BAICHUAN-OMNI-1.5 TECHNICAL REPORT

Baichuan Inc.*

https://github.com/baichuan-inc/Baichuan-Omni-1.5

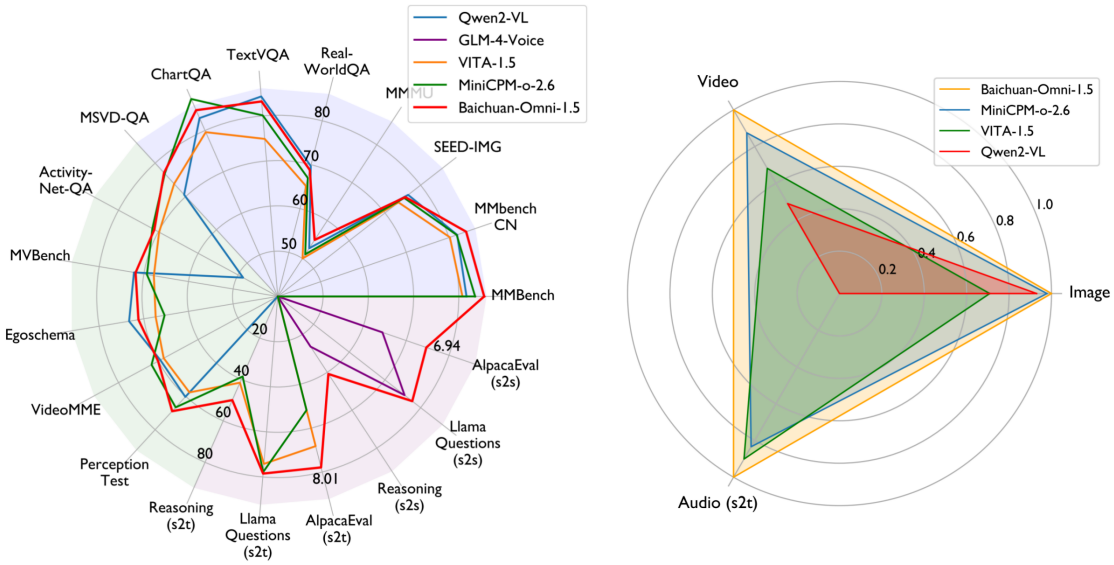

Figure 1: **Evaluation across image, video, and audio modalities. (Left)** Baichuan-Omni-1.5 covers more modalities than Qwen2 VL [143] and outperforms the current leading omni-modal model, VITA-1.5 [46] and MiniCPM-o 2.6[166]. **(Right)** Average scores across benchmarks for all modalities. All the scores are normalized by $x_{\text{norm}} = (x - x_{\text{min}} + 10)/(x_{\text{max}} - x_{\text{min}} + 10)$.

ABSTRACT

We introduce **Baichuan-Omni-1.5**, an omni-modal model that not only has omni-modal understanding capabilities but also provides end-to-end audio generation capabilities. To achieve fluent and high-quality interaction across modalities without compromising the capabilities of any modality, we prioritized optimizing three key aspects. First, we establish a comprehensive data cleaning and synthesis pipeline for multimodal data, obtaining about 500B high-quality data (text, audio, and vision). Second, an audio-tokenizer (Baichuan-Audio-Tokenizer) has been designed to capture both semantic and acoustic information from audio, enabling seamless integration and enhanced compatibility with MLLM. Lastly, we designed a multi-stage training strategy that progressively integrates multimodal alignment and multitask fine-tuning, ensuring effective synergy across all modalities. Baichuan-Omni-1.5 leads contemporary models (including GPT4o-mini and MiniCPM-o 2.6) in terms of comprehensive omni-modal capabilities. Notably, it achieves results comparable to leading models such as Qwen2-VL-72B across various multimodal medical benchmarks.

1 Introduction

Large language models (LLMs) have made great progress in solving various complex tasks [144, 153, 165], such as Qwen2.5 [163] and GPT4 [3]. Based on this, with the seamless connection of visual information and text information, the ability of multimodal large language models (MLLMs) [90, 150, 124, 167] in a wide range of multimodal tasks has

*See Contributions section for full author list.

also made breakthroughs, providing technical support in how machines understand and interact with the world. The advent of advanced proprietary MLLMs like GPT-4o [124], distinguished by their robust multimodal capabilities and inexhaustible interactive experiences, has not only highlighted the essential role of these technologies in real-world scenarios but also redefined the benchmarks for potential advancements in human-computer interaction.

However, current open-source multi-modal large language models (MLLMs) have typically focused on integrating visual and textual modalities, which limits their broader adoption in diverse applications and the quality of user interaction experiences, especially within multimodal dialogue systems. Some studies [45, 176] propose solutions that rely on separate modules for Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) tasks. This approach increases model latency and complexity, thereby limiting its real-time application scenarios. Other recent works have attempted to propose end-to-end solutions. For example, VITA-1.5 [46] and Mini-Omni2 [160] introduce a three-stage training strategy that progressively incorporates information from different modalities. However, these approaches still suffer from modality conflicts, which degrade omni-modal performance compared to unimodal performance, particularly in tasks such as pure text comprehension. Thus, integrating various modalities—such as text, audio, and vision—into a unified model has emerged as a crucial and urgent research topic.

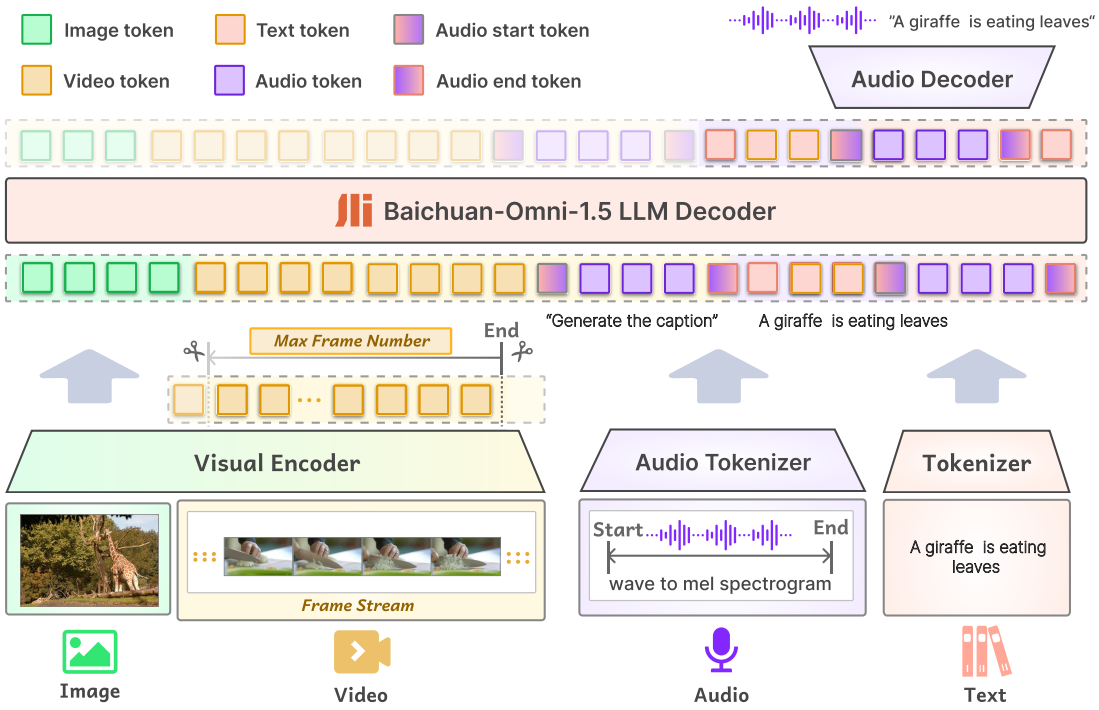

Figure 2: **Architecture of Baichuan-Omni-1.5 .** Our model is designed to process both pure text/audio inputs and combinations of video/image with text/audio. In terms of interactivity, the model initially predicts the start and end of audio inputs. During this period, incoming images and videos are encoded into features and fed into the MLLM in a streaming fashion to calculate attention. The audio features are then fed into the MLLM for inference following the end of the audio input, facilitating streaming input of audio and video.

Compared to the open-sourced counterparts, Baichuan-Omni-1.5 demonstrates significant improvements in the understanding of text, image, audio and video inputs. Notably, the model showcases impressive capabilities in controllable real-time voice interactions and collaborative real-time understanding across various modalities. In addition to its general capabilities, Baichuan-Omni-1.5 stands out as the most outstanding MLLM in the medical domain. This opens up exciting new possibilities for AGI to contribute to the well-being of human society. The architecture of Baichuan-Omni-1.5 is shown in Fig. 2. Based on the evaluation results, we summarize the key advantages and contributions of **Baichuan-Omni-1.5**:

- **Omni-modal Interaction**: Baichuan-Omni-1.5 is designed to process text, image, audio, and video inputs, delivering high-quality text and speech outputs. It is capable of achieving seamless, high-quality cross-modal interactions without compromising the capabilities of any modality.

- **Excellent Vision-Language Capability**: Baichuan-Omni-1.5 scores an average of 73.3 across ten image-understanding benchmarks, which surpasses GPT-4o-mini by an average of 6 points.
- **Unified and Outstanding Speech Capabilities**: We design an 8-layer RVQ audio tokenizer (Baichuan-Audio-Tokenizer) achieves an optimal balance between capturing semantic and acoustic information with 12.5 Hz frame rate, which supports high-quality controllable bilingual (Chinese and English) real-time conversations. At the same time, we have also open-sourced the audio understanding and generation benchmark(**OpenAudio-Bench**) to evaluate the end-to-end capabilities of audio.
- **Leading Medical Image Understanding**: To contribute to AGI research in the medical domain, we open-source a comprehensive medical understanding benchmark: **OpenMM-Medical**. Our model achieves state-of-the-art performance on GMAI-MMBench and OpenMM-Medical. Specifically, on OpenMM-Medical, Baichuan-Omni-1.5 scores 83.8% using a 7B LLM, surpassing Qwen2-VL-72B's score of 80.7%.

2 Related works

2.1 Multimodal Large Language Models (MLLMs)

In recent years, the rapid development of large language models (LLMs) such as Baichuan [162, 36], GPTs [3], LLaMA [40], and Qwen [7, 163] has demonstrated powerful capabilities in natural language understanding and generation. By integrating multimodal alignment and instruction tuning techniques, LLMs have advanced AI into a new phase, where these models can comprehensively understand and generate content across images, audio, and video. The rise of open-source MLLMs has significantly propelled the development of multimodal processing, spurring a new wave of technological innovation. Visual language models like LLaVA [99], Qwen2-VL [150], MiniCPM-V 2.5 [166], DeepSeek-VL2 [156], and Video-LLaVA [96, 188] have made important strides in image and video understanding, cross-modal association, and reasoning. Meanwhile, audio language models such as Qwen-Audio [27, 26], SALMONN [168], and SpeechGPT [176] have shown great potential in tasks such as the simulation of natural dialogue, markedly improving the quality of speech recognition and synthesis. Although most open-source models have progressed in handling images and text, they lag behind proprietary models like GPT-4o in supporting comprehensive multimodal interaction. To further address this gap, we introduce Baichuan-Omni-1.5, an MLLM with robust multimodal interaction capabilities. This model excels in data perception and processing in three modalities (text, audio, and vision), achieving more efficient cross-modal understanding and generation.

2.2 Omni Models with MLLMs

The rapid advancement of MLLMs has propelled the progress of omni models [150, 143], which integrate diverse modalities, such as text, vision, and audio. By processing and fusing information streams from different sensory modalities, these omni models can learn and reason within richer contexts, thereby providing a more comprehensive and profound understanding capability. This not only enhances performance on single-modality tasks, but also opens up new possibilities for cross-modal tasks. Several omni models have significantly improved the system's ability to understand and respond to various forms of information through innovative technical solutions and optimizations of existing methods. EMOVA [18] maintains leading performance in visual-linguistic and speech tasks while introducing emotionally rich omni-modal dialogue capabilities. VITA [45] achieves immediate response to user commands via non-wake-word interactions and audio interruption mechanisms. VITA 1.5 [46] deepens multimodal content generation and analysis by enhancing comprehension of complex scenarios. Mini-Omni [159] supports real-time voice input and output, improving the fluidity of the interaction. Mini-Omni2 [160] combines command interruption techniques to optimize data utilization efficiency and enhance dialogue control flexibility. These studies have substantially advanced multimodal interaction technologies, laying a solid technical foundation for achieving more natural human-machine communication.

2.3 Medical with MLLMs

The development of MLLMs in the medical field has also progressed rapidly, revolutionizing diagnostic processes and medical research by integrating various types of medical data. Technological advancements have enabled MLLMs not only to process complex visual information but also to combine image and text data, offering more comprehensive medical insights. As research has deepened, efforts have shifted toward more effective utilization of cross-modal data. For example, Biomed-GPT [179] stands out for its support of multiple biomedical modalities. Med-Flamingo [119] focuses on few-shot learning for medical visual question answering. LLAVA-Med [79] enhances model performance through extensive use of biomedical image-text pairs. These developments highlight the potential of multimodal integration to improve accuracy in medical tasks. To enhance practical application, many studies have expanded medical

instruction datasets and increased model parameter sizes. For instance, Med-PaLMs [146] and Med-Dr [54] adapt general-purpose multimodal models to meet specific medical needs, thereby improving both precision and clinical applicability. Notably, Med-PaLM fine-tunes the PaLM-E model with millions of samples, optimizing it for medical contexts.

3 Baichuan-Omni-1.5

In this section, we will further provide a comprehensive overview of Baichuan-Omni-1.5 , including high-quality data, model architecture and multi-stage multimodal training strategy.

3.1 High-Quality Multimodal Pretrain Data

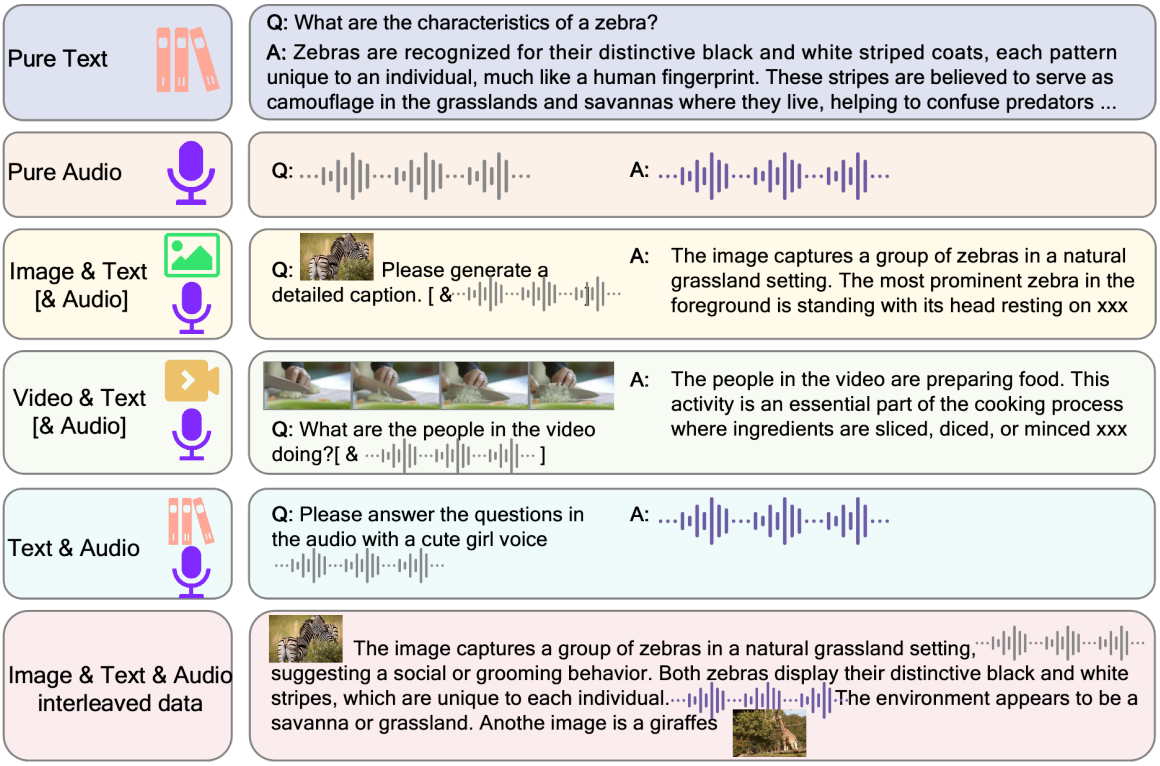

Figure 3: **Pretrain Data illustration of Baichuan-Omni-1.5 .** We construct an extensive omni-modal dataset, including text, image-text, video-text, audio-text, and their interactions. Our collection also contains interleaved image-audio-text and video-audio-text data.

To train our powerful Baichuan-Omni-1.5 , we construct comprehensive and high-quality cross-modal datasets that contain text, image-text, video-text, audio-text, and their interactions. We illustrate our data cases in Fig. 3 and show the statistic in Table 1, Table 2, and Table 3.

Image Data. We divide the image training data into three types: Interleaved image-text data, Caption data, and

Table 1: Detailed statistics of the training data of image pretrain.

Phase	Type	Public Datasets	Public	In-House
Pretrain	Pure-Text	-	-	150.7M
	Caption	[87][68][190][23]	33.2M	49.1M
	Interleaved	[72]	19.1M	28.7M
	OCR	[58]	12.4M	7.8M
Total	-	-	71.3M	238.2M

Question-Answer data. Specifically, we first collect various open-source datasets, including DenseFusion-1M [87], Synthdog [68], DreamLIP [190], InternVL-SA-1B-Caption [22, 23], PIN-14M [149], MINT-1T [6], LAION-5B [130], OBELIC [72], Cauldron [75], Monkey [94], ArxivQA [84], TGDoc [152], MM-Self-Instruct (Train split) [184], MMTab [191], AnyWord-3M [147], TinyChartData [58], and DocStruct4M [58], etc. These publicly available open-source datasets originate from a wide variety of sources. Thus, we carefully design sampling techniques to construct different data ratios within our data pipeline.

Second, to improve data diversity and improve model performance, we have the following two strategies for synthesizing image data: 1) We utilize in-house collected books and papers and parse them to generate Interleaved image-text, OCR data, and Chart data. These data are highly complete, specialized, and knowledge intensive. 2) Following [19], we also train a dedicated caption model that can produce desired image captions, such as ocr hints. These captions offer in-depth descriptions of the image content. 3) Currently, a large amount of open source dataset is mainly in English. In order to avoid the decline of the Chinese ability of the model, we synthesize a large amount of Chinese captions and interleaved data.

Video Data. The video dataset consists of a wide variety of publicly accessible resources that cover numerous tasks such as video classification [172, 1], action recognition [57], and temporal localization [154]. The video-text sources can be divided into video caption data and video question-answering (QA) data.

For video caption data, we utilize the open-sourced ShareGPT4Video [20], Koala [151], and WebVid [8]. Besides, we employ GPT-4o to produce high-quality captions for videos collected from YouTube. For video QA data, we collect ActivityNet-QA (Train split) [169], VideoChatGPT-Plus [108], ShareGemini [132], and NExTVideo [188].

Table 2: Detailed statistics of the training data of video pretrain.

QA Type	Dataset Name	Public Datasets	Questions
Description	Synthetic Data	-	300K
	ShareGPT-4o	[29]	2K
	Koala	[151]	30M
QA	Synthetic Data	[81][83][158]	164K
	VideoChatGPT-Plus	[108]	318K
	ShareGemini	[132]	205K
Total	-	-	31M

Audio Data. Audio data can be broadly categorized into two primary types: audio understanding data and audio generation data. Audio understanding data includes Automatic Speech Recognition (ASR), Audio Question Answering (AQA), Speech-to-Text Translation, and Audio-Text Interleave data. Audio generation data encompasses Text-to-Speech (TTS), Interleaved Text-to-Speech data, and pure audio data. Interleaved data consists of alternating text and audio modalities, segmented by punctuation marks to facilitate cross-modal knowledge transfer. The interleaved aligned generation data composed of fully aligned text and audio content, designed to enhance the model's ability to generate audio tokens under text supervision. The audio-text paired data (e.g., ASR and TTS data) improve the performance on fundamental speech tasks. Pure audio data, on the other hand, enhances the capability to independently process audio modalities.

Table 3: Detailed statistics of the training data of audio pretrain.

Type	Task	Data Format	Hours (k)
Audio Understanding	Automatic Speech Recognition (ASR)	`<prompt, audio, transcript>`	185
	Audio Query Answer (AQA)	`<prompt, audio, response>`	21
	Speech-to-Text Translation (S2TT)	`<prompt, audio, translated_text>`	15
	Audio-Text Interleaved (INTLV)	`<audio_1, text_2, audio_3, text_4, ...>`	393
Audio Generation	Text-to-Speech (TTS)	`<text, audio>`	51
	Interleaved Text-to-Speech (ITTS)	`<text_1, audio_1, text_2, audio_2, ...>`	142
	Pure Audio	`<audio>`	80
Total	-	-	887

Text Data. To construct a high-quality text corpus, we aggregated data from a wide range of sources, including web pages, books, academic papers, code, and other sources. Adhering to established data processing guidelines from earlier research [36, 104], we adopted a rigorous selection methodology aimed at boosting both the diversity and the quality of

our text corpus. This diversity ensures that the training corpus encompasses a broad spectrum of topics and linguistic styles, making it suitable for diverse applications. Meanwhile, our high-quality processing techniques are designed to eliminate redundancies and filter out noise, thereby enriching the dataset's informational density and overall utility. Finally, we obtain 150.7 million entries of pure text data.

Cross-Modal Interaction Data. To enhance the cross-modal interaction capabilities of our model, we synthesized a series of cross-modal interaction datasets encompassing image-audio-text and video-audio-text formats. The source of the image-text data comprises two types: image-text caption data and image-text interleaved data. Specifically, textual data are first segmented at the sentence level. Then, a random quarter of the text was converted into audio elements using our in-house text-to-speech (TTS) interface. Subsequently, we utilize the generated audio elements to replace the corresponding textual sentences in the original image-text data. This methodology facilitates an enriched cross-modal interaction framework by integrating diversified audio elements into the existing textual content. Our audio data contains 44 distinct voice types, ensuring a diversity in intonation. This setup is complemented with task prompts, such as "Please listen to the following audio describing the content of the image. Your task is to supplement additional information by combining the audio with the image upon completion of listening", aiming at predicting the remaining three-quarters of the textual descriptions. For the video-text data set, the audio components are directly extracted from the orignal videos to serve as the cross-modal audio element. In total, we generate 100B tokens of data for cross-modal interaction.

3.2 Model Architecture

Our Baichuan-Omni-1.5 is a unified omni-modal model composed of the visual branch, the audio branch and a pre-trained large language model (LLM) backbone, which supports text, audio, visual input as well as end-to-end text and audio output.

3.2.1 The Visual Branch

Like the current mainstream MLLM, the visual branch is designed to process image and video input into visual tokens, which are fed into the LLM along with the text tokens. We utilize NaViT [33] as the visual encoder architecture and load the weights of Qwen2-VL [150] as the initialization. NaViT can dynamically process images and videos of arbitrary resolution and aspect ratio. We then apply a visual projector composed of a two-layer MLP to compress the visual feature by a 2×2 factor, which strikes a balance between performance and efficiency.

3.2.2 The Audio Branch

The audio branch extends the LLM to enable end-to-end speech input and output. This is achieved by introducing the Baichuan-Audio-Tokenizer and a flow matching based decoder [98], which are responsible for transforming audio signals into discrete tokens and decoding audio tokens into speech waveform, respectively. We show the detail in Fig. 4.

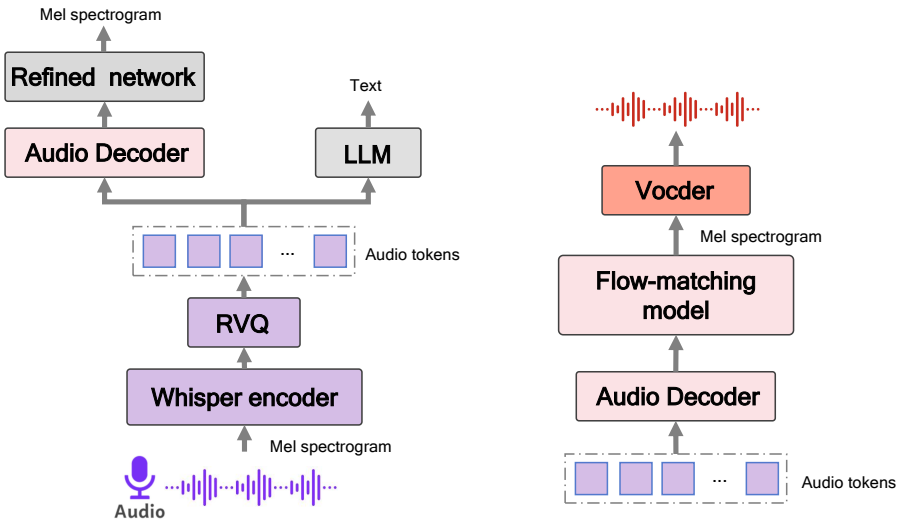

Figure 4: Audio tokenizer and audio decoder based on flow matching model.

The Baichuan-Audio-Tokenizer is based on Residual Vector Quantization (RVQ) [31] and multi-objective training [86, 116], with 12.5 Hz frame rate. After extracting high-level features from Mel spectrogram features using Whisper Large Encoder [127], the residual convolutional network performs downsampling to obtain low frame rate sequence features. An 8-layer residual vector quantizer is then used to quantize these features to generate audio tokens. These tokens are subsequently fed into both an audio decoder and the pretrained LLM to perform Mel spectrogram reconstruction and transcript prediction, respectively. The Audio Decoder adopts a structure symmetrical to the Whisper Encoder and employs a multi-scale Mel loss [116] to enhance the quality of sound reconstruction. During training, the parameters of pretrained LLM are fixed to ensure the semantic alignment between the audio tokenizer and the text space. In addition to traditional tasks such as ASR, AQA and S2TT, a proportion of interleaved text-audio data is incorporated to improve the ability of the VQ module to model complex contextual scenarios.

To further enhance the quality and perceptual fidelity of synthesized audio, the audio decoder module is refined using a flow matching model. Following the designs of Matcha-TTS [115] and CosyVoice [38], the U-Net includes a single down-sampling block, a single up-sampling block, and 12 intermediate blocks. Specifically, the flow-matching decoder is trained on 24 kHz audio data to generate target Mel spectrograms, which are then converted into speech waveforms using a HiFi-GAN [70, 38] vocoder[2].

3.3 Omni-Modal Training Strategies

In this section, we will further illustrate the omni-modal training strategies that cross image, audio, video and text data, which can gradually align different modalities into the language space. We show the training pipeline of Baichuan-Omni-1.5 in Fig. 5.

3.3.1 Image-Text Pretrain

The Image-Text Pretrain stage extends an LLM to process and understand visual input using 300 billion image-text samples, which can be divided into two stages.

- **Stage I:** In the first stage, we train the visual projector to establish the initial alignment between image representations and text using open source image captioning data, such as the LAION-5B dataset[130]. During this phase, we freeze the LLM and the visual encoder, only training the visual projector with a learning rate of $1e-3$.
- **Stage II:** In the second stage, we unfreeze the visual encoder and LLM to promote better alignment between image and text representations. In detail, we train the LLM and the visual projector with a learning rate of $1e-5$, and train the visual encoder with a lower learning rate of $1e-6$. We use public- and in-house image text data that contain interleaved data and image caption data to enhance visual-language performance. Specifically, we collect and caption high-quality ocr data and chart data to enhance the text/chart recognition and understanding ability at this stage. In addition, we use high-quality pure text data, which accounts for 40% of the total data, to better maintain the original capabilities of the language model.

3.3.2 Image-Audio-Text Pretrain

The Image-Audio-Text Pretrain stage extends an LLM pre-trained on visual data to understand audio data in an end-to-end manner using 887k hours of speech-text data, which incorporates our Baichuan-Audio-Tokenizer, a newly introduced audio embedding layer and an independent audio head.

Specifically, the audio tokens from Baichuan-Audio-Tokenizer are first transformed into audio embeddings through audio embedding layers. The audio LLM alternately generates aligned text tokens and audio tokens, with a special token enabling modality switching between text and audio. The generated audio tokens are processed by the independent audio head, which is designed based on prior works [76, 32] and consists of 3 layers of depth transformers and 8 classification heads.

To mitigate conflicts arising from the significant differences between speech and text features, we refer to previous works [69, 174] and utilize a method of interleaving audio and text data for pretraining. Additionally, a two-stage training strategy is adopted to preserve the original LLM's textual knowledge while integrating audio modality effectively.

- **Stage I:** During the first stage, we freeze the parameters of LLM, visual modules and audio tokenizer, and only the parameters of the audio embedding layer and the audio head are updated with a learning rate of $1e-4$. We use audio data including ASR, TTS, INTLV and ITTS data in this stage.

[2]https://www.modelscope.cn/models/iic/CosyVoice2-0.5B

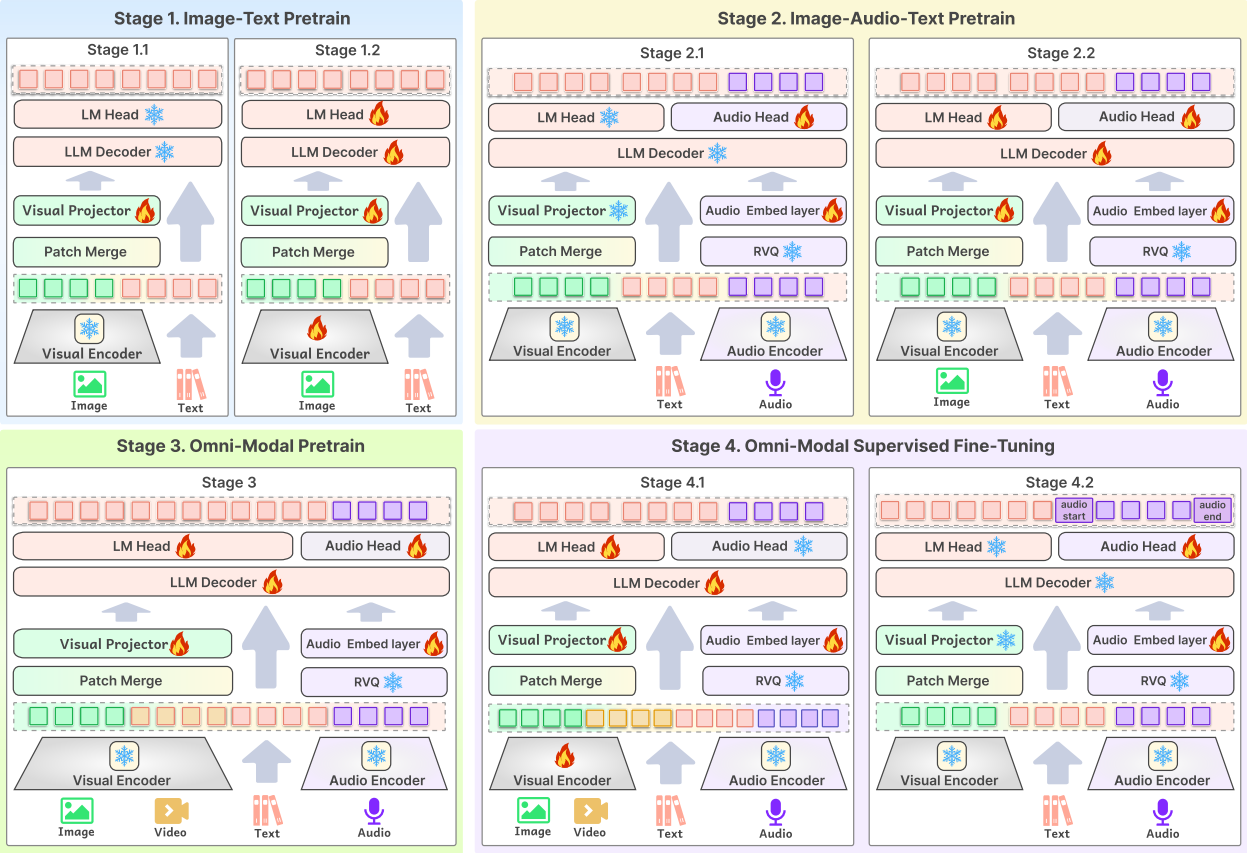

Figure 5: **Training Pipeline of Baichuan-Omni-1.5** . The pretraining phase is divided into three stages to incrementally incorporate vision and audio into the LLM while relieving modality conflicts. Stage 1 focuses on image-text training, which extends an LLM to process and understand visual input. Stage 2 extends an LLM pre-trained on visual data to understand audio input in end-to-end manner by incorporating our Baichuan-Audio-Tokenizer, a newly introduced audio embedding layers and an independent audio head. Stage 3 focuses on training Baichuan-Omni-1.5 using high-quality cross-modal interaction datasets encompassing image-audio-text and video-audio-text format, and extends the maximum sequence length to 64k to support long audio and video stream. Stage 4 enhances the model's instruction following and audio capabilities through supervised fine-tuning with omni-modal data. Stage 4.1: Freeze the Audio Head using omni-modal understanding data to boost modality interactivity and multitasking comprehension. Stage 4.2: Activate only the Audio Head and Audio Embed layer, with audio generation data to improve speech generation capabilities.

- **Stage II:** In the second stage, training is extended to all parameters except for the visual encoder and the audio tokenizer with a lower learning rate of $1e-5$. Specifically, We use audio data, image data and pure text, accounting for 0.2, 0.4, and 0.4, respectively, which can better improve audio capabilities while maintaining visual and language capabilities.

3.3.3 Omni-Modal Pretrain

Based on the visual and audio capabilities acquired from previous pretraining stages, we continue to train all the parameters using high-quality cross-modal interaction datasets encompassing image-audio-text and video-audio-text format, and we extend the maximum sequence length to 64k to support long voice and video streams. Specifically, the input video frames are sampled at a rate of 1 frame per second, with a maximum of 64 frames per video. Each input frame is resized to a maximum resolution of 560×1120 pixels to maintain optimal quality and detail. This thoughtful configuration strikes a balance between performance and efficiency, facilitating effective model training while managing the computational load. This training process uses a low learning rate of $4e-6$ to refine alignment with language modality and cross-modal interaction.

3.4 Multimodal Supervised Fine-Tuning

In this section, we describe the omni-modal supervised fine-tuning (SFT) phase, which is designed to enhance the model's capability to follow complex omni-modal instructions across a range of tasks. We collect comprehensive datasets encompassing open-source, synthetic, and in-house annotated data. These datasets span multiple tasks and contain approximately 17 million data pairs across various modalities, including text, audio, image-text, video-text, and image-audio combinations. Detailed information regarding the types and quantities is provided in Table 4.

Table 4: **Omni-modal SFT data statistics for Baichuan-Omni-1.5** . Here we summarize the category and quantities of our SFT dataset.

Category	Text	Image	Video	Audio	Image-Audio
Quantity	400K	16M	100K	282K	60K

Table 5: **Image SFT data for Baichuan-Omni-1.5** . This table summarizes the iamge SFT dataset categories, their sources, and proportions for various tasks.

Scene	Source	Proportion
GeneralQA	Leopard-Instruct [62], LLaVA-OneVision-Data [77], MMInstruct-GPT4V [100], the Cauldron [73], GeoGPT4V-1.0 [15], MMDU [103], Lova3 [189], CaD-Inst [14], VisionArena-Battle [25], Q-Instruct-DB [155], MultipanelVQA [42], ConMe [59], FABAInstruct [91], ScienceQA [129], MapQA [16], Others	32.26%
OCR	MathWriting [49], WebSight [74], ST-VQA [11], GQA [61], HME100K [170], UberTextQA [10], OCR-VQA [118], TallyQA [2], SlideVQA [140], VizWiz [10], NorHand-v3 [141], LLaVAR [187], Textualization [41], PViT [183], Others	26.51%
Graphical	DVQA [65], TinyChart [180], Chart2Text [67], ArxivQA [85], ChartLlama [53], InfographicVQA [113], FlowVQA [138], MultiChartQA [195], ChartGemma [112], UniChart [110], TAT-DQA [194], PlotQA [117], FigureQA [66], MMTab [192], Others	9.04%
Mathematics	MathV-360K [133], Geo170k [47], R-COT [35], A-OKVQA [131], Super-CLEVR [95], CLEVR-Math [97], TabMWP [106], GeoQA+ [5], MAVIS [182], Iconqa [107], UniGeo [17], PUMA_VarsityTutors [196], Others	10.31%
Spatiotemporal	CCTSDB2021 [178], SODA10M [52], EmbSpatial [37], LLaVA-VSD [63], SpatialSense [164], SpatialMM [134], Whatsup [12], VSR [181], SpatialSense [164], Others	2.63%
Captioning	TextCaps [135], MMsci [93], Synthetic Data, Others	8.23%
Medical	PubMed [114], HAM10000 [145], PMC-VQA [185], PathVQA [55], AIROGS [30]), MedFMC [148], Kvasir-VQA [48], IU X-ray [34], VQA-RAD [71], DME VQA [142], and other specialized medical datasets	11.02%

3.4.1 Image Data

Our Image SFT dataset comprises millions of examples collected from a wide range of public sources. It covers diverse visual domains, including natural scenes, structured documents, graphical data (e.g., charts), and specialized medical imagery. The data spans multiple languages, with Chinese and English as major components. It encompasses both single-image and multi-image tasks, featuring a mix of real-world photographs and synthetically generated visuals. Data quality is ensured through rigorous filtering, GPT-based regeneration of low-quality answers, and manual validation. Dataset proportions are carefully allocated to ensure comprehensive coverage of competencies.

Table 5 categorizes the image SFT datasets based on task-specific competencies:

GeneralQA Tasks: Datasets, such as Leopard-Instruct [62] and MMInstruct-GPT4V [100], are used to train the model, enabling it to understand and describe images. Notably, LLaVA-OneVision-Data [77] and the Cauldron data [73] encompass data of various types from multiple sources, placed under the umbrella of composite data, utilizing portions of their data beyond proprietary capabilities. Many of the datasets within the following specialized capabilities also originate from them.

OCR Tasks: For OCR tasks, the model needs to accurately recognize and understand the textual content within images, and further, to respond based on this understanding. We have collected a substantial amount of OCR data, such as NorHand-v3 [141], MathWriting [49], WebSight [74], HME100K [170], UberTextQA [10], OCR-VQA [118], TallyQA [2], SlideVQA [140]. It is worth noting that we have found the proportion of OCR data significantly impact the overall performance of the model. It necessitates multiple attempts at adjustment in conjunction with different models. Ultimately, we have set the OCR data to constitute 26.51% of all image data.

Graphical Tasks: Tasks related to data visualisation require the model to not only recognize content within graphs but also perform complex reasoning. Comprehensive and diverse chart data, such as DVQA [65], ArxivQA [85], TinyChart [180], Chart2Text [67], FlowVQA [138], MultiChartQA [195], UniChart [110], are selected, processed, filtered, and sampled.

Mathematics Tasks: Mathematical capabilities determine the upper limit of the model's ability to handle complex tasks. We have collected some of the most recently open-sourced, renowned, and beneficial datasets, including MathV-360K [133], Geo170k [47], R-COT [35], which contain a large number of Chain of Thought (CoT) processes and detailed calculation steps, thereby enhancing the model's mathematical and reasoning abilities.

Spatiotemporal Tasks: The CCTSDB2021 [178], EmbSpatial [37], SpatialMM [134], and VSR [181] datasets encompass real-world contextual reasoning tasks, covering object interactions and spatial relationships across various environments, from traffic scenarios to natural landscapes. These datasets provide a robust foundation for improving model performance in practical applications.

Captioning Tasks: TextCaps [135] and synthetic datasets include paired captions for both natural and synthetic images. In contrast, the mmsci dataset [93] integrates multimodal scientific literature, enabling models to learn how to describe complex technical charts and experimental results. Furthermore, synthetic datasets expand the diversity of the training corpus by simulating a wide array of potential visual scenarios.

Medical Tasks: PubMed [114] is an extensive database of medical literature that provides rich textual references for model training. Specialized medical datasets such as dermatology datasets (e.g., HAM10000 [145]), pathology datasets (e.g., PathVQA [55]), ophthalmology datasets (e.g., AIROGS [30]) contribute annotations with specialized knowledge.

3.4.2 Video Data

To enhance the model's ability to address video understanding challenges in complex real-world scenarios, we initially collected a substantial amount of open-source data. These include datasets on general video understanding [120, 139, 132, 20], action recognition [13, 136], temporal understanding [158], and other related tasks. The collected video data were systematically classified and analyzed to identify task types. To ensure balanced representation, we adjusted the proportions of each task type, resulting in a curated collection of 100K high-quality video SFT datasets. These datasets cover a wide range of tasks, including video classification across diverse scenarios, action recognition, and temporal localization. Furthermore, we utilized GPT-4o for fine-grained classification of all video data. The distribution of the video SFT data was meticulously adjusted based on factors such as scene type, task difficulty, answer accuracy, and video quality.

3.4.3 Audio Data

The audio SFT data are derived from a large collection of textual instructions. High-quality instructions are selected using a filtering strategy based on instruction type, diversity, and overall quality. Audio instructions are synthesized using a curated dataset of 10,000 distinct voice tones. Corresponding text responses are generated and segmented at natural conversational pauses before being converted into audio using the designated voice tones.

To ensure the quality of the synthesized audio, Automatic Speech Recognition (ASR) is applied to the generated audio files. The ASR outputs are compared against the original text to validate quality. This process results in the creation of high-quality end-to-end conversational datasets. Synthesized audio files with errors are added to the Text-to-Speech (TTS) dataset, while cases with ASR errors are incorporated into the ASR training dataset. This iterative approach of incorporating challenging examples enhances both TTS and ASR performance.

Special attention is required to address cases where text-to-audio conversion makes the original textual response unsuitable as an audio reply. This issue arises due to differences in tone, speed, and expression between text and audio. Some textual content may fail to convey the intended meaning or introduce ambiguity when converted into audio. Consequently, careful review and adjustment of such cases are essential during the generation process. This ensures that the synthesized data accurately reflects real-world voice interaction scenarios, enhancing data reliability and improving the model's practical applicability.

4 Experiment

In this section, we evaluate a range of MLLMs and LLMs, including proprietary models (GPT4o mini and GPT4o [124]), open-source general models (MAP-Neo [177], Qwen1.5-Chat [7], Llama3-Instruct [4], OLMo [50]), and open-source omni-modal models (VITA-1.0 [45], VITA-1.5 [46], Baichuan-Omni [89], and MiniCPM-o 2.6 [166]), across text, image, video, audio, medical, and omni benchmarks. Note that unless otherwise specified, the parameter numbers marked in brackets in the experimental tables indicate the parameter numbers of the LLM. Besides, unless otherwise specified, the results GPT-4o-mini and other open-source omni-modal models (VITA-1.5 and MiniCPM-o 2.6) are reproduced by ourselves with the same settings for fair comparison.

4.1 Performance in Pure Language Tasks

Evaluation Benchmarks. To assess the knowledge and reasoning capabilities of Baichuan-Omni-1.5, we utilize 4 comprehensive benchmarks, incuding MMLU [56], CMMLU [80], AGIEval [193], C-Eval [60] and GAOKAO-Bench [186]. MMLU comprises 57 specially designed tasks, consisting of multiple-choice questions, spanning various domains of knowledge including the humanities, social sciences, and natural sciences. CMMLU is specifically tailored to evaluate the complex knowledge and reasoning abilities of LLMs within the context of Chinese language and culture. AGIEval aims to assess the general cognitive and problem-solving capabilities of foundational models, using official, public, and qualification tests designed for human participants. C-EVAL offers a comprehensive Chinese evaluation suite intended to gauge the advanced knowledge and reasoning skills of LLMs in a Chinese context, which encompasses 13,948 multiple choice questions across 52 distinct disciplines ranging from the humanities to science and engineering. GAOKAO-Bench is an evaluation framework that assesses large models' language and reasoning skills using questions from China's National College Entrance Examination (GAOKAO) from 2010 to 2022. It includes a total of 2,811 questions covering a wide range of academic disciplines. For all evaluations, we employ zero-shot measurements.

Table 6: **Results on comprehensive pure text benchmarks.** $*$: Officially reported results. $\diamondsuit$: Retrieved results from official leaderboard or recent papers. Other unlabeled results are reproduced by ourselves.

Model	Comprehensive Tasks				
	MMLU (Acc.)	CMMLU (Acc.)	AGIEval (Acc.)	C-Eval (Acc.)	GAOKAO (Acc.)
Proprietary Models					
GPT-4o	**88.0**$^\diamondsuit$	**78.3**$^\diamondsuit$	**62.3**$^\diamondsuit$	**86.0**$^\diamondsuit$	-
GPT-4o-mini	82.0	67.6	52.2	63.6	70.8
Open-source Models (Pure text)					
MAP-Neo (7B)	58.2	55.1	33.9	57.5	-
Qwen1.5-Chat (7B)	61.5	68.0	39.3	68.8	-
Llama3-Instruct (8B)	67.1	51.7	38.4	50.7	-
OLMo (7B)	28.4	25.6	19.9	27.3	-
Open-source Models (Omni-modal)					
VITA (8x7B)	71.0*	46.6	46.2*	56.7*	-
VITA-1.5 (7B)	71.0	75.1	47.9	65.6	57.4
Baichuan-Omni (7B)	65.3	72.2	47.7	68.9	-
MiniCPM-o 2.6 (7B)	65.3	63.3	50.9	61.5	56.3
Baichuan-Omni-1.5 (7B)	72.2	75.5	54.4	73.1	**73.5**

Results. As shown in Table 6, Baichuan-Omni-1.5 demonstrates impressive performance on pure-text benchmarks, particularly when compared to open-source LLMs that focus solely on the language modality. For instance, on the general MMLU benchmark, Llama3-Instruct achieves 67.1%, while Baichuan-Omni-1.5 reaches 72.2%. The success of Baichuan-Omni-1.5 in the language modality can largely be attributed to our adjustments in the training strategy and the balanced ratio of multimodal training data, where a certain proportion of pure text data is maintained. The results demonstrate that our data synthesis and balancing methods, along with the multi-stage training strategy, can effectively address the issue of performance degradation in pure language tasks during multimodal training. Besides, compared to the latest open-source multimodal model MiniCPM-o 2.6, Baichuan-Omni-1.5 demonstrates a substantial advantage in Chinese benchmarks, such as CMMLU (63.3% v.s 75.5%) and C-Eval (61.5% v.s 73.1%), and largely surpasses

MiniCPM-o 2.6 in general benchmarks, MMLU (65.3% v.s 72.2%) and AGIEval (50.9% v.s 54.4%). These results show that compared to the current omni-modal models, which have a degenerate ability of text understanding after training with non-text modal data, while our model's ability to understand pure text remains strong.

4.2 Performance in Image Understanding Tasks

Baselines. We utilize the following baselines: proprietary models (GPT4o mini and GPT4o [124]), open-source models for vision-language (MiniCPM-Llama3-V 2.5 [166] and Qwen2-VL [143]), and open-source models for omni-modal (VITA-1.0 [45], VITA-1.5 [46], Baichuan-omni [89], and MiniCPM-o 2.6 [166]).

Evaluation Benchmarks. Here we perform evaluation on representative vision-language benchmarks to assess the image perception and understanding capabilities of Baichuan-Omni-1.5. The following benchmarks are utilized: MMBench-EN, MMBench-CN [101], SEEDBench [78], RealWorldQA [157], MMMU [171], MathVista [105], TextVQA [137], OCRBench [102], ChartQA [111], and HallusionBench [51]. To ensure consistent and reproducible evaluation results, we consistently utilize VLMEvalKit [39] across all assessments. All evaluations are executed in a zero-shot manner, adhering rigorously to the initial settings of the models. This setting guarantees that comparisons between different models and benchmarks remain unbiased and fair.

Table 7: **Results on Multi-choice benchmarks and Yes-or-No benchmarks.** $*$: Officially reported results. $\diamondsuit$: Retrieved results from official leaderboard or recent papers. Other unlabeled results are reproduced by ourselves.

Model	Multi-choice & Yes-or-No Question				
	MMBench-EN (Acc.)	MMBench-CN (Acc.)	SEED-IMG (Acc.)	MMMU (val) (Acc.)	HallusionBench (Acc.)
Proprietary Models					
GPT-4o	83.4$^\diamondsuit$	82.1$^\diamondsuit$	-	**69.1**$^\diamondsuit$	**55.0**$^\diamondsuit$
GPT-4o-mini	77.7	76.9	72.3	59.3	45.8
Open-source Models (Vision-language)					
Qwen2 VL (7B)	81.7	81.9	**76.5**	52.7	50.6*
MiniCPM-Llama3-V 2.5 (8B)	76.7	73.3	72.4	45.8*	42.5
Open-source Models (Omni-modal)					
VITA (8x7B)	74.7	71.4	72.6	45.3	39.7*
VITA-1.5 (7B)	80.8	80.2	74.2	50.8	44.8
Baichuan-Omni (7B)	76.2	74.9	74.1	47.3	47.8
MiniCPM-o 2.6 (7B)	83.6	81.8	75.4	51.1	50.1
Baichuan-Omni-1.5 (7B)	**85.6**	**83.6**	75.7	53.9	49.7

Results. As shown in Table 7 and Table 8, obviously, our model outperforms the latest open-source model, VITA-1.5 and MiniCPM-o 2.6, on most of the benchmarks. For example, compared with the recent MiniCPM-o 2.6, our model has higher performance in six out of ten benchmarks, which shows that our omni-modal model is already at the forefront of open source models. Besides, compared to other non-omni-modal models, Baichuan-Omni-1.5 achieves comparable or even superior performance. For example, compared with MiniCPM-Llama3-V 2.5, our model demonstrates better results across the majority of visual question answering (VQA) tasks, including MMBench, SEED-IMG, MME, HallusionBnech and MMMU which requres expert-level perception and reasoning. In general, compared with Qwen2-VL-7B, our model has comparable performance on various image understanding benchmarks. Our model gets better performance on MMBench-CN (81.9% v.s 83.6%), MMMU (52.7% v.s 53.9%), MathVista-mini (58.2% v.s 63.6%), and ChartQA (83.0% v.s 84.9%). In addition, it is worth noting that on MMBench-EN/CN and OCRBench, our model has surpassed the closed-source model GPT4o.

4.3 Performance in Video Understanding Tasks

Baselines. We compare Baichuan-Omni-1.5 with the following baselines: proprietary models (Gemini 1.5 Pro [128], GPT 4V [123], GPT-4o-mini, and GPT-4o [124]), open-source models for vision-language (Qwen2-VL [143], AnyGPT [175], VideoLLaMA 2 [24], VideoChat2 [82], LLaVA-NeXT-Video [188], and Video-LLaVA [96]), and open-source models for omni-modal (VITA-1.0 [45], VITA-1.5 [46], Baichuan-omni [89], and MiniCPM-o 2.6 [166]).

Table 8: **Results on image VQA benchmarks.** $*$: Officially reported results. $\diamondsuit$: Retrieved results from official leaderboard or recent papers. Other unlabeled results are reproduced by ourselves.

Model	Visual Question Answering				
	RealWorldQA (Acc.)	MathVista-mini (Acc.)	TextVQA (val) (Acc.)	ChartQA (Acc.)	OCRBench (Acc.)
Proprietary Models					
GPT-4o	**75.4**$^\diamondsuit$	63.8$^\diamondsuit$	-	85.7$^\diamondsuit$	73.6$^\diamondsuit$
GPT-4o-mini	66.3	53.4	66.8	-	77.4
Open-source Models (Vision-language)					
Qwen2 VL (7B)	69.7	58.2*	**84.3***	83.0*	84.5*
MiniCPM-Llama3-V 2.5 (8B)	63.5	54.3*	76.6	72.0	72.5
Open-source Models (Omni-modal)					
VITA (8x7B)	59.0	44.9*	71.8	76.6	68.5*
VITA-1.5 (7B)	66.8	**66.5**	74.9	79.6	73.3
Baichuan-Omni (7B)	62.6	51.9	74.3	79.6	70.0
MiniCPM-o 2.6 (7B)	67.7	64.6	80.1	**87.6**	**89.7***
Baichuan-Omini-1.5 (7B)	68.8	63.6	83.2	84.9	84.0

Evaluation Benchmarks. To assess the video understanding capabilities of Baichuan-Omni-1.5, we conduct a thorough evaluation on general video understanding tasks (General VQA) and open-ended video question answering (Open-ended VQA) tasks. For general video understanding tasks, the following benchmarks are utilized: Perception-Test [126], MVBench [82], VideoMME [44], and EgoSchema [109]. We report top-1 accuracy for all benchmarks. For open-ended video question answering tasks, we utilize ActivityNet-QA [169] and MSVD-QA [161] as evaluation benchmarks. We utilize GPT4-0125-preview to assess the quality of the response snippets. Specifically, we use GPT4-0125-preview to provide a "Yes-or-No" decision on the correctness of answers and a rating scaled from 0 to 5. We report the percentage of "Yes" responses as Accuracy and the average rating as Score.

Results. As shown in Table 9 and Table 10, our Baichuan-Omni-1.5 performs excellently on the two video tasks. **1) Video General VQA.** Baichuan-Omni-1.5 demonstrates comparable performance over proprietary models on benchmarks like Egoschema and VideoMME, and achieves strong performance across open-source multimodal models, which shows comprehensive video understanding capabilities of Baichuan-Omni-1.5. Specifically, on the four general VQA benchmarks, Baichuan-Omni-1.5 gets 63.8% average score and the recent omni-modal model VITA-1.5 and MiniCPM-o 2.6 achieve 56.3% and 59.8%, respectively. **2) Open-ended VQA.** Baichuan-Omni-1.5 demonstrates SOTA performance (both Accuracy and Score) on ActivityNet-QA and MSVD-QA across all open-source general models and omni-modal models, such as the most recent omni-modal models MiniCPM-o 2.6 and Qwen2 VL, and outperforms the proprietary model GPT-4o-mini (62.1%) on ActivityNet-QA.

4.4 Performance in Audio Understanding Tasks

Baselines. We compare Baichuan-Omni-1.5 with the following baselines: proprietary model (GPT-4o-Audio [124]), open-source voice model (GLM-4-Voice [173]), and open-source models for omni-modal (VITA-1.5 [46], MiniCPM-o 2.6 [166]).

Evaluation Benchmarks. To assess the audio understanding capabilities of Baichuan-Omni-1.5, we have built and open-sourced an OpenAudioBench and use GPT-4o [124] to evaluate the results, including Reasoning QA(self-constructed), Spoken Llama Questions [121], Web Questions [9], TriviaQA [64], and AlpacaEval [88]. For AlpacaEval, we select two subsets `helpful base` and `vicuna` from the original AlpacaEval dataset and remove questions related to math and code. This process follows Llama-Omni [43], with the aim of obtaining questions more suitable for speech scenarios, and the final AlpacaEval benchmark in our report comprises 199 questions in total. Considering the substantial size of the Web Questions and TriviaQA datasets, a full evaluation is impractical. Therefore, we randomly sample 1,000 questions from each original dataset. The instructions for these three benchmarks were synthesized using our TTS model.

Table 9: **Results on general video VQA benchmarks.** `max`: Maximum number of sampling frames. $*$: Officially reported results. $\diamondsuit$: Retrieved results from official leaderboard or recent papers. Other unlabeled results are reproduced by ourselves. Note that we use the "no subtitles" evaluation setting in VideoMME.

Model	# Frames	General VQA			
		MVBench (Acc.)	Egoschema (Acc.)	VideoMME (Acc.)	Perception-Test (Acc.)
Proprietary Models					
Gemini 1.5 Pro	-	**81.3**$^\diamondsuit$	63.2*	**75.0**$^\diamondsuit$	-
GPT-4o-mini	1 fps (max 32)	55.2	58.5	63.6	48.2
GPT-4o	-	-	**77.2***	71.9$^\diamondsuit$	-
GPT 4V	-	43.7$^\diamondsuit$	55.6*	59.9$^\diamondsuit$	-
Open-source Models (Vision-language)					
Qwen2 VL (7B)	2 fps (max 768)	67.0* | 64.4	66.7* | 66.6	63.3* | 59.0	62.3* | 60.3
AnyGPT (8B)	48	33.2	32.1	29.8	29.1
VideoLLaMA 2 (7B)	16	54.6*	51.7*	46.6*	51.4*
VideoChat2 (7B)	16	51.1*	42.1$^\diamondsuit$	33.7$^\diamondsuit$	47.3$^\diamondsuit$
LLaVA-NeXT-Video (7B)	32	46.5$^\diamondsuit$	43.9$^\diamondsuit$	33.7$^\diamondsuit$	48.8$^\diamondsuit$
Video-LLaVA (7B)	8	41.0$^\diamondsuit$	38.4$^\diamondsuit$	39.9$^\diamondsuit$	44.3$^\diamondsuit$
Open-source Models (Omni-modal)					
VITA (8x7B)	1 fps (max 32)	53.4	53.9	56.1	56.2
VITA-1.5 (7B)	1 fps (max 32)	55.5	54.7	57.3	57.6
Baichuan-Omni (7B)	1 fps (max 32)	60.9	58.8	58.2	56.8
MiniCPM-o 2.6 (7B)	1 fps (max 64)	58.6	50.7	63.4	66.6
Baichuan-Omini-1.5 (7B)	1 fps (max 32)	63.7	62.4	60.1	**68.9**

For Reasoning QA, we use GPT-4o to evaluate the score of the answers based on the given reference answers, and then calculate the accuracy rate. For Llama Questions, Web Questions, and TriviaQA, we provide reference answers and use GPT-4o to assess the correctness of the model's responses. Specifically, the score for Llama Questions is the percentage of answers judged as correct, while for Web Questions and TriviaQA, we scale the scores by dividing by 10 to normalize them to a range of 0 to 10. For AlpacaEval, we employ GPT-4o to rate responses on a scale of 1 to 10, with the final score being the average of these ratings.

For all audio benchmarks, we consider two different settings: 1) speech-to-speech generation in a non cascaded manner (denoted as s→s), where the input is audio and the output is interleaved text and audio. The output text is then merged and used for evaluation. 2) speech-to-text generation (denoted as s→t), where the input is audio and the output is text, which is used for evaluation.

Results. As shown in Table 11, our model performs excellently on audio understanding benchmarks, outperforming the latest open-source models. In the s→t setting, Baichuan-Omni-1.5 significantly outperforms models of the same size in Reasoning QA and AlpacaEval, achieving scores of 50 and 7.79, respectively. In the s→s setting, Baichuan-Omni-1.5 surpasses GLM-4-Voice across the board, particularly leading by 14.4 and 2.05 in Reasoning QA and AlpacaEval.

4.5 Performance in Omni Tasks

Baselines. We utilize the following baselines: proprietary models (GPT-4o-mini [124]) and recent open-source models for omni-modal (VITA-1.0 [45], VITA-1.5 [46], Baichuan-omni [89], and MiniCPM-o 2.6 [166]).

Evaluation Benchmarks. OmniBench [92] is an innovative benchmark specifically designed to rigorously assess a model's ability to simultaneously recognize, interpret, and reason across a diverse array of inputs, including visual, acoustic, and textual data. This benchmark is designed to provide a comprehensive evaluation of multi-modal processing capabilities, ensuring that models are effectively tested on their ability to integrate and analyze information from multiple sources concurrently. There are four common evaluation setups: 1) *Image & Audio*: use the original image and original audio as input. 2) *Image Caption & Audio*: use the image caption and original audio as input. 3) *Image & Audio Transcript*: use the original image and audio transcripts as input. 4) *Image Caption & Audio Transcript*: use the image caption and audio transcripts as input.

Table 10: **Results on open-ended video VQA benchmarks.** `max`: Maximum number of sampling frames. $*$: Officially reported results. Other unlabeled results are reproduced by ourselves.

Model	# Frames	Open-ended VQA			
		ActivityNet-QA		MSVD-QA	
		(Acc.)	(Score)	(Acc.)	(Score)
Proprietary Models					
Gemini 1.5 Pro	-	56.7*	-	-	-
GPT-4o-mini	1 fps (max 32)	62.1	3.1	67.5	3.3
GPT-4o	-	61.9*	-	-	-
GPT 4V	-	59.5*	-	-	-
Open-source Models (Vision-language)					
Qwen2 VL (7B)	2 fps (max 768)	17.4	1.9	61.1	3.5
VideoLLaMA 2 (7B)	16	50.2*	3.3*	70.9*	3.8*
VideoChat2 (7B)	16	49.1*	3.3*	70.0*	3.9*
LLaVA-NeXT-Video (7B)	32	53.5*	3.2*	67.4	3.4
Video-LLaVA (7B)	8	45.3*	3.3*	70.7*	3.9*
Open-source Models (Omni-modal)					
VITA (8x7B)	1 fps (max 32)	55.0	3.5	63.9	3.7
VITA-1.5 (7B)	1 fps (max 32)	59.6	3.0	67.6	3.3
Baichuan-Omni (7B)	1 fps (max 32)	58.6	**3.7**	72.2	**4.0**
MiniCPM-o 2.6 (7B)	1 fps (max 64)	**63.0**	3.1	73.7	3.6
Baichuan-Omni-1.5 (7B)	1 fps (max 32)	62.0	3.1	**74.2**	3.6

Table 11: **Results on audio understanding benchmarks.** ∇: The modalities parameter is set to ["text", "audio"], evaluation based on the output text. $\diamondsuit$: Supports only text-audio interleaved output. $\square$: Cascade output method, evaluation based on the output text.

Model	Audio Comprehensive Capacity									
	Reasoning QA		Llama Questions		Web Questions		TriviaQA		AlpacaEval	
	$s \rightarrow t$	$s \rightarrow s$	$s \rightarrow t$	$s \rightarrow s$	$s \rightarrow t$	$s \rightarrow s$	$s \rightarrow t$	$s \rightarrow s$	$s \rightarrow t$	$s \rightarrow s$
Proprietary Models										
GPT-4o-Audio$^{\nabla}$	**55.6**	-	**88.4**	-	**8.10**	-	**9.06**	-	**8.01**	-
Open-source Models (Pure Audio)										
GLM-4-Voice (9B)$^{\diamondsuit}$	-	26.5	-	71.0	-	5.15	-	4.66	-	4.89
Open-source Models (Omni-modal)										
VITA-1.5 (7B)$^{\square}$	41.0	-	74.2	-	5.73	-	4.68	-	6.82	-
MiniCPM-o 2.6 (7B)$^{\square}$	38.6	-	77.8	-	6.86	-	6.19	-	5.18	-
Baichuan-Omni-1.5 (7B)	50.0	**40.9**	78.5	**75.3**	5.91	**5.52**	5.72	**5.31**	7.79	**6.94**

Results. As shown in Table 12, we find that no matter what model is evaluated, the results of using audio transcripts are better than those of using the original audio. Taking Baichuan-Omni-1.5 as an example, the results of *Image & Audio* and *Image & Audio Transcript* are 42.9 and 47.9, respectively. The results of *Image Caption & Audio* and *Image Caption & Audio Transcript* are 37.7 and 46.9, respectively. This shows that the audio recognition and understanding capabilities of current omni-modal models still have a lot of room for improvement. Compared to the latest released omni-modal model MiniCPM-o 2.6 [166], our model outperforms it in three of the four settings, that is, 42.9 v.s 40.5, 37.7 v.s 30.8, and 46.9 v.s 46.3.

Table 12: **Overall Omni-Undesratnding Results.** All the results are reproduced by ourselves. GPT-4o-mini does not support audio input, we use its audio API and transcribe the audio and then input it.

Model	Omni-Understanding			
	Image & Audio (Acc.)	Image Caption & Audio (Acc.)	Image & Audio Transcript (Acc.)	Image Caption & Audio Transcript (Acc.)
Proprietary Models				
GPT-4o-mini	-	-	37.0	37.7
Open-source Models (Omni-modal)				
VITA (8x7B)	33.1	31.8	42.0	44.2
VITA-1.5 (7B)	33.4	29.6	48.5	**47.2**
Baichuan-Omni (7B)	32.2	26.5	42.6	44.2
MiniCPM-o 2.6 (7B)	40.5	30.8	**53.2**	46.3
Baichuan-Omni-1.5 (7B)	**42.9**	**37.7**	47.9	46.9

Table 13: **Results on medical benchmarks.** All the results are reproduced by ourselves.

Model	Medical Understanding	
	GMAI-MMB-VAL (Acc.)	OpenMM-Medical (Acc.)
Proprietary Models		
GPT-4o-mini	46.4	74.3
Open-source Models (Vision-Language)		
Qwen2 VL (7B)	46.3	76.9
Qwen2 VL (72B)	**50.7**	80.7
Open-source Models (Omni-modal)		
VITA-1.5 (7B)	36.7	67.1
MiniCPM-o 2.6 (7B)	41.5	73.6
Baichuan-Omni-1.5 (7B)	49.9	**83.8**

4.6 Performance in Medical Tasks

Baselines. We compare Baichuan-Omni-1.5 with the following baselines: proprietary models (GPT-4o-mini [124]), recent open-source models for omni-modal (VITA-1.5 [46] and MiniCPM-o 2.6 [166]).

Evaluation Benchmarks. We utilize GMAI-MMBench [21] and OpenMM-Medical as the evaluation benchmark. GMAI-MMBench is meticulously designed to evaluate the capabilities of MLLMs within real-world clinical settings, characterized by several distinctive features. It encompasses comprehensive medical knowledge, incorporating 284 diverse clinical datasets sourced globally and spanning 38 different modalities. The data structure is well-categorized, featuring an organized framework of 18 clinical VQA tasks and 18 clinical departments, systematically arranged in a lexical tree for ease of navigation and analysis. Additionally, the benchmark supports multi-perceptual granularity, offering interactive methods that range from the image level down to the region level, thereby providing a nuanced evaluation of perceptual detail across varying degrees of specificity.

In addition, we also construct a more diverse medical evaluation dataset named OpenMM-Medical. The images in OpenMM-Medical are sourced from 42 publicly available medical image datasets, such as ACRIMA [125] (fundus photography), BioMediTech [122] (microscopy images), and CoronaHack [28] (X-Ray). OpenMM-Medical comprises a total of 88,996 images, each designed to be paired with multiple-choice VQA. This evaluation dataset will be made openly available to the research community.

Results. As shown in Table 13, Baichuan-Omni-1.5 achieves the highest performance in both GMAI-MMBench [21] and OpenMM-Medical. In GMAI-MMBench validation, GPT4o-mini achieves 46.3% while Baichuan-Omni-1.5 gets 49.9%. On OpenMM-Medical, the recent omni-modal model MiniCPM-o 2.6 gets 73.6%, while our Baichuan-Omni-1.5 gets a large margin, 83.8%. From the previous experimental conclusions, our model has strong omni-modal

understanding capabilities, namely pure text, audio, images, and videos. In addition, we have also verified our strong capabilities in medical images. Therefore, we believe that our model has taken a big step towards the real-time consultation.

5 Conclusion

In this work, we introduce Baichuan-omni-1.5, an omni-modal model that represents a significant stride towards developing a comprehensive framework encompassing all human senses. Using high-quality multimodal data and multistage omni-modal pre-training and fine-tuning strategies, Baichuan-omni-1.5 achieves excellent performance in processing video, image, text, and audio understanding. The key features of Baichuan-omni-1.5 include: (1) robust capabilities in both pure text and multimodal understanding; (2) end-to-end parallel processing of omni-modal inputs (text, image, video, text) and dual-modal outputs (text and audio); (3) excellent performance in medical scenarios; and (4) high-quality controllable audio generation.

Despite these promising results, there remains substantial room for improvement in the foundational capabilities of each modality. That is, (1) enhance text understanding capabilities; (2) support longer video frame understanding; and (3) improve audio understanding and generation to not only recognize human voices but also natural environmental sounds such as flowing water, bird songs, and collision noises, among others.

Our future research will focus on refining these areas to ensure more sophisticated and versatile models capable of comprehending and interacting with complex environments. We anticipate that continued advancements in these domains will contribute significantly to the broader goal of achieving Artificial General Intelligence.

6 Contributors

Project Leads

Zenan Zhou, Weipeng Chen

Senior Leads

Jianhua Xu, Haoze Sun, Mingan Lin

Contributors

* indicates core contributors with equal contributions.

Yadong Li*, Jun Liu*, Tao Zhang*, Tao Zhang*, Song Chen*, Tianpeng Li*, Zehuan Li*, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi Ma, Kegeng Wu, Jia li, Aiyuan Yang, Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin Huang, Kun Li, Wenjing Luo, Yifei Duan, Lingling Zhu, Ran Xiao, Zhe Su, Jiani Pu, Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yujing Qiao, Lei Zhang, Yanjun Shen, Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao, Hao Liang, Youquan Li, Yanzhao Qin, Linzhuang Sun

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] Manoj Acharya, Kushal Kafle, and Christopher Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8076–8084, 2019.

[3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[4] AI@Meta. Llama 3 model card, 2024.

[5] Avinash Anand, Raj Jaiswal, Abhishek Dharmadhikari, Atharva Marathe, Harsh Popat, Harshil Mital, Ashwin R Nair, Kritarth Prasad, Sidharth Kumar, Astha Verma, et al. Geovqa: A comprehensive multimodal geometry dataset for secondary education. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 102–108. IEEE, 2024.

[6] Anas Awadalla, Le Xue, Oscar Lo, Manli Shu, Hannah Lee, Etash Kumar Guha, Matt Jordan, Sheng Shen, Mohamed Awadalla, Silvio Savarese, et al. Mint-1t: Scaling open-source multimodal data by 10x: A multimodal dataset with one trillion tokens. *arXiv preprint arXiv:2406.11271*, 2024.

[7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[8] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021.

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

[10] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010.

[11] Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R Manmatha. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16548–16558, 2022.

[12] Antoine Boutet, Davide Frey, Rachid Guerraoui, Arnaud Jégou, and Anne-Marie Kermarrec. Whatsup: A decentralized instant news recommender. In *2013 IEEE 27th International Symposium on Parallel and Distributed Processing*, pages 741–752. IEEE, 2013.

[13] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[14] LMM CaD. Comparison visual instruction tuning. *arXiv preprint*, 2021.

[15] Shihao Cai, Keqin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, and Bo Zheng. Geogpt4v: Towards geometric multi-modal large language models with geometric image generation. *arXiv preprint arXiv:2406.11503*, 2024.

[16] Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022.

[17] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022.

[18] Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, et al. Emova: Empowering language models to see, hear and speak with vivid emotions. *arXiv preprint arXiv:2409.18042*, 2024.

[19] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2025.

[20] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *arXiv preprint arXiv:2406.04325*, 2024.

[21] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361*, 2024.

[22] 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.

[23] 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.

[24] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[25] Christopher Chou, Lisa Dunlap, Koki Mashita, Krishna Mandal, Trevor Darrell, Ion Stoica, Joseph E Gonzalez, and Wei-Lin Chiang. Visionarena: 230k real world user-vlm conversations with preference labels. *arXiv preprint arXiv:2412.08687*, 2024.

[26] 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*, 2024.

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

[28] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv 2006.11988*, 2020.

[29] E. Cui, Y. He, Z. Ma, Z. Chen, H. Tian, W. Wang, K. Li, Y. Wang, W. Wang, X. Zhu, L. Lu, T. Lu, Y. Wang, L. Wang, Y. Qiao, and J. Dai. Sharegpt-4o: Comprehensive multimodal annotations with gpt-4o. `https://sharegpt4o.github.io/`, 2024.

[30] Coen De Vente, Koenraad A Vermeer, Nicolas Jaccard, He Wang, Hongyi Sun, Firas Khader, and et al. Airogs: Artificial intelligence for robust glaucoma screening challenge. *IEEE Transactions on Medical Imaging*, 43(1):542–557, 2023.

[31] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[32] 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.

[33] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, and et al. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[35] Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, et al. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv preprint arXiv:2410.17885*, 2024.

[36] Guosheng Dong, Da Pan, Yiding Sun, Shusen Zhang, Zheng Liang, Xin Wu, Yanjun Shen, Fan Yang, Haoze Sun, Tianpeng Li, et al. Baichuanseed: Sharing the potential of extensive data collection and deduplication by introducing a competitive large language model baseline. *arXiv preprint arXiv:2408.15079*, 2024.

[37] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024.

[38] 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*, 2024.

[39] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models, 2024.

[40] 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.

[41] Balz Engler. Textualization. In *Literary Pragmatics (Routledge Revivals)*, pages 179–189. Routledge, 2014.

[42] Yue Fan, Jing Gu, Kaiwen Zhou, Qianqi Yan, Shan Jiang, Ching-Chen Kuo, Yang Zhao, Xinze Guan, and Xin Wang. Muffin or chihuahua? challenging multimodal large language models with multipanel vqa. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6845–6863, 2024.

[43] 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*, 2024.

[44] Chaoyou Fu, Yuhan 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.

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

[46] 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.

[47] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023.

[48] Sushant Gautam, Andrea M Storås, Cise Midoglu, and et al. Kvasir-vqa: A text-image pair gi tract dataset. In *Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications*, pages 3–12, 2024.

[49] Philippe Gervais, Asya Fadeeva, and Andrii Maksai. Mathwriting: A dataset for handwritten mathematical expression recognition. *arXiv preprint arXiv:2404.10690*, 2024.

[50] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

[51] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024.

[52] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Jiageng Mao, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, et al. Soda10m: A large-scale 2d self/semi-supervised object detection dataset for autonomous driving. *arXiv preprint arXiv:2106.11118*, 2021.

[53] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023.

[54] Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv preprint arXiv:2404.15127*, 2024.

[55] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[56] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[57] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and vision computing*, 60:4–21, 2017.

[58] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Chen Li, Ji Zhang, Qin Jin, Fei Huang, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

[59] Irene Huang, Wei Lin, M Jehanzeb Mirza, Jacob A Hansen, Sivan Doveh, Victor Ion Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuhene, Trevor Darrel, et al. Conme: Rethinking evaluation of compositional reasoning for modern vlms. *arXiv preprint arXiv:2406.08164*, 2024.

[60] 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. *Advances in Neural Information Processing Systems*, 36, 2024.

[61] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[62] Mengzhao Jia, Wenhao Yu, Kaixin Ma, Tianqing Fang, Zhihan Zhang, Siru Ouyang, Hongming Zhang, Meng Jiang, and Dong Yu. Leopard: A vision language model for text-rich multi-image tasks. *arXiv preprint arXiv:2410.01744*, 2024.

[63] Yizhang Jin, Jian Li, Jiangning Zhang, Jianlong Hu, Zhenye Gan, Xin Tan, Yong Liu, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Llava-vsd: Large language-and-vision assistant for visual spatial description. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11420–11425, 2024.

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

[65] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2018.

[66] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

[67] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*, 2022.

[68] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022.

[69] Heeseung Kim, Soonshin Seo, Kyeongseok Jeong, Ohsung Kwon, Jungwhan Kim, Jaehong Lee, Eunwoo Song, Myungwoo Oh, Sungroh Yoon, and Kang Min Yoo. Unified speech-text pretraining for spoken dialog modeling. *arXiv preprint arXiv:2402.05706*, 2024.

[70] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

[71] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):1–10, 2018.

[72] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36, 2024.

[73] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

[74] Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*, 2024.

[75] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

[76] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11523–11532, 2022.

[77] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[78] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

[79] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

[80] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*, 2023.

[81] 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.

[82] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[83] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2023.

[84] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[85] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024.

[86] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6706–6713, 2019.

[87] Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*, 2024.

[88] Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`, 5 2023.

[89] 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), 2024.

[90] 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.

[91] Yifan Li, Anh Dao, Wentao Bao, Zhen Tan, Tianlong Chen, Huan Liu, and Yu Kong. Facial affective behavior analysis with instruction tuning. In *European Conference on Computer Vision*, pages 165–186. Springer, 2025.

[92] Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*, 2024.

[93] Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design-Vienna 2024*, 2024.

[94] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

[95] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14963–14973, 2023.

[96] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

[97] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

[98] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[99] 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.

[100] Yangzhou Liu, Yue Cao, Zhangwei Gao, Weiyun Wang, Zhe Chen, Wenhai Wang, Hao Tian, Lewei Lu, Xizhou Zhu, Tong Lu, et al. Mminstruct: A high-quality multi-modal instruction tuning dataset with extensive diversity. *Science China Information Sciences*, 67(12):1–16, 2024.

[101] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023.

[102] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, 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*, 2024.

[103] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024.

[104] Keer Lu, Zheng Liang, Xiaonan Nie, Da Pan, Shusen Zhang, Keshi Zhao, Weipeng Chen, Zenan Zhou, Guosheng Dong, Wentao Zhang, et al. Datasculpt: Crafting data landscapes for llm post-training through multi-objective partitioning. *arXiv preprint arXiv:2409.00997*, 2024.

[105] 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.

[106] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022.

[107] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021.

[108] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024.

[109] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

[110] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023.

[111] 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.

[112] Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. Chartgemma: Visual instruction-tuning for chart reasoning in the wild. *arXiv preprint arXiv:2407.04172*, 2024.

[113] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

[114] Johanna McEntyre and David Lipman. Pubmed: bridging the information gap. *Cmaj*, 164(9):1317–1319, 2001.

[115] Shivam Mehta, Ruibo Tu, Jonas Beskow, Éva Székely, and Gustav Eje Henter. Matcha-tts: A fast tts architecture with conditional flow matching. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11341–11345. IEEE, 2024.

[116] Lingwei Meng, Long Zhou, Shujie Liu, Sanyuan Chen, Bing Han, Shujie Hu, Yanqing Liu, Jinyu Li, Sheng Zhao, Xixin Wu, et al. Autoregressive speech synthesis without vector quantization. *arXiv preprint arXiv:2407.08551*, 2024.

[117] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020.

[118] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952. IEEE, 2019.

[119] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[120] Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*, 2023.

[121] 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.

[122] Loris Nanni, Michelangelo Paci, Florentino Luciano Caetano dos Santos, Heli Skottman, Kati Juuti-Uusitalo, and Jari Hyttinen. Texture descriptors ensembles enable image-based classification of maturation of human stem cell-derived retinal pigmented epithelium. *PLoS One*, 11(2):e0149399, 2016.

[123] OpenAI. GPT-4V(ision) system card. https://openai.com/index/gpt-4v-system-card/, 2023.

[124] OpenAI. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`, 2024.

[125] Silvia Ovreiu, Elena-Anca Paraschiv, and Elena Ovreiu. Deep learning & digital fundus images: Glaucoma detection using densenet. In *2021 13th international conference on electronics, computers and artificial intelligence (ECAI)*, pages 1–4. IEEE, 2021.

[126] Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, et al. Perception test: A diagnostic benchmark for multimodal video models. *arXiv preprint arXiv:2305.13786*, 2023.

[127] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[128] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[129] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

[130] 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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[131] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer, 2022.

[132] Share. Sharegemini: Scaling up video caption data for multimodal large language models, June 2024.

[133] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024.

[134] Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Gholamreza Haffari, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. *arXiv preprint arXiv:2411.06048*, 2024.

[135] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020.

[136] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-ego: A large-scale dataset of paired third and first person videos. *arXiv preprint arXiv:1804.09626*, 2018.

[137] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019.

[138] Shubhankar Singh, Purvi Chaurasia, Yerram Varun, Pranshu Pandya, Vatsal Gupta, Vivek Gupta, and Dan Roth. Flowvqa: Mapping multimodal logic in visual question answering with flowcharts. *arXiv preprint arXiv:2406.19237*, 2024.

[139] Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024.

[140] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645, 2023.

[141] Solène Tarride, Yoann Schneider, Marie Generali-Lince, Mélodie Boillet, Bastien Abadie, and Christopher Kermorvant. Improving automatic text recognition with language models in the pylaia open-source library. In *International Conference on Document Analysis and Recognition*, pages 387–404. Springer, 2024.

[142] Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. Consistency-preserving visual question answering in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 386–395. Springer, 2022.

[143] Qwen Team. Qwen2-VL: To See the World More Clearly. *Qwen*, August 2024.

[144] Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, 2024.

[145] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1):1–9, 2018.

[146] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024.

[147] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv*, 2023.

[148] Dequan Wang, Xiaosong Wang, Lilong Wang, and et al. A real-world dataset and benchmark for foundation model adaptation in medical image classification. *Scientific Data*, 10(1):574, 2023.

[149] Junjie Wang, Yin Zhang, Yatai Ji, Yuxiang Zhang, Chunyang Jiang, Yubo Wang, Kang Zhu, Zekun Wang, Tiezhen Wang, Wenhao Huang, et al. Pin: A knowledge-intensive dataset for paired and interleaved multimodal documents. *arXiv preprint arXiv:2406.13923*, 2024.

[150] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[151] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. *arXiv preprint arXiv:2410.08260*, 2024.

[152] Yonghui Wang, Wengang Zhou, Hao Feng, Keyi Zhou, and Houqiang Li. Towards improving document understanding: An exploration on text-grounding via mllms. *arXiv preprint arXiv:2311.13194*, 2023.

[153] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[154] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision*, pages 3164–3172, 2015.

[155] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Kaixin Xu, Chunyi Li, Jingwen Hou, Guangtao Zhai, et al. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25490–25500, 2024.

[156] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[157] x.ai. Grok-1.5 vision preview. `https://x.ai/blog/grok-1.5v`, 2024.

[158] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9777–9786, June 2021.

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

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

[161] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[162] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023.

[163] 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*, 2024.

[164] Kaiyu Yang, Olga Russakovsky, and Jia Deng. Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2051–2060, 2019.

[165] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, page 100211, 2024.

[166] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[167] 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.

[168] Wenyi Yu, Siyin Wang, Xiaoyu Yang, Xianzhao Chen, Xiaohai Tian, Jun Zhang, Guangzhi Sun, Lu Lu, Yuxuan Wang, and Chao Zhang. Salmonn-omni: A codec-free llm for full-duplex speech understanding and generation. *arXiv preprint arXiv:2411.18138*, 2024.

[169] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019.

[170] Ye Yuan, Xiao Liu, Wondimu Dikubab, Hui Liu, Zhilong Ji, Zhongqin Wu, and Xiang Bai. Syntax-aware network for handwritten mathematical expression recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4553–4562, 2022.

[171] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

[172] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

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

[174] Aohan Zeng, Zhengxiao Du, Mingdao Liu, Lei Zhang, Shengmin Jiang, Yuxiao Dong, and Jie Tang. Scaling speech-text pre-training with synthetic interleaved data. *arXiv preprint arXiv:2411.17607*, 2024.

[175] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, and Xipeng Qiu. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

[176] 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, 2023.

[177] Ge Zhang, Scott Qu, Jiaheng Liu, Chenchen Zhang, Chenghua Lin, Chou Leuang Yu, Danny Pan, Esther Cheng, Jie Liu, Qunshu Lin, et al. Map-neo: Highly capable and transparent bilingual large language model series. *arXiv preprint arXiv:2405.19327*, 2024.

[178] Jianming Zhang, Xin Zou, Li-Dan Kuang, Jin Wang, R Simon Sherratt, and Xioafeng Yu. Cctsdb 2021: a more comprehensive traffic sign detection benchmark. *Human-centric Computing and Information Sciences*, 12, 2022.

[179] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13, 2024.

[180] Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning. *arXiv preprint arXiv:2404.16635*, 2024.

[181] Peng Zhang, Can Li, Liang Qiao, Zhanzhan Cheng, Shiliang Pu, Yi Niu, and Fei Wu. Vsr: a unified framework for document layout analysis combining vision, semantics and relations. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I 16*, pages 115–130. Springer, 2021.

[182] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, et al. Mavis: Mathematical visual instruction tuning with an automatic data engine. *arXiv preprint arXiv:2407.08739*, 2024.

[183] Tianhao Zhang, Zhixiang Chen, and Lyudmila S Mihaylova. Pvit: Prior-augmented vision transformer for out-of-distribution detection. *arXiv preprint arXiv:2410.20631*, 2024.

[184] Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, et al. Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model. *arXiv preprint arXiv:2407.07053*, 2024.

[185] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[186] Xiaotian Zhang, Chunyang Li, Yi Zong, Zhengyu Ying, Liang He, and Xipeng Qiu. Evaluating the performance of large language models on gaokao benchmark. *arXiv e-prints*, pages arXiv–2305, 2023.

[187] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

[188] 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.

[189] Henry Hengyuan Zhao, Pan Zhou, Difei Gao, Zechen Bai, and Mike Zheng Shou. Lova3: Learning to visual question answering, asking and assessment. *arXiv preprint arXiv:2405.14974*, 2024.

[190] Kecheng Zheng, Yifei Zhang, Wei Wu, Fan Lu, Shuailei Ma, Xin Jin, Wei Chen, and Yujun Shen. Dreamlip: Language-image pre-training with long captions. In *ECCV*, 2024.

[191] Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. Multimodal table understanding. *arXiv preprint arXiv:2406.08100*, 2024.

[192] Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. Multimodal table understanding. *arXiv preprint arXiv:2406.08100*, 2024.

[193] 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.

[194] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866, 2022.

[195] Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. Multichartqa: Benchmarking vision-language models on multi-chart problems. *arXiv preprint arXiv:2410.14179*, 2024.

[196] Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. *arXiv preprint arXiv:2408.08640*, 2024.