90	35.45	**1.70**	4.00	26.12
Omni-modal LLMs					
Mini-Omini2-0.5B(Xie & Wu, 2024a)		117.71	9.80	4.70	54.69
VITA-1.5-7B(Fu et al., 2025)	-	127.34	7.20	**3.40**	122.71
MinMo-7B(Chen et al., 2025b)	4.89/4.76/4.96	-	1.74	3.89	-
MiniCPM-o2.6-7B	-	12.00	**1.70**	3.89	18.00
Our Models					
Nexus	3.35	**11.82**	3.50	5.29	**12.32**

S2TT task As for the S2TT, we exploit CoVoST2 (Wang et al., 2020), which provides massive multilingual S2TT datasets. As shown in Table 5, Nexus demonstrates superior performance compared to specialised speech LLM, Qwen2-Audio-7B-Instruct, in both translation tasks. Nevertheless, our model still falls short of the performance of its contemporary competitor, MiniCPM-o2.6-7B. However, we pre-train our model on a small audio dataset (about 30k hours) at the pre-training stage **(Finding3)**.

TTS task Finally, we evaluate Nexus on Seed-TTS benchmark Anastassiou et al. (2024)[5]. As shown in Table 6, our experiments demonstrate that when employing pretrained vocoders (e.g.,

[5]https://github.com/BytedanceSpeech/seed-tts-eval

Table 5: **Evaluation on Speech-to-Text Translation Benchmarks.** Nexus demonstrated superior performance compared to Qwen2-Audio-Instruct-7B in both translation tasks, however, it still falls short of the performance of its contemporary competitor due to the relatively modest scale of our pre-training speech dataset (about 30k hours).

Model	CH-EN (BLEU↑)	EN-CH (BLEU↑)
	CoVoST2	CoVoST2
Speech LLMs		
Qwen2-Audio-7B (Chu et al., 2024a)	24.40	45.20
Qwen2-Audio-Instruct-7B (Chu et al., 2024a)*	22.90	39.50
MinMo-7B (Chen et al., 2025a)*	<u>25.95</u>	<u>46.68</u>
Omni-modal LLMs		
MiniCPM-o2.6-7B	**27.20**	**48.20**
Our Models		
Nexus	23.17	40.21

Fishspeech1.4 or CosyVoice2.0), Nexus achieves performance comparable to its competitors (e.g., CosyVoice1.0 or CosyVoice2.0) **(Finding4)**.

Table 6: **Evaluation on Text-to-Speech (TTS) Benchmarks.** Nexus demonstrated lower performance in both Mandarin (CH) and English (EN) TTS tasks. Notice: MinMo-7B's result is based on their manually filtered samples as mentioned in their report Chen et al. (2025a), therefore, we do not compare with it.

Model	CH	EN
	CER↓	WER↓
CosyVoice1.0 (Du et al., 2024a)	8.64	3.59
CosyVoice2.0 (Du et al., 2024c)	4.61	**2.43**
CosyVoice2.0-SFT (Du et al., 2024c)	**2.06**	<u>3.19</u>
MinMo-7B (Chen et al., 2025a)	2.48	2.90
Nexus (Fishspeech1.4)	7.55	13.02
Nexus (CosyVoice2.0)	<u>4.53</u>	4.11

Thus far, our evaluations demonstrate that incorporating the audio modality enables the model to perform audio understanding and generation without compromising its vision comprehension and reasoning capabilities. In the following section, we provide an in-depth evaluation of the vision-language alignment from the perspective of representation space.

4.2 Representational Alignment Analysis

We formalise MLLM within the framework of an ***unembedding-embedding*** architecture. In this framework, the unembedding stage is responsible for learning transformations between observations (e.g., text, vision, audio) and latent spaces through encoders, while the embedding stage captures the complex interactions among latent variables within the latent space of the hidden layers in MLLMs. Each stage serves distinct functions and yields representations with different properties. Consequently, by focusing on each stage independently, we can have a systematical evaluation of model behaviours in representation spaces.

To assess the representational alignment among vision-language modalities at each stage, we first extract embeddings for each sample's modalities. Our evaluation focuses on two aspects: (1) whether our model inherently promotes superior vision-language alignment, and (2) whether incorporating the audio modality into the vision modality further enhances this alignment.

Quantitative evaluation of vision-language alignment is performed using kernel-alignment metrics. The kernel of a space encapsulates its underlying geometrical structure (i.e., distance metric), and the similarity between the kernels of different spaces serves as an indicator of the alignment between these spaces—in our case, across different modalities. Those metrics involve cka (Kornblith et al., 2019), cknna (Huh et al., 2024), svcca (Raghu et al., 2017), and k-nearest neighbours (knn)-based metrics: cycle knn, mutual knn, lcs knn, and edit knn. More information for those metrics can be found in (Huh et al., 2024).

Unembedding stage In the unembedding stage, we evaluate representational alignment using ImageNet (Deng et al., 2009), which offers a fine granularity in language-vision alignment, enabling a detailed assessment of representational performance. To get the audio input for each language-vision pair, we apply the pretrained Text-to-Speech LLMs, cosyvoice-300M-SFT checkpoint (Du et al., 2024b). As for the baseline, we compare our model with MLLMs after SFT, including Qwen2.5-VL (Bai et al., 2025), LLaMA3.2-Vision (Meta., 2024), LLaVA family Liu et al. (2023), Phi-3.5-Vision Abdin et al. (2024), and InternVL-8B (Chen et al., 2024).

As shown in Table 7, we can observe that (1) our model without audio inputs can lead to better vision-language alignment than the backbone model, Qwen2.5-VL-7B, on 5 out of 7 metrics , and (2) by incorporating audio modalities, our model outperforms others on 4 out of 7 metrics **(Finding5)**. These results indirectly demonstrate that our modular formalism design concept can yield an efficient tri-modal alignment at the encoding stage.

Table 7: Kernel alignment analysis for unembedding space. The representation for each sample is the averaged token embeddings. The best two values are shown in **bold** and <u>underlined</u>, we can observe our tri-modal model with audio inputs generally outperforms others on both datasets.

Baseline	cycle knn	mutual knn	lcs knn	cka	cknna	svcca	edit knn
ImageNet: concepts							
Qwen2.5-VL-7B	0.66053	**0.04278**	<u>1.47807</u>	0.07987	0.02483	0.09168	0.00100
Llama3.2-Vision-Instruct-11B	0.08608	<u>0.04205</u>	**1.52788**	0.06079	0.01403	0.11651	0.00061
LLaVA1.5-7B	0.48673	0.04164	1.43367	0.03363	0.01213	0.13553	0.00088
LLaVA1.6-7B	0.57173	0.02077	0.81645	0.08024	0.01577	<u>0.15240</u>	0.00037
Phi3.5-Vision-Instruct	0.02761	0.01257	0.52355	0.08614	0.00714	0.12118	0.00019
InternVL2-8B	0.08175	0.01637	0.72495	<u>0.09185</u>	0.00062	0.12148	0.00044
Nexus	**0.67839**	0.03418	1.25284	0.08758	<u>0.02822</u>	0.11963	<u>0.00112</u>
vision+audio	<u>0.66432</u>	0.03935	1.39794	**0.09706**	**0.02833**	**0.16667**	**0.00139**
Wikipedia Caption: short descriptive sentences							
Qwen2.5-VL-7B	0.52148	<u>0.07005</u>	<u>2.09765</u>	<u>0.10194</u>	<u>0.04112</u>	0.19668	<u>0.00173</u>
Llama3.2-Vision-Instruct-11B	0.31347	0.03623	1.29980	0.00968	0.00779	0.22120	0.00050
LLaVA1.5-7B	<u>0.57714</u>	0.02886	1.03808	0.04632	0.01586	<u>0.25615</u>	0.00046
LLaVA1.6-7B	**0.57812**	0.03935	1.36523	0.07933	0.03998	0.23114	0.00082
Phi3.5-Vision-Instruct	0.04980	0.03027	1.14843	0.01669	0.03890	0.18183	0.00066
InternVL2-8B	0.36914	0.04132	1.55761	0.04732	0.01658	0.21739	0.00093
Nexus	0.48828	0.06990	2.07714	0.08924	0.04101	0.20311	0.00167
vision+audio	0.54000	**0.10636**	**2.79003**	**0.12403**	**0.06329**	**0.25624**	**0.00230**

Embedding stage Next, we focus on the vision-language alignment at the embedding stage. We restrict our attention to the Wikipedia-caption corpus (Srinivasan et al., 2021). We present a comparative analysis between Qwen2.5-VL-7B and our model across all hidden layers. We feed both image and its audio into the model and consider the last token embedding as the final fused representation. As illustrated in Figure 5, the incorporation of the audio modality significantly enhances vision-language alignment, particularly in the intermediate hidden layers **(Finding6)**. This finding underscores the efficacy of our training strategy, whereby pretraining exclusively on audio data yields a robust tri-modal alignment.

5 Conclusion and Future Work

In this work, we propose a comprehensive, industry-level omni-modal LLM pipeline that comprises a modularised end-to-end framework, a lightweight training strategy, and an audio data synthesis pipeline to overcome the current challenges, including limited accessibility to tri-modal datasets,

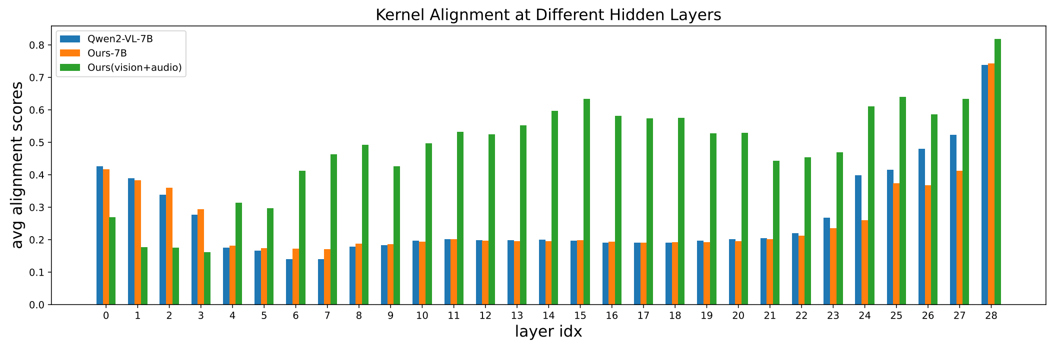

Figure 5: Averaged kernel-alignment score across different hidden layers. vision+audio: both visual and auditory modalities are concurrently fed into the model. The final fused representation is the last token in the sequence as the comprehensive summary of the integrated modalities. we can observe that incorporating audio modality (green bar) can result in better vision-language alignment at most hidden layers.

higher computational resource requirements, and the complexity of aligning features across three modalities. To this end, we propose an industry-level omni-perceive and -interactive tri-modal LLM, **Nexus**.

Extensive experiments validate the usefulness of our proposed pipeline and Nexus. In the visual understanding task, Nexus exhibits superior performance compared with its backbone model - Qwen2.5-VL-7B, validating the efficiency of our training strategy. In the Text-to-Speech task, based on pretrained vocoder (e.g., Fishspeech1.4 or CosyVoice2.0), Nexus is comparable to its backbone vocoder on Seed-TTS benchmark. In our real-world ASR testset, Nexus achieves outstanding performance, indicating its robustness in real scenarios.

In the future, we intend to link Nexus with vision generative models under our modularised framework to allow the model to seamlessly generate high-quality, contextually coherent visual outputs. Integrating the Artificial Intelligence-Generated Content (AIGC) capabilities into the current model can further broaden its applicability in creative and industrial domains, such as Embodied Intelligence.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-tts: A family of high-quality versatile speech generation models, 2024. URL `https://arxiv.org/abs/2406.02430`.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline, 2017. URL `https://arxiv.org/abs/1709.05522`.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, Yabin Li, Xiang Lv, Jiaqing Liu, Haoneng Luo, Bin Ma, Chongjia Ni, Xian Shi, Jialong Tang, Hui Wang, Hao Wang, Wen Wang, Yuxuan Wang, Yunlan Xu, Fan Yu, Zhijie Yan, Yexin Yang, Baosong Yang, Xian Yang, Guanrou Yang, Tianyu Zhao, Qinglin Zhang, Shiliang Zhang, Nan Zhao, Pei Zhang, Chong Zhang, and Jinren Zhou. Minmo: A multimodal large language model for seamless voice interaction, 2025a. URL `https://arxiv.org/abs/2501.06282`.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024.

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. Qwen2-audio technical report, 2024a. URL `https://arxiv.org/abs/2407.10759`.

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

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805. IEEE, 2023.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2023.

Yong Dai, Duyu Tang, Liangxin Liu, Minghuan Tan, Cong Zhou, Jingquan Wang, Zhangyin Feng, Fan Zhang, Xueyu Hu, and Shuming Shi. One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code. *arXiv preprint arXiv:2205.06126*, 2022.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. Aishell-2: Transforming mandarin asr research into industrial scale. *arXiv preprint arXiv:1808.10583*, 2018a.

Jiayu Du, Xingyu Na, Xuechen Liu, and Hui Bu. Aishell-2: Transforming mandarin asr research into industrial scale, 2018b. URL `https://arxiv.org/abs/1808.10583`.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens, 2024a. URL `https://arxiv.org/abs/2407.05407`.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, et al. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*, 2024b.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. Cosyvoice 2: Scalable streaming speech synthesis with large language models, 2024c. URL `https://arxiv.org/abs/2412.10117`.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*, 2024a.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models, 2024b. URL `https://arxiv.org/abs/2409.06666`.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024a. URL `https://arxiv.org/abs/2306.13394`.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024b. URL `https://arxiv.org/abs/2405.21075`.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024c.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025.

Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning, 2024d. URL `https://arxiv.org/abs/2501.00321`.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage, 2021. URL `https://arxiv.org/abs/2111.09344`.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*, 2023.

GE Hinton. A fast learning algorithm for deep belief nets. *Neural Computation*, 2006.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/huh24a.html`.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL `https://arxiv.org/abs/2001.08361`.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*, 2(3), 2024a.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnibench: Towards the future of universal omni-language models, 2024b. URL `https://arxiv.org/abs/2409.15272`.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis, 2024. URL `https://arxiv.org/abs/2411.01156`.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

Run Luo, Ting-En Lin, Haonan Zhang, Yuchuan Wu, Xiong Liu, Min Yang, Yongbin Li, Longze Chen, Jiaming Li, Lei Zhang, Yangyi Chen, Hamid Alinejad-Rokny, and Fei Huang. Openomni: Large language models pivot zero-shot omnimodal alignment across language with real-time self-aware emotional speech synthesis, 2025. URL `https://arxiv.org/abs/2501.04561`.

Meta. Llama3.2 technical report. 2024.

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

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm, 2024. URL `https://arxiv.org/abs/2305.15255`.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. Spgispeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition, 2021. URL `https://arxiv.org/abs/2104.02014`.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015a. doi: 10.1109/ICASSP.2015.7178964.

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

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411, 2020a.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*, 2020b.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL `https://arxiv.org/abs/2212.04356`.

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

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 2443–2449, 2021.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Changhan Wang, Anne Wu, and Juan Pino. Covost 2 and massively multilingual speech-to-text translation, 2020. URL `https://arxiv.org/abs/2007.10310`.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL `https://arxiv.org/abs/2503.20215`.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024b.

Rong Ye, Chengqi Zhao, Tom Ko, Chutong Meng, Tao Wang, Mingxuan Wang, and Jun Cao. Gigast: A 10,000-hour pseudo speech translation corpus, 2023. URL https://arxiv.org/abs/2204.03939.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024. URL https://arxiv.org/abs/2308.02490.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi, 2024. URL https://arxiv.org/abs/2311.16502.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Shengmin Jiang, Yuxiao Dong, and Jie Tang. Scaling speech-text pre-training with synthetic interleaved data, 2024. URL https://arxiv.org/abs/2411.17607.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP*, 2022.

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

Yingji Zhang, Marco Valentino, Danilo Carvalho, Ian Pratt-Hartmann, and Andre Freitas. Graph-induced syntactic-semantic spaces in transformer-based variational AutoEncoders. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 474–489, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.32. URL https://aclanthology.org/2024.findings-naacl.32/.

CONTRIBUTIONS

- Che Liu: Exploration of network architecture, code development, and ablation studies.
- Yingji Zhang: Representation analysis and report writing.
- Dong Zhang: Data organization, code development, partial ablation studies, model evaluation and report writing.
- Weijie Zhang: Experiments and training optimization.
- Chenggong Gong: Speech data and speech evaluation.
- Haohan Li: Testing code and model evaluation.
- Yu Lu: Survey of multimodal work and writing of related sections.
- Shilin Zhou: Lead of Text-to-Speech (TTS) section.
- Yue Lu: Responsible for video data and algorithms.
- Ziliang Gan: Testing and optimization of the model in the financial sub-domain.
- Ziao Wang: Assisted in data cleaning and dataset construction.
- Junwei Liao: Assisted in speech-related strategies and manuscript writing.
- Haipang Wu: Provided organizational support and departmental coordination.
- Ji Liu: Manuscript revision, participated in discussions, and provided feedback and suggestions.
- André Freitas: Participated in methodological discussions and provided feedback and suggestions.
- Qifan Wang: Provided consultation and discussion on training strategies.
- Zenglin Xu: Manuscript revision and suggestions on training strategies.
- Rongjuncheng Zhang: Responsible for the testing platform, strategy suggestions, and resource coordination.
- Yong Dai: Oversaw all aspects of the project, and major contributor to manuscript writing.

A EXPERIMENTAL DETAILS

The following tables describe the template we used for each pre-training task:

ASR

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|>audio.wav<|audio_eos|>Please help me transcribe my current speech into text.
<|im_end|>
<|im_start|>assistant
[text] <|im_end|>
```

Speech-to-Text Chat

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|>audio.wav<|audio_eos|> <|im_end|>
<|im_start|>assistant
[response] <|im_end|>
```

Speech-to-Text Translation

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|>audio.wav<|audio_eos|> What is the English equivalent of this Chinese speech?
<|im_end|>
<|im_start|>assistant
[response] <|im_end|>
```

Speech Synthesis

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|>text<|audio_eos|> <|im_end|>
<|im_start|>assistant
[audio.wav] <|im_end|>
```

Speech-to-Speech Chat

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|>audio.wav<|audio_eos|> <|im_end|>
<|im_start|>assistant
<|audio_bos|>audio.wav<|audio_eos|> <|im_end|>
```

Automatic Audio Caption

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|>audio.wav<|audio_eos|> please describe the speech. <|im_end|>
<|im_start|>assistant
[text] <|im_end|>
```

Speech Emotion Recognition

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|audio_bos|>audio.wav<|audio_eos|> what is the emotion in this speech? <|im_end|>
<|im_start|>assistant
[text] <|im_end|>
```