Ming-Omni: A Unified Multimodal Model for Perception and Generation

Inclusion AI, Ant Group*

*See Contributions section (Sec. 6) for full author list.

We propose Ming-Omni, a unified multimodal model capable of processing images, text, audio, and video, while demonstrating strong proficiency in both speech and image generation. Ming-Omni employs dedicated encoders to extract tokens from different modalities, which are then processed by Ling, an MoE architecture equipped with newly proposed modality-specific routers. This design enables a single model to efficiently process and fuse multimodal inputs within ***a unified framework****, thereby facilitating diverse tasks without requiring separate models, task-specific fine-tuning, or structural redesign. Importantly, Ming-Omni extends beyond conventional multimodal models by supporting audio and image generation. This is achieved through the integration of an advanced audio decoder for natural-sounding speech and Ming-Lite-Uni for high-quality image generation, which also allow the model to engage in context-aware chatting, perform text-to-speech conversion, and conduct versatile image editing. Our experimental results showcase Ming-Omni offers a powerful solution for unified perception and generation across all modalities. Notably, our proposed Ming-Omni is the first open-source model we are aware of to match GPT-4o in modality support, and we release all code and model weights to encourage further research and development in the community.*

Date: May 21, 2025
Project Homepage: Link
Code: https://github.com/inclusionAI/Ming/tree/main/Ming-omni

1 Introduction

Humans effortlessly integrate visual and auditory cues to express ideas and generate vivid mental imagery from descriptions, supporting creativity, problem-solving, and communication as core aspects of intelligence. The ultimate goal of Artificial General Intelligence (AGI) is to emulate this form of human-like multimodal intelligence, gradually evolving from a tool into a highly capable agent that enhances and liberates human productivity. The recent advance of Large Language Models (LLMs), coupled with extensive training on vast multi-modal datasets, have catalyzed the emergence of strong perceptual capabilities in both vision (Chen et al., 2024e; Bai et al., 2025b; KimiTeam et al., 2025) and audio (Ding et al., 2025; Xu et al., 2025), as well as generative capabilities in these two paradigms (Huang et al., 2025; Ding et al., 2025; OpenAI, 2025; Tong et al., 2024; Pan et al., 2025). Nevertheless, it remains challenging how to effectively blend both modalities in a single understanding model. Beyond comprehension tasks, another critical hurdle is integrating robust generation capabilities into these models to produce coherent, context-aware outputs across modalities while maintaining semantic consistency.

Despite these advancements, constructing a unified Omni-MLLM is hindered by representational disparities across modalities and divergent convergence rates during training. To address these

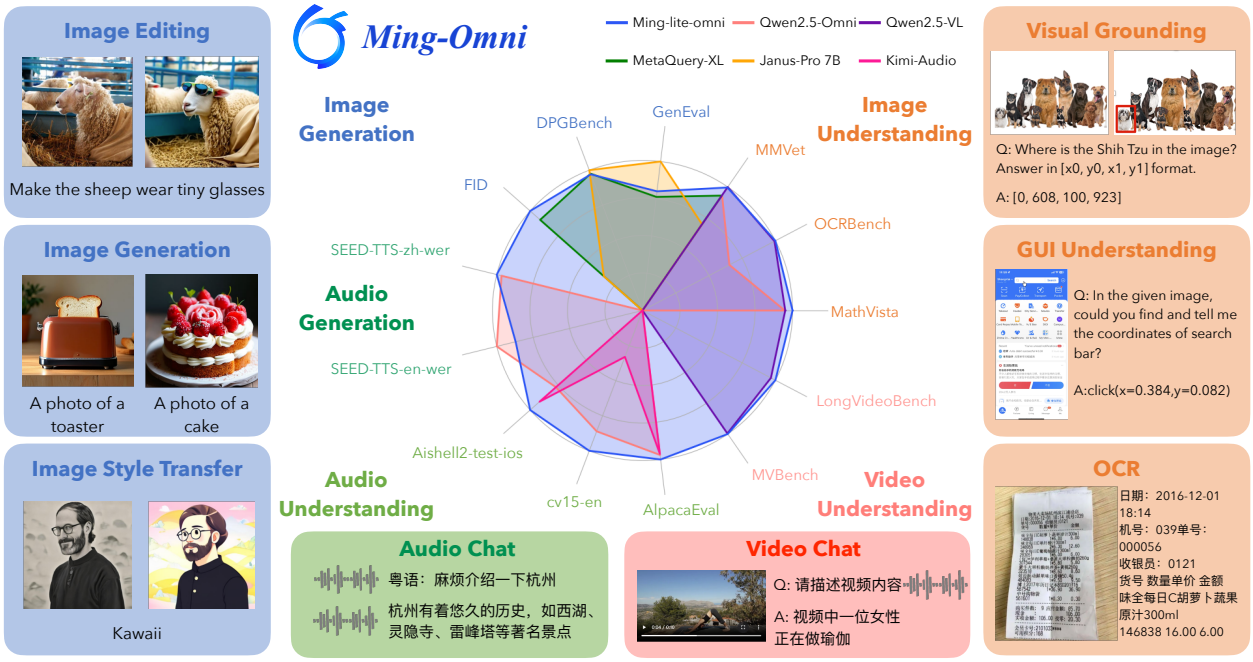


Figure 1 Ming-Omni is a versatile, unified end-to-end model capable of processing images, text, video, and audio, and generating text, speech, and images. These capabilities enable the lite version of our model, Ming-Lite-Omni, to support a broad range of tasks, including visual perception, audio-visual interaction, and image generation, among others.

challenges, Ming-Omni employs a language model with an integrated Mixture-of-Experts (MoE) architecture (LingTeam et al., 2025), where modality-specific routing is enabled through dedicated mechanisms for each type of token, allowing for tailored routing distributions. Furthermore, following (Guo et al., 2025), we apply a stepwise balancing strategy during pre-training to mitigate cross-modal data imbalance, and employ a dynamically adaptive approach during instruction tuning to align training progress across modalities, leading to better convergence and model performance. Through the optimization of model architecture and training strategies, Ming-Omni achieves robust perception and understanding across multiple modalities.

Building on this robust perceptual foundation, we extend Ming-Omni with an audio decoder and Ming-Lite-Uni (InclusionAI et al., 2025), enabling joint audio and image generation. For speech generation, we address challenges like prosodic naturalness, real-time response, context awareness, and complex acoustic environments (*e.g.*, dialects or accents) via two innovations: 1) Ming-Omni uses Byte Pair Encoding (BPE) to improve prosody and reduce token frame rate by 35%, boosting real-time performance. 2) A two-stage training strategy that prevents audio understanding and generation tasks from influencing each other, where the first stage focuses on understanding capabilities and the second stage concentrates on generation quality. This also enables Ming-Omni to balance efficiency and naturalness across diverse linguistic scenarios. Turning to visual generation, reconciling disparate visual feature representations is a fundamental challenge in unified multi-modal models. While existing methods like TokenFlow (Qu et al., 2024) and Janus (Wu et al., 2024a) achieve impressive results, they compromise semantic fidelity due to pixel-centric optimization. In contrast, Ming-Omni adopts a lightweight bridging framework that keeps the MLLM frozen and generates images progressively from coarse to fine using multi-scale learnable tokens guided by an alignment strategy. A dedicated connector integrates latent representations produced by the MLLM with the diffusion decoder, leveraging its semantic understanding for image generation.

These architectural innovations empower Ming-Omni to deliver exceptional cross-modal performance, as validated across image perception, audio-visual interaction, and image generation tasks. Specifically, in the image perception task, Ming-Omni attained performance comparable to that of Qwen2.5-VL-7B (Bai et al., 2025a) by activating only 2.8B parameters. Ming-Omni delivers superior performance in end-to-end speech understanding and instruction following, surpassing Qwen2.5-Omni (Xu et al., 2025) and Kimi-Audio (Ding et al., 2025). It also supports native-resolution image generation, editing, and style transfer, achieving a GenEval score of 0.64, outperforming mainstream models such as SDXL (Podell et al., 2023). In terms of FID, Ming-Omni reaches 4.85, setting a new SOTA across existing methods.

The key features of Ming-Omni can be summarized as follows.

- **Unified Omni-Modality Perception:** Ming-Omni, built on Ling (LingTeam et al., 2025), an MoE architecture LLM, resolves task conflicts and ensures coherent integration of tokens from different modalities through modality-specific routers.
- **Unified Perception and Generation:** Ming-Omni achieves unified understanding and generation, enabling the model to interpret multimodal instructions and user intent during generation, thereby improving generation quality and usability across multiple tasks.
- **Innovative Generation Capabilities:** Ming-Omni can perceive all modalities and generate high-quality text, real-time speech, and vivid images simultaneously, delivering exceptional cross-modal performance across diverse tasks including image perception, audio-visual interaction, and image generation.

2 Approach

As illustrated in Figure 2, Ming-Omni is a unified model capable of supporting inputs from multiple modalities, including images, audio, video, and text, while also facilitating speech generation and image generation. The overall training process of Ming-Omni is divided into two distinct phases: perception training and generation training. During the perception training phase, the focus is on training the language interface Ling (LingTeam et al., 2025) to effectively perceive and understand visual and audio tokens across different modalities. In the generation training phase, the emphasis shifts to training the audio decoder and the DiT module to enhance the model's generative capabilities. In the subsequent sections, we will explore how Ming-Omni achieves unified understanding and generation across all modalities.

2.1 Unified Understanding Across Modalities

The cornerstone of Ming-Omni lies in its state-of-the-art (SOTA) capability for comprehensive multimodal understanding. To achieve this, Ming-Omni integrates the Qwen2.5 (Bai et al., 2025a) vision backbone as its visual encoder, which supports arbitrary resolutions and demonstrates superior performance in both image and video processing tasks. Additionally, Ming-Omni utilizes Whisper (Radford et al., 2023) as its audio encoder, which has been proven to exhibit robust performance in ASR and speech understanding tasks. The embeddings generated by these encoders are projected to align with the dimensionality of the language model. These projected embeddings are then concatenated with the tokenized text inputs and fed into the language model Ling, which is based on a MoE architecture. Given the challenges of training an omni-modal large language model (omni-MLLM) due to the incongruence in representational spaces of different modalities and the disparity in convergence rates across modalities, we optimized the architecture of Ling by designing

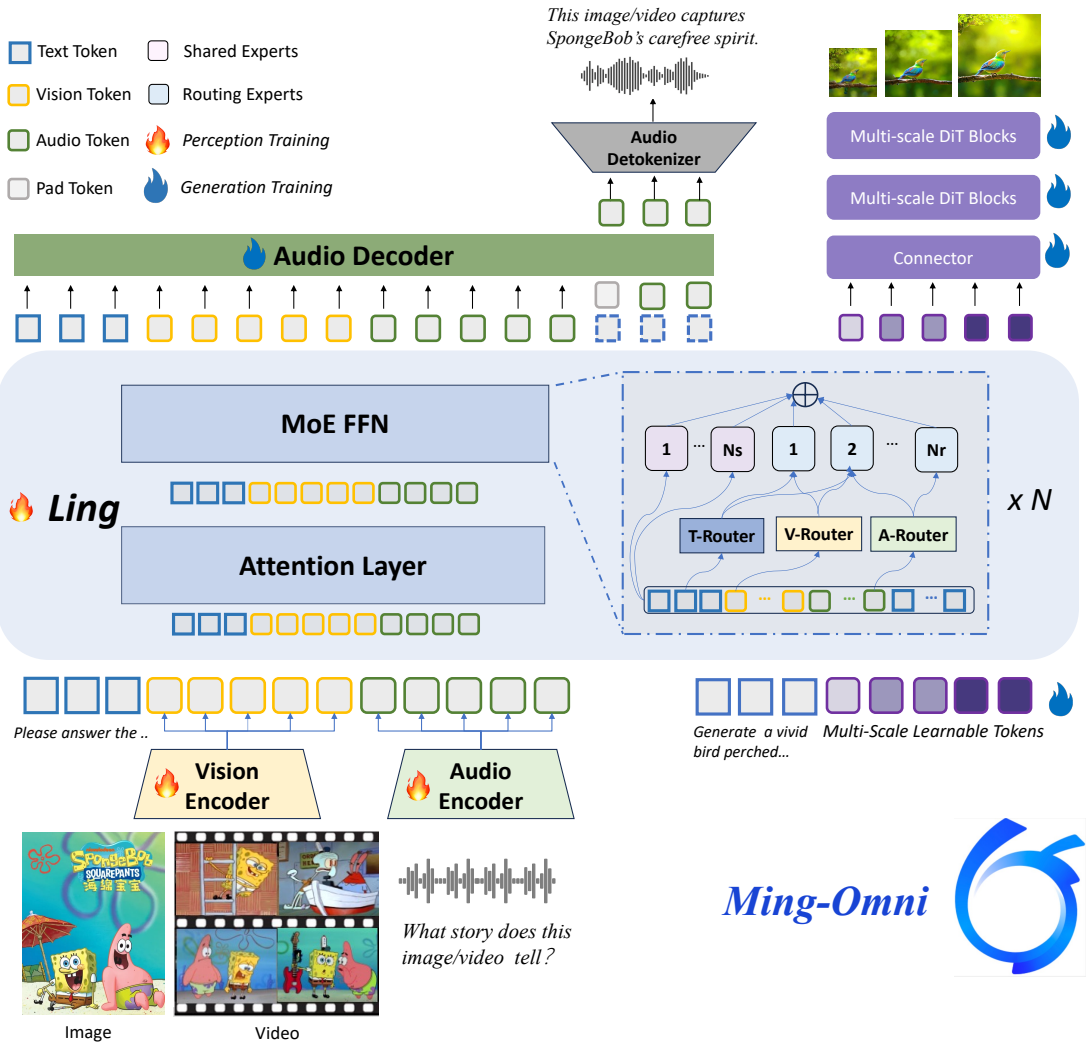


Figure 2 The overall framework of Ming-Omni. Ming-Omni extracts visual and audio tokens with dedicated encoders. These tokens are then combined with text tokens and processed through Ling (MoE architecture with modality-specific routers). Subsequently, it generates speech through an audio decoder and enables image generation via a diffusion model.

distinct routers for different modalities. This design enables tokens from each modality to be routed to specialized experts, thereby facilitating more precise and efficient routing of modality-specific information. To further address these challenges, we adopted a step-wise balance strategy during the pre-training stage and introduced a dynamic adaptive balance strategy during the instruction tuning stage, following (Guo et al., 2025). The dynamic adaptive balance strategy dynamically adjusts the loss weights based on the convergence rates of each modality, thereby mitigating conflicts between different modalities and ensuring optimal training progress across all modalities.

2.2 Unified Speech Understanding and Generation

In the Ming-Omni framework, following Qwen-Audio (Chu et al., 2024), we adopt Whisper as the audio encoder for its strong capabilities in audio modeling, which are not limited to human speech. We employ a combination of a linear layer and a convolutional downsampling layer to transform the audio encoder output features to the latent space of the Ling language model. To unleash the pre-existing world knowledge and the question answering capability in pre-trained language model in audio processing, a highly diverse audio data corpus is collected with metadata

attributes (such as conversational or command scenarios, ambient environment, *etc.*) labeled by an audio labeler (see section 3.3). On top of that, we find it critical to incorporate these additional metadata attributes into the instruction prompts of the speech understanding tasks in an optional manner, thereby offering the model supplementary contextual cues to enhance its comprehension performance. Furthermore, we instruct the model to first predict the language identifier of the input audio before performing downstream tasks such as speech recognition. Our experiments show that this strategy significantly improves overall performance, especially in dialectal speech recognition. Overall, our training framework—which incorporates contextual and language information—delivers substantial improvements over conventional ASR-inspired paradigms.

To generate audio contents, we follow CosyVoice and connect an audio decoder to the output of the language model, where the audio decoder is an autoregressive architecture that generates discrete audio tokens extracted by an audio tokenizer. For the generation of audio contents in the paradigm of end-to-end multimodal LLMs (MLLMs), two key challenges are encountered for high-quality speech generation: (i) the gap in the sequence length between text and audio tokens in autoregressive modeling; and (ii) the difficulty in generating speech responses closely aligned with the input context containing various input modalities.

To address the difference in the sequence length between the text and audio tokens, instead of using the discrete audio tokens directly as the target for the audio decoder, we apply Byte Pair Encoding (BPE) to the discrete tokens extracted by the audio tokenizer. The BPE efficiently compresses the token length of the discrete tokens by 36% (from 50Hz to approximately 32Hz), which effectively improves the training and inference efficiency. Since the BPE encoding process is entirely reversible, the efficiency gain is obtained without any loss on the quality of generated audio content. Further, similar to the BPE in the language domain, it also encourages learning of the combinatorial information within the audio content and thus enhances the prosodic performance.

To promote the relevance between the generated audio content and the input multi-modal context, we follow Qwen2.5-Omni (Xu et al., 2025) and feed the hidden state of the original input from the MLLMs to the audio decoder. This encourages the audio decoder to capture paralinguistic information in the original input, such as emotions and environmental details. However, we find joint training of MLLMs and the audio decoder in Qwen2.5-Omni brings the difficulty in optimizing both understanding and generation tasks, as is observed in (Shi et al., 2025). Hence, during the training of audio generation modules, we freeeze the entire MLLMs and only train the audio decoder using text-to-speech data and multi-modal context-aware triplet data. This preserves the multimodal understanding capabilities of the MLLMs while achieving strong speech generative performance.

2.3 Unified Image Understanding and Generation

A key challenge in unifying image understanding and generation lies in maintaining high generation quality without compromising visual understanding capabilities. This balance is typically reflected in performance trade-offs across benchmarks from both domains. A critical obstacle is enabling the image generator to make effective use of the multimodal language model's internal representations, which encode rich world knowledge but are not directly aligned with visual outputs.

In pursuit of this goal, many existing approaches attempt to model visual tokens within a shared feature space for both understanding and generation. However, achieving a meaningful balance often requires complex design choices, including architectural modifications, loss balancing, and multi-stage training strategies. Without such mechanisms, improvements in one task often come at the cost of the other. As a result, tokens optimized for understanding tend to degrade generation

Figure 3 Instruction based image style transfer results outputted by Ming-Omni.

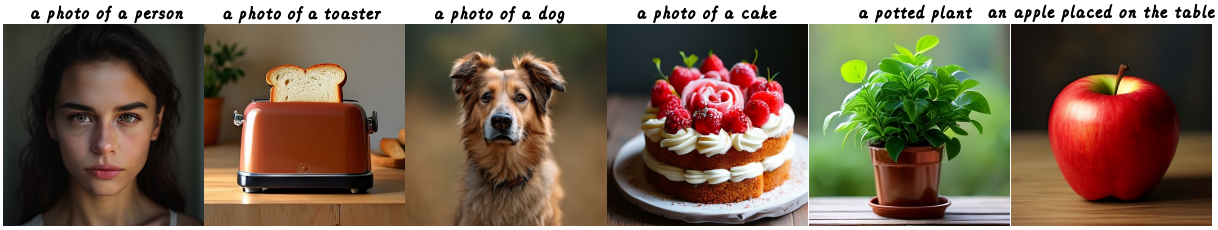


Figure 4 Instruction based T2I results outputted by Ming-Omni.

quality, while tokens tuned for generation may reduce understanding accuracy. We consider this a challenging yet potentially promising direction. Our ongoing efforts in this path will be presented in detail in a future version of this work.

At this released version, we propose a lightweight bridging approach which leverages multi-scale learnable tokens and multi-scale representation alignment. This design enables the model to adaptively benefit from tokens specialized for either understanding or generation. To avoid interference with image tokens for understanding, we introduce a dedicated image diffusion model during fine-tuning, ensuring high-quality generation while preserving semantic alignment and understanding performance. Specifically, for understanding, we adopt the Qwen2.5-VL (Bai et al., 2025b) vision backbone with around 675 million parameters, retaining its architecture but retraining it jointly on images and videos. On the generation side and the bridging component, we introduce a novel multi-scale learnable token scheme, coupled with the representation-level alignment and the connector to enhance cross-task consistency. Details of key modules are described below.

Figure 5 Instruction based image editing results outputted by Ming-Omni.

Multi-Scale Learnable Tokens Fusion and Processing Given an input image x, we define a set of scales $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$, where each s_k corresponds to a spatial resolution, e.g., $s_k \in \{4 \times 4, 8 \times 8, 16 \times 16\}$. Each scale s_k is associated with a dedicated set of learnable query tokens $Q_{s_k} \in \mathbb{R}^{N_{s_k} \times d}$, where N_{s_k} is the number of tokens for scale s_k, and d is the hidden dimension size. Each Q_{s_k} is designed to capture information at different granularity levels, including global layout and color distribution, major objects and mid-level structures, and fine textures and detailed patterns. To preserve semantics at different spatial scales, we introduce explicit boundary markers by adding learnable start and end tokens around each scale's sequence. Each token is also assigned a scale-specific positional encoding derived from the spatial grid of that resolution. All scale-level sequences and their positional encodings are concatenated to form the input to the transformer encoder, which then processes this combined sequence to produce the final hidden representations.

Multi-Scale Representation Alignment Beyond bridging with multi-scale learnable tokens, we further achieve implicit unification of understanding and generation through feature-level alignment. Specifically, we align the intermediate hidden states from the DiT backbone with the final semantic representations by minimizing the mean squared error between them. This alignment loss encourages consistency between hierarchical representations and outputs through native-resolution optimization.

As shown in Figure 3 / 4 / 5, Ming-Omni supports both text-to-image generation and instruction-based image editing tasks, expanding its applicability across a broader range of creative and practical applications. For more details, please refer to Ming-Lite-Uni (InclusionAI et al., 2025).

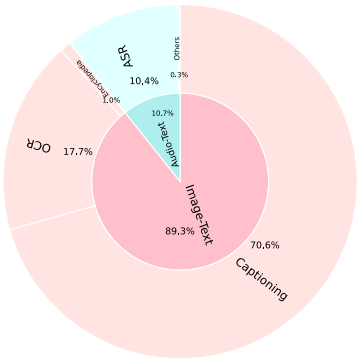
2.4 Overall Training Procedure

The training procedure of Ming-Omni comprises two stages: perception and generation training. The perception stage, consisting of pre-training, instruction tuning, and alignment tuning, is consistent with the M2-omni (Guo et al., 2025) training pipeline. Specifically, both the pre-training and instruction tuning stages are divided into three sub-stages, with each sub-stage designed to incrementally incorporate additional tasks. After the perception training stage, we further expand Ming-Omni's multi-modal generation capabilities through the generation training stage. In particular, this stage consists of two parallel training tasks: text-to-speech and image generation, where we freeze the multimodal perception LLM and train additional generation components. For image generation training, only the added connector, multiscale learnable queries, and the DiT blocks are trained. For text-to-speech training, we train the audio decoder to support multi-modal context-aware dialogue. Through the generation training stage, Ming-Omni expands image and speech generation capabilities without compromising its multi-modal understanding capabilities.

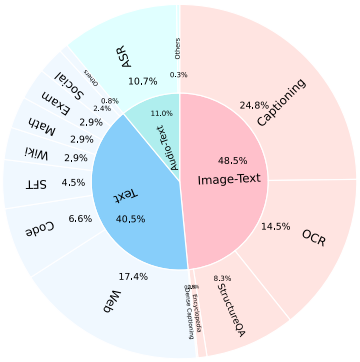
3 Data Construction

3.1 Data Overall

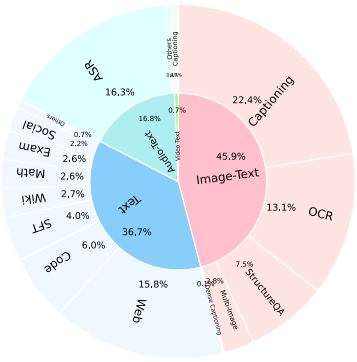
We collect a large amount of training data covering different modalities and tasks. We built this diverse training data by collecting open source data and constructing multiple data production pipelines. The training objectives, modalities, and tasks of each training stage are different. Therefore, we construct different data configurations for each stage, as shown in Fig. 6. Pre-training data, as shown in Fig. 6 (a)-(c), is constructed adhering to two criteria: (1) aligning different modalities and (2) facilitating concept learning of world knowledge. Consequently, the resulting pre-training dataset exhibits diversity, aggregating multi-source data across various modalities. For instruction tuning, we aim to equip the model with the ability to address a broad spectrum of multimodal tasks. According to the tasks, we collect both open-source and in-house multimodal instruction tuning data for training, as shown in Fig. 6 (d)-(f). The training data of alignment tuning, text-to-speech generation, and image generation are also constructed according to the corresponding tasks. The detailed data sources and data construction methods are further elaborated in this section.



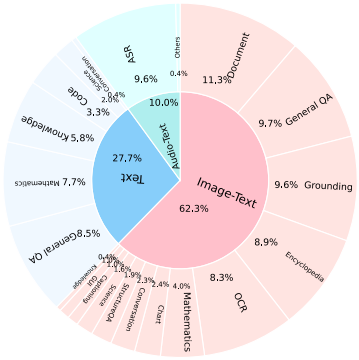
(a) PT-Stage-1-Data
Image-Text: 885M
Audio-Text: 389k hour



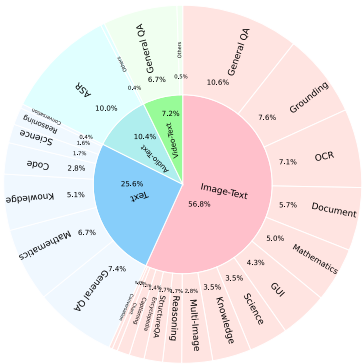
(b) PT-Stage-2-Data
Image-Text: 576M
Audio-Text: 389k hour
Text: 494M



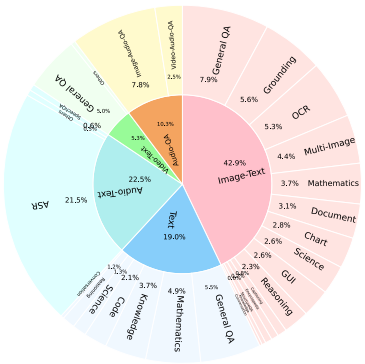
(c) PT-Stage-3-Data
Image-Text: 613M
Audio-Text: 266k hour
Text: 494M, Video-Text: 9M



(d) IT-Stage-1-Data
Image-Text: 31.7M
Audio-Text: 330k hour
Text: 14.6M



(e) IT-Stage-2-Data
Image-Text: 35.7M
Audio-Text: 400k hour
Text: 15.5M, Video-Text: 4.3M



(f) IT-Stage-3-Data
Image-Text: 36.3M, Text: 15.5M
Audio-Text: 521k hour
Video-Text: 4.3M, Audio QA: 8.3M

Figure 6 Overview of the data configurations during pre-training and instruction tuning. Note that the "Others" type of audio data includes AST, AAC, AAT, and SER data. The "Others" type of pre-training text data in (b) and (c) includes books, academic, news, and domain data. And the "Others" type of video-text data in (e) and (f) includes knowledge, captioning, and reasoning data.

3.2 Image Data

Image data serves as the cornerstone of our multi-modal data corpus. We integrate image understanding and generation datasets to enable our MLLM to derive unified perception-generation capabilities. Notably, we propose an iterative self-evolving framework that effectively improves the data quality and reduces the data volume; we also incorporate structured data and encyclopedia data that both empower our MLLM with fine-grained expert-level knowledge.

3.2.1 Image Understanding Data

Caption Data. Caption data provides fundamental multi-modal alignment abilities. Our caption corpus is aggregated from a diverse collection of public datasets (*e.g.*, Wukong (Gu et al., 2022), Laion (Schuhmann et al., 2022), DenseFusion (Li et al., 2024c), ZERO-250M (Xie et al., 2023), COYO-700M (Byeon et al., 2022), *etc.*) and in-house ones. Nevertheless, the presence of noisy, irrelevant, or incorrect image-caption pairs remains a critical challenge, leading to sub-optimal performance and hallucinatory MLLM responses. Inspired by DiverseEvol (Wu et al., 2023) and ASK-LLM (Sachdeva et al., 2024), we propose an iterative self-evolving data refinement framework for caption corpus optimization. Initially, we partition the full caption corpus $\mathcal{D}_{full}$ into T non-overlapping split $\{D_m\}_{m=1}^{T}$, and randomly select one split to train the initial model M_0, accompanied by an empty data pool $\mathcal{P}_0 = \varnothing$ and a pre-defined filtering threshold τ. At each iteration $t \in [1, T]$, we first employ the previous model M_{t-1} to evaluate all samples within another new split D_t. Samples whose inference scores exceed τ are grouped into D_t^{high}, while the others with low scores are discarded since they are often simple and trivial captions. Afterwards, we calculate quality scores on D_t^{high} by designing tailored prompts to guide available MLLMs in assessing whether each sample is suitable for model training, only retaining verified high-quality samples to form the informative sub-split $\ddot{D}_t^{high}$. At the conclusion of each iteration t, we append $\ddot{D}_t^{high}$ to the evolving data pool $\mathcal{P}_t = \mathcal{P}_{t-1} \cup \ddot{D}_t^{high}$, and training another new model M_t based on $\mathcal{P}_t$ for the next iteration. After completing all T iterations, we derive the refined caption corpus $\mathcal{P}_T = \{\ddot{D}_t^{high}\}_{t=1}^{T}$ to replace the initial set $\mathcal{D}_{full}$, thereby enhancing the quality and diversity of the caption corpus while significantly reducing the data volume.

Structured Data. Structured data enhances MLLM capabilities in addressing knowledge-intensive and information-seeking queries. Beyond conventional image understanding tasks that primarily require MLLM to recognize specific visual content like object detection and OCR, structured data presents challenges in querying fine-grained knowledge associated with specific visual entities. Inspired by InfoSeek (Chen et al., 2023a), we develop an effective structured data synthesis pipeline to generate large-scale information-seeking QA pairs. Specifically, we first extract semantically significant entities from image content and their corresponding descriptions by leveraging multiple available MLLMs to perform a cross verification. We then employ a well-trained linking model to organize all extracted entities into a structured knowledge base. Afterwards, we synthesize information-seeking QA pairs by carefully designed prompts based on knowledge triplets, which embeds informative fine-grained knowledge.

Encyclopedia Data. Encyclopedia data integrates advanced domain-specific expertise into MLLMs for expert-level comprehension and perception, *e.g.*, identifying rare or endangered species via Latin binomial nomenclature. Our encyclopedia corpus spans 8 specific domains across biological categories (*Plants and Animals*), cultural categories (*Celebrities, Anime characters, and Artworks*), and daily-life categories (*Ingredients, Dishes, and Vehicles*). To construct a high-quality expert-level corpus, we first collect a wide range of encyclopedia entities from academic databases and

institutional websites. We then employ these entities as search queries to collect semantically relevant images via search engines. Afterwards, we develop a progressive encyclopedia data filtering scheme, including clip consistency validation, MLLM-based binary verification, and manual refinement.

GUI Data. GUI (Graphical User Interface) data enables MLLM to address complex real-world Android interaction tasks. Our GUI corpus is mainly constructed from three public datasets: AitW (Rawles et al., 2023a), GUICourse (Chen et al., 2024b), and AndroidControl (Li et al., 2024b). Moreover, we leverage available MLLMs to optimize the reasoning process for each step within a human-like GUI interaction operation, which is subsequently reviewed by another MLLM to improve data quality.

Reasoning Data. Reasoning data activates the latent reasoning capabilities of MLLMs through the supervised long Chain-of-Thought (CoT) learning. Our reasoning corpus is mainly composed of two sources: textual reasoning data from Ling (LingTeam et al., 2025) and multi-modal reasoning data from R1-Onevision (Yang et al., 2025).

Preference Data. Preference data optimizes MLLM responses through enhanced improved alignment with human-centric interaction patterns in the alignment tuning stage. Our preference corpus is primarily constructed from three sources: user-generated conversations in applications, search queries from websites, and high-quality instruction datasets. Specifically, we first retrieve relevant web images via search engines to complement text-only user-generated dialogues or queries. We then leverage available MLLMs to generate diverse specialized questions and their corresponding answers. Afterwards, we engage MLLMs and skilled human annotators to assess the quality of these QA pairs. Ultimately, we organize those high-quality positive samples and the other negative ones to construct the preference corpus, which comprises 41 subcategories across 9 primary domains.

3.2.2 Image Generation Data

Image generation data extends MLLM capabilities beyond conventional image understanding tasks. Following Ming-Lite-Uni (InclusionAI et al., 2025), our image generation corpus mainly comes from two sources: High-quality images collected from public image generation datasets (*e.g.*, InstructPix2Pix-clip-filtered (Brooks et al., 2023), SEED-Data-Edit-part2/3 (Ge et al., 2024), Ultra-edit (Zhao et al., 2024), and *etc.*); and image style transfer data sampled from StyleBooth and WikiArt. Readers can refer to Ming-Lite-Uni (InclusionAI et al., 2025) for more construction details and data examples.

3.3 Audio Data

As previously mentioned, the diversity of the audio data is critical for unleashing the pre-existing knowledge and capabilities in the pretrained language model to audio processing. Hence, we use a large amount of data from open-source datasets (detailed in Table 14) as well as in-house datasets consisting of web data and synthetic data. The web data are obtained from a carefully designed data filtering process. **(i)** To begin with, we crawl a large set of audio data from the web based on a set of keywords that are expanded from handcrafted ones using hundreds of domain-specific lexical variations. **(ii)** VAD (Gao et al., 2023) is applied to obtain well-conditioned short audio clips. **(iii)** An audio labeler is trained iteratively, where the labeler is first trained with a high-quality dataset, before it is used to label the whole audio corpus. The labeled data are then further used to improve the precision of the audio labeler.

Based on the short audio clips obtained from **(ii)** and the audio labeler from **(iii)**, we acquire a

large number of high-quality audio clips with labels from different domains. Ablation studies are performed for an optimal data ratio among labels. Empirical findings indicate that a larger number of English audio clips compared to Chinese audio clips results in significant improvements in English understanding, without degrading the ability to understand Chinese. Furthermore, dialect data constitutes only 2% of the total data corpus, and the performance of dialect understanding plateaus as the ratio of dialect data increases.

3.4 Video Data

Video data enables MLLM to understand spatial-temporal content beyond sequences of static images. Our video corpus is mainly curated from open-source datasets (*e.g.*, VideoGPT+ (Maaz et al., 2024), Vript (Yang et al., 2024), Openvid-1M (Nan et al., 2024), Youku-mPLUG-10M (Xu et al., 2023), *etc.*) and public English / Chinese websites. Following LLaVA-Video (Zhang et al., 2024d), we adopt a hierarchical annotation pipeline to progressively generate dense video captions, open-ended question-answer pairs, and multi-choice question-answer pairs.

3.5 Text Data

Text data is essential for MLLM to maintain and improve language proficiency. Our text corpus is derived from two sources, *i.e.*, Ling (LingTeam et al., 2025) and M2-omni (Guo et al., 2025).

4 Evaluation

We benchmark Ming-Lite-Omni, a light version of Ming-Omni, which is derived from Ling-lite and features 2.8 billion activated parameters, and compare it against leading SoTA MLLMs with under 10B parameters.

4.1 Public Benchmarks

As shown in Table 1∼12, our holistic assessment covers more than 50 rigorously curated public benchmarks across the following five distinct multi-modal dimensions:

Image → Text (Understanding). Our evaluation of the image-to-text understanding capabilities primarily encompasses the following six tasks: 1) Fundamental image understanding capabilities evaluated on OpenCompass image-text comprehensive benchmarks (Contributors, 2023). 2) Image Reasoning capabilities evaluated on OpenCompass image-text reasoning benchmarks (Contributors, 2023). 3) Image Grounding capabilities evaluated on RefCOCO (Kazemzadeh et al., 2014), RefCOCO+ (Kazemzadeh et al., 2014), and RefCOCOg (Mao et al., 2016). 4) OCR capabilities evaluated on ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), OCRBenchV2 (Fu et al., 2024b), OmniDocBench (Ouyang et al., 2024), and TextVQA-VAL (Singh et al., 2019). 5) GUI capabilities evaluated on ScreenSpot (Cheng et al., 2024), ScreenSpot-V2 (Wu et al., 2024d), and AITZ(EM) (Zhang et al., 2024a). And 6) knowledge-based Question Answering capabilities evaluated on InfoSeek (Chen et al., 2023b).

Text → Image (Generation). We incorporate text-to-image generation capabilities to enable our MLLM with unified perception-generation abilities, which are evaluated on GenEval (Ghosh et al., 2024), DPG-Bench (Hu et al., 2024a), and FID.

Audio → Text (Understanding). Our evaluation of the audio-to-text understanding capabilities mainly includes the following two tasks: 1) Fundamental audio understanding capabilities evaluated

Table 1 Performance of Ming-Lite-Omni on **OpenCompass Image-Text Benchmarks** compared to leading MLLMs.

Benchmark	Ming-Lite Omni	Qwen2.5 Omni	Qwen2.5VL 7B-Instruct	InternVL2.5 8B-MPO	Gemma3 27B
AI2D	83.1	83.2	84.4	**84.5**	81.4
HallusionBench	55.0	-	**55.8**	51.7	48.8
MMBench-TEST-V11	80.8	81.8	**82.8**	82.0	78.9
MMMU	56.3	59.2	56.6	54.8	**64.8**
MMStar	64.7	64.0	**65.3**	65.2	59.6
MMVet	71.3	66.8	**71.6**	68.1	71.0
MathVista	**71.6**	67.9	68.1	67.9	67.6
OCRBench	**88.4**	57.8	87.8	88.2	75.3

Table 2 Performance of Ming-Lite-Omni on **PUBLIC Grounding Benchmarks** compared to leading MLLMs.

Benchmark	Split	Ming-Lite Omni	Qwen2.5-Omni 7B	InternVL2.5 8B	Grounding-DINO Large	Qwen2.5VL 7B-Instruct
RefCOCO	val	**90.6**	90.5	90.3	**90.6**	90.0
	testA	93.1	93.5	**94.5**	93.2	92.5
	testB	86.3	86.6	85.9	**88.2**	85.4
RefCOCO+	val	**85.4**	**85.4**	85.2	82.8	84.2
	testA	89.8	91.0	**91.5**	89.0	89.1
	testB	79.2	**79.3**	78.8	75.9	76.9
RefCOCOg	val	86.8	**87.4**	86.7	86.1	87.2
	test	87.5	**87.9**	87.6	87.0	87.2
Average		87.3	**87.7**	87.6	86.6	86.6

Table 3 Performance of Ming-Lite-Omni on **PUBLIC GUI Benchmarks** compared to leading MLLMs. The superscript "*" denotes the reproduced results.

Benchmark	Ming-Lite Omni	InternVL3 8B	Qwen2.5VL 7B-Instruct
ScreenSpot	**82.1**	79.5	78.9*
ScreenSpot-V2	**84.1**	81.4	-
AITZ(EM)	**66.6**	-	57.6*

Table 4 Performance of Ming-Lite-Omni on **PUBLIC Information-Seeking Benchmarks** compared to leading MLLMs.

Benchmark (InfoSeek)	Ming-Lite Omni	PaLI-X	Qwen2.5VL 32B
H-mean	**27.7**	22.1	19.4
Unseen-question	**30.4**	23.5	20.6
Unseen-entity	**25.4**	20.8	18.3

Table 5 Performance of Ming-Lite-Omni on **Text-to-Image Generation Benchmarks** compared to leading models. "*Gen.*" denotes models for pure image generation, while "*Uni.*" denotes models capable of both image understanding and generation. Note that the global best performance is highlighted by an underline, and the local best result in "*Gen.*" or "*Uni.*" is marked with **bold**.

Type	Model	GenEval							DPG-Bench	FID↓
		1-Obj.	2-Obj.	Count	Colors	Posit.	Color.	AVG		
Gen.	LlamaGen	0.71	0.34	0.21	0.58	0.07	0.04	0.32	-	-
	LDM	0.92	0.29	0.23	0.70	0.02	0.05	0.37	-	-
	SDv1.5	0.97	0.38	0.35	0.76	0.04	0.06	0.43	-	-
	PixArt-α	0.98	0.50	0.44	0.80	0.08	0.07	0.48	-	-
	SDv2.1	0.98	0.51	0.44	0.85	0.07	0.17	0.50	68.09	26.96
	Emu3-Gen	0.98	0.71	0.34	0.81	0.17	0.21	0.54	80.60	-
	SDXL	0.98	0.74	0.39	0.85	0.15	0.23	0.55	74.65	8.76
	DALL-E 3	0.96	0.87	0.47	0.83	**0.43**	0.45	0.67	-	-
	SD3-Medium	**0.99**	**0.94**	**0.72**	**0.89**	0.33	**0.60**	**0.74**	-	-
Uni.	LWM	0.93	0.41	0.46	0.79	0.09	0.15	0.47	-	-
	SEED-X	0.97	0.58	0.26	0.80	0.19	0.14	0.49	-	-
	Show-o	0.95	0.52	0.49	0.82	0.11	0.28	0.53	-	-
	TokenFlow-XL	0.95	0.60	0.41	0.81	0.16	0.24	0.55	-	-
	Janus	0.97	0.68	0.30	**0.84**	**0.46**	0.42	0.61	79.68	10.10
	JanusFlow	-	-	-	-	-	-	0.63	80.09	9.51
	JanusPro-7B	-	-	-	-	-	-	**0.80**	**84.19**	13.48
	Ming-Lite-Omni	**0.99**	**0.77**	**0.68**	0.78	0.31	0.29	0.64	81.72	**4.85**

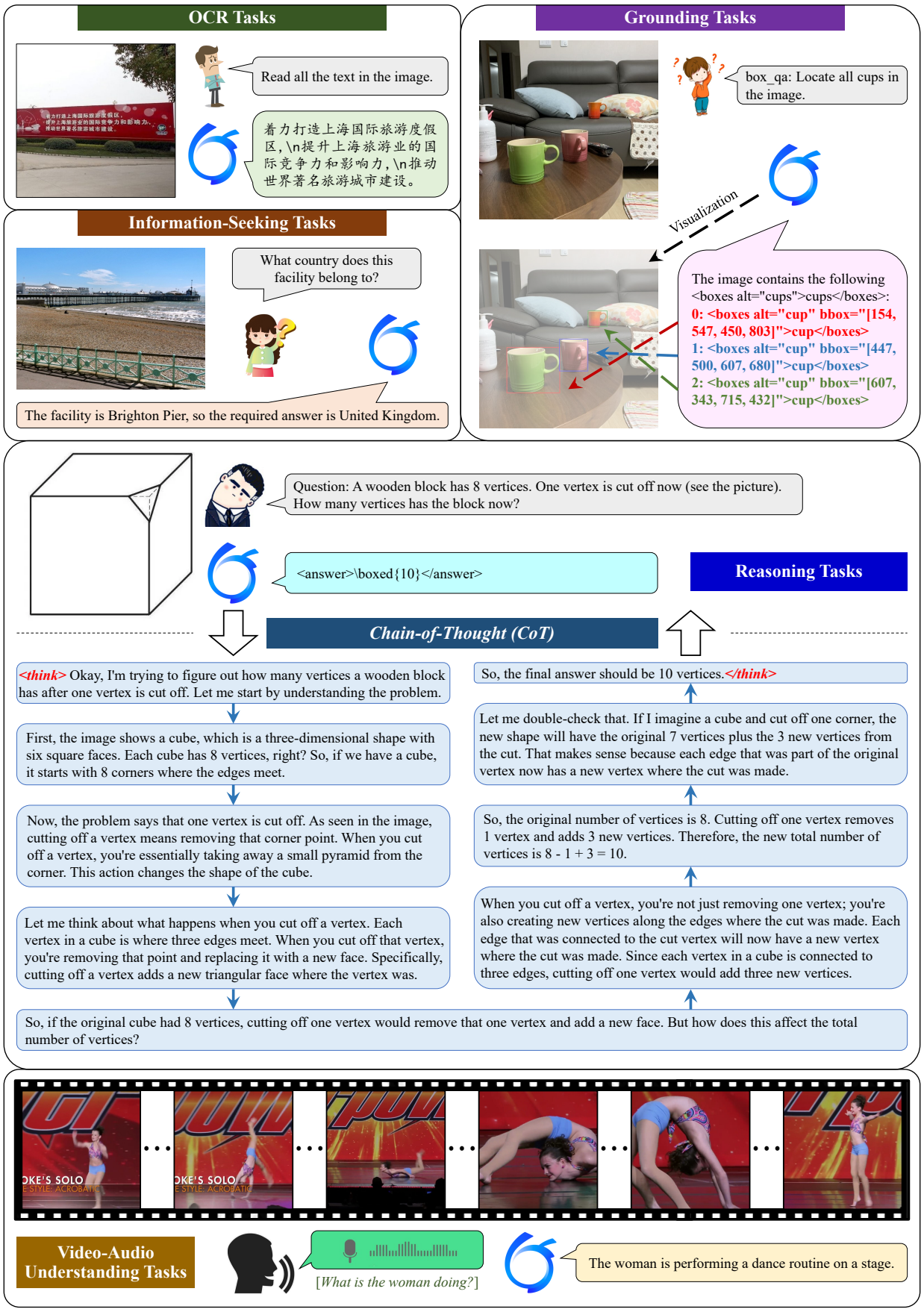


Figure 7 Visualization results of OCR, Grounding, Information-Seeking, Reasoning, and Video-Audio Understanding tasks.

Table 6 Performance of Ming-Lite-Omni on **PUBLIC OCR Benchmarks** compared to leading MLLMs. "·/·" denotes results under English (*En*) and Chinese (*Zh*) testing conditions as "*En*/*Zh*".

Benchmark	Ming-Lite Omni	InternVL3 8B	Gemma3 27B	Qwen2.5VL 7B-Instruct
ChartQA	85.1	86.6	83.4	**87.2**
DocVQA	93.0	92.7	89.5	**95.6**
OCRBench-v2	53.3/52.0	-	-	**56.3/57.2**
TextVQA	82.8	80.2	83.2	**85.1**
OmniDocBench↓	34.0/**34.4**	-	-	**30.8**/39.8

Table 7 Performance of Ming-Lite-Omni on **IN-HOUSE Encyclopedia Benchmarks** compared to leading MLLMs.

Benchmark (In-house)	Ming-Lite Omni	Qwen2.5VL 7B-Instruct	Ovis2 8B	Ola 7B
Plant	**54.96**	47.85	37.83	36.33
Animal	**56.70**	50.85	45.49	40.38
Vehicle	41.91	42.29	**46.01**	39.05
Ingredient	**62.28**	54.09	54.72	53.88
Dish	**44.30**	39.07	42.56	42.56
Average	**52.03**	46.83	45.32	42.44

Table 8 Performance of Ming-Lite-Omni on **PUBLIC Video Understanding Benchmarks** compared to leading MLLMs. All models are evaluated based on 128 uniformly sampled frames.

Benchmark	Ming-Lite Omni	Qwen2.5VL 7B-Instruct	LLaVA-One Vision-7B
MVBench	**67.7**	67.4	56.7
VideoMME	67.0	**67.3**	58.2
VideoMMMU	46.3	**47.4**	33.9
LongVideoBench	**56.6**	54.7	50.5
Average	**59.4**	59.2	49.8

Table 9 Performance of Ming-Lite-Omni on **IN-HOUSE Human Preference Benchmarks**.

Benchmark (In-house)	Ming-Lite Omni	Qwen2.5VL 7B-Instruct
Relevance	**4.479**	4.308
Fluency	**4.907**	4.765
Richness	3.498	**3.828**
Appropriateness	**4.740**	4.727
Correctness	**3.856**	3.741
Average Score	**4.296**	4.274

on a broad range of public benchmarks, including public Chinese benchmarks like Aishell1 (Bu et al., 2017) and Wenetspeech (Zhang et al., 2022a), and public English benchmarks like Librispeech (Zeyer et al., 2021) and Voxpopuli (Wang et al., 2021). And 2) audio question-answering capabilities evaluated on various benchmarks across five specific tasks, such as AlpacaEval and CommonEval from VoiceBench (Chen et al., 2024d) for open-ended QA tasks, and SD-QA for knowledge-based QA tasks.

Text → Audio (Generation). We incorporate text-to-audio generation capabilities to enable our MLLM with unified audio perception-generation abilities, which are evaluated on Seed-TTS-Eval (Anastassiou et al., 2024).

Video → Text (Understanding). Our evaluation of the video-to-text understanding capabilities contains the following four benchmarks: MVBench (Li et al., 2024a), VideoMME (Fu et al., 2024a), VideoMMMU (Hu et al., 2025), and LongVideoBench (Wu et al., 2024b).

4.2 In-house Benchmarks

In addition to public benchmarks, we also establish three in-house benchmarks to comprehensively evaluate multiple capabilities of MLLMs, including:

Encyclopedia Benchmark. We build an encyclopedia benchmark to adequately evaluate the expert-level comprehension capabilities of MLLMs. Following the encyclopedia data construction pipeline in Sec. 3.2.1, we first extract representative image-text pairs, and engage skilled human annotators to verify their correctness and manually write four negative answer candidates. Subsequently, another group of experienced human annotators review the human-generated multi-choice questions, ensuring that four incorrect options are semantically different from the correct answer, while sharing some visual similarities that could confuse less capable models. The final encyclopedia benchmark comprises a total of 29,924 samples across five encyclopedia categories: Plants (7,225 cases), Animals (7,192 cases), Ingredients (7,123 cases), Dishes (6,530 cases), and Vehicles (1,854 cases).

Table 10 Performance of Ming-Lite-Omni on **PUBLIC and IN-HOUSE Audio Understanding Benchmarks** compared to leading Models, including five different dimensions.

Type	Benchmark	Ming-Lite Omni	Qwen2.5 Omni	Qwen2 Audio	Kimi Audio
PUBLIC Chinese Benchmarks	Aishell1 ↓	1.47	1.18	1.53	**0.60**
	Aishell2-test-android ↓	**2.55**	2.75	2.92	2.64
	Aishell2-test-ios ↓	**2.52**	2.63	2.92	2.56
	Cv15-zh ↓	6.31	**5.20**	6.90	7.21
	Fleurs-zh ↓	2.96	3.00	7.50	**2.69**
	Wenetspeech-testmeeting ↓	5.95	**5.90**	7.16	6.28
	Wenetspeech-testnet ↓	5.46	7.70	8.42	**5.37**
	Average (Chinese) ↓	**3.89**	4.05	5.34	3.91
PUBLIC English Benchmarks	Llibrispeech-test-clean ↓	1.44	1.80	1.60	**1.28**
	Librispeech-test-other ↓	2.80	3.40	3.60	**2.42**
	Multilingual-librispeech ↓	**4.15**	7.56	5.40	5.88
	Cv15-en ↓	**6.89**	7.60	8.60	10.31
	Fleurs-en ↓	**3.39**	4.10	6.90	4.44
	Voxpopuli-v1.0-en ↓	**5.80**	**5.80**	6.84	7.97
	Average (English) ↓	**4.08**	5.04	5.49	5.38
IN-HOUSE Dialect Benchmarks	Hunan ↓	**7.88**	29.31	25.88	31.93
	Minnan ↓	**13.84**	53.43	123.78	80.28
	Guangyue ↓	**4.36**	10.39	7.59	41.49
	Chuanyu ↓	**4.33**	7.61	7.77	6.69
	Shanghai ↓	**10.49**	32.05	31.73	60.64
IN-HOUSE Domain Benchmarks	Chat ↓	**2.34**	3.68	4.29	2.96
	Government ↓	**1.77**	2.23	2.70	2.03
	Health ↓	3.31	4.02	4.18	**2.38**
	Knowledge ↓	3.69	3.17	3.33	**1.98**
	Local-live ↓	2.44	**2.03**	2.34	2.05
Average (All IN-HOUSE) ↓		**5.45**	14.79	21.36	23.24

Table 11 Performance of Ming-Lite-Omni on **PUBLIC Audio Question-Answering Benchmarks** compared to leading models. "*Audio.*" denotes models specialized in pure audio understanding tasks, while "*Omni.*" denotes models capable of multi-modal perception and generation capabilities beyond audio-centric tasks. Note that the global best performance is highlighted by an underline, and the local best result in "*Audio.*" and "*Omni.*" is marked with **bold**.

Type	Models	Open-ended QA		Knowledge	Multi-Choice QA		Instruction	Safety
		AlpacaEval	CommonEval	SD-QA	MMSU	OpenBookQA	IFEval	AdvBench
Audio.	Step-Audio-chat	3.99	2.99	46.84	31.87	29.19	**65.77**	86.73
	Qwen2-Audio-chat	3.69	3.40	35.35	35.43	49.01	22.57	98.85
	Baichuan-Audio	4.00	3.39	49.64	48.80	63.30	41.32	86.73
	GLM-4-Voice	4.06	3.48	43.31	40.11	52.97	24.91	88.08
	Kimi-Audio	**4.46**	**3.97**	**63.12**	**62.17**	**83.52**	61.10	**100.00**
Omni.	Megrez-3B-Omni	3.50	2.95	25.95	27.03	28.35	25.71	87.69
	DiVA	3.67	3.54	57.05	25.76	25.49	39.15	98.27
	Qwen2.5-Omni	4.49	3.93	55.71	**61.32**	**81.10**	52.87	**99.42**
	Baichuan-Omni-1.5	4.50	4.05	43.40	57.25	74.51	54.54	97.31
	MiniCPM-o	4.42	**4.15**	50.72	54.78	78.02	49.25	97.69
	Ming-Lite-Omni	**4.63**	4.06	**58.84**	47.53	61.98	**58.36**	99.04

Table 12 Performance of Ming-Lite-Omni on **PUBLIC Text-to-Speech Benchmarks** compared to leading MLLMs. Ming-Lite-Omni-context denotes the model trained with multi-modal context-aware audio triplet data.

Type	Benchmark (Seed-TTS-Eval)	Ming-Lite Omni	Ming-Lite Omni-context	Seed TTS	MaskGCT	E2 TTS	F5 TTS	CosyVoice2	Qwen2.5 Omni
Chinese	Zh-wer ↓	1.69	1.98	**1.11**	2.27	1.97	1.56	1.45	1.70
	Zh-sim ↑	0.68	0.68	**0.80**	0.77	0.73	0.74	0.75	0.75
English	En-wer ↓	4.31	5.10	2.24	2.62	2.19	**1.83**	2.57	2.72
	En-sim ↑	0.51	0.51	**0.76**	0.71	0.71	0.65	0.65	0.63

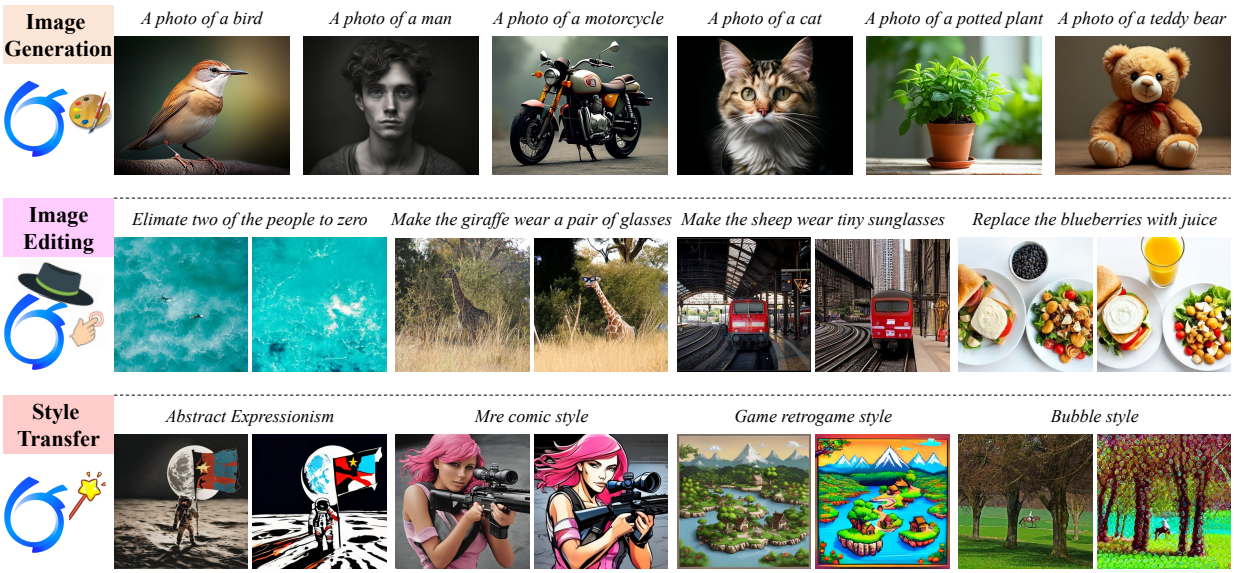


Figure 8 Visualization results of Text → Image tasks, including image generation task, image editing task, and style transfer task.

Human Preference Benchmark. We construct an in-house human preference benchmark to evaluate the human-centric interaction patterns exhibited in MLLM responses. Following the preference data construction pipeline in Sec. 3.2.1, all testing samples are reviewed by skilled human annotators to ensure the constructed benchmark is clean and reliable. Our benchmark contains a total of 1,025 samples. Inspired by SuperClue-V, we evaluate the quality of MLLM responses from five key dimensions: Relevance, Fluency, Information Richness, Format Appropriateness, and Correctness.

Multi-Dialect and Multi-Domain Audio Understanding Benchmark. We extend audio understanding benchmarks by incorporating two established testing sets to evaluate the capabilities of MLLMs in multi-dialect and multi-domain settings. Specifically, we collect audios from real users across five specific domains, and invite participants from five regions to record their native dialects. Additionally, we engage skilled human annotators to review all data to ensure the quality and reliability. The final multi-dialect benchmark comprises a total of 25,000 samples collected from five regions: Hunan, Minnan, Guangyue, Chuanyu, and Shanghai (each region has 5,000 cases); while the multi-domain benchmark includes 2,252 samples from five dimensions: Chat (443 cases), Government (462 cases), Health (450 cases), Knowledge (421 cases), and Local-live (476 cases).

4.3 Quantitative Results

We conduct comprehensive evaluations of Ming-Lite-Omni against state-of-the-art MLLMs on 14 different multimodal benchmarks, as illustrated in Table 1∼12. Extensive experiments demonstrate that Ming-Lite-Omni achieves comparable performance with leading MLLMs.

Image → Text (Understanding). In addition to comparable performance on various image understanding benchmarks like grounding (Table 2) and OCR (Table 6) tasks, Ming-Lite-Omni also outperforms the current leading MLLMs on a series of benchmarks, particularly in GUI (Table 3) and knowledge-bases QA (Table 4) tasks. Moreover, Ming-Lite-Omni also shows robust capabilities in addressing expert-level encyclopedia tasks (Table 7) and exhibits superior human-centric interaction patterns when communicating with real users (Table 9).

Specifically, as illustrated in Table 3, Ming-Lite-Omni achieves exceptional advancements on GUI

benchmarks. Ming-Lite-Omni outperforms InternVL3-8B by +2.6% on ScreenSpot and +2.7% on ScreenSpot-V2. Moreover, it achieves an accuracy of 66.6% on AITZ(EM), surpassing the reproduced results of Qwen2.5VL-7B-Insruct by +9.0%, demonstrating robust GUI grounding and action reasoning capabilities in Android environment.

In addition, Ming-Lite-Omni obviously surpasses current leading MLLMs by a large margin on knowledge-intensive benchmarks. As shown in Table 4, Ming-Lite-Omni demonstrates superior performance compared to models even with larger parameters, achieving an obvious performance gain of +8.3%/+9.8%/+7.1% across three dimensions of the InfoSeek benchmark. Additionally, as presented in Table 7, Ming-Lite-Omni achieves an average performance improvement of +5.20% over Qwen2.5VL-7B-Insruct on the in-house encyclopedia benchmark. These results indicate that Ming-Lite-Omni incorporates richer expert-level knowledge and presents superior capability in querying fine-grained information from images, highlighting the effectiveness of integrating high-quality structured data and encyclopedia data.

Lastly, Ming-Lite-Omni exhibits better human-centric interaction patterns when communicating with real users. As illustrated in Table 9, compared to Qwen2.5VL-7B-Insruct, Ming-Lite-Omni presents superior capability in generating human-centric responses that are more relevant, more fluent, better formatted, and less prone to hallucination.

Text → Image (Generation). As shown in Table 5, our experimental results demonstrate that the generation quality of Ming-Lite-Omni is on par with state-of-the-art diffusion models. Notably, Ming-Omni significantly outperforms all baselines in terms of FID, highlighting its top-performing performance in visual quality enhancement and artifact suppression. The slight drop in GenEval is attributed to a trade-off between the instruction-following ability of diffusion models and the artifact sensitivity of autoregressive models. Additionally, we observe that JanusPro achieves higher GenEval scores partly due to the use of rewritten prompts, which may not be a fair performance comparison since our results are evaluated with the original prompts. Lastly, all other baselines are tailored for basic image generation tasks with constrained generalization ability, which could hardly support the rich editing and style variation capabilities offered by our approach. This highlights the core advantage of our unified framework, which leverages the implicit understanding ability of MLLMs to enable more intelligent and controllable generation.

Audio → Text (Understanding). As illustrated in Table 10, Ming-Lite-Omni achieves superior performance on audio understanding tasks, yielding new SoTA results on two out of seven public Chinese benchmarks and setting new SoTA on four out of six public English benchmarks (reaching a total of 6/13 SoTA). Specifically, Ming-Lite-Omni outperforms Qwen2.5-Omni (Xu et al., 2025) on both public audio understanding benchmarks as well as in-house multi-dialect and multi-domain testing sets in terms of average performance, demonstrating its superior capabilities in handling diverse general audio understanding tasks. In addition, Ming-Lite-Omni also presents competitive performance in audio QA tasks compared to SoTA audio-centric methods and unified multi-modal ones in Table 11. These results reflect the effectiveness of incorporating a high-quality audio corpus and the staged training strategies proposed in Sec. 2.2.

Text → Audio (Generation). As shown in Table 12, Ming-Lite-Omni attains competitive results compared to SoTA text-to-speech (TTS) methods, revealing strong capabilities in handling unified audio perception-generation tasks.

Video → Text (Understanding). Following the conventional protocol, we uniformly sample 128 frames for each video during evaluation. As demonstrated in Table 8, Ming-Lite-Omni reaches

a new SOTA in terms of the average metric across four widely-used benchmarks, outperforming Qwen2.5VL-7B-Instruct (Bai et al., 2025a) by +0.2% and LLaVA-OneVision-7B by +9.6% in average performance. Moreover, Ming-Lite-Omni achieves a noticeable improvement of +1.9% on the LongVideoBench benchmark compared with Qwen2.5VL-7B-Instruct, revealing its superior capability in capturing and understanding informative spatial-temporal content particularly for long-duration videos.

4.4 Visualization Results

To further present the unified perception and generation capabilities of Ming-Lite-Omni in addressing various multi-modal tasks, we present a selection of qualitative examples through responses generated from various prompts. As illustrated in Figure 7 and 8, Ming-Lite-Omni is proficient in handling [*Image, Audio, Video, Text*] $\Longleftrightarrow$ [*Image, Audio, Text*] tasks. Specifically, we visualize the capabilities of *Image* $\rightarrow$ *Text* understanding tasks, including OCR, grounding, information-seeking, and reasoning in Figure 7. Furthermore, we also illustrate the capabilities of *Video*&*Audio* $\rightarrow$ *Text* understanding tasks in the bottom of the Figure 7. Lastly, we visualize the capabilities of *Text* $\rightarrow$ *Image* generation tasks in Figure 8, including image generation, image editing, and style transfer.

5 Conclusion

We introduce Ming-Omni, the first open-source model we are aware of to match GPT-4o in modality support. It can perceive text, images, videos and audio modalities and generate text, natural speech in real-time and images simultaneously. Ming-Omni is built on the Ling MoE architecture with modality-specific routers to mitigate modality conflicts, and excels in multi-modal interaction and generation. Ming-Omni attained performance on par with Qwen2.5-VL-7B by activating only 2.8B parameters, and demonstrated SOTA performance in end-to-end speech understanding and speech instruction following. In addition, Ming-Omni also support native-resolution image generation, editing, and style transfer, surpasses mainstream generation models like SDXL. Overall, Ming-Omni demonstrates robust adaptability and efficiency across multimodal perception and generation tasks, showcasing promising prospects for future research and industrial applications.

6 Contributors

Authors are listed **alphabetically by the first name**.

Ant Inclusion AI
Biao Gong
Cheng Zou
Chuanyang Zheng
Chunluan Zhou
Canxiang Yan
Chunxiang Jin
Chunjie Shen
Dandan Zheng
Fudong Wang
Furong Xu
GuangMing Yao
Jun Zhou
Jingdong Chen
Jianxin Sun
Jiajia Liu
Jianjiang Zhu
Jun Peng
Kaixiang Ji
Kaiyou Song
Kaimeng Ren
Libin Wang
Lixiang Ru
Lele Xie
Longhua Tan
Lyuxin Xue
Lan Wang
Mochen Bai
Ning Gao
Pei Chen
Qingpei Guo
Qinglong Zhang
Qiang Xu
Rui Liu
Ruijie Xiong
Sirui Guo
Tinghao Liu
Taisong Li
Weilong Chai
Xinyu Xiao
Xiaomei Wang
Xiaoxue Chen
Xiao Lu
Xiaoyu Li
Xingning Dong
Xuzheng Yu
Yi Yuan
Yuting Gao
Yunxiao Sun
Yipeng Chen
Yifei Wu
Yongjie Lyu
Ziping Ma
Zipeng Feng
Zhijiang Fang
Zhihao Qiu
Ziyuan Huang
Zhengyu He

References

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

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *CLEF 2021 Working Notes*, CEUR Workshop Proceedings, Bucharest, Romania, September 21-24 2021. CEUR-WS.org.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline. In *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment, O-COCOSDA*, pages 1–5. IEEE, 2017.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.130/.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek, editors, *Annual Conference of the International Speech Communication Association, Interspeech*, pages 3670–3674. ISCA, 2021.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression, 2022. https://arxiv.org/abs/2212.02746.

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*, 2024a.

Wentong Chen, Junbo Cui, Jinyi Hu, Yujia Qin, Junjie Fang, Yue Zhao, Chongyi Wang, Jun Liu, Guirong Chen, Yupeng Huo, Yuan Yao, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Guicourse: From general vision language models to versatile gui agents, 2024b.

Wentong Chen, Junbo Cui, Jinyi Hu, Yujia Qin, Junjie Fang, Yue Zhao, Chongyi Wang, Jun Liu, Guirong Chen, Yupeng Huo, Yuan Yao, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Guicourse: From general vision language models to versatile gui agents, 2024c.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.925. https://aclanthology.org/2023.emnlp-main.925/.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023b.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants, 2024d. https://arxiv.org/abs/2410.17196.

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

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Yantao Li, Jianbing Zhang, and Zhiyong Wu. Seeclick: Harnessing gui grounding for advanced visual gui agents, 2024.

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.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*, 2024.

Yuntian Deng, Anssi Kanervisto, Jeffrey Ling, and Alexander M. Rush. Image-to-markup generation with coarse-to-fine attention, 2017. https://arxiv.org/abs/1609.04938.

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

Longfei Ding, Mengbiao Zhao, Fei Yin, Shuiling Zeng, and Cheng-Lin Liu. A large-scale database for chemical structure recognition and preliminary evaluation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 1464–1470, 2022. doi: 10.1109/ICPR56361.2022.9956654.

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*, 2024a.

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

Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, et al. Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario. *arXiv preprint arXiv:2104.03603*, 2021.

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

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, and Shiliang Zhang. Funasr: A fundamental end-to-end speech recognition toolkit, 2023. https://arxiv.org/abs/2305.11013.

Yuying Ge, Sijie Zhao, Chen Li, Yixiao Ge, and Ying Shan. Seed-data-edit technical report: A hybrid dataset for instructional image editing. *arXiv preprint arXiv:2405.04007*, 2024.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022.

Qingpei Guo, Kaiyou Song, Zipeng Feng, Ziping Ma, Qinglong Zhang, Sirui Gao, Xuzheng Yu, Yunxiao Sun, Tai-Wei Chang, Jingdong Chen, et al. M2-omni: Advancing omni-mllm for comprehensive modality support with competitive performance. *arXiv preprint arXiv:2502.18778*, 2025.

Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation, 2024. https://arxiv.org/abs/2407.05361.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024a.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm, 2024b. https://arxiv.org/abs/2402.09181.

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

InclusionAI, Biao Gong, Cheng Zou, Dandan Zheng, Hu Yu, Jingdong Chen, Jianxin Sun, Junbo Zhao, Jun Zhou, Kaixiang Ji, Lixiang Ru, Libin Wang, Qingpei Guo, Rui Liu, Weilong Chai, Xinyu Xiao, and Ziyuan Huang. Ming-lite-uni: Advancements in unified architecture for natural multimodal interaction, 2025. https://arxiv.org/abs/2505.02471.

Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Video question answering with spatio-temporal reasoning. *IJCV*, 2019.

Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos, 2022. https://arxiv.org/abs/2210.03929.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. Libriheavy: A 50, 000 hours ASR corpus with punctuation casing and context. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 10991–10995. IEEE, 2024.

Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. Geomverse: A systematic evaluation of large models for geometric reasoning, 2023. https://arxiv.org/abs/2312.12241.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. https://aclanthology.org/D14-1086/.

KimiTeam, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Jacob Krantz, Erik Wijmans, Arjun Majundar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision and language navigation in continuous environments. In *European Conference on Computer Vision (ECCV)*, 2020.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Wei Li, Shufei Zhang, Mao Su, Wanli Ouyang, Yuqiang Li, and Dongzhan Zhou. Chemvlm: Exploring the power of multimodal large language models in chemistry area, 2025. https://arxiv.org/abs/2408.07246.

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, 2024a.

Wei Li, William Bishop, Alice Li, Chris Rawles, Folawiyo Campbell-Ajala, Divya Tyamagundlu, and Oriana Riva. On the effects of data scale on ui control agents, 2024b. https://arxiv.org/abs/2406.03679.

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*, 2024c.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning, 2024d. https://arxiv.org/abs/2406.11303.

Yupu Liang, Yaping Zhang, Cong Ma, Zhiyang Zhang, Yang Zhao, Lu Xiang, Chengqing Zong, and Yu Zhou. Document image machine translation with dynamic multi-pre-trained models assembling. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7084–7095, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.392. https://aclanthology.org/2024.naacl-long.392/.

LingTeam, Binwei Zeng, Chao Huang, Chao Zhang, Changxin Tian, Cong Chen, Dingnan Jin, Feng Yu, Feng Zhu, Feng Yuan, et al. Every flop counts: Scaling a 300b mixture-of-experts ling llm without premium gpus. *arXiv preprint arXiv:2503.05139*, 2025.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering, 2021. https://arxiv.org/abs/2102.09542.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos?, 2024. https://arxiv.org/abs/2403.00476.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning, 2021. https://arxiv.org/abs/2105.04165.

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.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20, 2016. doi: 10.1109/CVPR.2016.9.

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. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. https://aclanthology.org/2022.findings-acl.177/.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2199–2208, 2021. doi: 10.1109/WACV48630.2021.00225.

Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, Mikhail Sirotenko, Yukun Zhu, and Tobias Weyand. Neptune: The long orbit to benchmarking long video understanding, 2025. https://arxiv.org/abs/2412.09582.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. Spgispeech: 5, 000 hours of transcribed financial audio for fully formatted end-to-end speech recognition. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek, editors, *Annual Conference of the International Speech Communication Association, Interspeech*, pages 1434–1438. ISCA, 2021.

OpenAI. Introducing 4o image generation. https://openai.com/index/introducing-4o-image-generation/, 2025.

Linke Ouyang, Yuan Qu, Hongbin Zhou, Jiawei Zhu, Rui Zhang, Qunshu Lin, Bin Wang, Zhiyuan Zhao, Man Jiang, Xiaomeng Zhao, Jin Shi, Fan Wu, Pei Chu, Minghao Liu, Zhenxiang Li, Chao Xu, Bo Zhang, Botian Shi, Zhongying Tu, and Conghui He. Omnidocbench: Benchmarking diverse pdf document parsing with comprehensive annotations, 2024. https://arxiv.org/abs/2412.07626.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, et al. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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

Liao Qu, Huichao Zhang, Yiheng Liu, Xu Wang, Yi Jiang, Yiming Gao, Hu Ye, Daniel K Du, Zehuan Yuan, and Xinglong Wu. Tokenflow: Unified image tokenizer for multimodal understanding and generation. *arXiv preprint arXiv:2412.03069*, 2024.

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

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark, 2024. https://arxiv.org/abs/2405.08813.

Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Android in the wild: a large-scale dataset for android device control. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023a. Curran Associates Inc.

Christopher Rawles, Alice Li, Daniel Rodriguez, Oriana Riva, and Timothy Lillicrap. Android in the wild: A large-scale dataset for android device control, 2023b. https://arxiv.org/abs/2307.10088.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129, 2012.

Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed Huai hsin Chi, James Caverlee, Julian J. McAuley, and Derek Zhiyuan Cheng. How to train data-efficient llms. *ArXiv*, abs/2402.09668, 2024. https://api.semanticscholar.org/CorpusID:267682083.

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. *arXiv preprint arXiv:2210.08402*, 2022.

Minjoon Seo, Hannaneh Hajishirzi, Ali Farhadi, Oren Etzioni, and Clint Malcolm. Solving geometry problems: Combining text and diagram interpretation. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1466–1476, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1171. https://aclanthology.org/D15-1171/.

Jiatong Shi, Chunlei Zhang, Jinchuan Tian, Junrui Ni, Hao Zhang, Shinji Watanabe, and Dong Yu. Balancing speech understanding and generation using continual pre-training for codec-based speech llm, 2025. https://arxiv.org/abs/2502.16897.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv preprint arXiv:2010.11567*, 2020.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019.

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, et al. Kespeech: An open source speech dataset of mandarin and its eight subdialects. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

Changhan Wang, Juan Pino, Anne Wu, and Jiatao Gu. Covost: A diverse multilingual speech-to-text translation corpus. *arXiv preprint arXiv:2002.01320*, 2020a.

Changhan Wang, Anne Wu, and Juan Miguel Pino. Covost 2: A massively multilingual speech-to-text translation corpus. *CoRR*, abs/2007.10310, 2020b.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*, 2021.

Haoxu Wang, Fan Yu, Xian Shi, Yuezhang Wang, Shiliang Zhang, and Ming Li. Slidespeech: A large scale slide-enriched audio-visual corpus. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 11076–11080. IEEE, 2024a. https://doi.org/10.1109/ICASSP48485.2024.10448079.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Conghui He, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation, 2024b. https://arxiv.org/abs/2307.06942.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation, 2024a. https://arxiv.org/abs/2410.13848.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024b.

Shengguang Wu, Keming Lu, Benfeng Xu, Junyang Lin, Qi Su, and Chang Zhou. Self-evolved diverse data sampling for efficient instruction tuning. *arXiv preprint arXiv:2311.08182*, 2023.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, and Yu Qiao. Os-atlas: A foundation action model for generalist gui agents, 2024c. https://arxiv.org/abs/2410.23218.

Zhiyong Wu, Zhenyu Wu, Fangzhi Xu, Yian Wang, Qiushi Sun, Chengyou Jia, Kanzhi Cheng, Zichen Ding, Liheng Chen, Paul Pu Liang, et al. Os-atlas: A foundation action model for generalist gui agents, 2024d.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. Funqa: Towards surprising video comprehension, 2024. https://arxiv.org/abs/2306.14899.

Chunyu Xie, Heng Cai, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Jianfei Song, Henrique Morimitsu, Lin Yao, Dexin Wang, Xiangzheng Zhang, Dawei Leng, Baochang Zhang, Xiangyang Ji, and Yafeng Deng. Ccmb: A large-scale chinese cross-modal benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 4219–4227, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3611877. https://doi.org/10.1145/3581783.3611877.

Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, et al. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *arXiv preprint arXiv:2306.04362*, 2023.

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

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *arXiv preprint arXiv:2406.06040*, 2024.

Wentao Yang, Zhe Li, Dezhi Peng, Lianwen Jin, Mengchao He, and Cong Yao. Read ten lines at one glance: Line-aware semi-autoregressive transformer for multi-line handwritten mathematical expression recognition, 2023.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, Lei Xie, and Yonghong Yan. Open source magicdata-ramc: A rich annotated mandarin conversational(ramc) speech dataset. In Hanseok Ko and John H. L. Hansen, editors, *Annual Conference of the International Speech Communication Association, Interspeech*, pages 1736–1740. ISCA, 2022.

Kexin Yi*, Chuang Gan*, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020. https://openreview.net/forum?id=HkxYzANYDB.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 6167–6171. IEEE, 2022.

Wenwen Yu, Chengquan Zhang, Haoyu Cao, Wei Hua, Bohan Li, Huang Chen, Mingyu Liu, Mingrui Chen, Jianfeng Kuang, Mengjun Cheng, Yuning Du, Shikun Feng, Xiaoguang Hu, Pengyuan Lyu, Kun Yao, Yuechen Yu, Yuliang Liu, Wanxiang Che, Errui Ding, Cheng-Lin Liu, Jiebo Luo, Shuicheng Yan, Min Zhang, Dimosthenis Karatzas, Xing Sun, Jingdong Wang, and Xiang Bai. Icdar 2023 competition on structured text extraction from visually-rich document images, 2023. https://arxiv.org/abs/2306.03287.

Albert Zeyer, André Merboldt, Wilfried Michel, Ralf Schlüter, and Hermann Ney. Librispeech transducer model with internal language model prior correction. In Hynek Hermansky, Honza Cernocký, Lukás Burget, Lori Lamel, Odette Scharenborg, and Petr Motlícek, editors, *Annual Conference of the International Speech Communication Association, Interspeech*, pages 2052–2056. ISCA, 2021.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. WENETSPEECH: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, pages 6182–6186. IEEE, 2022a.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition, 2022b. https://arxiv.org/abs/2110.03370.

Jiwen Zhang, Jihao Wu, Yihua Teng, Minghui Liao, Nuo Xu, Xiao Xiao, Zhongyu Wei, and Duyu Tang. Android in the zoo: Chain-of-action-thought for gui agents. *arXiv preprint arXiv:2403.02713*, 2024a.

Jiwen Zhang, Jihao Wu, Yihua Teng, Minghui Liao, Nuo Xu, Xiao Xiao, Zhongyu Wei, and Duyu Tang. Android in the zoo: Chain-of-action-thought for gui agents, 2024b. https://arxiv.org/abs/2403.02713.

Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Ziyu Guo, Shicheng Li, Yichi Zhang, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, Peng Gao, Chunyuan Li, and Hongsheng Li. Mavis: Mathematical visual instruction tuning with an automatic data engine, 2024c. https://arxiv.org/abs/2407.08739.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024d. https://arxiv.org/abs/2410.02713.

Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37: 3058–3093, 2024.

Appendix

A Open-source image data

In addition to the open source image data utilized in (Guo et al., 2025), the newly added data are presented in Table 13.

Table 13 The list of newly added open-source image data used during our training.

Dataset
OS-ATLAS (Wu et al., 2024c)
M2E (Yang et al., 2023)
IM2LATEX-100K (Deng et al., 2017)
Mini-CASIA-CSDB (Ding et al., 2022)
CASIA-CSDB (Ding et al., 2022)
DoTA (Liang et al., 2024)
ICDAR23-SVRD (Yu et al., 2023)
AitZ (Zhang et al., 2024b)
AitW (Rawles et al., 2023b)
GUICourse (Chen et al., 2024c)
OmniMedVQA (Hu et al., 2024b)
SLAKE (Liu et al., 2021)
VQA-Med (Ben Abacha et al., 2021)
Geometry3K (Lu et al., 2021)
UniGeo (Chen et al., 2022)
MAVIS (Zhang et al., 2024c)
GeoS (Seo et al., 2015)
PixMo-count (Deitke et al., 2024)
Geoqa+ (Cao and Xiao, 2022)
GeomVerse (Kazemi et al., 2023)
ChemVLM (Li et al., 2025)

B Open-source audio data

We include the complete list of open source audio data we used during our training in Table 14.

C Open-source video data

The comprehensive list of open-source video datasets utilized during the training process is provided in Table 15.

Table 14 The complete list of open-source audio data used during our training.

Dataset	Audio Length (hrs)
WenetSpeech (Zhang et al., 2022b)	10518
KeSpeech (Tang et al., 2021)	1428
AliMeeting (Yu et al., 2022)	120
AISHELL-1 (Bu et al., 2017)	155
AISHELL-3 (Shi et al., 2020)	65
AISHELL-4 (Fu et al., 2021)	61
CoVoST (Wang et al., 2020a)	456
CoVoST2 (Wang et al., 2020b)	18
Magicdata (Yang et al., 2022)	747
Gigaspeech (Chen et al., 2021)	10288
Libriheavy (Kang et al., 2024)	51448
LibriSpeech (Zeyer et al., 2021)	960
SlideSpeech (Wang et al., 2024a)	473
SPGISpeech (O'Neill et al., 2021)	5000
TED-LIUM (Rousseau et al., 2012)	208
Emilla (He et al., 2024)	90305
Multilingual LibriSpeech (Pratap et al., 2020)	45000
Peoples Speech (Galvez et al., 2021)	30000

Table 15 The complete list of open-source video data used during our training.

Dataset
TGIF-Transition (Jang et al., 2019)
ShareGPT4Video (Chen et al., 2024a)
videogpt-plus (Maaz et al., 2024)
Llava-video-178k (Zhang et al., 2024d)
Video-Vista (Li et al., 2024d)
Neptune (Nagrani et al., 2025)
FunQA (Xie et al., 2024)
Temp-Compass (Liu et al., 2024)
EgoTask (Jia et al., 2022)
InternVid (Wang et al., 2024b)
CLEVRER (Yi* et al., 2020)
VLN-CE (Krantz et al., 2020)
Vript (Yang et al., 2024)
Cinepile (Rawal et al., 2024)
OpenVid-1M (Nan et al., 2024)