MGM-Omni: Scaling Omni LLMs to Personalized Long-Horizon Speech

Chengyao Wang[1*], Zhisheng Zhong[1*], Bohao Peng[1*], Senqiao Yang[1], Yuqi Liu[1], Haokun Gui[2], Bin Xia[1], Jingyao Li[1], Bei Yu[1], and Jiaya Jia[2,3]

[1] The Chinese University of Hong Kong
[2] Hong Kong University of Science and Technology
[3] SmartMore

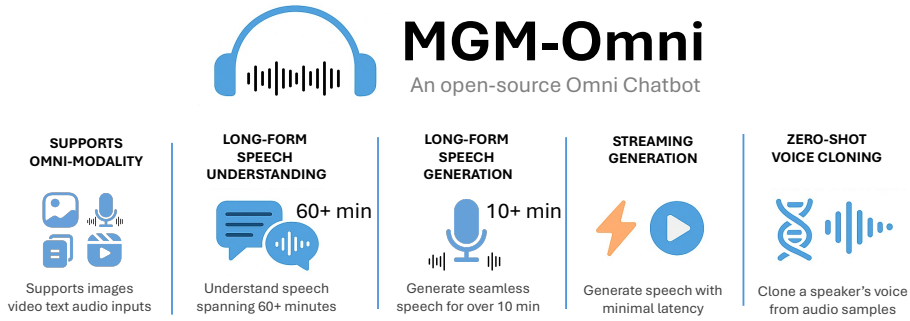

Fig. 1: MGM-Omni is an advanced Omni LLM for omnimodal understanding, long-form understanding, long-form speech generation and zero-shot voice clone. It can comprehend audio inputs exceeding 60 minutes and produce consistent, high-quality speech outputs longer than 10 minutes.

Abstract. We present MGM-Omni, an Omni MLLM for omni-modal understanding and expressive, long-horizon speech generation. MGM-Omni adopts a "brain–mouth" design with a dual-track, token-based architecture that cleanly decouples multimodal reasoning from real-time speech generation. This design enables efficient cross-modal interaction and low-latency, streaming speech generation. For multimodal understanding, a unified training strategy coupled with a dual audio encoder design enables long-form omni modal perception across diverse acoustic conditions. For speech generation, a chunk-based parallel decoding scheme narrows the text–speech token-rate gap, accelerating inference and supporting streaming zero-shot voice cloning with stable timbre over extended durations. Compared to concurrent work, MGM-Omni achieves these capabilities with markedly data-efficient training. Extensive experiments demonstrate that MGM-Omni outperforms existing open source models in preserving timbre identity across extended sequences, producing natural and context-aware speech, and achieving superior long-form audio and omnimodal understanding. MGM-Omni establishes an efficient, end-to-end paradigm for omnimodal understanding and controllable, personalized long-horizon speech generation. Code and models are available at `https://github.com/dvlab-research/MGM-Omni`.

* Equal contribution.

1 Introduction

The evolution of large language models (LLMs) from purely text-based systems [38, 52] to multimodal frameworks has marked a significant paradigm shift in artificial intelligence. Vision language models (VLMs) such as GPT [39, 40], Gemini [11, 50] and LLaVA [26, 32] have demonstrated remarkable capabilities in understanding and processing visual information, effectively bridging the gap between vision and language. Audio serves as a bridge between humans and AI. However, integration of audio, particularly understanding and generating long-form and expressive audio, remains a significant challenge in multimodal systems. Most existing approaches can only understand and synthesize short audio segments lasting a few minutes, which severely limits the practical utility of fully multimodal models. Robust long-audio understanding can dramatically accelerate information access, while the ability to generate long-form audio in arbitrary timbres would enable more flexible and efficient information expression.

Model	VU	AU	LAU	SG	LSG	VC
CosyVoice2 [14]				✓		✓
Higgs-Audio-v2 [8]				✓	✓	✓
Qwen2.5-VL [6]	✓					
Qwen2.5-Omni [58]	✓	✓		✓		
Lyra [64]	✓	✓	✓	✓		
MGM-Omni	✓	✓	✓	✓	✓	✓

Table 1: Model Comparison. VU, AU, LAU, SG, LSG, and VC denote visual understanding, audio understanding, long audio understanding, speech generation, long speech generation, and zero-shot voice cloning.

The integration of audio in multimodal systems is hindered by the disparity between audio and text modalities. Audio token sequences are significantly longer and operate at finer temporal resolution than their corresponding text token sequences [46, 53]. This disparity creates three challenges. First, existing systems lack robust long-form audio understanding, struggling to maintain coherence and semantic accuracy across extended inputs. Second, in generation, a one-to-many alignment problem complicates mapping semantic words or units to long acoustic sequences, leading to misaligned prosody and unnatural pacing in long-form speech. Third, autoregressive generation is prone to error accumulation, where minor inaccuracies cascade, degrading timbre consistency and audio quality. Despite recent progress [8, 20, 51, 58], these systems do not address the intertwined issues of long-form understanding, alignment, and generation.

To address these limitations, we introduce MGM-Omni, an Omni LLM that unifies vision, language, and audio in an any-to-any framework for low-latency multimodal understanding and generation. MGM-Omni adopts a dual-track architecture, separating multimodal reasoning (MLLM, the brain) from speech synthesis (SpeechLM, the mouth), enabling efficient cross-modal processing and real-time audio generation. For audio understanding, we employ a dual-encoder design that fuses acoustic and semantic features, enabling unified inference across short and long audio. For speech generation, we introduce Chunk-Based Parallel Decoding, which mitigates the text–speech token-rate gap by segmenting text and predicting multiple speech tokens in parallel, improving alignment, reducing

long-sequence error accumulation, and boosting speed by up to 3×. Trained on approximately 400k hours of audio, MGM-Omni supports zero-shot voice cloning from arbitrary reference voices. We also propose Long-TTS-Eval to systematically assess long-form and complex speech generation. Overall, MGM-Omni enables expressive, personalized long-horizon speech with consistent timbre and robust text–speech alignment. Our main contributions are threefold:

- We propose MGM-Omni, an Omni LLM featuring a novel dual-track design that unifies omni-modal understanding and expressive speech generation, moving beyond cascaded systems.
- We introduce a Chunk-Based Parallel Decoding mechanism that mitigates the token-rate mismatch between text and speech, enabling efficient, high-fidelity, and context-aware long-form audio synthesis with customized voice.
- Through extensive experiments, we demonstrate that MGM-Omni achieves leading performance in long audio understanding, zero-shot voice cloning and natural, context-aware long-form speech generation.

2 Related Work

Multi-modal Large Language Models. The advent of large language models (LLMs) [38, 50, 52] has revolutionized natural language processing, paving the way for multimodal extensions that integrate diverse data modalities such as text, image, video and audio [6, 27, 28, 32, 34, 58]. Early multimodal models centered on vision–language alignment via contrastive learning. CLIP [44] demonstrated the efficacy of zero-shot image classification through joint embedding spaces. Building on this foundation, vision language models (VLMs) like Flamingo [2], LLaVA [32] and MiniGPT-4 [66] adapted frozen visual encoders (e.g., CLIP-ViT) to instruction-tuned LLMs to enable general-purpose multimodal understanding. Subsequent works such as Mini-Gemini [28], the LLaVA series [26,31], and the Qwen-VL series [5,6,54] further advance VLMs with high-resolution image comprehension, video understanding, visual grounding and visual reasoning. Despite this progress, most MLLMs remain vision-centric, with limited support for audio modalities. Recent efforts start to incorporate the audio modality into LLMs [20, 24, 56] and MLLMs [1, 35, 57, 58, 64], enhancing the interactivity between humans and AI. However, these approaches still struggle to understand and generate long-form audio, and cannot control the timbre of the generated speech. MGM-Omni addresses these limitations with a dual-track, token-based architecture that natively fuses language and audio, enabling omni-modal understanding and expressive, controllable long-form speech generation.

Speech Generation. In recent years, driven by the emergence of large language models (LLMs) and large-scale speech-text pre-training, zero-shot text-to-speech generation (TTS) has advanced markedly [3, 8, 14, 21, 65]. CosyVoice2 [14] builds a TTS system with chunk-aware flow matching and LLMs, enabling streaming multilingual speech synthesis with zero-shot voice cloning. Qwen2.5-Omni [58] incorporates this design with a thinker–talker pipeline for end-to-end perception

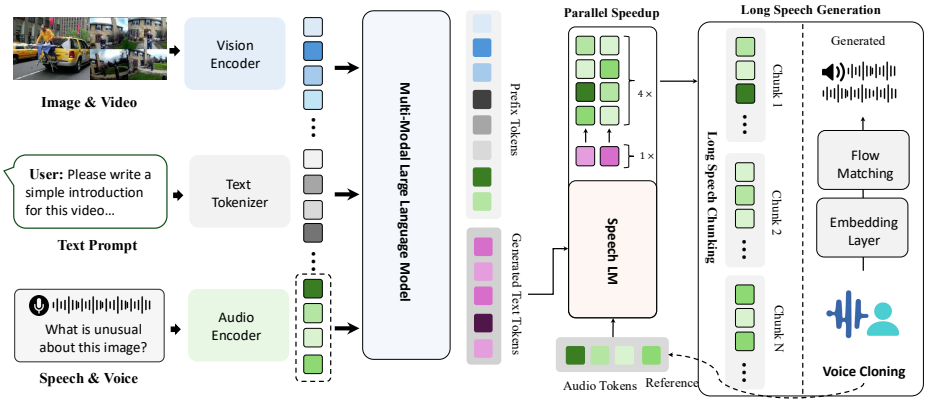

Fig. 2: The overview of MGM-Omni. MGM-Omni decouples omni-modal understanding and speech generation into an MLLM and a SpeechLM. The MLLM processes text, images, video, and audio to produce text, while the SpeechLM generates speech from the MLLM's output in real time.

and generation across text, images, audio, and video. However, these systems still struggle with long-form speech generation. More recent efforts such as MOSS-TTSD [51], Higgs-Audio-v2 [8] and VibeVoice [43], support expressive bilingual dialogue generation with personalized voice, yet challenges remain in maintaining timbre consistency over long sequences, ensuring real-time cross-modal fidelity, and achieving low latency. MGM-Omni addresses this issue via a chunk-based parallel decoding approach, enabling expressive long-form speech generation with consistent timbre and low latency.

3 MGM-Omni

MGM-Omni is capable of processing text, images, video and speech, and generate both textual and spoken outputs. To support high-quality, long-form speech synthesis without compromising the efficiency and effectiveness of omnimodal understanding, MGM-Omni decouples multimodal understanding and speech generation into two components: MLLM, serving as the "brain" for multimodal understanding and text generation, and SpeechLM, serving as the "mouth" for real-time speech generation. For input in different modalities, we employ modality-specific encoders to extract features, which are subsequently passed to the MLLM. The MLLM generates text tokens and passes them to SpeechLM, which produces speech tokens in real-time via a Chunk-Based Parallel Decoding strategy. These speech tokens are further converted into Mel-spectrograms through a flow-matching model [29], and the final audio is synthesized using a vocoder. It's worth noting that MLLM and SpeechLM are trained separately, each with two stages. The overall framework is illustrated in Figure 2.

3.1 Omni Understanding

MGM-Omni is built upon Qwen2.5-VL [6], a state-of-the-art open-source Vision-Language Model (VLM) that supports image and video understanding with a native-resolution ViT [12]. Based on Qwen2.5-VL, MGM-Omni attempts to extend towards Omni-Understanding, especially by incorporating audio understanding capabilities.

Dual Audio Encoder. MGM-Omni adopts a dual audio encoder design to capture both acoustic and semantic audio features. The primary encoder, Qwen2-Audio [10], is an audio encoder continually trained on Whisper-large-v3 [45] for enhanced general sound perception. To strengthen semantic understanding, we incorporate Belle-Whisper-large-v3 [7], another Whisper-based encoder specialized in Chinese and English speech recognition. This dual encoder setup yields two complementary representations: the main audio feature X_{main} and the auxiliary audio feature X_{aux}.

Information Mining. To effectively integrate these complementary features, we design an audio information mining approach inspired by Mini-Gemini [28]. Specifically, X_{main} serves as the query $Q \in \mathbb{R}^{N \times C}$, while X_{aux} provides the key-value pair: $K \in \mathbb{R}^{N \times C}$ and $V \in \mathbb{R}^{N \times C}$, allowing the model to retrieve semantically relevant cues from X_{aux} under the guidance of X_{main}. Formally, information mining can be defined as:

$$T_A = \text{MLP}(Q + \text{Softmax}(\phi(Q) \times \phi(K)^\top) \times \phi(V)), \tag{1}$$

where ϕ denotes a projection layer and MLP represents a multi-layer perceptron. This approach enhances audio representation by making it acoustically and semantically aware, yielding audio tokens T_A for subsequent LLM processing.

Training Strategy. Following Lyra [64], we build a two-stage training pipeline to integrate audio understanding capabilities to MLLM. In the first stage, we conduct audio-to-text pre-training to align the audio encoder to LLM with information mining. In the second stage, we perform unified omni-modal training to equip the model with the capability of omni-modal understanding. The first stage primarily uses audio transcription data, while the second stage comprises audio transcription, audio QA, audio-instruct VQA, and text instruction tuning data. This training strategy enables omni-cognition and robust cross-modal reasoning. Both stages are trained with Lora [19].

Omni Length Understanding. MGM-Omni aims to support both long and short sequence input. However, training with sequences of diverse lengths is inefficient: large batch sizes cause long sequence samples to run out of memory, while small batch sizes waste memory on short sequence samples. To address this issue, we propose a unified training pipeline. First, we group audio of similar lengths into the same batch. Second, we dynamically adjust the batch size, smaller for long-context inputs and larger for short-context inputs. This strategy significantly improves training efficiency, enabling a single model to possess comprehension capabilities across sequences of different lengths.

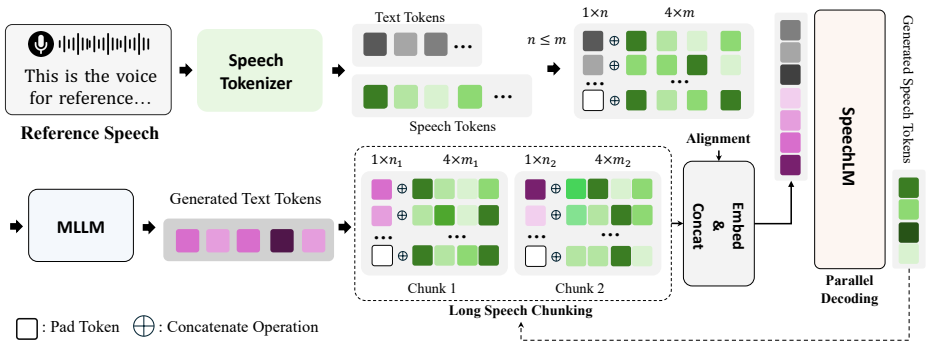

Fig. 3: The overview of SpeechLM in MGM-Omni. Conditioned on MLLM-generated text and the reference audio clip, SpeechLM generate speech with Chunk-based Parallel Decoding.

3.2 Omni Generation

MGM-Omni can generate both long-form text and speech. The textual output is autoregressively produced by the Omni-MLLM. The generated text, together with the personalize reference audio, is subsequently served as the conditioning for SpeechLM to synthesize speech via a Chunk-based Parallel Decoding method. The overall speech generation pipeline is depicted in Figure 3.

Speech Generation. SpeechLM takes text tokens from Omni-MLLM as input and generates speech tokens in an autoregressive manner. It is initialized from the Qwen3 [59] language model, with an additional TTS-Adapter appended to its output. TTS-Adapter consists of six randomly initialized Qwen3 blocks, designed to transform text representations into speech representations. The speech tokens produced by SpeechLM are then converted into Mel-spectrograms through a Flow-Matching model, and finally synthesized into audio via HiFi-GAN [25] vocoder. We used the flow-matching model from CosyVoice2 [14], which supports chunk-aware streaming decoding.

Speech Tokenizer. We employ the CosyVoice2 finite scalar quantization (FSQ) speech tokenizer to obtain discrete speech representations for speech generation. The tokenizer operates at a rate of 25 Hz, meaning that 25 tokens represent one second of audio. In comparison, humans typically express only two or three words per second. This discrepancy highlights that for a given utterance, the number of speech tokens is substantially larger than the number of text tokens. This leads to the following two issues:

- As speech length increases, the gap between text and speech tokens widens, weakening their correlation and degrading long-form generation quality.
- The much higher number of speech tokens compared to text tokens slows inference and harms streaming efficiency.

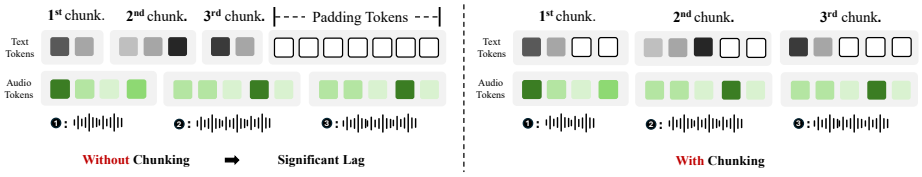

Fig. 4: Speech token decoding. Chunk-based decoding narrows the gap between text and its corresponding speech, making it easier to scale to long-form speech generation.

To address these two challenges, we propose a Chunk-Based Parallel Decoding for efficient long-form speech generation.

Chunk-based Decoding. To improve text–speech alignment in long-form speech generation, we introduce Chunk-based Decoding for speech token generation. As shown in Figure 4, the input text is divided into smaller chunks that are sequentially processed by SpeechLM, with each chunk producing a corresponding speech segment. During decoding, we adopt a token delay strategy: speech token generation within a chunk is initiated only after the first four text tokens, which are replaced by padding tokens in the speech sequence. This design ensures that every speech token is aligned with its corresponding text token while avoiding early mis-synchronization. By reducing the alignment distance between modalities, Chunk-based Decoding enhances cross-modal correspondence and improves the robustness of long-form speech synthesis. In contrast to naive chunking approach, our approach preserves both the previously generated text and speech as context, thereby maintaining global fluency and coherence in the final output. Notably, Chunk-based Decoding is highly compatible with our dual-track "brain–mouth" design, preserving omnimodal understanding and text generation speed while improving speech synthesis quality.

Chunking. SpeechLM autoregressively appends MLLM-generated text tokens to a buffer. When it detects the end of a sentence and the buffer contains more than 50 tokens, it groups the buffered tokens into a chunk. If the buffer exceeds 150 tokens, it is forcibly split into a chunk to prevent overly long synthesized audio. As a result, each chunk typically corresponds to about 20–40 seconds of audio. Sentence boundaries are detected using periods, question marks, exclamation marks, and line breaks.

Parallel Decoding. To improve efficiency, we introduce a parallel decoding strategy for efficient speech token generation. Specifically, we extend the vocabulary so that SpeechLM can decode both modalities in a single step. Let V_{text} denote the text vocabulary, V_{speech} denote the speech tokenizer vocabulary, and k denote the parallel size. The extended vocabulary size is thus defined as:

$$|V| = |V_{\text{text}}| + k|V_{\text{speech}}|. \tag{2}$$

For speech tokenization, the input for each decoding step t consists of one text token x_t and k speech tokens $\{s_t^1, s_t^2, \ldots, s_t^k\}$. We use $f(\cdot)$ to denote the embedding function, and the hidden features h_t^{in} for LLM input can be averaged as:

$$h_t^{\text{in}} = \frac{1}{k+1}\left(f(x_t) + \sum_{i=1}^{k} f(s_t^i)\right). \tag{3}$$

For speech detokenization, we employ a TTS-Adapter to project the LLM output hidden state h_t^{out} into the speech representation space, after which the lm_head predicts the next set of speech tokens:

$$\{\hat{s}_{t+1}^1, \ldots, \hat{s}_{t+1}^k\} = \text{lm\_head}\big(\text{TTS-Adapter}(h_t^{\text{out}})\big). \tag{4}$$

While parallel decoding is commonly used with RVQ speech tokenizers [51,56], it is rarely applied to FSQ speech tokenizers. We found that using parallel decoding with FSQ speech tokenizers not only maintains speech synthesis performance but also significantly improves efficiency. Additionally, it further shortens the distance between text and speech tokens, enhancing their correlation.

3.3 Omni Voice

MGM-Omni is capable of generating long-form speech in any personalized voice. To support this capability, we collect over 400k hours of speech data and carefully designed both the data pipeline and the training strategy.

Training Data. To enable zero-shot voice cloning, we curated around 400k hours of audio: around 300k hours of raw speech and around 100k hours of TTS in Chinese and English. For the raw speech portion, we aggregates diverse open-source corpora, including Emilia [18], Libri-heavy [22], Common Voice [4], and Aishell series [9,13,47]. For TTS synthesis portion, we sampled Chinese dialogues from Belle-10M [7] and English dialogues from Lamini-Instruct [55], uniformly selecting 900k Chinese and 700k English conversations by length. Because these sources are somewhat outdated, we regenerated all responses with Qwen2.5-72B [60] and synthesized audio using megatts3 [21], randomly assigning a pre-processed reference voice per sample. Notably, we use roughly 400k hours of audio in total, significantly less than concurrent works [8, 43, 51].

Pre-training. The SpeechLM consists of a pre-trained Qwen3 [59] LLM paired with a randomly initialized TTS-Adapter. The model is trained to generate speech from given text and reference audio through a next speech token prediction objective. The goal of the pre-training stage is to align the speech and text modalities. At this stage, the parameters of the pre-trained Qwen3 LLM remain frozen, while only the TTS-Adapter is updated. Both raw and synthesized speech data are leveraged in pre-training to ensure robustness across diverse speaker timbres.

Model	LibriSpeech Test		CommonVoice		AISHELL
	clean WER ↓	other WER ↓	EN WER ↓	ZH CER ↓	CER ↓
Audio LLMs					
Whisper-large-v3 [45]	1.8	3.6	9.3	12.8	
Qwen2-Audio [10]	**1.3**	**3.4**	**8.6**	**5.2**	
Omni LLMs					
Lyra [64]	2.0	4.0			
VITA-1.5 [17]	3.4	7.5			2.2
Ola [35]	1.9	4.3			
Qwen2.5-Omni [58]	1.6	3.5	**7.6**	5.2	**1.5**
MGM-Omni-7B	1.7	3.6	8.8	4.5	1.9
MGM-Omni-32B	**1.5**	**3.2**	8.0	**4.0**	1.8

Table 2: Omni-comparison on ASR benchmarks. We use Common-Voice, LibriSpeech and AISHELL to evaluate the ASR capability on Chinese and English.

Post-training. The post-training phase aims to enhance SpeechLM's capacity for fluent and accurate speech generation. During this phase, the parameters of both the LLM and the TTS-Adapter are jointly optimized with different learning rates. The TTS-Adapter is trained at a rate five times higher than that of the LLM. The training corpus is primarily composed of high-fidelity TTS-synthesised speech, supplemented with a smaller portion of raw speech data.

4 Experiments

In this section, we present a comprehensive evaluation covering audio understanding, omni-modality understanding, and speech generation, to demonstrate the main properties of MGM-Omni, with particular emphasis on its capacity for long audio understanding and long audio generation and zero-shot voice cloning.

4.1 Multimodoal Understanding

Audio Understanding MGM-Omni can process speech and sound at various lengths. In this section, we assess MGM-Omni on audio benchmarks to evaluate its comprehension, reasoning, and long-context robustness.

Short Audio Understanding. We compare the audio understanding ability (audio → text) of MGM-Omni against leading Audio and Omni LLMs on two primary tasks: automatic speech recognition (ASR) and general audio question answering. First, we evaluate the ASR ability on LibriSpeech [42], CommonVoice [4] and AISHELL [9]. As shown in Table 2, MGM-Omni delivers competitive or superior performance for English and Chinese ASR. For general audio understanding, we evaluate audio QA on AIR-Bench [61], a comprehensive benchmark that

Model	Speech ↑	Sound ↑	Music ↑	Mix ↑	Average ↑
Audio LLMs					
SpeechGPT [62]	1.6	1.0	1.0	4.1	1.9
SALMONN [49]	6.2	6.3	6.0	6.1	6.1
Qwen2-Audio [10]	**7.2**	**7.0**	**6.8**	**6.8**	**6.9**
Omni LLMs					
LLaMA-Omni [15]	5.2	5.3	4.3	4.0	4.7
Mini-Omni2 [56]	3.6	3.5	2.6	3.1	3.2
IXC2.5-OmniLive [63]	1.6	1.8	1.7	1.6	1.7
VITA-1.5 [17]	4.8	5.5	4.9	2.9	4.5
Qwen2.5-Omni [58]	6.8	5.7	4.8	5.4	5.7
Ola [35]	**7.3**	6.4	5.9	6.0	6.4
MGM-Omni-7B	**7.3**	**6.5**	**6.3**	6.1	**6.5**
MGM-Omni-32B	7.1	**6.5**	6.2	**6.2**	**6.5**

Table 3: Omni-comparison on Audio QA benchmarks. We use AIR-Bench for audio QA evaluation.

covers speech, sound, and music inputs. As summarized in Table 3, MGM-Omni outperforms all open source Omni LLMs, including Qwen2.5-Omni [58].

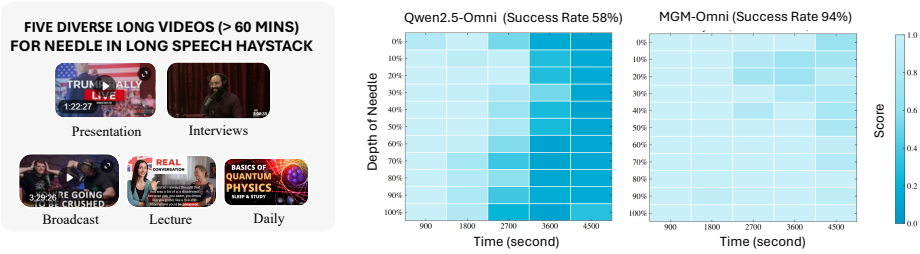

Fig. 5: Omni-comparison for Long Audio Understanding. We adopt needle-in-the-haystack evaluation and report the average success rate over five materials.

Long Audio Understanding. Unlike many open-source Audio and Omni LLMs, MGM-Omni is capable of processing audio inputs exceeding one hour in length. To evaluate its ability on long-form audio understanding, we conducted a needle-in-the-haystack test. Specifically, we select five long audio materials from different topics, each longer than three hours, and evaluate target durations up to 4,500 seconds by inserting open-ended audio question-answer needles at different timestamps. After receiving the full audio input, the model is asked to retrieve the inserted information; exact or semantically equivalent answers are counted as correct. As illustrated in Figure 5, MGM-Omni signifi-

Model	AI2D ↑	ChartQA ↑	MMBench ↑	MME ↑	VideoMME ↑
Qwen2.5-VL-7B [6]	**85.8**	87.3	**82.6**	**2347**	65.1
Qwen2.5-Omni-7B [58]	83.2	85.3	81.8	2340	64.3
MGM-Omni-7B	83.2	**87.8**	81.5	2285	**66.3**

Table 4: Omni-comparison on image and video benchmarks. MGM-Omni largely preserves competitive visual understanding while adding audio understanding and speech generation.

Model	TextVQA^S ↑	DocVQA^S ↑	ChartQA^S ↑	AI2D^S ↑
Intern-Omni-9B [41]	69.1	80.0	56.1	54.0
Lyra-9B [64]	80.0	85.5	61.0	63.1
MGM-Omni-7B	**81.7**	87.4	69.3	70.4
MGM-Omni-32B	78.2	**88.4**	**72.1**	**71.3**

Table 5: Omni-comparison on vision-speech benchmarks. We convert the textual questions in multiple visual question answering benchmarks into synthesized speech to evaluate the multimodal understanding ability.

cantly outperforms Qwen2.5-Omni [58], and the success rate is averaged over the five materials. Moreover, we provide a quantitative comparison in the appendix.

Omni-Modality Understanding MGM-Omni processes text, image, video, and audio inputs. We evaluate MGM-Omni on visual and omni-modal understanding to verify the preservation of visual representations and cross-modal synergy after incorporating audio.

Visual Understanding. We evaluate MGM-Omni on a suite of image and video VQA benchmarks [16, 23, 33, 36] to assess the visual understanding ability (vision → text). As shown in Table 4, MGM-Omni largely preserves competitive visual capability while adding audio understanding and speech generation. Compared with the Qwen2.5-VL base model [6], MGM-Omni improves on ChartQA and VideoMME, remains close on MMBench, and shows a modest drop on AI2D and MME. It is also competitive with Qwen2.5-Omni [58], an open source Omni-MLLM trained with enterprise-scale compute and data. We attribute this to the dual-track brain-mouth design and introducing audio capacity via LoRA [19] fine-tuning.

Omni Modal Understanding. Following Lyra [64], we further evaluate its omni-modal understanding (multimodality → text) by comparing MGM-Omni against other omni-modal LLMs on several speech-instructed VQA benchmarks [23, 36, 37, 48]. As shown in Table 5, MGM-Omni shows a strong ability to follow speech instructions.

4.2 Speech Generation

MGM-Omni supports long-form synthesis (exceeding 10 minutes) with customizable voices. Here, we assess the speech generation capabilities (text → speech) in both short- and long-form settings.

Model	Model Size	EN WER ↓	EN SIM ↑	ZH CER ↓	ZH SIM ↑
CosyVoice2 [14]	0.5B	2.57	0.652	1.45	0.748
Qwen2.5-Omni-3B [58]	0.5B	2.51	0.635	1.58	0.744
Qwen2.5-Omni-7B [58]	2B	2.33	0.641	1.42	0.754
Higgs-Audio-v2 [8]	6B	2.44	0.677	1.66	0.743
MGM-Omni-TTS-0.6B	0.6B	2.48	0.670	1.42	0.750
MGM-Omni-TTS-2B	2B	2.28	0.684	1.28	0.755
MGM-Omni-TTS-4B	4B	**2.22**	**0.686**	**1.18**	**0.758**

(a) Zero-shot short TTS comparison of error rate and speaker similarity in Seed-TTS-Eval. For Qwen2.5-Omni, size indicates the talker module size.

Model	Model Size	RTF ↓	EN WER ↓	ZH CER ↓	EN^H WER ↓	ZH^H CER ↓
CosyVoice2 [14]	0.5B	0.34	14.80	**5.27**	42.48	32.76
MOSS-TTSD-v0.5 [51]	2B	0.23	8.69	6.82	62.61	62.97
Higgs-Audio-v2 [8]	6B	0.33	27.09	31.39	98.61	98.85
VibeVoice [43]	–	0.45	19.23	15.92	76.05	65.58
MGM-Omni-TTS-2B	2B	**0.19**	**4.98**	5.58	**26.26**	**23.58**

(b) Long-form TTS comparison of error rate and inference speed in our collected Long-TTS-Eval. EN-H and ZH-H refers to hardcase evaluation. CosyVoice2 is evaluated with a chunk mode.

Model	EN SIM (%) ↑	EN Std. ↓	ZH SIM (%) ↑	ZH Std. ↓
CosyVoice2 [14]	83.88	1.74	92.23	1.82
MOSS-TTSD-v0.5 [51]	80.68	3.60	88.93	2.23
Higgs-Audio-v2 [8]	76.29	9.72	89.00	4.88
VibeVoice [43]	84.22	4.19	88.56	6.06
MGM-Omni-TTS-2B	**84.66**	**1.63**	**94.32**	**0.97**

(c) Long-form timbre consistency on Long-TTS-Eval. We split each generation into 30-second segments and report the mean and standard deviation of segment-level speaker similarity.

Table 6: Omni-comparison TTS benchmarks. We evaluate short-form TTS with Seed-TTS-Eval (top) and long-form TTS with Long-TTS-Eval (middle and bottom).

Short Speech Generation. We evaluated MGM-Omni against state-of-the-art zero-shot TTS systems and Omni LLMs to assess speech generation. As shown in Table 6a, MGM achieves lower error rates and higher speaker similarity than

open-source TTS models and Omni LLMs on seed-tts-eval [3], demonstrating strong text-to-speech performance and robust zero-shot voice cloning.

Long-TTS-Eval benchmark. MGM-Omni supports more than 10 minutes of long-form speech generation with personalized voice. However, most TTS benchmarks focus on short, simple utterances and rarely stress-test long context or tricky text like formulas, URLs, or classical Chinese poetry. To address this, we introduce **Long-TTS-Eval**, a benchmark designed to evaluate long-form text-to-speech generation systematically. Long-TTS-Eval has two subsets, *long* and *hard*, for English and Chinese TTS evaluation. The *long* subset targets long-horizon narration and includes six text domains, including literature, news, knowledge, talks, comments, and academic papers. These texts are sourced from public materials such as news outlets, Wikipedia, YouTube transcripts, and arXiv papers, resulting in 360 English and 356 Chinese samples. The *hard* subset focuses on complex text normalization and robustness, consisting of URLs, emails, phone numbers, large numbers, and math formulas, resulting in 262 English and 265 Chinese samples. We report WER and CER for English and Chinese with the same pipeline as Seed-TTS-Eval [3]. To reduce penalization from equivalent verbalizations (e.g., "5%" vs. "five percent"), we additionally provide normalized references generated by GPT-5 [40] for evaluation. The final evaluation metric is the minimum value of WER and CER with the original reference text and normalized reference text. Beyond intelligibility, long-form speech generation also requires stable speaker identity across the whole utterance. For long-form timbre consistency, we split each generated speech into 30-second segments and compute speaker similarity between each segment and the reference audio. We report the mean and standard deviation of segment-level similarity, where a higher mean and lower deviation indicate more stable speaker preservation. This design separates intelligibility, text-normalization robustness, and timbre consistency, making the benchmark sensitive to both transcription accuracy and speaker preservation over long horizons. We leave more detailed information on the benchmark in the appendix.

Long Speech Generation. We compare MGM-Omni with two categories of open-source TTS systems: (1) Native long TTS models, represented by MOSS-TTSD-v0.5 [51], Higgs-Audio-v2 [8], and VibeVoice [43]; and (2) Non-native models extended via chunking, represented by CosyVoice2 [14]. We report WER for English TTS, CER for Chinese TTS, and RTF for inference efficiency. The RTF is evaluated by a subset of Long-TTS-Eval. As shown in Table 6b, MGM-Omni achieves lower error rates in most scenarios, along with the lowest RTF. Moreover, as shown in Table 6c, MGM-Omni obtains the highest segment-level speaker similarity and the lowest similarity variance on both English and Chinese long-form subsets, indicating more stable timbre preservation over extended synthesis. Notably, MGM-Omni's two-stage training uses under 400k hours of audio, far fewer than the 1M or even 10M hours used in concurrent long-form TTS works. The gains are particularly important for long-form narration, where small local pronunciation errors, unstable prosody, or gradual speaker drift can accu-

mulate across many segments and noticeably degrade listening quality. These results demonstrate the effectiveness, robustness, and data efficiency of MGM-Omni.

4.3 Ablation Study

Audio Encoder	EN WER ↓	ZH CER ↓
Qwen2-Audio	13.0	3.9
Belle-Whisper	21.7	5.0
Info Mining	**9.1**	**3.5**

(a) Audio Encoder

Parallel	RTF	EN WER ↓	ZH CER ↓
1	0.57	**1.86**	**1.15**
2	0.32	2.02	1.23
4	**0.19**	2.28	1.28

(b) Parallel Decoding

Chunk	EN WER ↓	ZH CER ↓
	31.84	8.97
✓	**4.98**	**5.64**

(c) Chunk-Based Decoding

Table 7: Ablation study. We conduct ablation studies on the audio encoder, parallel decoding, and chunk-based decoding.

Audio Encoder. We ablate audio encoder designs on CommonVoice ASR [4]. As shown in Table 7a, combining the Qwen2-Audio encoder [10] and the Belle-Whisper-large-v3 encoder [7] with information mining yields the best overall performance in audio understanding. The result suggests that the two encoders provide complementary information: the speech-recognition-oriented encoder contributes fine-grained linguistic cues, while the general audio encoder retains broader acoustic and paralinguistic evidence that is useful for robust perception. Note that, compared with the final model, we do not use the long audio QA data here for training.

Chunk-based Decoding. We evaluate long-form speech generation on our Long-TTS-Eval to assess the impact of chunk-based decoding. As shown in Table 7c, removing chunk-based decoding leads to a substantially higher error rate, exceeding that of concurrent works. This shows that chunking is not merely an engineering shortcut for handling long sequences; it is also central to stabilizing generation because each segment is decoded under a manageable context length while remaining conditioned on the global reference voice. Given that concurrent methods typically use millions to tens of millions of hours of audio, we attribute MGM-Omni's data efficiency primarily to its use of chunk-based decoding.

Parallel Decoding. We ablate the impact of parallel decoding by comparing both TTS performance and inference speed. TTS performance is measured on Seed-TTS-Eval [3], while inference speed is assessed using 16 Chinese and 16 English samples drawn from Long-TTS-Eval. We report the real-time factor (RTF) on a single H800 GPU to compare the inference speed. As shown in Table 7b, increasing the parallel size slightly raises the audio error rate but substantially accelerates inference by 3×. This trade-off is favorable for interactive and long-form scenarios, where latency strongly affects usability and minor token-level deviations are less noticeable than long waiting time. To balance quality and

speed, we set the parallel size to 4. We anticipate that incorporating more advanced Multi-Token Prediction (MTP) techniques [30] will further improve audio quality at larger parallel sizes.

5 Conclusion

We present MGM-Omni, an Omni LLM that supports long-form omnimodal understanding and robust long-duration speech generation with personalized voices. We propose a "brain–mouth" dual-track design that decouples real-time speech synthesis (SpeechLM) from multimodal understanding and reasoning (MLLM), enabling efficient cross-modal interaction in an any-to-any framework without degrading single-modal performance. For understanding, we employ a dual audio encoder that fuses acoustic and semantic cues to deliver robust long-form audio perception. For generation, we introduce Chunk-Based Parallel Decoding to bridge the token-rate gap between text and speech, enabling efficient, low-latency synthesis. Conditioning SpeechLM on reference audio further enables zero-shot voice cloning with consistent timbre. In short, with academic-scale data and compute, MGM-Omni achieves accurate long-form audio understanding, stable real-time long-duration speech synthesis, and consistent speaker identity across diverse acoustic conditions, paving the way toward scalable any-to-any human-like multimodal intelligence.

Limitations and future work. MGM-Omni still has limitations in extremely long interactive sessions, fine-grained speaking-style control, and rare acoustic conditions. Its visual understanding is not yet as strong as specialized vision-language models, and the current system does not support agentic tool use or fully duplex speech interaction. Future work will further improve visual understanding, long-context memory, controllable speech generation, tool-augmented interaction, and real-time duplex voice communication.

Responsible use and misuse prevention. As with other systems that support voice-conditioned speech generation, MGM-Omni should be used with care in applications where synthesized speech may be confused with a real speaker. Our training relies heavily on open-source datasets, and we will follow the corresponding licenses, terms of use, and data governance requirements. We assume reference voices are used with speaker consent; practical deployments should consider consent checks, provenance metadata, watermarking, and synthetic-speech detection when releasing voice-cloning capabilities.

Acknowledgements. This work was supported in part by the Research Grants Council under the Areas of Excellence scheme grant AoE/E-601/22-R.

References

1. AI, I.: Ming-omni: A unified multimodal model for perception and generation (2025), https://arxiv.org/abs/2506.09344
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
3. Anastassiou, P., Chen, J., Chen, J., Chen, Y., Chen, Z., Chen, Z., Cong, J., Deng, L., Ding, C., Gao, L., et al.: Seed-tts: A family of high-quality versatile speech generation models. arXiv preprint arXiv:2406.02430 (2024)
4. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G.: Common voice: A massively-multilingual speech corpus. arXiv preprint arXiv:1912.06670 (2019)
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 **1**(2), 3 (2023)
6. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
7. BELLEGroup: Belle: Be everyone's large language model engine. https://github.com/LianjiaTech/BELLE (2023)
8. Boson AI: Higgs Audio V2: Redefining Expressiveness in Audio Generation. https://github.com/boson-ai/higgs-audio (2025), gitHub repository. Release blog available at https://www.boson.ai/blog/higgs-audio-v2
9. Bu, H., Du, J., Na, X., Wu, B., Zheng, H.: Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In: 2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA). pp. 1–5. IEEE (2017)
10. Chu, Y., Xu, J., Yang, Q., Wei, H., Wei, X., Guo, Z., Leng, Y., Lv, Y., He, J., Lin, J., Zhou, C., Zhou, J.: Qwen2-audio technical report. arXiv preprint arXiv:2407.10759 (2024)
11. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
12. Dehghani, M., Mustafa, B., Djolonga, J., Heek, J., Minderer, M., Caron, M., Steiner, A., Puigcerver, J., Geirhos, R., Alabdulmohsin, I.M., et al.: Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. Advances in Neural Information Processing Systems **36**, 2252–2274 (2023)
13. Du, J., Na, X., Liu, X., Bu, H.: Aishell-2: Transforming mandarin asr research into industrial scale. arXiv preprint arXiv:1808.10583 (2018)
14. Du, Z., Wang, Y., Chen, Q., Shi, X., Lv, X., Zhao, T., Gao, Z., Yang, Y., Gao, C., Wang, H., et al.: Cosyvoice 2: Scalable streaming speech synthesis with large language models. arXiv preprint arXiv:2412.10117 (2024)
15. Fang, Q., Guo, S., Zhou, Y., Ma, Z., Zhang, S., Feng, Y.: Llama-omni: Seamless speech interaction with large language models. arXiv preprint arXiv:2409.06666 (2024)
16. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark

of multi-modal llms in video analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24108–24118 (2025)
17. Fu, C., Lin, H., Long, Z., Shen, Y., Dai, Y., Zhao, M., Zhang, Y.F., Dong, S., Li, Y., Wang, X., et al.: Vita: Towards open-source interactive omni multimodal llm. arXiv preprint arXiv:2408.05211 (2024)
18. He, H., Shang, Z., Wang, C., Li, X., Gu, Y., Hua, H., Liu, L., Yang, C., Li, J., Shi, P., et al.: Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In: 2024 IEEE Spoken Language Technology Workshop (SLT). pp. 885–890. IEEE (2024)
19. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
20. Huang, A., Wu, B., Wang, B., Yan, C., Hu, C., Feng, C., Tian, F., Shen, F., Li, J., Chen, M., et al.: Step-audio: Unified understanding and generation in intelligent speech interaction. arXiv preprint arXiv:2502.11946 (2025)
21. Jiang, Z., Ren, Y., Li, R., Ji, S., Ye, Z., Zhang, C., Jionghao, B., Yang, X., Zuo, J., Zhang, Y., et al.: Sparse alignment enhanced latent diffusion transformer for zero-shot speech synthesis. arXiv preprint arXiv:2502.18924 (2025)
22. Kang, W., Yang, X., Yao, Z., Kuang, F., Yang, Y., Guo, L., Lin, L., Povey, D.: Libriheavy: A 50,000 hours asr corpus with punctuation casing and context. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 10991–10995. IEEE (2024)
23. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
24. KimiTeam, Ding, D., Ju, Z., Leng, Y., Liu, S., Liu, T., Shang, Z., Shen, K., Song, W., Tan, X., Tang, H., Wang, Z., Wei, C., Xin, Y., Xu, X., Yu, J., Zhang, Y., Zhou, X., Charles, Y., Chen, J., Chen, Y., Du, Y., He, W., Hu, Z., Lai, G., Li, Q., Liu, Y., Sun, W., Wang, J., Wang, Y., Wu, Y., Wu, Y., Yang, D., Yang, H., Yang, Y., Yang, Z., Yin, A., Yuan, R., Zhang, Y., Zhou, Z.: Kimi-audio technical report (2025), https://arxiv.org/abs/2504.18425
25. Kong, J., Kim, J., Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in neural information processing systems **33**, 17022–17033 (2020)
26. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
27. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: European Conference on Computer Vision. pp. 323–340. Springer (2024)
28. Li, Y., Zhang, Y., Wang, C., Zhong, Z., Chen, Y., Chu, R., Liu, S., Jia, J.: Mini-gemini: Mining the potential of multi-modality vision language models. arXiv preprint arXiv:2403.18814 (2024)
29. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
30. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
31. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv:2310.03744 (2023)
32. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)

33. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
34. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)
35. Liu, Z., Dong, Y., Wang, J., Liu, Z., Hu, W., Lu, J., Rao, Y.: Ola: Pushing the frontiers of omni-modal language model. arXiv preprint arXiv:2502.04328 (2025)
36. Masry, A., Do, X.L., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In: Findings of the association for computational linguistics: ACL 2022. pp. 2263–2279 (2022)
37. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
38. OpenAI: Chatgpt. https://openai.com/blog/chatgpt/ (2023)
39. OpenAI: Gpt-4 technical report. arXiv:2303.08774 (2023)
40. OpenAI: Gpt-5 system card (2025)
41. OpenGVLab: InternOmni: Extending internvl with audio modality. https://internvl.github.io/blog/2024-07-27-InternOmni/ (2024)
42. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: an asr corpus based on public domain audio books. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5206–5210. IEEE (2015)
43. Peng, Z., Yu, J., Wang, W., Chang, Y., Sun, Y., Dong, L., Zhu, Y., Xu, W., Bao, H., Wang, Z., et al.: Vibevoice technical report. arXiv preprint arXiv:2508.19205 (2025)
44. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
45. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022). https://doi.org/10.48550/ARXIV.2212.04356, https://arxiv.org/abs/2212.04356
46. Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al.: Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 4779–4783. IEEE (2018)
47. Shi, Y., Bu, H., Xu, X., Zhang, S., Li, M.: Aishell-3: A multi-speaker mandarin tts corpus and the baselines. arXiv preprint arXiv:2010.11567 (2020)
48. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
49. Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Zhang, C.: Salmonn: Towards generic hearing abilities for large language models. arXiv preprint arXiv:2310.13289 (2023)
50. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
51. Team, O.: Text to spoken dialogue generation (2025)

52. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. arXiv:2302.13971 (2023)
53. Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., Kavukcuoglu, K., et al.: Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499 **12**, 1 (2016)
54. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
55. Wu, M., Waheed, A., Zhang, C., Abdul-Mageed, M., Aji, A.F.: Lamini-lm: A diverse herd of distilled models from large-scale instructions. CoRR **abs/2304.14402** (2023), https://arxiv.org/abs/2304.14402
56. Xie, Z., Wu, C.: Mini-omni: Language models can hear, talk while thinking in streaming. arXiv preprint arXiv:2408.16725 (2024)
57. Xie, Z., Wu, C.: Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. arXiv preprint arXiv:2410.11190 (2024)
58. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al.: Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215 (2025)
59. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
60. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., et al.: Qwen2.5 technical report. arXiv preprint arXiv:2502.13923 (2025)
61. Yang, Q., Xu, J., Liu, W., Chu, Y., Jiang, Z., Zhou, X., Leng, Y., Lv, Y., Zhao, Z., Zhou, C., et al.: Air-bench: Benchmarking large audio-language models via generative comprehension. arXiv preprint arXiv:2402.07729 (2024)
62. Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., Qiu, X.: Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. arXiv preprint arXiv:2305.11000 (2023)
63. Zhang, P., Dong, X., Cao, Y., Zang, Y., Qian, R., Wei, X., Chen, L., Li, Y., Niu, J., Ding, S., et al.: Internlm-xcomposer2. 5-omnilive: A comprehensive multimodal system for long-term streaming video and audio interactions. arXiv preprint arXiv:2412.09596 (2024)
64. Zhong, Z., Wang, C., Liu, Y., Yang, S., Tang, L., Zhang, Y., Li, J., Qu, T., Li, Y., Chen, Y., et al.: Lyra: An efficient and speech-centric framework for omni-cognition. arXiv preprint arXiv:2412.09501 (2024)
65. Zhou, S., Zhou, Y., He, Y., Zhou, X., Wang, J., Deng, W., Shu, J.: Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. arXiv preprint arXiv:2506.21619 (2025)
66. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)