Ming-Flash-Omni: A Sparse, Unified Architecture for Multimodal Perception and Generation

Inclusion AI, Ant Group*

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

*We propose Ming-Flash-Omni, an upgraded version of Ming-Omni, built upon a sparser Mixture-of-Experts (MoE) variant of Ling-Flash-2.0 with 100 billion total parameters, of which only 6.1 billion are active per token. This architecture enables **highly efficient scaling** (dramatically improving computational efficiency while significantly expanding model capacity) and empowers **stronger unified multimodal intelligence** across vision, speech, and language, representing a key step toward Artificial General Intelligence (AGI). Compared to its predecessor, the upgraded version exhibits substantial improvements across multimodal understanding and generation. We significantly advance speech recognition capabilities, achieving state-of-the-art performance in **contextual ASR** and highly competitive results in **dialect-aware ASR**. In image generation, Ming-Flash-Omni introduces **high-fidelity text rendering** and demonstrates marked gains in **scene consistency** and **identity preservation** during image editing. Furthermore, Ming-Flash-Omni introduces **generative segmentation**, a capability that not only achieves strong standalone segmentation performance but also enhances spatial control in image generation and improves editing consistency. Notably, Ming-Flash-Omni achieves state-of-the-art results in text-to-image generation and generative segmentation, and sets new records on all 12 contextual ASR benchmarks, all within a single unified architecture.*

Date: Oct 28, 2025
Project Homepage: https://github.com/inclusionAI/Ming

1 Introduction

In everyday life, humans naturally integrate visual and auditory cues to express ideas through speech or writing, while also forming vivid mental images from descriptions or concepts. This ability to visualize enhances creativity, problem-solving, and communication, serving as a core aspect of human intelligence and interaction. The ultimate goal of Artificial General Intelligence (AGI) is to replicate this human-like multimodal intelligence, evolving from a mere tool into a powerful agent that augments and liberates human productivity.

Driven by advances in Large Language Models (LLMs) and extensive training on large-scale multimodal datasets, Multi-modal Large Language Models (MLLMs) have demonstrated remarkable perceptual capabilities in both vision (Chen et al., 2024d; Bai et al., 2025b; KimiTeam et al., 2025; Xu et al., 2025c) and audio (Ding et al., 2025a; Xu et al., 2025a,c), as well as generative capabilities in these two modalities (Huang et al., 2025; Ding et al., 2025a; OpenAI, 2025; Tong et al., 2024b; Pan et al., 2025; Xu et al., 2025c). Nevertheless, effectively integrating comprehension and generation across multiple modalities into a unified model remains challenging. While humans naturally learn by combining multiple modalities, leveraging their complementary strengths and interactions to enhance overall learning efficiency, building a unified Omni-MLLM is hindered by representational disparities and modality imbalances.

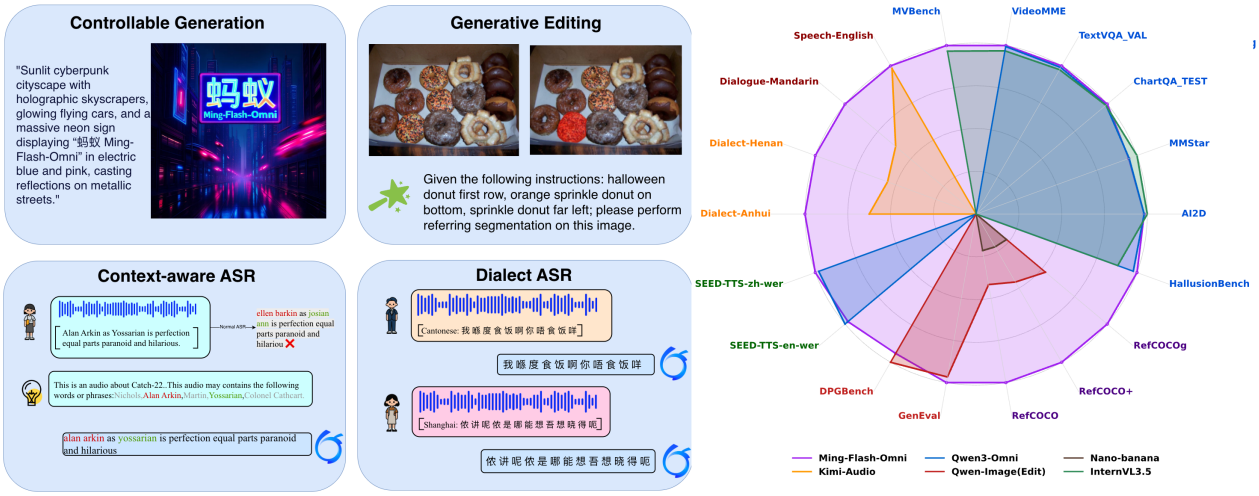

Figure 1 Ming-Flash-Omni generally demonstrates highly competitive performance across various domains, including vision-text understanding, controllable image generation, speech recognition, and speech synthesis. Specifically, in image generation, Ming-Flash-Omni introduces a novel generative segmentation paradigm to achieve fine-grained spatial and semantic control over the generated images. Moreover, Ming-Flash-Omni significantly enhances Context-Aware Speech Recognition (ContextASR) and Chinese dialect recognition, thereby broadening its applicability in real-world scenarios.

In this paper, we introduce Ming-Flash-Omni, which builds upon the Ming-Omni architecture with a redesigned foundation and targeted enhancements across multimodal understanding and generation. At its core, Ming-Flash-Omni adopts Ling-Flash-2.0 lin (2025) (a scaled-up, highly sparse Mixture-of-Experts architecture) where an increased sparsity ratio enables substantial model capacity while maintaining bounded inference latency, striking a favorable trade-off between performance and efficiency.

On the understanding side, the model introduces two key advances. First, Ming-Flash-Omni upgrade the positional encoding to VideoRoPE Wei et al. (2025), a refined variant specifically designed to better capture temporal dynamics in video sequences, thereby enhancing the model's ability to understand complex visual events. Second, Ming-Flash-Omni focus on improving the context-aware ASR capability itself, enhancing the model's ability to leverage surrounding linguistic context during speech recognition and thereby achieving more accurate transcription in context-dependent scenarios.

On the generation side, Ming-Flash-Omni introduces three key advancements: 1) in speech synthesis, discrete acoustic tokens are replaced with continuous representations, effectively circumventing quantization-induced artifacts and yielding more natural and expressive TTS outputs; 2) the model supports generative semantic segmentation, enabling pixel-level semantic content generation conditioned on multimodal inputs; and 3) it enables fine-grained controllable image generation with improved identity preservation and in-image text generation capabilities.

These architectural innovations empower Ming-Flash-Omni to deliver exceptional cross-modal performance in both comprehension and generation tasks. Specifically, in the image perception task, Ming-Flash-Omni attained performance comparable to that of Qwen3-Omni (Xu et al., 2025c). Ming-Flash-Omni also delivers superior performance in end-to-end speech understanding and generation.For instance, it achieves SOTA on all 12 metrics on ContextASR-Bench.

The remainder of this paper is organized as follows. Section 2 presents the detailed architecture of Ming-Flash-Omni. Sections 3 describes the pretraining and post-training datasets. Section

4 reports the evaluation results and compare Ming-Flash-Omni with recent multimodal models. Sections 5 is conclusion.

2 Ming-Flash-Omni

As illustrated in Figure 2, Ming-Flash-Omni retains the unified two-stage pipeline of Ming-Omni AI et al. (2025a), where perception supports multimodal understanding and generation targets speech and image synthesis, while markedly advancing long-context modeling, reasoning, and controllable generation. At the core is Ling-Flash-2.0 lin (2025), a sparse MoE LM (100B; 6.1B per token) with a dual balancing scheme that stabilizes training and improves efficiency. On the perception side, Ming-Flash-Omni employs VideoRoPE Wei et al. (2025) for temporal modeling, context-aware ASR for more reliable speech understanding, and a think mode for deeper multi-step reasoning. On the generation side, we replace discrete speech tokens with continuous acoustic latents, avoiding quantization loss and improving fidelity; for images, we upgrade to a synergistic training paradigm that enables generative segmentation-as-editing, facilitating fine-grained and controllable generation. Overall, Ming-Flash-Omni advances the unified model with stable expert routing and scalable long-context modeling, yielding more reliable multimodal understanding and controllable generation.

2.1 Unified Understanding Across Modalities

The cornerstone of Ming-Flash-Omni is enhanced multimodal understanding. We retain the established visual and audio encoders (Qwen2.5 (Bai et al., 2025a) and Whisper (Radford et al., 2023)) and feed their projected embeddings, concatenated with tokenized text, into Ling-flash-2.0, a sparse MoE language model with distinct routers per modality. Beyond this, Ming-Flash-Omni incorporates VideoRoPE to maintain temporal coherence over long-range frame sequences, thus emphasizing temporal modeling. Furthermore, Ming-Flash-Omni adopt a context-aware ASR training paradigm that conditions decoding on task or domain context, addressing common shortcomings of conventional ASR in real-world, multi-domain scenarios (limited world knowledge and unreliable proper-noun recognition) and yielding more accurate, context-consistent transcripts. To stabilize training in the more sparse Ling-flash-2.0, we employ a hybrid expert-balancing scheme that combines an auxiliary load-balancing loss (as in Ming-Omni) with per-router bias updates (as in Ling-flash-2.0), promoting uniform expert activation and improved convergence.

2.2 Unified Speech Understanding and Generation

Ming-Flash-Omni retains the same audio encoder as Ming-Omni (AI et al., 2025a), with optimization efforts focused on improving contextual automatic speech recognition (ASR) performance by incorporating preceding textual context and hotword lists during training. For generation, we replace discrete speech tokens with continuous acoustic latents, avoiding quantization loss and improving fidelity. Specifically, we utilize a fixed, pre-trained audio head based on Qwen2.5 (0.5B parameters), which takes LLM-generated text tokens and downsampled VAE latents as input and autoregressively predicts the conditioning signals for the flow-matching head—following the paradigm of Jia et al. (2025).

It is worth noting that the continuous features used for our generation tasks are derived from our unified continuous speech tokenizer, which is based on a VAE-GAN architecture, with the overall training objective comprising two main components: a generator loss and a discriminator loss. The generator loss consists of (i) a multi-scale mel-spectrogram reconstruction loss, (ii) an adversarial

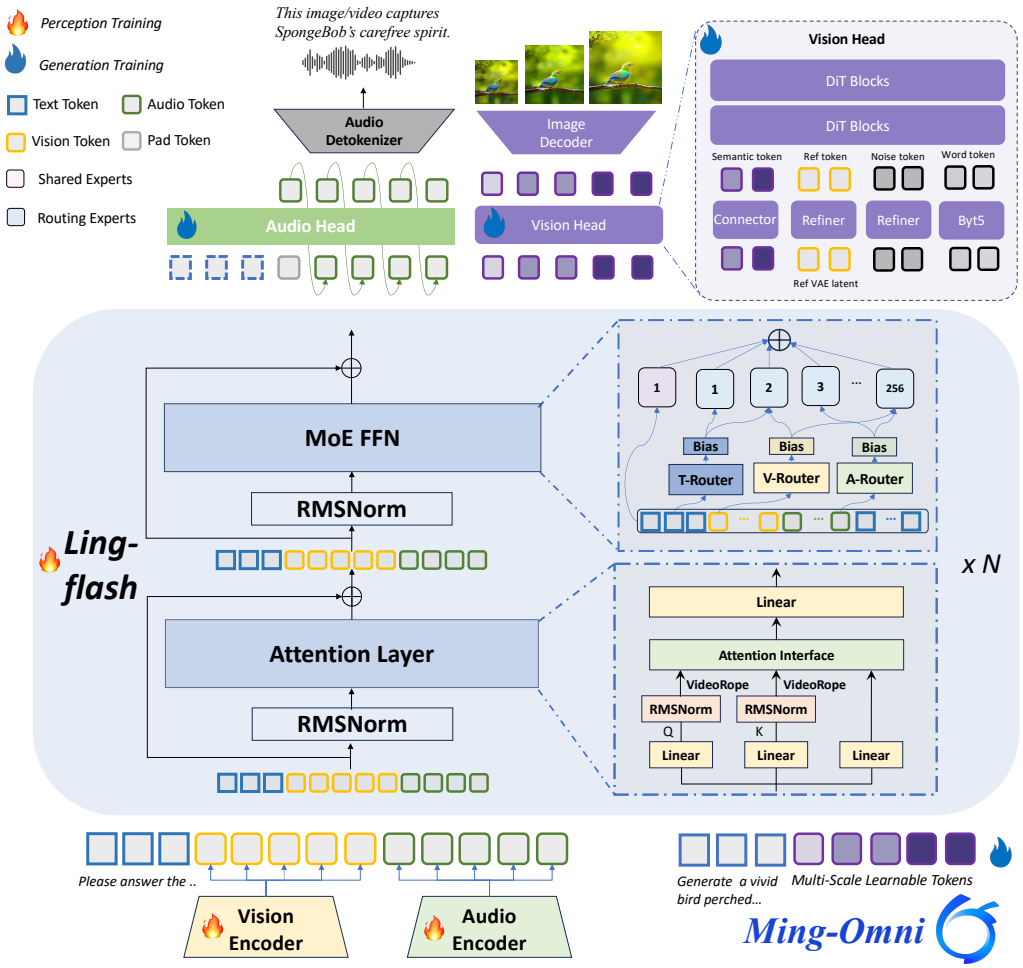

Figure 2 The overall framework of Ming-Flash-Omni. This version features a sparser LLM based on Ling-flash-2.0 MoE architecture, and integrates VideoRoPE to enhance temporal modeling. Speech generation now uses continuous features instead of discrete tokens, and image generation has been upgraded with support for segmentation.

loss, (iii) a feature matching loss, and (iv) a KL divergence regularization term. The discriminator, in turn, is trained primarily with the adversarial loss. For more implementation details, please refer to Canxiang et al. (2025).

2.3 Unified Image Understanding and Generation

A core challenge in building unified multimodal models is the effective fusion of image understanding and generation. While our Ming-Omni model injects hierarchical semantics via multi-scale query tokens, its language pathway remains frozen during training to prevent interference from the generative objective. This freezing approach, while ensuring stability, creates a critical bottleneck: an inherent discrepancy between the learning objectives of understanding and generation. Consequently, even with injected hierarchical semantics, fine-grained visual knowledge—such as object attributes and spatial relationships—cannot be efficiently transferred to high-precision generation and editing tasks, limiting the model's final quality and controllability.

Generative Segmentation as an Editing Task To overcome this bottleneck, Ming-Flash-Omni propose a synergistic training paradigm that reframes image segmentation as a generative editing

task. Instead of producing an abstract binary mask (e.g., for "segment the banana"), the model performs a semantics-preserving edit (e.g., "color the banana purple"). This reformulation forcibly binds the learning objectives of understanding and generation: successful generation requires a precise understanding of the object's contours and boundaries. Understanding thus becomes a *prerequisite* for editing, and the edit's quality provides a direct supervisory signal for the model's comprehension, fundamentally unifying their optimization objectives.

Crucially, this training cultivates a more fundamental, generalizable skill: **fine-grained spatio-semantic control**, which indirectly resolves the compositionality problem in pure text-to-image generation. **In our evaluation on the GenEval benchmark,** `Ming-Flash-Omni` **achieved a score of 0.90, surpassing leading non-Reinforcement Learning (non-RL) methods.** This result suggests that the foundational skill of spatio-semantic control can effectively generalize to pure text-to-image generation tasks.

Empowering Advanced Controllable Capabilities This mastery of spatio-semantic control provides a solid foundation for a suite of advanced functions:

Identity (ID) Preservation. We use a VAE-encoded identity vector and a composite loss (global semantic and local pixel) to maintain subject consistency. The model's learned boundary perception is crucial for accurately isolating non-edited regions, ensuring high faithfulness.

High-Fidelity Text Rendering. By integrating a specialized Glyph-byT5 text encoder, our model leverages its learned pixel-level control to accurately place text, ensuring seamless contextual integration and high-quality results.

Multi-Image Fusion and Style Transfer. The model can deconstruct and recombine elements from multiple images (e.g., ID from A, background from B) using distinct **Concept Vectors**. This complex fusion directly relies on the delineation skill acquired through our core training paradigm.

2.4 Overall Training Procedure

The training procedure of `Ming-Flash-Omni` retains a two-stage pipeline: perception and generation.

The perception stage, consistent with the `Ming-Omni`, includes pre-training, instruction tuning, alignment tuning, and an additional coherent RL phase. The coherent RL phase proceeds sequentially: Dynamic-GRPO (D-GRPO) AI et al. (2025b) followed by Unified-DPO (U-DPO). First, we use D-GRPO, applying RL with verifiable rewards on datasets with checkable answers (via on-policy GRPO with balanced sampling, dynamic hyperparameter adjustment, and task-specific accuracy rewards) to reinforce correct reasoning across multimodal tasks. Next, we employ U-DPO for preference alignment, extending standard DPO with an auxiliary instruction-tuning loss term on the chosen samples to enhance instruction adherence and stylistic coherence. Concretely, we employ a replay strategy with instruction-stage multimodal data and interleave updates on DPO and instruction-tuning data to stabilize optimization and maintain balanced gradient flow.

After perception, we freeze the perception MLLM and optimize only the image generator, while leveraging the pre-trained audio generator from Canxiang et al. (2025). For image generation, the training procedure contains two sequential stages. In the first stage, we pre-train a diffusion-based image generator using a flow matching objective, while keeping the perception MLLM frozen. The generator is equipped with multi-scale learnable queries to capture hierarchical visual semantics from textual inputs. In the second stage, we extend the model to support image editing by conditioning

the denoising process on reference images: the VAE-encoded representations of input images are concatenated with the noisy latent to enforce structural and semantic consistency with the original content. Additionally, input word-level captions are encoded via ByteT5 embeddings to enrich textual conditioning.

2.5 Infrastructure

Compared to large language models (LLMs), training multimodal foundation models presents several key challenges, primarily stemming from **data heterogeneity** and **model heterogeneity**. First, data heterogeneity arises from the need to dynamically switch between diverse input modalities (text, images, audio, and video) during training. These modalities exhibit significant differences in tensor shape, most notably in the form of dynamic batch sizes and variable-length sequences. This variability complicates the design of a unified parallel computation layout. As a result, computational workloads become unevenly distributed across processing ranks, leading to load imbalance. Moreover, the frequent allocation and deallocation of GPU memory buffers for inputs of varying shapes induce severe memory fragmentation, substantially degrading training efficiency and hardware utilization. Second, in contrast to large language models (LLMs), which are predominantly based on homogeneous, decoder-only Transformer architectures, multimodal foundation models typically employ modality-specific encoders at the input stage, introducing model heterogeneity. Although these encoders are relatively lightweight in terms of parameter count, they are highly sensitive to parallelization strategies. If not carefully partitioned across devices, they can induce substantial pipeline bubbles (*i.e.*, idle computation cycles) during pipeline-parallel execution, thereby constraining overall training throughput.

To address these challenges, `Ming-Flash-Omni` is trained on an enhanced version of the Megatron-LM Shoeybi et al. (2019) framework with two key extensions tailored for multimodal workloads:

- Sequence Packing for Data Heterogeneity: To handle dynamic input shapes, we integrate sequence packing, which densely packs multiple variable-length sequences into fixed-length batches. This significantly improves memory utilization and computational density.
- Flexible Encoder Sharding for Model Heterogeneity: To mitigate pipeline bubbles caused by encoders, we extend Megatron-LM to support fine-grained encoder sharding, enabling flexible partitioning across data parallelism (DP), pipeline parallelism (PP), and tensor parallelism (TP). This achieves more balanced workloads and higher end-to-end training efficiency.

These optimizations collectively achieve more than twice the training throughput of the baseline Megatron-LM implementation.

3 Data Construction

We have collected a large and diverse set of training data to enable models to process and understand information from multiple modalities, including text, images, audio and videos. The majority of this data comes from `Ming-Omni` (AI et al., 2025a). In addition, we develop several data processing pipelines to ensure data quality, diversity and deduplication. Establishing an effective multimodal data strategy is essential for the joint multi-modal training, as it facilitates seamless alignment of knowledge across diverse modalities. We categorize the training data based on the core modalities they are designed to enhance, including image, audio, video, and text. The detailed sources and construction methods for each type of data are elaborated in this section.

3.1 Image Data

Image data serves as the cornerstone of our multi-modal corpus. Following Ming-Omni (AI et al., 2025a), we integrate both image-understanding and image-generation datasets to enable the MLLM to acquire unified perception and generation capabilities. Additionally, we further design novel pipelines to synthesize high-quality datasets across diverse dimensions to improve model's capabilities and user interaction quality.

3.1.1 Image Understanding Data

OCR Data: Text recognition and document understanding capabilities are crucial for MLLM. We construct a large-scale heterogeneous training dataset with millions of samples, consisting of three data sources: open-source data, expert-collaborative pseudo-labeled data, and human-annotated enhancement data. The expert-collaborative pseudo-labeled data is generated by diagnosing model weaknesses, and using expert models to label targeted data. In addition, to enhance the model's capability in text–visual analysis and logical reasoning, we incorporate the Chain-of-Thought (CoT) paradigm into the training data. We incorporate the open-source ChartQA-PoT dataset to enhance the model's numerical reasoning ability on charts and pioneeringly use executable Python code as the intermediate reasoning representation.

Reasoning Data: In the reasoning training of Ming-Flash-Omni, we enrich the CoT data to enhance the model's reasoning capabilities. These data are primarily constructed around three key themes: mathematical logic, spatial reasoning, and GUI reasoning. We design an efficient CoT generation and filtering pipeline to construct high-quality, well-structured reasoning data that enhances the model's multi-step reasoning capability. (1) We sample long CoT data using state-of-the-art multimodal reasoning models (*e.g.*, Gemini (Gemini et al., 2023)) to build an initial CoT pool. (2) We evaluate the accuracy and quality of the synthesized CoT data and filter out the low-quality data. Based on this pipeline, we construct 1.5M multimodal long CoT samples, with a maximum length of 16K tokens. Besides, to overcome the limitations of the "direct answer" pattern in text-based QA data, we use reinforcement learning models to generate multi-step reasoning traces and final answers. Experimental results demonstrate that this data significantly improves Ming-Flash-Omni's performance on complex reasoning tasks.

Preference alignment Data: To further enhance user interaction and response quality, we introduce preference alignment data, focusing on three main aspects: (1) Instruction intent understanding: To enhance the model's ability to understand the true intent behind user instructions, we design a novel instruction intent reasoning paradigm where the model first determines if the instruction is clear. For clear instructions, it provides structured reasoning and final answers. For ambiguous ones, it infers the likely intent, generates clarification prompts, and aligns responses accordingly. By incorporating experience-alignment training, we significantly improve the model's intent understanding and enhance the overall user experience. (2) Multi-turn conversation: Users often engage in repeated questioning on the same context. To maintain semantic consistency across turns, we decompose complex instructions into multi-turn conversations and generate high-quality responses using expert-annotated LLM pipelines (Laban et al., 2025). This ensures context retention and reduces performance degradation in multi-turn scenarios. (3) Complex multimodal instruction following: Users often provide sequential, interdependent commands. We design a multimodal instruction generation pipeline to generate SFT and DPO-style data to cover basic and complex instruction types (Zhang et al., 2024; Qian et al., 2024; Ding et al., 2025b).

Structured Data: Structured data enhances MLLM capabilities in querying fine-grained knowledge

associated with specific visual entities. To enhance the MLLM's ability to handle knowledge-intensive and information-seeking queries, we design two pipelines to generate large-scale entity-relation and encyclopedic QA data. (1) *Entity-relation QA data:* We design a multi-stage pipeline to construct a high-quality entity-relation QA corpus. The pipeline integrates scene graph extraction, visual grounding, entity–relation verification, and diverse QA generation to produce pairs covering entities, attributes, and relations. To ensure quality, we filter source datasets for complex multi-entity scenes and apply two-stage validation before and after QA generation. (2) *Encyclopedia QA Data:* We design an automated pipeline to generate encyclopedia entity QA data from Chinese Baidu Encyclopedia and English Wikidata. It constructs <image, entity, knowledge> triplets by extracting entities from images or retrieving images for knowledge-base entities, filtering out invalid ones, and validating image–text consistency. These triplets are then converted into VQA data using LLMs.

3.1.2 Image Generation Data

Image generation data enhances MLLM capabilities beyond traditional image understanding tasks. In addition to the image generation data used in Ming-Omni (AI et al., 2025a), we specifically integrate image segmentation data, text generation data and portrait preservation data to further improve the user experience.

Image segmentation data: To improve the model's generative segmentation capability, we construct two types of data: (1) We use the open-source referring segmentation datasets RefCOCO/+/g (Kazemzadeh et al., 2014; Mao et al., 2016) to construct image editing data. The original image serves as the reference, and binary masks highlight target regions with specified colors to create edited images. (2) For semantic and panoptic segmentation, samples are built from COCO-Panoptic (Lin et al., 2014; Kirillov et al., 2019), where each class or instance is assigned a unique color via a predefined colormap to generate edited images.

Portrait preservation data: The portrait preservation data consist of two data sources: (1) ID Photo Dataset: We collected and constructed 200k paired lifestyle-ID photos. And filter the data using four criteria, *e.g.*, face similarity, face size and confidence, face angle and manual review. (2) Landmark Check-In Portrait Dataset: We collect 20K high-quality portraits from 225 landmarks.The original images serve as edited images, while using advanced segmentation model like SAM2 (Ravi et al., 2024) to segments the foreground person and places them onto 1,000 manually collected daily background scenes as pre-edited images. And then use LLM generate diverse prompts for each landmark.

Text generation data: We build a Chinese-English text generation dataset across three difficulty levels: (1) Monotonic background text rendering: Text is rendered directly by setting background color, font type, size, color, and position. (2) Text rendering on existing images: texts are rendered on suitable smooth regions obtained by Felzenszwalb algorithm. (3) Text-image integrated rendering: Using the SOTA LLMs to generate text rendering prompts, the advanced generation models (*e.g.*, Qwen-image (Wu et al., 2025) and Nano Banana) are used for image generation, followed by OCR for consistency checks, resulting in a high-quality dataset.

3.2 Audio Data

For audio data, we mainly use the data from Ming-Omni (AI et al., 2025a). In addition, we incorporate the following three datasets to further enhance the model's audio understanding and generation capabilities.

Context ASR Data: Current ASR systems face challenges in recognizing homophones or phonetically similar words when the context is limited, pronunciations are unclear, or accents are present. ContextASR addresses these issues by leveraging the preceding context. We propose to synthesize a large-scale dataset using LLMs to endow models with ContextASR capabilities. We extract named entities and construct context passages using LLMs based on existing ASR text, producing 3 million Chinese and English samples in the format <audio, text, context, entity_list>. During training, we further filter and sample the data to reduce keyword density, remove keywords that are absent from the text, and generate negative samples to enhance the model's discriminative ability.

TTS Data: The diversity of TTS data is essential for fully leveraging the pretrained language model's capabilities in audio generation. In addition to the open-source data used in Ming-Omni, we develop a data generation pipeline to create large-scale TTS data. Specifically, (1) we crawl extensive audio data using keywords expanded from handcrafted seeds through domain-specific lexical variations, (2) apply VAD (Gao et al., 2023) to segment well-conditioned short clips, and (3) iteratively train an audio labeler—initially on high-quality data, then using its predictions to label the corpus and refine its accuracy. Based on the short audio clips and the audio labeler, we acquire a large number of high-quality audio clips with labels from different domains.

3.3 Video Data

Video streaming conversation constitutes a fundamental capability for MLLMs in video understanding. However, acquiring large-scale streaming multi-turn conversation data is prohibitively expensive. In this work, we propose a pipeline to systematically synthesize diverse, balanced, and high-quality multi-turn conversation video datasets. We collect 8.2M videos from the internet, ranging from 90 seconds to 10 minutes, and first filter out low-quality videos with high speech density, high shot density, irregular aspect ratios, or low resolution. We then filter out low-information, incoherent, or overly simple videos using SOTA MLLMs. To ensure a balanced dataset, we use advanced embedding models (*e.g.*, Qwen3-Embedding (Zhang et al., 2025) and M3-Embedding (Chen et al., 2024a)) to extract embeddings and then cluster the videos to suppress high-frequency data while preserving long-tail content. Finally, we use SOTA video understanding models to generate high-quality video conversations. This produces 1.2M conversation turns across 5-minute average videos, balanced across various task categories.

3.4 Text Data

For text data, we utilize corpus from Ling (LingTeam et al., 2025), M2-Omni (Guo et al., 2025), and Ming-Omni (AI et al., 2025a) to preserve and further enhance the model's language proficiency.

4 Evaluation

In this section, we present the evaluation details and quantitative examples of Ming-Flash-Omni on both public and in-house benchmarks.

4.1 Public Benchmarks

The details of the public benchmarks are provided in Appendix A. As shown in Table 1∼9, our holistic assessment covers more than 50 rigorously curated public benchmarks across the following seven distinct multi-modal dimensions: Image → Text (Understanding), Text → Image (Generation),

Table 1 Performance of Ming-Flash-Omni on **Vision-to-Text Benchmarks** compared to leading models.* denotes metrics tested using the official benchmark prompts.

Type	Benchmark	Ming-Flash Omni	Qwen3-Omni 30B-A3B	Qwen3-VL 30B-A3B	InternVL3.5 30B-A3B
General	MMStar	68.3	68.5	**72.1**	72
	AI2D	85.2	85.2	**86.9**	86.8
	HallusionBench	61.1	59.7	**61.5**	53.8
	CV-bench	**81.6**	-	-	77.3
	MathVista_MINI	**81.9**	80.0	**81.9**	80.9
	CRPE	78.4	-	**80.0***	77.6
OCR	ChartQA	**88.4**	87.5	87.1	87.4
	DocVQA	94.8	95.0	**95.2**	94.2
	OCRBench	879	860	**903**	880
	TextVQA	**82.6**	81.7	81.7	80.5
Complex instruction	MIA-Bench	93.8	**94.5**	94.4	-
Multi-image	MMTBench_val_mi	**68.0**	-	65.7*	-
	MuirBench	61.5	-	**62.9**	-
	Llava_interleave_Bench	**63.3**	-	51.1*	-
Video	MVBench	**74.6**	-	72.3	72.1
	VideoMME	70.9	70.5	**74.5**	68.7
	VideoMME_w_subtitle	**73.0**	-	-	71.8
	LongVideoBench	61.7	-	-	**63.8**

Table 2 Performance of Ming-Flash-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
		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
	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
	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**	**84.08**
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
	JanusFlow	0.97	0.59	0.45	0.83	0.53	0.42	0.63	80.09
	JanusPro-7B	**0.99**	0.89	0.59	0.90	0.79	0.66	0.80	84.19
	UniWorld-V1	0.98	0.93	0.81	0.89	0.74	0.71	0.84	81.38
	OmniGen2	**0.99**	**0.96**	0.74	**0.98**	0.71	0.75	0.86	83.57
	BAGEL	**0.99**	0.94	0.81	0.88	0.64	0.63	0.82	-
	Qwen-Image	**0.99**	0.92	**0.89**	0.88	0.76	**0.77**	0.87	**88.32**
	Qwen-Image-RL	1.00	0.95	0.93	0.92	0.87	0.83	0.91	-
	Ming-Flash-Omni	**0.99**	0.94	0.80	0.91	**0.95**	**0.77**	**0.90**	83.08

Table 3 Performance of Ming-Flash-Omni on **Image-to-Image Editing Benchmarks** compared to leading models. All metrics are evaluated by GPT-4.1. "*Edit.*" denotes models specifically trained for image editing, while "*Unified.*" denotes models capable of image understanding, generation and editing.

Type	Model	GEdit-Bench-EN (Full set)↑			GEdit-Bench-CN (Full set)↑		
		G_SC	G_PQ	G_O	G_SC	G_PQ	G_O
Edit.	Instruct-Pix2Pix	3.58	5.49	3.68	-	-	-
	AnyEdit	3.18	5.82	3.21	-	-	-
	MagicBrush	4.68	5.66	4.52	-	-	-
	Step1X-Edit	7.09	6.76	6.70	7.20	6.87	6.86
	Qwen-Image-Edit	8.00	7.86	7.56	7.82	7.79	7.52
Unified.	UniWorld-v1	4.93	**7.43**	4.85	-	-	-
	OmniGen	5.96	5.89	5.06	-	-	-
	OmniGen2	7.16	6.77	6.41	-	-	-
	BAGEL	**7.36**	6.83	6.52	**7.34**	6.85	6.50
	Ming-Flash-Omni	7.32	7.27	**6.67**	7.26	**7.20**	**6.51**

Table 4 Performance of Ming-Flash-Omni on **Image-to-Mask Segmentation Benchmarks** compared to leading models. Model types are denoted as: *Vision.* for vision-only models, *SAM.* for models equipped with an additional SAM-like segmentation head, and *Unified.* for unified MLLMs capable of both understanding and generation. Results with "*" are obtained by evaluating on 500 images sampled from each dataset via the official API.

Type	Model	RefCOCO (val)↑	RefCOCO+ (val)↑	RefCOCOg (val)↑
Vision.	VLT	67.5	56.3	55.0
	CRIS	70.5	62.3	59.9
	LAVT	72.7	62.1	61.2
	PolyFormer-B	74.8	67.6	67.8
SAM.	LISA-7B	74.1	62.4	66.4
	PixelLM-7B	73.0	66.3	69.3
	OMG-LLAVA	75.6	65.6	70.7
Unified.	Nano-banana*	15.7	13.9	14.9
	Qwen-Image-Edit*	30.3	28.8	34.0
	Ming-Flash-Omni	72.4	62.8	64.3

Table 5 Performance of Ming-Flash-Omni on **PUBLIC Text-to-Speech Benchmarks** compared to leading MLLMs.

Type	Benchmark (Seed-TTS-Eval)	Ming-Flash Omni	Ming-Lite Omni	Qwen3 Omni	Seed TTS	F5 TTS	CosyVoice2	Qwen2.5 Omni
Chinese	Zh-wer ↓	**0.99**	1.69	1.07	1.11	1.56	1.45	1.70
	Zh-sim ↑	0.74	0.68	-	**0.80**	0.74	0.75	0.75
English	En-wer ↓	1.59	4.31	**1.39**	2.24	1.83	2.57	2.72
	En-sim ↑	0.68	0.51	-	**0.76**	0.65	0.65	0.63

Table 6 Performance of Ming-Flash-Omni on Context ASR benchmarks.

Model	**Performance**			
	Speech-English	**Dialogue-English**	**Speech-Mandarin**	**Dialogue-Mandarin**
	WER | NE-WER | NE-FNR	**WER | NE-WER | NE-FNR**	**WER | NE-WER | NE-FNR**	**WER | NE-WER | NE-FNR**
Qwen2-Audio	11.49 | 27.27 | 35.08	13.99 | 33.02 | 32.92	9.92 | 24.10 | 30.02	7.00 | 22.76 | 26.17
Baichuan-Audio	7.52 | 5.87 | 4.55	5.66 | 10.01 | 3.64	2.16 | 6.65 | 2.35	2.96 | 11.48 | 3.94
Kimi-Audio	2.90 | 6.68 | 8.01	4.67 | 13.50 | 11.31	1.95 | 11.13 | 15.28	2.90 | 15.91 | 16.68
Baichuan-Omni-1.5	8.16 | 7.69 | 6.53	9.91 | 14.40 | 5.54	2.98 | 8.39 | 4.71	5.00 | 16.83 | 7.84
Qwen2.5-Omni-3B	3.99 | 7.80 | 9.69	4.83 | 14.36 | 12.85	2.13 | 10.55 | 14.11	3.12 | 15.07 | 15.17
Qwen2.5-Omni-7B	3.96 | 7.38 | 8.72	5.32 | 11.83 | 9.24	1.84 | 9.80 | 12.19	2.40 | 14.06 | 13.17
Ming-Flash-Omni	**2.85** | **2.63** | **2.36**	**3.76** | **7.06** | **1.90**	**1.31** | **5.62** | **1.59**	**1.78** | **8.47** | **0.85**

Table 7 Performance of Ming-Flash-Omni on **PUBLIC and IN-HOUSE Audio Understanding Benchmarks**.

Type	Benchmark	Ming-Flash Omni	Qwen3 Omni	Qwen2 Audio	Kimi Audio
PUBLIC Chinese Benchmarks	Aishell1 ↓	1.22	1.04	1.53	**0.60**
	Aishell2-test-android ↓	**2.51**	2.64	2.92	2.64
	Aishell2-test-ios ↓	**2.46**	2.55	2.92	2.56
	Cv15-zh ↓	5.84	**4.31**	6.90	7.21
	Fleurs-zh ↓	2.80	**2.20**	7.50	2.69
	Wenetspeech-testmeeting ↓	**5.58**	5.89	7.16	6.28
	Wenetspeech-testnet ↓	5.05	**4.69**	8.42	5.37
	SpeechIO ↓	2.51	**2.18**	3.01	2.23
	Average (Chinese) ↓	3.50	**3.19**	5.05	3.70
PUBLIC English Benchmarks	Llibrispeech-test-clean ↓	1.24	**1.22**	1.60	1.28
	Librispeech-test-other ↓	**2.29**	2.48	3.60	2.42
	Multilingual-librispeech ↓	4.04	**3.67**	5.40	5.88
	Cv15-en ↓	6.63	**6.05**	8.60	10.31
	Fleurs-en ↓	3.07	**2.72**	6.90	4.44
	Voxpopuli-v1.0-en ↓	**5.99**	6.02	6.84	7.97
	Average (English) ↓	3.88	**3.69**	5.49	5.38
IN-HOUSE Dialect Benchmarks	Hunan ↓	**6.98**	20.82	25.88	31.93
	Minnan ↓	**13.03**	24.60	123.78	80.28
	Guangyue ↓	**4.07**	27.43	7.59	41.49
	Chuanyu ↓	**3.75**	5.54	7.77	6.69
	Shanghai ↓	**9.92**	30.96	31.73	60.64
	Anhui ↓	**5.38**	5.70	5.72	8.61
	Dongbei ↓	4.38	**3.92**	4.87	4.40
	Henan ↓	**7.93**	8.94	12.31	14.40
	Hubei ↓	**7.90**	14.55	16.39	20.07
	Jiangsu ↓	**11.20**	13.19	12.80	17.25
	Kejiahua ↓	**15.43**	27.88	22.33	29.78
	Shaanxi ↓	6.03	**5.31**	6.32	6.09
	Shandong ↓	**12.63**	17.04	15.00	18.66
	Tianjin ↓	**14.61**	19.70	21.78	34.57
	Yunnan ↓	**15.91**	19.78	21.57	32.79
IN-HOUSE Domain Benchmarks	Noisy ↓	10.69	**9.25**	12.46	24.40
	Chat ↓	3.17	**2.09**	4.29	2.96
	Government ↓	1.89	**1.55**	2.70	2.03
	Health ↓	3.11	**2.22**	4.18	2.38
	Knowledge ↓	3.08	11.52	3.33	**1.98**
	Local-live ↓	1.74	**1.37**	2.34	2.05
	Average (All IN-HOUSE) ↓	**7.75**	13.02	17.39	21.12

Table 8 Performance of Ming-Flash-Omni on **PUBLIC Audio Question-Answering Benchmarks**.

Models	Mean	Open-ended QA		Knowledge	Multi-Choice QA		Instruction	Safety
		AlpacaEval	CommonEval	SD-QA	MMSU	OpenBookQA	IFEval	AdvBench
Step-Audio-chat	57.10	79.80	59.80	46.84	31.87	29.19	65.77	86.73
Qwen2-Audio-chat	54.70	73.80	68.00	35.35	35.43	49.01	22.57	98.85
Baichuan-Audio	62.50	80.00	67.80	49.64	48.80	63.30	41.32	86.73
GLM-4-Voice	57.20	81.20	69.60	43.31	40.11	52.97	24.91	88.08
Kimi-Audio	76.90	89.20	79.40	63.12	62.17	83.52	61.10	**100.00**
Megrez-3B-Omni	46.20	70.00	59.00	25.95	27.03	28.35	25.71	87.69
DiVA	55.70	73.40	70.80	57.05	25.76	25.49	39.15	98.27
Qwen2.5-Omni	74.10	89.80	78.60	55.71	61.32	81.10	52.87	99.42
Qwen3-Omni-Flash-Instruct	**85.40**	**95.40**	91.00	**76.80**	68.40	**91.40**	**75.20**	99.40
Baichuan-Omni-1.5	71.10	90.00	81.00	43.40	57.25	74.51	54.54	97.31
MiniCPM-o	71.70	88.40	83.00	50.72	54.78	78.02	49.25	97.69
Ming-Flash-Omni	83.50	94.80	**92.60**	71.72	**68.70**	84.84	72.49	99.62

Table 9 Performance of Ming-Flash-Omni on **StreamingMultiturnBench** compared to leading omni-MLLMs.

Model	Accuracy	Completeness	Relevance	Conciseness	Naturalness	Proactivity	Average
Qwen2.5-Omni	54.03	53.68	78.00	77.38	98.28	**46.25**	67.93
Ming-Lite-Omni	44.63	40.58	69.28	93.88	95.65	28.78	62.13
Ming-Flash-Omni	**57.03**	**57.10**	**80.40**	**94.13**	**99.40**	41.38	**71.57**

Figure 3 Visualization results of Ming-Flash-Omni on understanding tasks, including world knowledge, multi-image understanding, mathematical reasoning, OCR, contextual ASR, and dialect-aware ASR.

Figure 4 Visualization results of Ming-Flash-Omni on Text/Image → Image tasks, including image generation task, image editing task, and image segmentation task.

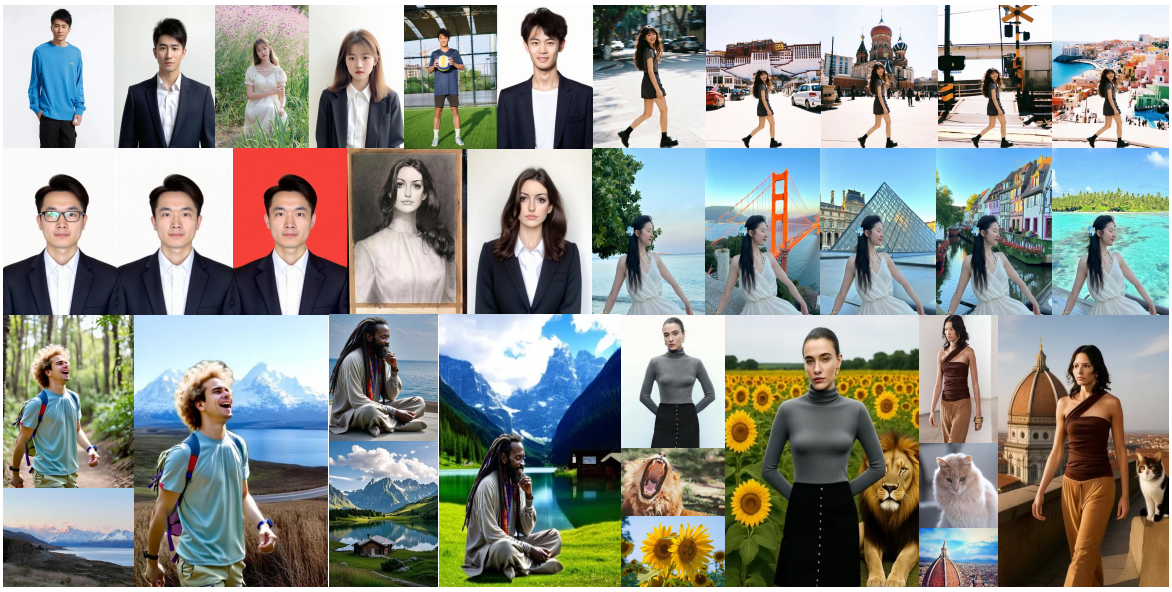

Figure 5 Visualization results of Ming-Flash-Omni on Image → Image tasks, including ID photo generation, ID photo editing, background replacement, and multi-image editing.

Image → Image (Editing), Image → Image (Segmentation), Audio → Text (Understanding), Text → Audio (Generation), and Video → Text (Understanding).

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:

Multi-Dialect and Multi-Domain Audio Understanding Benchmark. To extend audio understanding benchmarks into multi-dialect and multi-domain settings, we constructed two specialized datasets. The multi-dialect dataset was created from 15 regions, while the multi-domain one was curated from six domains. All samples were manually verified for quality by trained annotators. The final datasets comprise 51,986 multi-dialect samples and 10,397 multi-domain samples, with the latter distributed across: Noisy (8,145), Chat (443), Government (462), Health (450), Knowledge (421), and Local Services (476).

Video Streaming Multi-turn Benchmark. The evaluation of video streaming multi-turn dialogue capabilities requires quantifying not only the model's understand capability but also assessing its interactive experience, including proactivity and naturalness. Previous streaming dialogue datasets, such as StreamBench (Lin et al., 2024) and OvO-Bench (Niu et al., 2025), have primarily focused on the understanding aspect while lacking a thorough evaluation of the interactive experience. To address this gap, we introduce StreamingMultiturnBench. To construct StreamingMultiturnBench, we manually selected 380 videos, carefully ensuring coverage of multiple key domains including life recording, education, TV shows, video games, and documentaries. Then we use SOTA closed-source modelfor machine annotation. Subsequently, a team of 10 human annotators revise and double-check the dialogue content to ensure it aligns with human conversational preferences. This process yielded 2,200 video question-answer pairs. During evaluation, we use advanced closed-source model, *e.g.* GPT-4o (OpenAI, 2025), to compare the model's output against the human-annotated answers, scoring it on a scale of 1 to 5 across the six dimensions: accuracy, completeness, relevance,

naturalness, conciseness, and proactivity. The final score is the average for each dimension. To align our metrics with other video benchmarks, we linearly scale the results to a 100-point scale. We commit to open-sourcing and publicly maintaining this benchmark to ensure reproducibility.

4.3 Quantitative Results

We conduct comprehensive evaluations of Ming-Flash-Omni against state-of-the-art MLLMs on over 50 different multimodal benchmarks, as illustrated in Table 1∼9. Extensive experiments demonstrate that Ming-Flash-Omni achieves comparable performance with leading MLLMs.

Vision → Text (Understanding) As shown in Table 1, Ming-Flash-Omni demonstrates strong and competitive performance across a wide range of vision–language benchmarks. Specifically, on general-purpose understanding task, it achieves performance on par with leading omni-models (most notably scoring 81.9% on MathVista), though it still exhibits a slight gap compared to state-of-the-art vision–language models. Similarly, on OCR-centric benchmarks, Ming-Flash-Omni attains state-of-the-art results among omni-modal models, yet remains marginally behind the best proprietary counterparts. In multi-image understanding, Ming-Flash-Omni outperforms the leading open-source vision–language model Qwen3-VL-30B-A3B on MMT-Bench (68.0) and LLaVA-Interleave Bench (63.3), while showing a minor performance gap on MuirBench, suggesting opportunities for further improvement. In video understanding, Ming-Flash-Omni achieves state-of-the-art performance on MVBench with a score of 74.6, demonstrating robust general video reasoning capabilities. It further exhibits specialized proficiency in processing linguistic content within videos, attaining a leading score of 73.0 on VideoMME with subtitles. Although its performance on long-form video understanding is slightly lower, its consistently strong results across most metrics underscore its advanced video comprehension capabilities.

Text → Image (Generation). As shown in Table 2, our experimental results demonstrate that the generation quality of Ming-Flash-Omni is on par with state-of-the-art diffusion models. Notably, on the Geneval benchmark, our model surpasses all non-Reinforcement Learning methods, demonstrating exceptional controllability. This advantage is particularly pronounced in the "Position" and "Color." sub-categories. On the DPG-Bench benchmark, Ming-Flash-Omni achieves an overall score of 83.08, a performance level comparable to pure image generation models like SD3-Medium (84.08) and leading unified models like OmniGen2 (83.57).

Image → Image (Editing). As shown in Table 3, Ming-Flash-Omni demonstrates impressive image editing performance, surpassing all other unified models. Specifically, Ming-Flash-Omni supports editing instructions in Chinese, achieving performance comparable to that with English instructions. Compared to Qwen-Image-Edit which utilizes a 20B DiT head, Ming-Flash-Omni achieves comparable semantic consistency and perceptual quality with a much more efficient 2B DiT head—only one-tenth the parameters. This efficiency also translates to remarkable inference speeds, typically between 1-2 seconds per generation.

Image → Image (Segmentation). As shown in Table 4, Ming-Flash-Omni is capable of performing segmentation tasks, achieving performance comparable to that of specialized models designed explicitly for this purpose. Compared to other unified MLLMs, Ming-Flash-Omni demonstrates a significant advantage in segmentation. For instance, Qwen-Image-Edit often struggles to accurately localize the target object, while Nano-banana frequently misinterprets user intent during inference. In contrast, Ming-Flash-Omni exhibits superior robustness and a more accurate understanding of spatial and semantic instructions.

Audio → Text (Understanding). Our model sets a new state-of-the-art (SOTA) on all 12 subtasks of the ContextASR-Bench (Table 6), underscoring its superior ability to leverage context—a vital skill for real-world applications like multi-turn dialogue and hotword enhancement. It also exhibits highly competitive performance across various ASR benchmarks, with notable strengths in dialect recognition (Table 7). In audio question answering, Ming-Flash-Omni surpasses all open-source audio-centric and other Omni models, with the exception of Qwen3-Omni-Flash-Instruct(Xu et al., 2025b) (Table 8). Taken together, these findings demonstrate the robust and versatile audio understanding capabilities of Ming-Flash-Omni.

Text → Audio (Generation). As shown in Table 5, leveraging advancements in speech representation and model architecture, Ming-Flash-Omni achieves SOTA performance among open-source models on the test-zh subset of the SEED-TTS-Eval benchmark(Anastassiou et al., 2024b). Furthermore, its WER on the test-en subset is surpassed only by that of Qwen3-Omni.

Video + Audio → Text (Video Streaming Conversation). As shown in Table 9, benefiting from the introduction of high-quality and diverse streaming video multiturn data, Ming-Flash-Omni has achieved significant improvements in all dimensions compared to Ming-Lite-Omni. Ming-Flash-Omni also outperforms Qwen2.5-Omni (Xu et al., 2025a) in the dimensions of accuracy, completeness, relevance, and conciseness, providing better experience in streaming video conversation scenarios.

4.4 Visualization Results

In this section, we present several visualization examples to better illustrate the capabilities of Ming-Flash-Omni. First, as shown in the figure, Ming-Flash-Omni demonstrates strong visual understanding across multiple dimensions: it leverages rich world knowledge to accurately infer geographic locations from visual cues; excels at multi-image understanding and generates creative, coherent text grounded in multiple images; solves complex mathematical problems through clear, step-by-step reasoning; and exhibits robust document understanding by accurately parsing intricate formulas and answering questions about sophisticated charts and diagrams within images. Turning to speech recognition, Ming-Flash-Omni achieves strong performance on contextual ASR tasks. By leveraging contextual information, it effectively resolves many challenging cases where conventional ASR systems tend to fail—such as ambiguous homophones, domain-specific terminology, or noisy conversational speech. Moreover, this version also supports multiple Chinese dialects, significantly broadening its applicability in real-world multilingual and regional speech scenarios. Lastly, we visualize the capabilities of *Text/Image →Image* generation tasks in Figure 4 and Figure 5, covering a wide range of applications including image generation, image editing, image segmentation, multi-image editing, ID photo generation, ID photo editing, and background replacement. As can be seen, Ming-Flash-Omni not only supports a broader set of generative capabilities but also achieves higher output quality and greater controllability compared to previous versions.

5 Conclusion

In this paper, we present Ming-Flash-Omni, built upon Ling-Flash-2.0 with 100 billion parameters, where only 6.1B parameters are activated per token. Ming-Flash-Omni demonstrates advanced multimodal perception and generation capabilities with improved computational efficiency while scaling model capacity. It achieves SOTA performance across a broad spectrum of tasks, including multi-image and video processing, image generation, generative segmentation, Contextual Automatic Speech Recognition (ContextASR), and multi-dialect recognition, outperforming omni models

of comparable scale. We believe the open-sourcing of our models and code will facilitate the development of AGI by advancing multimodal intelligence research and enabling broader real-world applications.

6 Contributors

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

Ant Inclusion AI
Bowen Ma
Cheng Zou
Canxiang Yan
Chunxiang Jin
Chunjie Shen
Chenyu Lian
Dandan Zheng
Fudong Wang
Furong Xu
GuangMing Yao
Jun Zhou
Jingdong Chen
Jianing Li
Jianxin Sun
Jiajia Liu
Jian Sha
Jianjiang Zhu
Jianping Jiang
Jun Peng
Kaixiang Ji
Kaimeng Ren
Libin Wang
Lixiang Ru
Longhua Tan
Lu Ma
Lan Wang
Mochen Bai
Ning Gao
Qingpei Guo
Qinglong Zhang
Qiang Xu
Rui Liu
Ruijie Xiong
Ruobing Zheng
Sirui Gao
Tao Zhang
Tianqi Li
Tinghao Liu
Weilong Chai
Xinyu Xiao
Xiaomei Wang
Xiaolong Wang
Xiao Lu
Xiaoyu Li
Xingning Dong
Xuzheng Yu
Yi Yuan
Yuting Gao
Yuting Xiao
Yunxiao Sun
Yipeng Chen
Yifan Mao
Yifei Wu
Yongjie Lyu
Ziping Ma
Zhiqiang Fang
Zhihao Qiu
Ziyuan Huang
Zizheng Yang
Zhengyu He

References

Ling-flash-2.0, 2025. https://huggingface.co/inclusionAI/Ling-flash-2.0. Accessed: 2025-09-17.

Inclusion AI, Biao Gong, Cheng Zou, Chuanyang Zheng, Chunluan Zhou, Canxiang Yan, Chunxiang Jin, Chunjie Shen, Dandan Zheng, Fudong Wang, et al. Ming-omni: A unified multimodal model for perception and generation. *arXiv preprint arXiv:2506.09344*, 2025a.

Inclusion AI, Fudong Wang, Jiajia Liu, Jingdong Chen, Jun Zhou, Kaixiang Ji, Lixiang Ru, Qingpei Guo, Ruobing Zheng, Tianqi Li, et al. M2-reasoning: Empowering mllms with unified general and spatial reasoning. *arXiv preprint arXiv:2507.08306*, 2025b.

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, 2024a.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024b.

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.

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.

Canxiang, Chunxiang Yan, Dawei Jin, Haibing Huang, Han Yu, Hui Peng, Jie Zhan, Jing Gao, Jingdong Peng, Jun Chen, Kaimeng Zhou, Ming Ren, Mingxue Yang, Qiang Yang, Qin Xu, Ruijie Zhao, Shaoxiong Xiong, Xuezhi Lin, Yi Wang, Yifei Yuan, Yongjie Wu, Zhengyu Lyu, He, Zhihao , Zhiqiang Qiu, Ziyuan Fang, and Huang. Ming-uniaudio: Speech llm for joint understanding, generation and editing with unified representation. *arXiv preprint arXiv:https://arxiv.org/submit/6926109/view*, 2025.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024b.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants, 2024c. 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, 2024d.

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*, 2025a.

Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. *arXiv preprint arXiv:2504.07957*, 2025b.

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.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024b.

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.

Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 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 (CVPR)*, pages 14375–14385, June 2024.

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.

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*, 2024.

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.

Dongya Jia, Zhuo Chen, Jiawei Chen, Chenpeng Du, Jian Wu, Jian Cong, Xiaobin Zhuang, Chumin Li, Zhen Wei, Yuping Wang, et al. Ditar: Diffusion transformer autoregressive modeling for speech generation. *arXiv preprint arXiv:2502.03930*, 2025.

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/.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016.

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.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9404–9413, 2019.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*, 2025.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024a.

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, 2024b.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), December 2024. ISSN 1869-1919. doi: 10.1007/s11432-024-4235-6. http://dx.doi.org/10.1007/s11432-024-4235-6.

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. In *International Conference on Learning Representations (ICLR)*, 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.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18902–18913, 2025.

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

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.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*, 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.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

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.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms, 2024a. https://arxiv.org/abs/2406.16860.

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*, 2024b.

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.

Fei Wang, Xingyu Fu, James Y.Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Yan Lorena, Jacky Wwenjie Mo, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024a.

He Wang, Linhan Ma, Dake Guo, Xiong Wang, Lei Xie, Jin Xu, and Junyang Lin. Contextasr-bench: A massive contextual speech recognition benchmark, 2025. https://arxiv.org/abs/2507.05727.

Weiyun Wang, Yiming Ren, Haowen Luo, Li Tiantong, Yan Chenxiang, Chen Zhe, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, et al. The all-seeing project v2: Towards general relation comprehension of the open world. *arXiv preprint arXiv:2402.19474*, 2024b.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, et al. Videorope: What makes for good video rotary position embedding? *arXiv preprint arXiv:2502.05173*, 2025.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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, 2024.

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*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report, 2025b. https://arxiv.org/abs/2509.17765.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025c.

Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, et al. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024.

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, 2022.

Xinghua Zhang, Haiyang Yu, Cheng Fu, Fei Huang, and Yongbin Li. Iopo: Empowering llms with complex instruction following via input-output preference optimization. *arXiv preprint arXiv:2411.06208*, 2024.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Appendix

A Public Benchmarks

Image → Text (Understanding). Our evaluation of the image-to-text understanding capabilities primarily encompasses the following six tasks: 1) general image understanding capabilities evaluated on MMStar (Chen et al., 2024b), AI2D (Kembhavi et al., 2016), HallusionBench (Guan et al., 2024), CV-Bench (Tong et al., 2024a), MathVista (Lu et al., 2024), and CRPE (Wang et al., 2024b). 2) OCR capabilities evaluated on ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2021), OCRBench (Liu et al., 2024), and TextVQA-VAL (Singh et al., 2019). 3) multi-image capabilities evaluated on MMTBench (Ying et al., 2024), MuirBench (Wang et al., 2024a), and LLaVA-interleave Bench (Li et al., 2024a). 4) Complex instruction capabilities evaluated on MIA-Bench (Qian et al., 2024). 5) video understanding capabilities evaluated on MVBench (Li et al., 2024b), VideoMME (Fu et al., 2024b), and LongVideoBench (Wu et al., 2024).

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., 2024), and FID.

Image → Image (Editing). Our evaluation of image-to-image editing capabilities is conducted on the GEdit-Bench benchmark (Liu et al., 2025).

Image → Image (Segmentation). We evaluate the segmentation capability of our MLLM on the standard referring expression segmentation (RES) benchmarks RefCOCO/+ (Kazemzadeh et al., 2014) and RefCOCOg (Mao et al., 2016).

Audio → Text (Understanding). Our evaluation of the audio-to-text understanding capabilities mainly includes the following three tasks: 1) Fundamental audio understanding capabilities evaluated on a broad range of public benchmarks, including public Chinese benchmarks like Aishell1 (Bu et al., 2017) and Wenetspeech (Zhang et al., 2022), 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., 2024c) for open-ended QA tasks, and SD-QA for knowledge-based QA tasks. Finally 3) evaluates the model's ability to utilize context on ContextASR-Bench(Wang et al., 2025).

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., 2024a).

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

B The prompt used in data generation

This section presents the prompt used for data generation in Sec. 3.

Prompt for portrait preservation data

```
Based on the person's age and gender in the image, generate a detailed
description of a realistic life scene photo.
Requirements:
1. Must include: clothing, location, action, style.
2. Strictly within 50 words.
3. Change the scene (real scenes: indoor, outdoor, residential areas, homes,
companies, kitchens, parks, etc.) and character costumes to ensure that even
similar images have different descriptions.
4. Clothing must match age and gender, vary descriptions.
5. Scenario details must include at least 3 realistic elements, be vivid and
lifelike.
Output only the description in English, no additional text.
Example: A young woman with medium-length dark hair tied back in a neat
ponytail stands outdoors near a park bench, dressed in a crisp white button-
up shirt under a fitted black blazer. She wears a subtle headband, her brow
knitted as she runs her fingers through her hair, her lips slightly
downturned and eyes glistening with unshed tears. The scene captures her from
 a distant angle, with trees and a blurred pathway in the soft-focus
background, suggesting contemplation or distress amidst nature.
```

Prompt for text generation data

```
Describe according to the following requirements:
1. Randomly select a theme from {theme}, and generate text content in Chinese
, English, and numbers, with 3-5 characters.
2. Use your imagination; the theme is not fixed. Keep the description under
100 characters.
3. Generate a suitable image description based on the text content.
4. Refer to the {text style} and choose appropriate font, color, and layout.

Output format:
Description, Text: "content", Font style, Font color, and Image layout

Examples:
In the image, there is a gym scene, Text: "Power 3.0", font style is broken
art font with cracks and shattering effects, font color is metallic gray with
 dark red gradient, text layout is centered-right, background includes
dumbbells, treadmills, and reflective glass walls.
```

Prompt for video multi-turn conversation data

```
# Role: Expert Video Analyst
You are an expert video analyst. Your task is to analyze the provided video
 and output a structured JSON object containing your analysis. You must
adhere strictly to the format and rules described below.

# Instructions:
Analyze the video and generate a single JSON object with the following keys
. Your response should ONLY be the JSON object, enclosed in a single
markdown code block (```json ... ```). Do not include any other text,
explanations, or introductory phrases.

# JSON Fields Definition:
1. `"category"`: (String) Select ONLY ONE category from the following list
 that best describes the main subject of the video.
If the video category is not among the listed below, output "Others".
* **List of 34 Categories**: [
     "Others", "LifeRecord-TravelLog", "LifeRecord-DailyLife", "LifeRecord-
     HouseTour", "LifeRecord-Reaction", "LifeRecord-AnimalPet", "LifeRecord
     -Cooking", "LifeRecord-Fashion", "LifeRecord-Workout", "Education-
     Lecture", "Education-Finance", "Education-Multilingual", "Education-
     Handifraft", "Education-Science", "Education-Art", "Education-
     OnlineTutorial", "TVShow-TVSeries", "TVShow-News", "TVShow-TalkShow",
     "TVShow-Celebration", "TVShow-CommentaryProgram", "Competition-
     Football", "Competition-Athletics", "Competition-Basketball", "
     Competition-Snooker", "Competition-Boxing", "Competition-Car", "
     VideoGames-Sandbox", "VideoGames-OpenWorld", "Documentary-Nature", "
     Documentary-Science", "Documentary-Culture", "Documentary-Kids", "
     Movie-Comedy", "Movie-Adventure"]

2. `"caption"`: (String) A concise but descriptive summary of the video's
content in English, not exceeding 30 words. Describe what is happening, who
 is involved, and the main subject.

3. `"content_score"`: (Integer) An integer from 1 to 10. This score
evaluates the video's potential as a prompt for a synthetic AI-user
conversation.
* **Score 1-3 (Low)**: The video content is sparse, repetitive, or features
 a single person's monologue with little interaction or environmental
detail. It's difficult to ask questions or start a conversation based on it
.
* **Score 4-7 (Medium)**: The video has some interesting elements but may
lack depth or variety. It can support a few questions but may not lead to
an extended, rich conversation.
* **Score 8-10 (High)**: The video is rich in content, detail, and action.
It shows a process, an interaction, or a complex scene that naturally
invites questions and allows for a deep, extended conversation (e.g., a
detailed cooking tutorial, a complex assembly process, a travel vlog with
multiple activities).

4. `"fov"`: (Integer) Field of View.
* **1**: The video is shot from a first-person perspective (FPV), where the
 camera acts as the viewer's eyes.
* **0**: The video is shot from a third-person or static perspective.
```

````
5. `"task_complexity"`: (Integer) An integer from 1 to 10, representing
the complexity of the primary task shown in the video. If no specific task
is shown, rate the complexity of the main activity.
* **Score 1-3 (Low)**: Simple, everyday actions that require minimal skill
(e.g., opening a bottle, pouring water, petting a cat).
* **Score 4-7 (Medium)**: Tasks that require some skill, knowledge, or
multiple steps (e.g., cooking a simple dish, assembling IKEA furniture,
basic makeup application).
* **Score 8-10 (High)**: Highly complex, specialized, or professional tasks
 that require significant expertise, precision, or effort (e.g., performing
 surgery, building a car engine, professional programming, playing a
complex musical piece).

# Required Output Format:
    ```json
    {
      "category": "...",
      "caption": "...",
      "content_score": ...,
      "fov": ...,
      "task_complexity": ...
    }
```
````