

VITA: Towards Open-Source Interactive Omni Multimodal LLM

Chaoyou Fu[♣], Haojia Lin³, Zuwei Long¹, Yunhang Shen¹, Meng Zhao¹
Yifan Zhang, Xiong Wang¹, Di Yin¹, Long Ma¹, Xiawu Zheng³
Ran He⁴, Rongrong Ji³, Yunsheng Wu¹, Caifeng Shan², Xing Sun^{1,†}

¹Tencent Youtu Lab, ²NJU, ³XMU, ⁴CASIA

♣ Project Leader † Corresponding Author

<https://vita-home.github.io>

Abstract

The remarkable multimodal capabilities and interactive experience of GPT-4o underscore their necessity in practical applications, yet open-source models rarely excel in both areas. In this paper, we introduce **VITA**, the first-ever open-source Multimodal Large Language Model (MLLM) adept at simultaneous processing and analysis of **Video, Image, Text, and Audio** modalities, and meanwhile has an advanced multimodal interactive experience. Starting from Mixtral 8×7B as a language foundation, we expand its Chinese vocabulary followed by bilingual instruction tuning. We further endow the language model with visual and audio capabilities through two-stage multi-task learning of multimodal alignment and instruction tuning. VITA demonstrates robust foundational capabilities of multi-lingual, vision, and audio understanding, as evidenced by its strong performance across a range of both unimodal and multimodal benchmarks. Beyond foundational capabilities, we have made considerable progress in enhancing the **natural multimodal human-computer interaction** experience. To the best of our knowledge, we are the first to exploit non-awakening interaction and audio interrupt in MLLM. We design additional state tokens, and corresponding training data and strategies to perceive various interaction scenarios. The deployment of VITA employs a duplex scheme, where one model is responsible for generating responses to user queries, and the other continuously tracks environmental inputs, selectively outputting new responses with updated interactions. This allows VITA to feature impressive human-machine interaction functionalities such as non-awakening interaction and audio interrupt interaction. VITA is the **first step** for the open-source community to explore the seamless integration of multimodal understanding and interaction. While there is still lots of work to be done on VITA to get close to close-source counterparts, we hope that its role as a pioneer can serve as a cornerstone for subsequent research.

1 Introduction

Large Language Models (LLMs) have undergone significant evolution [39, 21, 9, 48, 12] and recently, we have witnessed the flourishing of Multimodal Large Language Models (MLLMs) [52, 29, 31, 30, 40], which exhibit surprising multimodal capabilities. Particularly, GPT-4o [40] has significantly advanced the field of MLLMs, distinguishing itself with two key attributes: (1) a unified framework that processes text, vision, and audio signals in an end-to-end manner, and (2) the capability to enable natural multimodal human-computer interaction. These breakthroughs have established a new standard in the discipline. However, there is a conspicuous absence of open-source

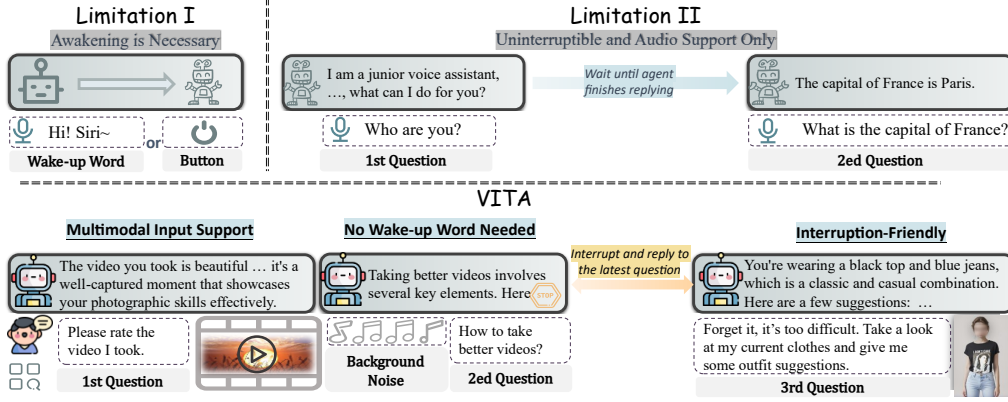


Figure 1: **Interaction of VITA.** Traditional audio interaction requires a pre-defined wake-up word, e.g., saying “Hi! Siri~” every time you ask a question, or a button to control the input question audio (**Limitation 1**). The human-computer interaction is always blocked when the model generates output, as the previous system can only respond to input queries sequentially (**Limitation 2**). By contrast, on the one hand, unlike previous methods where audio, text, and vision are always separated, VITA supports these modalities end-to-end. On the other hand, VITA makes two contributions to multimodal interaction. **Non-awakening Interaction:** VITA automatically filters background noise like non-query human voices, thereby eliminating the need for the wake-up word and the button. **Audio Interrupt Interaction:** If the user interrupts with another question, the generation process is paused, and the model immediately responds to the latest query.

models that echo these highlights. The pressing need for open-source community to further promote development in this field cannot be overstated.

This paper introduces VITA, as a pioneer that has initially achieved the both two attributes, by integrating architectural innovations with advanced training and development strategies. The functionality and architecture of VITA are depicted in Fig. 1 and Fig. 2, respectively. The implementation of VITA mainly comprises three steps:

Bilingual Instruction Tuning of LLM. The official Mixtral 8×7B [22] always lacks proficient Chinese language expression and comprehension. To tackle this, we expand the vocabulary of the base model and continued with further instruction tuning using the collected high-quality bilingual text corpus. This makes the LLM proficient in both Chinese and English.

Multimodal Alignment and Instruction Tuning. To align the text feature space with video, image, and audio, we collect massive high-quality multimodal data to align individual encoders and connectors, which process different modalities respectively, to the LLM. Multimodal instruction tuning data are meticulously constructed. While giving VITA a powerful multimodal foundational capability, we teach it to recognize the type of input queries end-to-end by introducing a state token. This makes it possible to interact without audio awakening while inference.

Development with Duplex Pipeline. In terms of model deployment, we introduce a duplex scheme. As shown in Fig. 2, two VITA models are deployed simultaneously: one is responsible for generating responses to the current audio query, and the other continuously monitors for the new one. If any, the current generation is interrupted, and the model outputs the response to the new query. In order to improve the efficiency of the interaction, we have carried out massive engineering optimizations, such as adapting multimodal vLLM [26].

The contributions of this paper are as follows:

- We develop an open-source high-performance multimodal base model that simultaneously supports video, image, text, and audio inputs in both English and Chinese. The model accepts either pure text/audio inputs or video/image combined with text/audio inputs. We design a comprehensive training process, which includes enhancing the Chinese capabilities of the LLM, constructing multimodal training data, and a multi-stage training pipeline.

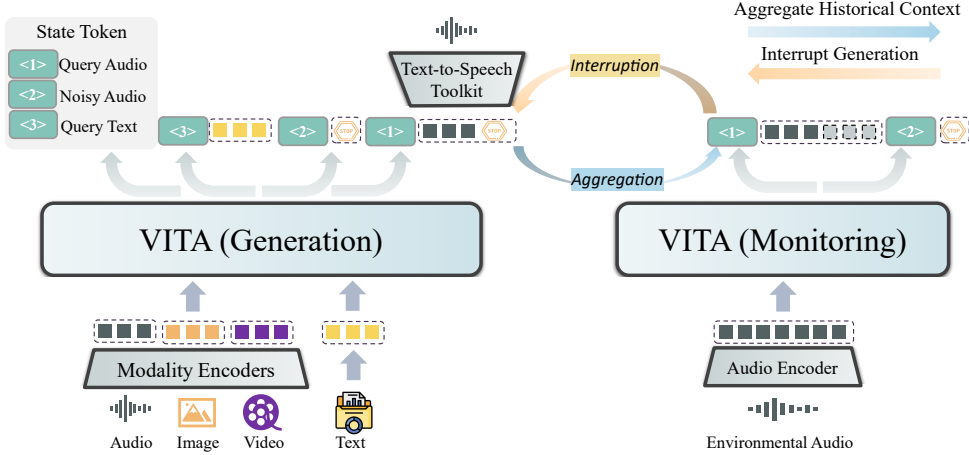


Figure 2: **Architecture of VITA.** VITA is capable of processing inputs in the form of pure text/audio, as well as video/image combined with text/audio. Besides, we set different state tokens for different query inputs. <1> corresponds to the effective query audio, such as “what is the biggest animal in the world?”, for which we expect a response from the model. <2> corresponds to the noisy audio, such as someone in the environment calls me to eat, for which we expect the model not to reply. <3> corresponds to the query text, i.e., the question given by the user in text form. During the training phase, we try to teach the model to automatically distinguish different input queries. During the deployment phase, with <2> we can implement non-awakening interaction. Based on this, we further introduce a duplex scheme for the audio interrupt interaction. Two models are running at the same time, where the generation model is responsible for handling user queries. When the generation model starts working, the other model monitors the environment. If the user interrupts with another effective audio query, the monitoring model aggregates the historical context to respond to the latest query, while the generation model is paused and tune to monitor, i.e., the two models swap identities.

- As a pioneer, we make preliminary explorations in the field of natural multimodal human-computer interaction. By introducing a state token, the model can automatically identify the type of the input audio to achieve non-awakening interaction. Meanwhile, the duplex scheme makes it possible to realize audio interrupt interaction.
- We fully open-source our model, training code, and inference deployment framework, aiming at promoting the advancements of the research community. As a cutting-edge research, we will continue to contribute to the multimodal foundation models and interactions.

2 Related Work

Leveraging advanced LLMs such as GPTs [39, 2], LLaMA [47, 48], Alpaca [43], Vicuna [9], and Mistral [21], MLLMs exhibit enhanced multimodal capabilities, particularly through end-to-end training techniques. Recent open-source MLLMs, such as Otter [27], mPLUG-Owl [51], LLaVA [31], Qwen-VL [1], Cambrian-1 [46], Mini-Gemini [29], MiniCPM-V 2.5 [19], DeepSeek-VL [33], SliME [56], and Bunny [17], have made progress in solving multimodal fundamental problems, such as vision-language alignment and instruction following.

Among them, some representative open-source models like InternLM-XComposer-2.5 [55] and InternVL-2 [8] have been rapidly advancing, demonstrating strong multimodal understanding capabilities and closely rivaling proprietary models in various multimodal benchmarks. However, compared to proprietary models such as GPT-4o [40] and Gemini-Pro 1.5 [44], which support more than two modalities like audio, image, and text, most open-source models focus on image-text modalities [53]. Furthermore, open-source models rarely concentrate on user interaction capabilities, leaving this area relatively unexplored. In comparison, the proposed VITA not only exhibits impressive performance in perceiving data across four modalities, i.e., video, image, text, and audio, but also makes preliminary strides in enhancing user interaction capabilities. Through the comprehensive open-sourcing of VITA, we hope to accelerate developments in this field.

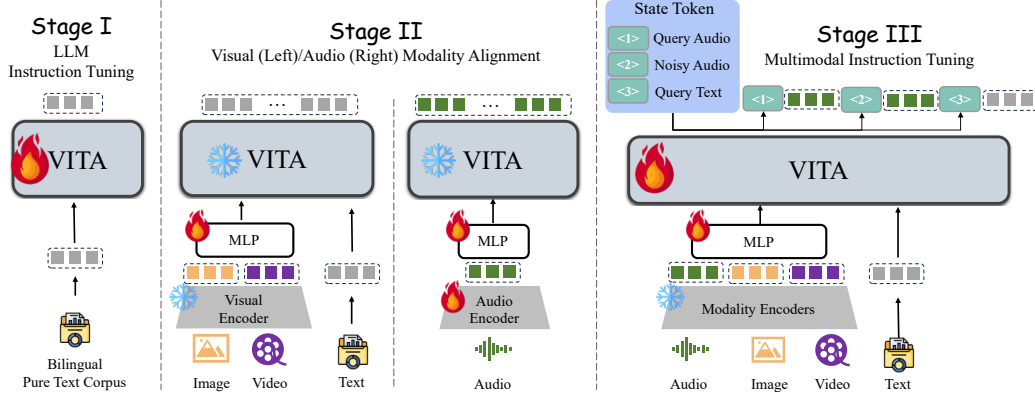


Figure 3: **Training pipeline of VITA.** The first stage **LLM Instruction Tuning** enhances the language model Mixtral $8 \times 7B$ by expanding its vocabulary size and fine-tuning it with a high-quality bilingual text corpus, thereby achieving proficiency in both Chinese and English. The second stage **Multimodal Alignment** connects individual encoders with the LLM to process various modalities. By amassing a substantial collection of high-caliber multimodal data, we synchronize the text feature space with that of video, image, and audio. The last stage **Multimodal Instruction Tuning** allows the model to follow text or audio instructions to understand the image or video. A specially designed state token is used to distinguish the type of input query, facilitating subsequent multimodal human-computer interaction.

3 VITA

As depicted in Fig. 3, the overall training pipeline of VITA consists of three stages: LLM instruction tuning, multimodal alignment, and multimodal instruction tuning. The development of VITA is also an important part.

3.1 LLM Instruction Tuning

Mixtral $8 \times 7B$ ¹ [22] is a representative LLM with an architecture of sparse mixture of experts (SMoE). Its performance is among the top-tier open-source LLMs, making it an ideal starting point of our work. Nonetheless, we observe that the official Mixtral model exhibits limited proficiency in understanding Chinese. To infuse bilingual (Chinese and English) comprehension capabilities, we broaden the vocabulary of the base model with Chinese, increasing the vocabulary size from 32,000 to 51,747. This extension can also reduce the number of tokens under the same text, thus improving inference efficiency. With the extended vocabulary in place, we use 5 million synthetic bilingual corpus for pure-text instruction tuning.

3.2 Multimodal Alignment

In this stage, we aim to bridge the representation gap between text and other modalities, thereby laying the groundwork for multimodal understanding.

3.2.1 Visual Modality

Visual Encoder. We employ InternViT-300M-448px as the visual encoder², which accepts a 448×448 image as input, generating 256 tokens after using a visual connector that is a simple two-layer MLP. For high-resolution image input, we implement the dynamic patching strategy [7] to capture local details. Videos are treated as special cases of images. If the video length is less than 4 seconds, we uniformly sample 4 frames. If the video length is between 4 and 16 seconds, we sample one frame per second. For videos longer than 16 seconds, we uniformly sample 16 frames. To prevent

¹<https://huggingface.co/mistralai/Mixtral-8x7B-v0.1>

²<https://huggingface.co/OpenGVLab/InternViT-300M-448px>

Table 1: Training data of multimodal instruction tuning. The images of the synthetic data come from open-source datasets Wukong [15], LAION [41], and CC12M [3].

Data Scenario	QA Type	Dataset Name	# Concatenated Entries (K)	# Total Entries (K)	# Audio Questions (K)	# Text Questions (K)	Language
General Image	Description	ShareGPT4V	50.7	99.5	49.7	49.7	Eng
		Allava-Caption	362.2	708.1	354.6	353.6	Eng
		ShareGTP4o-Image	39.2	57.3	28.6	28.7	Eng
		Synthetic Data	304.6	594.5	297.0	297.5	CN
	QA	LLaVA-150K	57.9	99.8	109.0	109.1	CN
		LLaVA-Mixture-sample	308.5	308.5	1103.0	920.9	Eng
		Lvis-Instruct	110.4	110.4	562.1	466.0	Eng
		ScienceQA	12.7	12.7	0.0	12.7	Eng
		Synthetic Data	14.1	21.8	106.0	105.3	CN
OCR & Diagram	Description & QA	Anyword-3M	47.5	770.3	384.5	385.8	CN
		ICDAR2019-LSVT	233.2	233.2	680.0	583.1	CN
	QA	ICDAR2017-RCTW	6.6	7.7	3.7	4.0	CN
		Open-Chart	32.3	41.5	229.2	229.0	CN
		Synthetic Data	97.3	156.0	418.8	345.2	CN
General Video	Description	ShareGemini	777.7	777.7	104.1	673.7	CN & Eng
	QA	Synthetic Data	160.5	160.5	280.6	179.5	CN & Eng
Pure Text	QA	Synthetic Data	134.5	800.8	398.5	1113.9	CN & Eng
Total			2749.9	4960.3	5109.4	5857.7	CN & Eng

the introduction of an excessive number of visual tokens, we do not perform dynamic patching on individual frames of the video.

Visual Alignment. We only train the visual connector during the visual alignment phase. Table 1 summarizes the used training data, except the pure text part. In addition, in this phase we do not use audio questions. (1) For the general image description task, we introduce the GPT-4V generated part from ShareGPT4V [6] to ensure data quality. We also introduce Allava-Caption [4] and ShareGTP4o-Image³, and supplement these with Chinese image descriptions generated by existing MLLMs. (2) For the general image Question-Answering (QA) task, we initially gather three datasets: LLaVA-Mixture-sample [31], Lvis-Instruct [50], and ScienceQA [34]. We then use existing MLLMs to generate an additional 21.8K Chinese QA data. Besides, we remove the caption subset from LLaVA-150K [31] and translate the rest into Chinese. (3) For the OCR & Diagram tasks, we include Anyword-3M [49], ICDAR2019-LSVT⁴, ICDAR2017-RCTW⁵, Open-Chart (a collection of ChartQA [36], DVQA [23], InfoVQA [37], Pew [25], and OpenCQA [24]), as well as some synthetic data generated by existing MLLMs from other open-source data with text. For Anyword-3M, we select data where the answer’s corresponding text length is between 20 and 50, with the question asking to identify the text in the image.

For ICDAR2019-LSVT, ICDAR2017-RCTW, and Open-Chart, we generate detailed descriptions and QA pairs using existing MLLMs. (4) For general image descriptions, we use the ShareGemini [42] dataset. (5) For general video QA, we re-label open-source data from Video-ChatGPT [35] and VideoChat2 [28] using existing MLLMs. We sample 800k entries from 5 million pure text data to maintain the LLM’s text comprehension abilities. It is used during multimodal instruction tuning instead of visual alignment, because the LLM parameters of the latter are frozen.

Data Concatenation. For pure text data and image data, we aim to concatenate the context length to 6K tokens as illustrated in Fig. 4 (the amount of concatenated data is indicated in the Concatenated

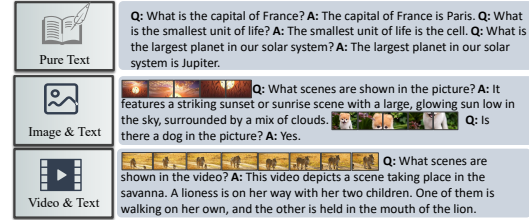


Figure 4: **Data Concatenation.** Different text data is directly concatenated to 6K tokens. Images are first divided into local patches and then different image-text pairs are concatenated. Video data is directly sampled frame by frame as input, without the need for concatenation. In this way, we could unify the length of data in different training batches, thus improving training efficiency.

³<https://sharegpt4o.github.io/>

⁴<http://icdar2019.org/>

⁵<https://iapr.org/icdar2017>

Table 2: System prompts for **image input**, **video input**, and **pure text input**.

System prompt for image data
<p>You are an AI robot and your name is VITA.</p> <p>You are a multimodal large language model developed by the open-source community. Your aim is to be helpful, honest, and harmless.</p> <p>You support the ability to communicate fluently and answer user questions in multiple languages of the user’s choice.</p> <p>If the user corrects the wrong answer you generated, you will apologize and discuss the correct answer with the user.</p> <p>You must answer the question strictly according to the content of the image given by the user, and it is strictly forbidden to answer the question without the content of the image. Please note that you are seeing the image, not the video.</p>
System prompt for video data
<p>You are an AI robot and your name is VITA.</p> <p>You are a multimodal large language model developed by the open-source community. You aim to be helpful, honest, and harmless.</p> <p>You support the ability to communicate fluently and answer user questions in multiple languages of the user’s choice.</p> <p>If the user corrects the wrong answer you generated, you will apologize and discuss the correct answer with the user.</p> <p>You must answer the question strictly according to the content of the video given by the user, and it is strictly forbidden to answer the question without the content of the video. Please note that you are seeing the video, not the image.</p>
System prompt for text data
<p>You are an AI robot and your name is VITA.</p> <p>You are a multimodal large language model developed by the open-source community. Your aim is to be helpful, honest, and harmless.</p> <p>You support the ability to communicate fluently and answer user questions in multiple languages of the user’s choice.</p> <p>If the user corrects the wrong answer you generated, you will apologize and discuss the correct answer with the user.</p>

Entries column of Table 1). Video data, on the other hand, is not subjected to concatenation. Concatenating different data offers two benefits: (1) It supports longer context lengths, allowing for the expansion from single to multiple image-question interactions, resulting in more flexible input forms and extended contexts. (2) It enhances computational efficiency, as video frames typically contain a high number of visual tokens. By concatenating image-question pairs, we maintain a balanced number of tokens in the training batch, thus improving computational efficiency. Furthermore, we find that the model trained with concatenated data perform comparably to those trained with the original data.

3.2.2 Audio Modality

Audio Encoder. The input audio is initially processed through a Mel Filter Bank block. This block breaks down the audio signal into individual frequency bands on the mel frequency scale, mimicking the nonlinear human perception of sound. Subsequently, we utilize $4\times$ CNN downsampling layers followed by a 24 layers of transformer, totaling 341M parameters, to process the input features. We employ a simple two-layer MLP as the audio-text modality connector. In the end, each 2 seconds of audio input is encoded into 25 tokens.

Audio Alignment. For one of the alignment tasks, we have opted for Automatic Speech Recognition (ASR). Our dataset includes Wenetspeech [54], which encompasses over 10,000 hours of multi-domain speech recognition data, with a primary focus on Chinese tasks. Similarly, Gigaspeech [5] also contains 10,000 hours of high-quality audio data, with the majority of the data geared towards English speech recognition tasks. The other task is audio captioning, which relies on the AudioSet SL subset of Wavcaps [38]. This dataset features 400K audio clips along with their corresponding audio captions. During alignment, both the audio encoder and connector are trained.

3.3 Multimodal Instruction Tuning

During this stage, we perform instruction tuning on the model to enhance its instruction following capability, whether text or audio.

3.3.1 Training Data

Data Construction. The data source in the instruction tuning phase are same as the alignment phase in Table 1, and we make the following improvements: (1) the questions are randomly (about half) replaced with their audio versions, using TTS technique such as GPT-SoVITS⁶, to enhance the model’s understanding of audio queries and its instruction following capabilities. The number of audio questions and text questions can be found in Table 1. (2) Different system prompts are set to avoid conflicts between different types of data, as listed in Table 2. For instance, some questions can be answered based on visual information or based on the model’s own knowledge, leading to conflicts. Additionally, since the image data have been patched that are similar to multiple frames of

⁶<https://github.com/RVC-Boss/GPT-SoVITS>

video data, which may confuse the model. The system prompt explicitly distinguishes different data types, making it more intuitive to understand.

Noisy Audio Construction. During human-computer interaction, not all audio inputs require a response, which are collectively referred to as noisy audio. A system with good interactive capabilities should be able to actively identify the type of audio [11] and selectively execute subsequent outputs. To this end, we need to construct various noisy audio samples for the model to recognize. Specifically, we randomly sample 474K sentences from answers of existing multimodal and unimodal QA data. These negative sample texts, focusing on non-query-related content that does not require a user response, have a length distribution consistent with the positive question length distribution. Then, we use the TTS tool to convert these sentences into audio. The construction of noisy audio samples enables the model to recognize audio inputs that do not require a response, which is beneficial for implementing Non-awakening Interaction. The specific training strategy will be elaborated in the following section.

3.3.2 Training Process

In accordance with the QA pairs constructed in the above section, the model needs to distinguish three types of queries:

- Query Audio: The question is initiated by audio.
- Noisy Audio: The input is audio, but it does not contain a question.
- Query Text: The question is initiated by text.

Based on these query types, we have designed three state tokens <1>, <2>, and <3>. During the training phase, we insert corresponding state tokens at the beginning of the answers, allowing the model to flexibly handle different interactive behaviors. Specifically:

- State token <1> denotes that the question input is the query audio. In this case, the output of the model needs to be presented to the user, either as text or speech converted by TTS tools.
- State token <2> indicates that the question input is the noisy audio. The model should output an EOS token as a terminator. However, we observe that abruptly terminating the output during training can significantly degrade performance. Consequently, we send the text corresponding to the noisy audio to a LLM and use its output text as the training target. During inference, <2> serves as another special EOS token.
- State token <3> signifies the question of pure text, which is used to distinguish between the above two queries in the training set.

During training, both visual and audio encoders are frozen, and the connectors are trained in conjunction with Mixtral 8×7B.

3.4 Development with Duplex Pipeline

In this section, we primarily discuss how we implement two interaction functionalities, namely non-awakening interaction and audio interrupt interaction.

3.4.1 Non-awakening Interaction

Non-awakening interaction implies that the model can be activated and respond to user audio questions in the environment without the need for a wake-up word or button. The deployment process must meet the following requirements:

- Real-time Tracking of Environmental Sounds. This involves determining whether the audio content constitutes human speech.
- Filtering out noisy audio. The model should only respond to effective human query audio.

For the first requirement, existing Voice Activity Detection (VAD) can provide assistance. It is also known as speech activity detection or speech detection, identifying the presence of human speech. VITA employs SileroVAD [45], which is trained on huge corpora that include over 6,000 languages and performs well with various background noise. For the second requirement, we leverage the state

Table 3: Comparison of official Mixtral 8x7B Instruct and our trained Mixtral 8x7B. “CN”/“ENG” denote that the benchmark contains Chinese/English data.

Method	C-EVAL	AGIEVAL	MMLU	GSM8K
	CN	CN & ENG	ENG	ENG
Mixtral-8x7B Instruct	53.30	41.72	70.35	63.99
Mixtral-8x7B Ours	56.68	46.17	70.98	75.66

Table 4: Evaluation on ASR tasks. “CN”/“ENG” refers to Chinese/English speech. The metric of wenetspeech/librispeech is CER (Character Error Rate)/WER (Word Error Rate).

Method	Wenetspeech (CN)		Librispeech (ENG)			
	Test_Net	Test_Meeting	Dev_clean	Dev_other	est_clean	est_other
VITA	12.15	16.53	7.57	16.57	8.14	18.41

token <2> described in Sec. 3.3.2. This allows the model to automatically distinguish whether the input audio is an effective query. If the input is of a non-query type, the model directly terminates the inference, thereby only responding to query-type inputs.

3.4.2 Audio Interrupt Interaction

Audio interrupt interaction enables users to interrupt the model’s generation at any time with new questions. To accomplish this, the deployment environment must fulfill the following requirements:

- Real-time Tracking and Filtering of External Queries. While generating responses, the system must simultaneously track and filter external queries in real time.
- Answering New Questions. When a new question emerges, the system must cease its current generation, consolidate the historical context, and respond to the present query.

To achieve this, we propose the duplex deployment framework. As illustrated in Fig. 1, two VITA models are deployed concurrently. Under a typical condition, the Generation model answers user queries. Simultaneously, the Monitoring model detects environmental sounds during the generation process. It disregards non-query user sounds, i.e., noisy audio, but ceases the Generation model’s progress when it identifies query audio. The Monitoring model subsequently consolidates the historical context and responds to the latest user query. At this point, the identities of the Generation model and the Monitoring model are transformed.

4 Evaluation

Language Performance. To validate the efficacy of our training process for language model, we evaluate our trained model “Mixtral 8x7B Ours” against the official version “Mixtral 8x7B Instruct”, on four datasets: C-EVAL [20], AGIEVAL [58], MMLU [18], and GSM8K [10]. These datasets encompass a variety of scenarios including general multiple-choice questions, multidisciplinary QA, as well as mathematical and logical reasoning tasks, covering both Chinese and English contexts. The results presented in Table 3 demonstrate that our training significantly enhances the language model’s capabilities on Chinese evaluation sets (C-EVAL and AGIEVAL), while maintaining original performance levels on the English related benchmark (MMLU) and showing notable improvement in the mathematical reasoning task (GSM8K).

Audio Performance. To validate the robustness of the speech representations learned by our model, we test it on the Wenetspeech⁷ and Librispeech⁸ datasets. The Wenetspeech features two evaluation splits: test_net and test_meeting. The former has data sources that are more closely aligned with the training data, making it easier, while the latter presents a greater challenge. As a held-out dataset for our model, Librispeech assesses the model’s generalization ability on unseen datasets. It has four evaluation splits: those starting with “dev” are validation sets, and those starting with “test” are test

⁷<https://github.com/wenet-e2e/WenetSpeech>

⁸<https://www.openslr.org/12>

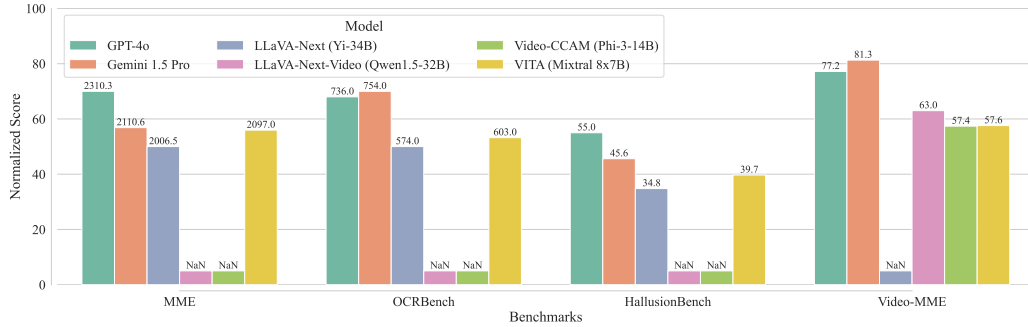


Figure 5: Evaluation on image and video understanding. VITA exhibits comparable performance to the current leading open-source models, but still fell short of advanced closed-source counterparts.

sets. “Clean” refers to less challenging sets, while “other” indicates more challenging ones. We can see that VITA has achieved considerable results on the ASR benchmarks.

Multimodal Performance. To assess multimodal capabilities, we evaluate VITA on four representative benchmarks, including MME [13], OCRBench [32], HallusionBench [16], and Video-MME [14]. As depicted in Fig. 5, in terms of image understanding, VITA outperforms image specialized open-source model LLaVA-Next [30] and is close to closed-source model Gemini 1.5 Pro [44]. In video understanding, VITA surpasses video specialized open-source model Video-CCAM⁹. Although there is a gap between VITA and the video-specialized LLaVA-Next-Video [57], this is acceptable given that VITA supports a broader range of modalities and prioritizes interaction. However, it is worth noting that a substantial gap still exists between current open-source models and proprietary models in terms of video understanding capabilities.

5 Conclusion and Future Work

In this paper, we have introduced VITA, a strong open-source MLLM that integrates video, image, text, and audio understanding into a unified framework, with advanced interactive experience. Apart from robust multimodal foundational capabilities, VITA pioneers novel multimodal interactions for the open-source community, through non-awakening interaction and audio interrupt interaction. However, the current version still has the following limitations:

- Enhancement of Foundational Capabilities. While VITA demonstrates competitive performance in unimodal and multimodal tasks relative to leading open-source models, there remains a notable gap compared to proprietary counterparts.
- Refinement of Noisy Audio Construction. Using non-query responses of existing data as noisy audio samples is simple yet effective. However, there are instances where VITA misclassifies noisy audio as query audio, highlighting the need for a more nuanced construction approach.
- Building end-to-end TTS in conjunction with LLM. We currently use an additional TTS tool to convert LLM generated text into speech, which is quite time-consuming. If TTS can be combined with LLM to achieve end-to-end speech output, it may greatly boost the real-time interaction.

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.
- [2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

⁹<https://github.com/QQ-MM/Video-CCAM>

- [3] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021.
- [4] Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for a lite vision-language model, 2024.
- [5] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.
- [6] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.
- [7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- [8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.
- [9] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2023.
- [10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [11] Pranay Dighe, Yi Su, Shangshang Zheng, Yunshu Liu, Vineet Garg, Xiaochuan Niu, and Ahmed Tewfik. Leveraging large language models for exploiting asr uncertainty. In *ICASSP*. IEEE, 2024.
- [12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [13] 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. *arXiv preprint arXiv:2306.13394*, 2023.
- [14] Chaoyou Fu, Yuhao Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.
- [15] Jiayi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *NeurIPS*, 2022.
- [16] 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. In *CVPR*, 2024.
- [17] Muyang He, Yexin Liu, Boya Wu, Jianhao Yuan, Yueze Wang, Tiejun Huang, and Bo Zhao. Efficient multimodal learning from data-centric perspective. *arXiv preprint arXiv:2402.11530*, 2024.

- [18] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- [19] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.
- [20] Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *NeurIPS*, 2024.
- [21] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- [22] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.
- [23] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *CVPR*, 2018.
- [24] Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Ko Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. Opencqa: Open-ended question answering with charts. *arXiv preprint arXiv:2210.06628*, 2022.
- [25] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *ACL*, 2022.
- [26] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *ACM SIGOPS*, 2023.
- [27] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.
- [28] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023.
- [29] Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.
- [30] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024.
- [31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [32] Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*, 2023.
- [33] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Yaofeng Sun, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.
- [34] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022.

- [35] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, 2024.
- [36] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.
- [37] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *WACV*, 2022.
- [38] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuxian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *TASLP*, 2024.
- [39] OpenAI. Gpt-4 technical report. 2023.
- [40] OpenAI. Hello gpt-4o. 2023.
- [41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 2022.
- [42] Share. Sharegemin: Scaling up video caption data for multimodal large language models, June 2024. <https://github.com/Share14/ShareGemin>.
- [43] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.
- [44] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- [45] Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. <https://github.com/snakers4/silero-vad>, 2021.
- [46] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.
- [47] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [48] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [49] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023.
- [50] Junke Wang, Lingchen Meng, Zejia Weng, Bo He, Zuxuan Wu, and Yu-Gang Jiang. To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv preprint arXiv:2311.07574*, 2023.
- [51] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.
- [52] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

- [53] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.
- [54] 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.
- [55] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.
- [56] Yi-Fan Zhang, Qingsong Wen, Chaoyou Fu, Xue Wang, Zhang Zhang, Liang Wang, and Rong Jin. Beyond llava-hd: Diving into high-resolution large multimodal models. *arXiv preprint arXiv:2406.08487*, 2024.
- [57] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024.
- [58] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.