2025-09-22

Qwen3-Omni Technical Report

Qwen Team

https://github.com/QwenLM/Qwen3-Omni
https://huggingface.co/spaces/Qwen/Qwen3-Omni-Demo
https://modelscope.cn/studios/Qwen/Qwen3-Omni-Demo

Abstract

We present Qwen3-Omni, a single multimodal model that for the first time maintains state-of-the-art performance across text, image, audio, and video without any degradation relative to single-modal counterparts. Qwen3-Omni matches the performance of same-sized single-modal models within the Qwen series and excels particularly on audio tasks. Across 36 audio and audio-visual benchmarks, Qwen3-Omni achieves open-source state-of-the-art (SOTA) on 32 benchmarks and overall SOTA on 22, outperforming strong closed-source models such as Gemini-2.5-Pro, Seed-ASR, and GPT-4o-Transcribe. Qwen3-Omni adopts a Thinker–Talker Mixture-of-Experts (MoE) architecture that unifies perception and generation across text, images, audio, and video, yielding fluent text and natural real-time speech. It supports text interaction in 119 languages, speech understanding in 19 languages and speech generation in 10 languages. The system can process audio recordings up to 40 minutes per instance for ASR and spoken-language understanding, enabling high-quality audio and audiovisual experiences across locales. It demonstrates strong instruction following and allows fine-grained customization of conversational tone and persona via user-defined system prompts. To reduce first-packet latency in streaming synthesis, the Talker autoregressively predicts discrete speech codecs using a multi-codebook scheme. Leveraging the representational capacity of these codebooks, we replace computationally intensive block-wise diffusion with a lightweight causal ConvNet, enabling streaming from the first codec frame. In cold-start settings (no prior context), Qwen3-Omni achieves a theoretical end-to-end first-packet latency of 234 ms. To further strengthen multimodal reasoning, we introduce a Thinking model that explicitly reasons over inputs from any modality. Since the research community currently lacks a general-purpose audio captioning model, we fine-tuned Qwen3-Omni-30B-A3B to obtain Qwen3-Omni-30B-A3B-Captioner, which produces detailed, low-hallucination captions for arbitrary audio inputs. Qwen3-Omni-30B-A3B, Qwen3-Omni-30B-A3B-Thinking and Qwen3-Omni-30B-A3B-Captioner are publicly released under the Apache 2.0 license.

1 Introduction

Humans perceive visual and auditory inputs in parallel, cognitively process these signals, and emit responses through textual expression, vocalization, and tool-mediated or bodily actions, facilitating information exchange with other organisms and demonstrating intelligence. Building on the rapid advances in the understanding and reasoning capabilities of unimodal large models (Brown et al., 2020; OpenAI, 2023; Gemini Team, 2024; Anthropic, 2023a;b; 2024; Bai et al., 2023a; Yang et al., 2024; 2025a; Touvron et al., 2023; Dubey et al., 2024; Li et al., 2023; Liu et al., 2023; Zhu et al., 2023; Bai et al., 2023b; 2025; Chu et al., 2023; 2024), natively multimodal systems have drawn substantial attention (OpenAI, 2024; Comanici et al., 2025; Xu et al., 2025). Human learning typically progresses through the coordinated use of multiple modalities, where complementary specialization and cross-modal synergy improve learning efficiency. However, contemporary LLM-centric multimodal models often exhibit modality trade-offs, with gains in one modality accompanied by degradation in others.

In this report, we take a step toward resolving this limitation by exploring integrated multimodal training within the prevailing LLM-based paradigm. We demonstrate that joint multimodal training can achieve parity across all modalities—i.e., no modality-specific performance degradation—while markedly enhancing cross-modal capabilities such as video understanding. A key ingredient is mixing unimodal and cross-modal data during the early stage of text pretraining. As evidenced by Qwen3-Omni-30B-A3B-Base, its text and vision performance is on par with same-sized single-modal text and vision base models across extensive benchmarks, while simultaneously exhibiting strong audio competence, audiovisual understanding, cross-modal “thinking”, and real-time audiovisual interaction. The development of non-degrading multimodal systems is an achievable objective. Such systems are characterized by two key properties: first, their ability to match the performance of specialized unimodal models in their respective

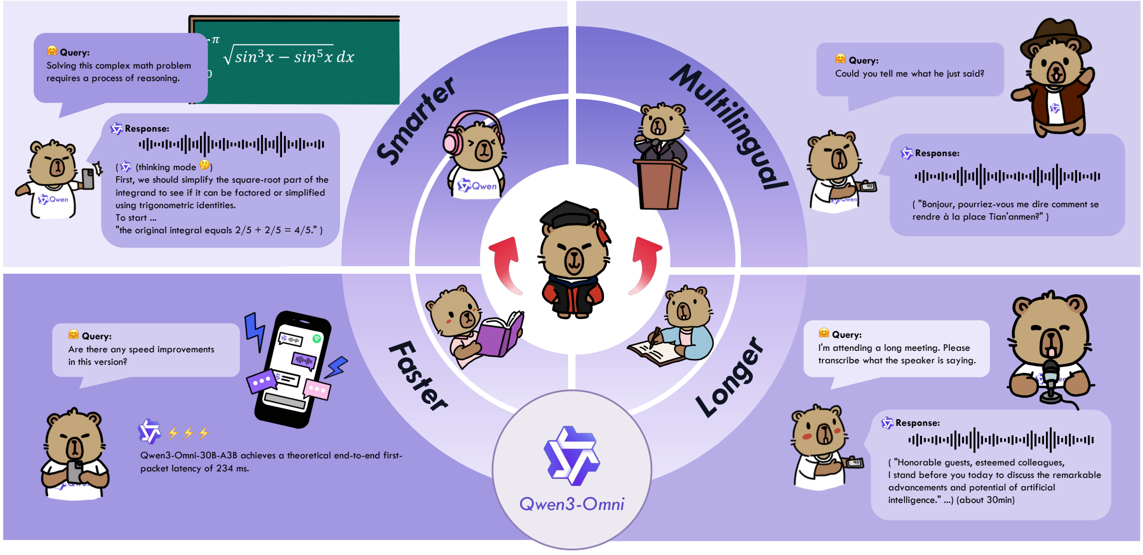
Query: Solving this complex math problem requires a process of reasoning.
$\int_0^{\pi} \sqrt{sin^3x - sin^5x}\,dx$
Response: ((thinking mode) First, we should simplify the square-root part of the integrand to see if it can be factored or simplified using trigonometric identities. To start ... "the original integral equals 2/5 + 2/5 = 4/5.")
Smarter
Multilingual
Query: Could you tell me what he just said?
Response: ("Bonjour, pourriez-vous me dire comment se rendre à la place Tian'anmen?")
Query: Are there any speed improvements in this version?
Qwen3-Omni-30B-A3B achieves a theoretical end-to-end first-packet latency of 234 ms.
Faster
Longer
Query: I'm attending a long meeting. Please transcribe what the speaker is saying.
Response: ("Honorable guests, esteemed colleagues, I stand before you today to discuss the remarkable advancements and potential of artificial intelligence." ...) (about 30min)
Qwen3-Omni

Figure 1: Qwen3-Omni is a unified end-to-end model capable of processing multiple modalities, such as text, audio, image and video, and generating real-time text or speech response. Based on these features, Qwen3-Omni supports a wide range of tasks, including but not limited to voice dialogue, video dialogue, and video reasoning.

tasks, and second, their capacity to facilitate novel cross-modal reasoning and interaction. These latter capabilities represent a significant advantage, as they are not present in traditional unimodal approaches.

Qwen3-Omni builds on the Thinker–Talker architecture introduced in Qwen2.5-Omni (Xu et al., 2025) and introduces **five key upgrades**: (1) both the Thinker and Talker are upgraded to Mixture-of-Experts (MoE) designs; (2) we replace Whisper audio encoder with our AuT (Audio Transformer) encoder, trained from scratch on 20 million hours of supervised audio, yielding stronger general-purpose audio representations. AuT employs block-wise window attention to enable real-time prefill caching; (3) on the speech generation side, we adopt a multi-codebook representation, whose increased capacity supports faithful modeling of diverse voices, paralinguistic cues, and acoustic phenomena; (4) the Talker shifts from single-track to multi-track codec modeling, autoregressively predicting multiple codebook layers via MTP modules, while the waveform stage (Code2Wav) replaces block-wise DiT with a lightweight convolutional network (ConvNet); and (5) the input and output audio code rates are reduced to 12.5 Hz, with the output codec enabling single-frame, immediate speech synthesis. Taken together, these changes enable low-latency speech interaction under high concurrency in industrial-scale deployments.

Compared with Qwen2.5-Omni, Qwen3-Omni introduces **four major improvements**: (1) support for audio understanding on inputs exceeding 40 minutes; (2) expanded language coverage to 119 written languages, 19 and 10 spoken languages for understanding and generation respectively ; (3) a Thinking model enabling full-modality reasoning, including audio–video and audio-only scenarios; and (4) improved streaming performance with end-to-end latency as low as 234 ms.

Critically, Qwen3-Omni maintains state-of-the-art performance on text and visual modalities without degradation relative to same-size single-model Qwen counterparts. Across 36 audio and audio-visual benchmarks, it achieves open-source SOTA on 32 and sets the SOTA on 22, outperforming strong closed-source systems such as Gemini 2.5 Pro, Seed-ASR, and GPT-4o-Transcribe.

The remainder of this paper is organized as follows. Section 2 presents the algorithms and architecture of Qwen3-Omni. Sections 3 and 4 describe the pretraining and post-training datasets and pipelines, respectively. Section 5 reports the experimental results. Section 6 compares Qwen3-Omni with recent Qwen models of comparable parameter scales, demonstrating multimodal performance without modality-induced degradation.

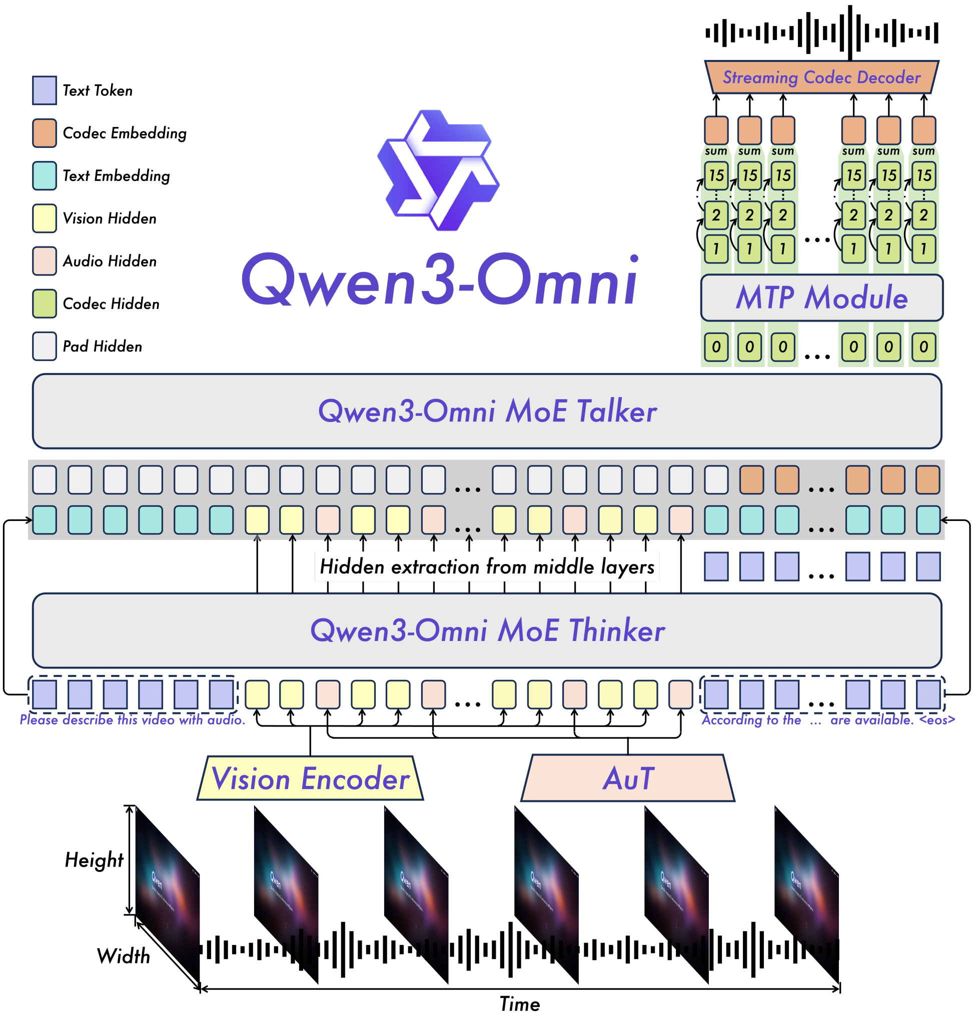

Figure 2: The overview of Qwen3-Omni. Qwen3-Omni adopts the Thinker-Talker architecture. Thinker is tasked with text generation while Talker focuses on generating streaming speech tokens by receives high-level representations directly from Thinker. To achieve ultra–low-latency streaming, Talker autoregressively predicts a multi-codebook sequence. At each decoding step, an MTP module outputs the residual codebooks for the current frame, after which the Code2Wav renderer incrementally synthesizes the corresponding waveform, enabling frame-by-frame streaming generation.

2 Architecture

2.1 Overview

As shown in Figure 2, Qwen3-Omni employs Thinker-Talker architecture (Xu et al., 2025). Compared with Qwen2.5-Omni, Qwen3-Omni introduces the following changes for greater scalability and control:

- Both the Thinker and Talker adopt Mixture-of-Experts (MoE) architectures to support high concurrency and fast inference.
- Talker no longer consumes the Thinker's high-level text representations and conditions only on audio and visual multimodal features. This design is motivated by: (i) for textual content, discrete tokens and embeddings are effectively information-equivalent; and (ii) multimodal conditioning is necessary for audio–video–coordinated speech generation such as preserving prosody/timbre in speech translation. Moreover, this decoupling allows external modules (e.g., RAG, function calling, safety filters) to intervene on the Thinker's textual output and, if desired, supply text to the Talker via controlled preprocessing for streaming synthesis.
- Since textual representations are decoupled, the Thinker and Talker can use distinct system prompts, independently controlling the Thinker's response style and the Talker's audio style.
- The Talker adopts a multi-codebook autoregressive scheme: Talker generates one codec frame

per step, while the MTP module produces the remaining residual codebooks.

- The Code2Wav is implemented as a lightweight causal ConvNet, simplifying the final stage of audio synthesis.

During training and inference, the Talker directly ingests high-dimensional multimodal features from the Thinker and shares access to the full conversational history. As a result, the system operates as a cohesive single model, enabling end-to-end training and unified inference.

In the following sections, we first introduce with our newly proposed AuT encoder, including its training methodology. Then, describe how Thinker processes various inputs. We then detail Talker's multi-codebook streaming speech generation. Finally, we highlight a series of improvements on both the understanding and generation modules aimed at achieving ultra–low-latency, end-to-end streaming audio inference.

2.2 Audio Transformer (AuT)

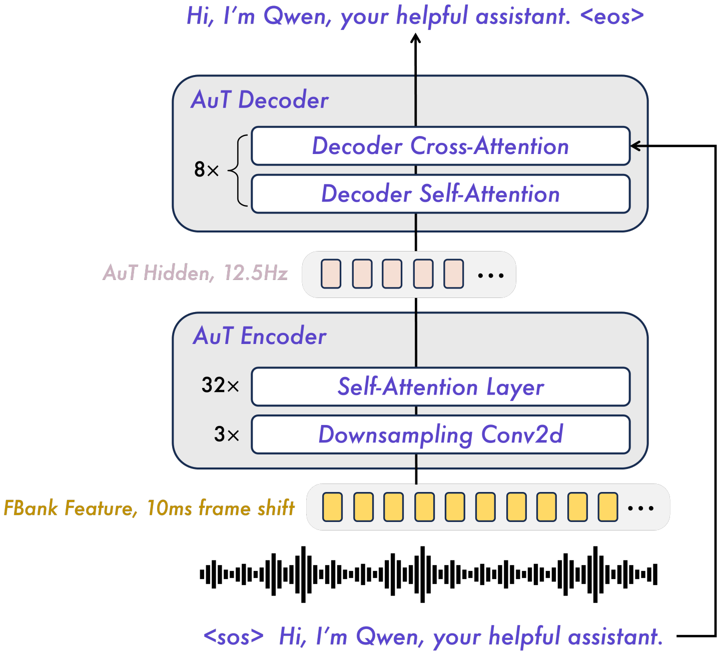

Figure 3: The overview of AuT. AuT is an attention-encoder-decoder based auto-regressive model, which is trained from scratch on 20 million hours of supervised audio. Qwen3-Omni employs the AuT encoder as the audio encoder to obtain general purpose audio representations at a token rate of 12.5Hz.

Audio Transformer (AuT) is an attention-encoder-decoder model, as is shown in Figure 3, trained from scratch on 20 million hours of supervised audio data. During training, the filter bank features of the audio are downsampled 8 times using Conv2D blocks before the attention layers, reducing the token rate to 12.5 Hz. To learn stronger and more general-purpose audio representations, AuT is trained on large-scale audio datasets with both speech recognition and audio understanding tasks. Specifically, the training data includes 80% Chinese and English pseudo-labeled ASR data, 10% ASR data from other languages, and 10% audio understanding data. To balance the efficiency of real-time prefill caching with the performance for offline audio tasks, AuT utilizes flash attention with dynamic attention window sizes, covering attention query patterns ranging from 1 to 8 seconds. In Qwen3-Omni, we employ the AuT encoder as the audio encoder, which contains approximately 0.6B parameters.

2.3 Perceivation

Text, Audio, Image and Video (w/o Audio). Thinker converts text, audio, image, and video (without audio) into a series of representations for input. For text inputs, we use Qwen's tokenizer (Yang et al., 2025a), which applies byte-level byte-pair encoding with a vocabulary of 151,643 regular tokens. For audio inputs and audio extracted from video, we resample to 16 kHz and convert the raw waveform into a 128 channel mel-spectrogram with a 25 ms window and a 10 ms hop. We adopt AuT encoder as our audio encoder, which is trained from scratch on 20 millions hours of audio data, and each frame of the audio representation corresponds to approximately an 80 ms segment of the original audio signal. Furthermore, we employ the vision encoder from Qwen3-VL, initialized from SigLIP2-So400m (Tschannen et al., 2025) with approximately 543 million parameters, enabling handling of both image and video inputs. The vision encoder is trained on a mixture of image and video data, ensuring strong image understanding and video comprehension. To preserve video information as completely as possible while aligning with the audio sampling rate, we sample video frames at a dynamic frame rate.

Video and Multimodal Position Embedding (TM-RoPE) Drawing inspiration from Qwen2.5-Omni, we employs a Time-aligned Multimodal Rotary Position Embedding (TM-RoPE), which extends the Multimodal Rotary Position Embedding (M-RoPE) (Bai et al., 2023b) by incorporating absolute temporal information. TM-RoPE factorizes the conventional rotary position embedding into three distinct dimensions: temporal, height, and width. In the original M-RoPE formulation, temporal dependencies are modeled using the initial 16 rotary angles, which correspond to higher frequencies and exhibit stronger oscillatory patterns. While this design is effective for capturing fine-grained local temporal variations, it can impede the model's ability to extrapolate over extended sequences. To address this limitation, we introduce a modified allocation of rotary angles. Specifically, the temporal, height, and width dimensions are interleaved and assigned 24, 20, and 20 rotary angles, respectively. This redistribution fosters a more balanced representation of both local semantics and long-range dependencies, thereby enhancing the model's overall performance. The application of TM-RoPE is tailored to the specific modality of the input data. For text inputs, the three components share identical position identifiers, rendering TM-RoPE functionally equivalent to a one-dimensional RoPE (Su et al., 2024). Similarly, audio inputs utilize shared position IDs but are further augmented with absolute temporal encodings, where each temporal ID corresponds to a duration of 80 ms. For image data, a constant temporal ID is assigned to all visual tokens, while their distinct row and column positions determine the height and width IDs.

In the context of multimodal audiovisual streams, the audio component is encoded with a temporal ID for every 80 ms. The video is treated as a sequence of frames with monotonically increasing temporal IDs that are dynamically adjusted based on their actual timestamps to ensure a consistent temporal resolution of 80 ms per ID. The height and width IDs for video frames are assigned in the same manner as for still images. To prevent positional conflicts when processing multiple modalities, the position numbering is made contiguous, with each subsequent modality commencing from one plus the maximum position ID of the preceding modality. This refined approach to positional encoding enables the model to effectively integrate and jointly model information from diverse modalities. In a departure from Qwen2.5-Omni, which segments audiovisual representations into fixed 2-second chunks, Qwen3-Omni directly aligns these representations using their temporal IDs, which are explicitly anchored to absolute time. This design choice affords the model the flexibility to support streaming inputs of arbitrary duration.

2.4 Speech Generation

For speech synthesis in multi-turn dialogues, our Talker module is conditioned on a rich context inherited from a "Thinker" component, comprising historical textual tokens, multimodal representations, and the current turn's streamed text. This reliance on long-context information is critical, as high-fidelity speech synthesis must adapt acoustic attributes like prosody, loudness, and emotion to the ongoing discourse, a principle well-established in context-aware generative models.

Architecturally, our approach departs from Xu et al. (2025) by operating directly on RVQ tokens. The Talker employs a hierarchical prediction scheme: the backbone ingests the aggregated codebook features of the current frame and uses a linear head to predict the zeroth codebook, after which a multi-token prediction (MTP) module generates all residual codebooks. This strategy enables the model to learn a complete representation of acoustic details, enhancing vocal expressivity. Consequently, waveform reconstruction is simplified to a lightweight causal ConvNet (Code2Wav), which significantly reduces inference latency and computational cost (FLOPs) while achieving superior audio fidelity compared to more complex DiT-based vocoders.

2.5 Designs for Streaming and Concurrency

In streaming audiovisual interaction scenarios, the first-packet latency is a critical factor affecting user experience, and the model's concurrency capability is key to reducing service costs and improving response speed. This section discusses how Qwen3-Omni enhances concurrency and reduces first-packet latency through algorithmic and architectural optimizations.

Chunked Prefilling and MoE Architecture. In Qwen3-Omni, we retain the chunked-prefilling mechanism as implemented in Qwen2.5-Omni, whose audio and vision encoders are capable of outputting chunks along the temporal dimension. During real-time interaction, Thinker and Talker modules perform

Table 1: **The architectural design of Qwen3-Omni-30B-A3B and the end-to-end first-packet latency for Audio/Video (ms).**

Module	Architecture	Params	Streaming
Audio Encoder	AuT	650M	✓
Vision Encoder	SigLIP2-So400M	540M	-
Thinker	MoE Transformer	30B-A3B	✓
Talker	MoE Transformer	3B-A0.3B	✓
MTP	Dense Transformer	80M	✓
Code2wav	ConvNet	200M	✓
End-to-End First-Packet Latency: **234/547ms**			

asynchronous prefilling: when Thinker completes prefilling the current chunk, its output high-level representations are immediately used to prefill the Talker's current chunk asynchronously, while Thinker prefills its next chunk. This approach significantly reduces the Time-To-First-Token (TTFT) for both the Thinker and the Talker. Architecturally, both Thinker and the Talker in Qwen3-Omni adopt the MoE design, which is highly effective for improving service throughput. Compared to dense models, the MoE architecture significantly decreases IO consumption arising from KV cache during processing of long sequences, thereby increasing tokens per second (TPS) during generation and enhancing concurrency.

Streaming Multi-Codebook Codec Generation. To minimize the user's waiting time for receiving the first generated packet, we propose a *left context only multi-codebook generation* mechanism. As shown in Figure 2, once Talker generates the first token, the MTP module predicts the remaining tokens for the current frame. These tokens are then decoded into waveform by a streaming multi-codebook codec decoder that only attends to the left context. Unlike Qwen2.5-Omni that requires waiting for sufficient block-context from the Talker before synthesis, Qwen3-Omni can output the waveform immediately after the Talker generates each token, significantly reducing first-packet latency.

Lightweight MTP module and ConvNet. Both the MTP module and codec decoder are lightweight modules, which have low computational FLOPs and support batched inference, making them well-suited for high-concurrency scenarios. The MTP Module is an ultra-lightweight fixed-step autoregressive dense transformer, with low memory bandwidth requirements on inference hardware, thereby naturally enabling efficient batch processing of high throughput requests. Its fixed-step autoregressive inference mechanism allows it to effectively leverage a fixed KV cache memory space for acceleration, achieving low inference latency. Meanwhile, the ConvNet-based codec decoder also achieves high throughput with low latency because its convolutional architecture enjoys extensive hardware acceleration support across diverse inference platforms, and it enables efficient batched inference.

Table 2: **Theoretical First-Packet Latency of Qwen3-Omni wit Different Concurrency.**

	Qwen3-Omni-30B-A3B		
	1 Concurrency	**4 Concurrency**	**6 Concurrency**
Thinker-Talker Tail Packet Preprocessing Latency	72/160ms	94/180ms	100/200ms
Thinker Time-to-First-Token (TTPT)	88/160ms	468/866ms	673/1330ms
Talker Time-to-First-Token (TTPT)	57/210ms	145/450ms	376/734ms
MTP Module Time Cost Per Token	14ms	16ms	18ms
Codec Decoder Time Cost Per Code	3ms	5ms	5ms
Overral Latency (Audio/Video)	**234/547ms**	**728/1517ms**	**1172/2284ms**
Thinker Token Generation Rate (TPS)	75 tokens/s	63 tokens/s	53 tokens/s
Talker Token Generation Rate (TPS)	140 tokens/s	125 tokens/s	110 tokens/s
Generation RTF(Real Time Factor)	**0.47**	**0.56**	**0.66**

Table 2 presents the theoretical first-packet latency for Qwen3-Omni under typical computational resources across varying concurrency scenarios. Experiments are conducted on the vLLM framework (Kwon et al., 2023) to process concurrent audiovisual streams, with optimizations applied via *torch.compile* and CUDA Graph acceleration to the MTP Module and codec decoder. Several factors influence the total first-packet latency. First, the model sizes of Thinker and Talker impact their tail packet preprocessing latency (multi-modal data preprocessing and inference for Audio and Vision Encoder) and Time-To-First-Token (TTPT). Second, the architectures and sizes of the MTP Module and Codec Decoder affect their inference latency. Due to the sequential dependency between these components, the total first-packet latency represents the sum of these individual latencies. As shown in the results, the MoE architecture of Thinker and Talker ensures that their prefill latency and TTPT remain largely unaffected under high concurrency. Meanwhile, the lightweight design of the MTP Module and Codec Decoder minimizes their computational overhead, resulting in a lower impact on first-packet latency. Furthermore, after the initial packet is output and the model starts streaming audio synthesis, the 12.5Hz token rate Talker requires only one token to synthesize 80ms audio. Consequently, the Generation Real Time Factor (RTF) is calculated by dividing the sum of: (1) the time taken by Thinker and Talker to generate one token; and (2) the processing time per token for the MTP Module and Codec Decoder by 80ms. As demonstrated, the RTF consistently remains below 1 across varying concurrency levels, ensuring that users receive continuously streaming audio responses.

3 Pretraining

Table 3: **Languages and dialects support of Qwen3-Omni-30B-A3B.**

Modality	# Langs	Languages
Text	119	See Qwen3 for the full list.
Speech Input	19	ar, de, en, es, fr, id, it, ja, ko, ms, nl, pt, ru, th, tr, ur, vi, yue, zh
Speech Output	10	de, en, es, fr, it, ja, ko, pt, ru, zh

Qwen3-Omni is pre-trained on a diverse dataset that encompasses multiple languages and dialects as shown in Table 3 and modalities, including image-text, video-text, audio-text, video-audio, video-audio-text, and pure text corpora. Unlike Qwen2.5-Omni, which uses a single prompt for each task, we employ a wider range of natural language prompts to enhance both the generalization ability and instruction-following capabilities. To achieve robust performance across all modalities, our training strategy incorporates both unimodal and cross-modal data from the early pretraining stage.

The pre-training of Qwen3-Omni is structured into three distinct stages. In the first stage, we lock the LLM parameters and focus on training the vision and audio encoders, utilizing a vast corpus of audio-text and image-text pairs to enhance semantic understanding within the LLM. In the second stage, we unfreeze all parameters and train with a wider range of multimodal data for more comprehensive learning. In the final stage, we use data with a sequence length of 32,768 to enhance the model's ability to understand complex long-sequence data:

(1) **Encoder Alignment Stage (S1)**: During the initial pretraining phase, the LLM component of Qwen3-Omni is initialized with parameters from Qwen3 (Yang et al., 2025a), while the vision encoder is adopted from Qwen3-VL, and the audio encoder is initialized with AuT. The two encoders are trained separately on the fixed LLM, with both initially focusing on training their respective adapters before training the encoders. We abandon the stage used in Bai et al. (2025); Xu et al. (2025) where the encoder and adapter are trained jointly while keeping the LLM frozen, because this approach may cause the encoder to compensate for the limitations of the frozen LLM, which can lead to degraded perception capabilities.

(2) **General Stage (S2)**: The second phase of pretraining utilizes a large-scale dataset containing approximately 2 trillion tokens, with the following distribution across modalities: text (0.57 trillion), audio (0.77 trillion), image (0.82 trillion), video (0.05 trillion), and video-audio (0.05 trillion). During this stage, the introduction of more diverse multimodal data and tasks enhances the model's understanding and interaction capabilities in auditory, visual, textual, and audiovisual information.

(3) **Long Context Stage (S3)**: In the final pre-training phase, we increased the maximum token length from 8,192 to 32,768 and also raised the proportion of long audio and long video in the training data. Experimental results indicate that these adjustments lead to significant improvements in the model's ability to understand long sequence data.

4 Post-training

4.1 Thinker

The post-training phase comprises a three-stage training process for Thinker, enabling Qwen3-Omni to possess instruction-following capabilities. The dataset, designed in the ChatML (OpenAI, 2022) format, includes pure text-based dialogue data, visual modality conversation data, audio modality conversation data, and mixed-modality conversation data.

In the first stage, we introduce a lightweight Supervised Fine-Tuning (SFT) to bridge the gap between pretrained representations and downstream task requirements through targeted instruction optimization. SFT deliberately diverges from the pretraining data schema while maintaining architectural consistency with the pretrained model, enabling efficient knowledge transfer and preserving the completeness of the pretrained features.

The second stage adopts the Strong-to-Weak Distillation pipeline as described in Qwen3 (Yang et al., 2025a) to further improve model performance. This distillation process consists of two main phases:

(1) **Off-policy Distillation**: In the initial phase, outputs generated by teacher models are combined to provide response distillation. This helps lightweight student models acquire fundamental reasoning abilities, establishing a strong foundation for subsequent on-policy training.

(2) **On-policy Distillation**: In the second phase, the student model generates the responses based on sampled prompts. These on-policy sequences are then used for fine-tuning, where the student's predicted logits are aligned with those of a teacher model (Qwen3-32B or Qwen3-235B-A22B) by minimizing the KL divergence.

Finally, we leverage GSPO (Zheng et al., 2025) to comprehensively enhance the model's capabilities and stability across various modalities, including text, image, video, and audio. To provide feedback for the aforementioned modalities, we employ two different types of rewards:

- **Rule-based Reward**: For verifiable multimodal tasks (e.g., mathematics, coding, instruction following), the reward signal is derived from a set of predefined rules. Well-designed rule-based rewards can assess the correctness of model outputs with high precision, preventing issues like reward hacking.
- **Model-based Reward**: To assess performance on multimodal tasks that lack objective, predefined evaluation metrics, we adopt an LLM-as-a-judge protocol. The role of the automated evaluator is filled by Qwen3 for general tasks, while the specialized vision-language model, Qwen2.5-VL, is used for visually-grounded tasks. To ensure a more robust and grounded assessment, the LLM evaluator is furnished with the corresponding ground-truth or reference answer for a given query, where applicable.

4.2 Talker

We introduce a four-stage training process for Talker, enabling Qwen3-Omni to generate speech response simultaneously with text. All training data is structured in the ChatML format to ensure consistency with Thinker.

In the first stage, we leverage hundreds of millions of speech data with multimodal context to train Talker, establishing a monotonic mapping from multimodal representation to speech. In the second stage, we perform Continual Pretraining (CPT) with high-quality data, which alleviates hallucinations caused by noisy data in the first stage and significantly improve the quality of generated speech. Concurrently, we perform long-context training that enhances Talker's ability to process extended and complex inputs and generate contextually appropriate speech response. In the third stage, to improve the generalization of multilingual speech generation and system stability, we construct preference pairs from diverse multilingual speech samples and optimize the model using Direct Preference Optimization (DPO) (Rafailov et al., 2023). Finally, we apply speaker fine-tuning on the aforementioned base model, enabling Talker to adopt specific voices while refining the naturalness, expressiveness, and controllability of its speech response.

4.3 Captioner

Captioning is a foundational task in multimodal understanding, integral to the training and evaluation of large multimodal models. However, the vast majority of existing research has concentrated on visual captioning, largely neglecting the audio modality. This omission is significant, as auditory perception is a crucial component of human sensory experience and interaction with the world. To address this gap and facilitate more comprehensive research in multimodal perception, we introduce the Qwen3-Omni-30B-A3B-Captioner. This model was developed by fine-tuning the Qwen3-Omni-30B-A3B on a large-scale dataset of detailed audio descriptions. The resulting system generates detailed, low-hallucination captions for arbitrary audio inputs. The **Appendix** 9.2 provides qualitative results that demonstrate our model's captioning capabilities across diverse acoustic scenarios.

5 Evaluation

A comprehensive evaluation was performed on a suite of models, including Qwen3-Omni-30B-A3B-Instruct, Qwen3-Omni-30B-A3B-Thinking, and two in-house developed variants, designated Qwen3-Omni-Flash-Instruct and Qwen3-Omni-Flash-Thinking. These "Flash" models were designed to improve both computational efficiency and performance efficacy, integrating new functionalities, notably the support for various dialects. The evaluation results are divided into two main categories: understanding (X→Text) and speech generation (X→Speech).

5.1 Evaluation of X→Text

In this section, we evaluate Qwen3-Omni's ability to comprehend various multimodal inputs (text, audio, vision, and audiovisual video) and generate textual responses.

Text→Text Our evaluation of Qwen3-Omni on text → text primarily focuses on general tasks, reasoning ability, coding ability, alignment tasks, agent, and multilingual tasks. Specifically, we utilize MMLU-Redux (Gema et al., 2024) and GPQA (Rein et al., 2023) for general tasks, AIME25 (AIME, 2025) and ZebraLogic (Lin et al., 2025) for reasoning evaluation, MultiPL-E (Cassano et al., 2023) for coding, IFEval (Zhou et al., 2023), Creative Writing V3 (Paech, 2024) and WritingBench (Wu et al., 2025b) for alignment tasks, BFCL-v3 (Yan et al., 2024) for agent evaluation, MultiIF (He et al., 2024) and PolyMath (Wang et al., 2025c) for multilingual tasks.

Audio→Text The evaluation can be categorized into basic audio tasks, including Automatic Speech Recognition (ASR), Speech-to-Text (S2TT), and Music Understanding, as well as advanced audio tasks, including Voice Chatting and Audio Reasoning. For music understanding, we use RUL-MuchoMusic (Zang et al., 2025) for a comprehensive evaluation of the music understanding capabilities of the model. We utilize MMAU (Sakshi et al., 2024) and MMSU (Wang et al., 2025a) for audio reasoning tasks, VoiceBench (Chen et al., 2024b) for voice-chatting tasks. We also employ multiple datasets including GTZAN (Tzanetakis & Cook, 2002), four subsets of MTG-Jamendo (MTG, Bogdanov et al. (2019)), and MagnaTagATune (Law et al., 2009) to evaluate the model's capabilities across various music information retrieval tasks including genre identification, emotion and theme recognition, instrument recognition and music keyword annotation. We follow the evaluation set composition in MARBLE (Yuan et al., 2023) for GTZAN, MTG-Jamendo and MagnaTagATune.

Vision→Text The evaluation of the model's vision-to-text capabilities encompasses a suite of benchmarks targeting diverse and challenging tasks. To assess performance in general visual question answering, the model is evaluated on MMStar (Chen et al., 2024a), HallusionBench (Guan et al., 2024), and MM-MT-Bench (Agrawal et al., 2024). For the specialized domain of mathematical and STEM reasoning, we utilize MathVista (Lu et al., 2024), MathVision (Wang et al., 2024a), MMMU (Yue et al., 2023), and MMMU-Pro (Yue et al., 2024). The model's proficiency in document understanding is measured using the AI2D (Kembhavi et al., 2016) and ChartQA (Masry et al., 2022) benchmarks. Furthermore, the model's numerical reasoning and counting abilities are specifically tested on CountBench (Paiss et al., 2023). To evaluate performance on dynamic visual data, we report results on three long video understanding benchmarks: Video-MME (Fu et al., 2024), LVBench (Wang et al., 2024b), and MLVU (Zhou et al., 2025a).

AudioVisual Video→Text To evaluate the model's ability to process dynamic multi-modal information, we first assessed its performance on the WorldSense benchmark (Hong et al., 2025). This benchmark is designed to measure the integration of visual and auditory signals, a foundational capability for operating in complex, open-world environments. To further examine the model's higher-order cognitive functions, we then evaluated its performance on two audiovisual reasoning benchmarks: DailyOmni (Zhou et al., 2025b) and VideoHolmes (Cheng et al., 2025).

5.1.1 Performance of Text→Text

We compare Qwen3-Omni with other leading large language models (thinking or instruct). According to Table 4 and 5, notably, despite a smaller parameter count, Qwen3-Omni-30B-A3B-Instruct surpasses the performance of the larger open-source model Qwen3-235B-A22B Non-Thinking and the formidable closed-source model GPT-4o-0327 across a suite of benchmarks, including GPQA, AIME25, ZebraLogic, WritingBench, and PolyMath. Concurrently, Qwen3-Omni-30B-A3B-Thinking demonstrates performance comparable to that of Gemini-2.5-Flash-Thinking and Qwen3-235B-A22B Non-Thinking. Furthermore, Qwen3-Omni-30B-A3B exhibits textual capabilities on par with its text-only counterparts, namely the Qwen3-30B-A3B-Instruct-2507 and Qwen3-30B-A3B-Thinking-2507.

Table 4: **Text → Text performance of Qwen3-Omni-Instruct and other non-reasoning baselines. The highest scores are shown in bold.**

		GPT-4o-0327	Qwen3-235B-A22B	Qwen3-30B-A3B	Qwen3-Omni-30B-A3B	Qwen3-Omni-Flash
			Non Thinking	-Instruct-2507	-Instruct	-Instruct
GeneralTasks	MMLU-Redux	**91.3**	89.2	89.3	86.6	86.8
	GPQA	66.9	62.9	**70.4**	69.6	69.7
Reasoning	AIME25	26.7	24.7	61.3	65.0	**65.9**
	ZebraLogic	52.6	37.7	**90.0**	76.0	76.1
Code	MultiPL-E	82.7	79.3	**83.8**	81.4	81.5
Alignment Tasks	IFEval	83.9	83.2	**84.7**	81.0	81.7
	Creative Writing v3	84.9	80.4	**86.0**	80.6	81.8
	WritingBench	75.5	77.0	**85.5**	82.6	83.0
Agent	BFCL-v3	66.5	**68.0**	65.1	64.4	65.0
Multilingual Tasks	MultiIF	**70.4**	70.2	67.9	64.0	64.7
	PolyMATH	25.5	27.0	**43.1**	37.9	39.3

Table 5: **Text → Text performance of Qwen3-Omni-Thinking and other reasoning baselines. The highest scores are shown in bold.**

		Gemini-2.5-Flash	Qwen3-235B-A22B	Qwen3-30B-A3B	Qwen3-Omni-30B-A3B	Qwen3-Omni-Flash
		Thinking	Thinking	-Thinking-2507	-Thinking	-Thinking
General Tasks	MMLU-Redux	92.1	**92.7**	91.4	88.8	89.7
	GPQA	**82.8**	71.1	73.4	73.1	73.1
Reasoning	AIME25	72.0	81.5	**85.0**	73.7	74.0
	LiveBench 20241125	74.3	**77.1**	76.8	71.8	70.3
Code	MultiPL-E	**84.5**	79.9	81.3	80.6	81.0
Alignment Tasks	IFEval	**89.8**	83.4	88.9	85.1	85.2
	Arena-Hard v2	56.7	**61.5**	56.0	55.1	57.8
	Creative Writing v3	**85.0**	84.6	84.4	82.5	83.6
	WritingBench	83.9	80.3	85.0	85.5	**85.9**
Agent	BFCL-v3	68.6	70.8	**72.4**	63.2	64.5
Multilingual Tasks	MultiIF	74.4	71.9	**76.4**	72.9	73.2
	PolyMATH	49.8	**54.7**	52.6	47.1	48.7

5.1.2 Performance of Audio→Text

We compare Qwen3-Omni with other leading specialist and generalist models on ASR & S2TT, voice-chatting, audio reasoning, and music understanding benchmarks. For brevity, we defer the results of the Qwen3-Omni-Thinking model on ASR & S2TT and music understanding to the **Appendix** 9.1.

Table 6: **Transcription performance for Audio→Text tasks (ASR & S2TT), comparing Qwen3-Omni-Instruct with the baselines. The highest scores are shown in bold.**

	Seed	Voxtral	Voxtral	GPT-4o	Gemini-2.5	Qwen2.5	Qwen3-Omni	Qwen3-Omni
	-ASR	-Mini	-Small	-Transcribe	-Pro	-Omni	-30B-A3B-Instruct	-Flash-Instruct
				EN & ZH ASR (wer)				
Wenetspeech *net* | *meeting*	4.66 | **5.69**	24.30 | 31.53	20.33 | 26.08	15.30 | 32.27	14.43 | 13.47	5.91 | 7.65	4.69 | 5.89	**4.62** | 5.75
Librispeech *clean* | *other*	1.58 | 2.84	1.88 | 4.12	1.56 | 3.30	1.39 | 3.75	2.89 | 3.56	1.74 | 3.45	**1.22** | 2.48	1.27 | **2.44**
CV15-en	-	9.47	7.79	10.01	9.89	7.61	6.05	**5.94**
CV15-zh	-	24.67	19.30	9.84	8.00	5.13	4.31	**4.28**
Fleurs-en	3.40	3.96	3.77	3.32	2.94	3.77	**2.72**	2.74
Fleurs-zh	2.69	12.22	7.98	2.44	2.71	2.54	2.20	**2.19**
				Multilingual ASR (wer)				
Fleurs-avg (19 lang)[a]	-	15.67	8.09	4.48	5.55	14.04	5.33	**5.31**
				Lyric ASR (wer)				
MIR-1K (vocal-only)[b]	6.45	23.33	18.73	11.87	9.85	8.15	5.90	**5.85**
Opencpop-test	2.98	31.01	16.06	7.93	6.49	2.84	**1.54**	2.02
				S2TT (BLEU)				
Fleurs-en2xx[c]	-	30.35	37.85	-	**39.25**	29.22	37.50	36.22
Fleurs-xx2en	-	27.54	32.81	-	**35.41**	28.61	31.08	30.71
Fleurs-zh2xx	-	17.03	22.05	-	**26.63**	17.97	25.17	25.10
Fleurs-xx2zh	-	28.75	34.82	-	**37.50**	27.68	33.13	31.19

[a] These 19 languages include Arabic, Cantonese, Chinese, Dutch, English, French, German, Indonesian, Italian, Japanese, Korean, Malay, Portuguese, Russian, Spanish, Thai, Turkish, Urdu, Vietnamese.

[b] Transcription is converted into Simplified Chinese.

[c] The results encompass translations across 15 languages: Arabic, Cantonese, Chinese, English, French, German, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Thai, Vietnamese. For notation, "en2xx" denotes translation from English into each of the other 14 target languages, where "xx" ranges over the remaining language codes.

As shown in Table 6, Qwen3-Omni-Instruct achieves state-of-the-art En & Zh ASR and lyric ASR performance on Librispeech, Wenetspeech, Fleurs, CommonVoice, Opencpop-test and MIR-1K (vocal). It also delivers better or comparable performance with other specialist or generalist models like Voxtral-Small and Gemini-2.5-Pro on Multilingual ASR and S2TT. These results show a strong performance of Qwen3-Omni in speech recognition and speech translation.

Additionally, on VoiceBench shown in Table 7, Qwen3-Omni-Thinking achieves an impressive average score of 89.5, surpassing all other audio language models except Gemini-2.5-Pro (89.6). This showcases our model's strong capabilities in speech interaction. Qwen3-Omni also demonstrates impressive performance in audio reasoning, outperforming the powerful closed-source models Gemini-2.5-Pro and Gemini-2.5-Flash on the MMAU benchmark, as well as Gemini-2.5-Flash and GPT-4o-Audio on MMSU. These results demonstrate the powerful capabilities of Qwen3-Omni in general audio understanding and reasoning.

Table 7: **Voice interaction and audio reasoning performance for Audio→Text tasks, comparing Qwen3-Omni with the baselines. The highest scores are shown in bold.**

	GPT-4o -Audio	**Gemini-2.5 -Flash**	**Gemini-2.5 -Pro**	**Qwen2.5 -Omni**	**Qwen3-Omni -30B-A3B-Instruct**	**Qwen3-Omni -30B-A3B-Thinking**	**Qwen3-Omni -Flash-Instruct**	**Qwen3-Omni -Flash-Thinking**
				VoiceBench				
AlpacaEval	95.6	96.1	94.3	89.9	94.8	96.4	95.4	**96.8**
CommonEval	89.8	88.3	88.4	76.7	90.8	90.5	**91.0**	90.9
WildVoice	91.6	92.1	93.4	77.7	91.6	90.5	**92.3**	90.9
SD-QA	75.5	84.5	**90.1**	56.4	76.9	78.1	76.8	78.5
MMSU	80.3	66.1	71.1	61.7	68.1	83.0	68.4	**84.3**
OpenBookQA	89.2	56.9	92.3	80.9	89.7	94.3	91.4	**95.0**
BBH	84.1	83.9	**92.6**	66.7	80.4	88.9	80.6	89.6
IFEval	76.0	83.8	**85.7**	53.5	77.8	80.6	75.2	80.8
AdvBench	98.7	98.9	98.1	99.2	**99.3**	97.2	**99.4**	98.9
Overall	86.8	83.4	**89.6**	73.6	85.5	88.8	85.6	89.5
				Audio Reasoning				
MMAU-v05.15.25	62.5	71.8	77.4	65.5	77.5	75.4	**77.6**	76.5
MMSU	56.4	70.2	**77.7**	62.6	69.0	70.2	69.1	71.3

For music understanding, we compare Qwen3-Omni-Instruct with both generalist audio language models and specialist models in Table 8. For multi-label classification tasks on MTG-Jamendo and MagnaTagATune, we use micro F1 to compare with BERT-like music specialists instead of AP/AUROC, as language models output discrete label sets without calibrated per-label probabilities/scores required by ranking-based metrics. It is shown in Table 8 that Qwen3-Omni-Instruct achieve state-of-the-art performance on RUL-MuchoMusic. On GTZAN, MTG-Jamendo, and MagnaTagATune, the scores of Qwen3-Omni-Instruct also significantly surpass other audio language models, including Gemini-2.5-Pro and GPT-4o-Audio, as well as self-supervised music specialist models probed on the respective datasets. These results demonstrate the superior capabilities of Qwen3-Omni-Instruct across a variety of music understanding tasks.

Table 8: **Music understanding performance for Audio→Text tasks, comparing Qwen3-Omni-Instruct with baselines. The highest scores are shown in bold.**

	Best Specialist Models	**GPT-4o -Audio**	**Gemini-2.5 -Pro**	**Qwen2.5 -Omni**	**Qwen3-Omni -30B-A3B-Instruct**	**Qwen3-Omni -Flash-Instruct**
RUL-MuchoMusic	47.6 (Audio Flamingo 3) (Goel et al., 2025)	36.1	49.4	47.3	52.0	**52.1**
GTZAN *Acc.*	87.9 (CLaMP 3) (Wu et al., 2025a)	76.5	81.0	81.7	93.0	**93.1**
MTG Genre *Micro F1*	35.8 (MuQ-MuLan) (Zhu et al., 2025)	25.3	32.6	32.5	39.0	**39.5**
MTG Mood/Theme *Micro F1*	10.9 (MuQ-MuLan) (Zhu et al., 2025)	11.3	14.1	8.9	21.0	**21.7**
MTG Instrument *Micro F1*	39.8 (MuQ-MuLan) (Zhu et al., 2025)	34.2	33.0	22.6	40.5	**40.7**
MTG Top50 *Micro F1*	33.2 (MuQ-MuLan) (Zhu et al., 2025)	25.0	26.1	21.6	36.7	**36.9**
MagnaTagATune *Micro F1*	41.6 (MuQ) (Zhu et al., 2025)	29.2	28.1	30.1	44.3	**46.8**

5.1.3 Performance of Vision → Text

To comprehensively evaluate the capabilities on Vision → Text, we compare Qwen3-Omni-Instruct with the Qwen2.5-VL-72B and other good-performing closed-source vision-language models. As illustrated in

Table 9, Qwen3-Omni-Instruct demonstrates comparable performance to Qwen2.5-VL-72B, and attains better results on Math & STEM related tasks like MMMU-Pro$_{\text{overall}}$, MathVista$_{\text{mini}}$, and MATH-Vision$_{\text{full}}$, than other vision language models including GPT4-o and Gemini-2.0-Flash. These results reveal the excellent capability of our model on image understanding and reasoning tasks.

To assess its capabilities, we evaluated the performance of Qwen3-Omni-Thinking against several state-of-the-art reasoning models. The comparative results, summarized in Table 10, indicate that our proposed model achieves significant advancements. For instance, on Math and STEM benchmarks, it outperforms the Qwen3-Omni-Instruct baseline by 4.4 points. It is also noteworthy that our Qwen3-Omni-30B-A3B-Thinking model attains a performance level on par with substantially larger baselines, which highlights its excellent balance of effectiveness and computational efficiency. A limitation of the current model is its suboptimal performance on long video benchmarks. This deficiency stems from two architectural constraints: a limited capacity for positional extrapolation and a restricted context length. Addressing these constraints is a key objective for future work.

Table 9: **Vision → Text performance of Qwen3-Omni-Instruct and other non-reasoning baselines. The highest scores are shown in bold.**

Datasets	**GPT4-o**	**Gemini-2.0-Flash**	**Qwen2.5-VL 72B**	**Qwen3-Omni-30B-A3B -Instruct**	**Qwen3-Omni-Flash -Instruct**
General Visual Question Answering					
MMStar	64.7	**71.4**	70.8	68.5	69.3
HallusionBench	55.0	56.3	55.2	**59.7**	58.5
MM-MT-Bench	**7.7**	6.7	7.6	7.4	7.6
Math & STEM					
MMMU$_{\text{val}}$	69.1	**71.3**	70.2	69.1	69.8
MMMU-Pro$_{\text{overall}}$	51.9	56.1	51.1	57.0	**57.6**
MathVista$_{\text{mini}}$	63.8	71.4	74.8	75.9	**77.4**
MATH-Vision$_{\text{full}}$	30.4	48.6	38.1	56.3	**58.3**
Documentation Understanding					
AI2D$_{\text{w.M.}}$	84.6	86.7	**88.7**	85.2	86.4
ChartQA$_{\text{test Avg.}}$	86.7	64.6	**89.5**	86.8	87.1
Counting					
CountBench	87.9	91.2	**93.6**	90.0	90.0
Video Understanding					
Video-MME$_{\text{w/o sub}}$	71.9	72.4	**73.3**	70.5	71.4
LVBench	30.8	**57.9**	47.3	50.2	51.1
MLVU	64.6	71.0	74.6	75.2	**75.5**

Table 10: **Vision → Text performance of Qwen3-Omni-Thinking and other reasoning baselines. The highest scores are shown in bold.**

Datasets	**Gemini-2.5-Flash -Thinking**	**InternVL-3.5-241B-A28B**	**Qwen3-Omni-30B-A3B -Thinking**	**Qwen3-Omni-Flash -Thinking**
General Visual Question Answering				
MMStar	75.5	**77.9**	74.9	75.5
HallusionBench	61.1	57.3	62.8	**63.4**
MM-MT-Bench	7.8	–	**8.0**	**8.0**
Math & STEM				
MMMU$_{\text{val}}$	76.9	**77.7**	75.6	75.0
MMMU-Pro$_{\text{overall}}$	**65.8**	–	60.5	60.8
MathVista$_{\text{mini}}$	77.6	**82.7**	80.0	81.2
MATH-Vision$_{\text{full}}$	62.3	**63.9**	62.9	63.8
Documentation Understanding				
AI2D$_{\text{w.M.}}$	**88.6**	87.3	86.1	86.8
ChartQA$_{\text{test Avg.}}$	–	88.0	**89.5**	89.3
Counting				
CountBench	88.6	–	88.6	**92.5**
Video Understanding				
Video-MME$_{\text{w/o sub}}$	**79.6**	72.9	69.7	69.8
LVBench	**64.5**	–	49.0	49.5
MLVU	**82.1**	78.2	72.9	73.9

5.1.4 Performance of AudioVisual Video→Text

As is shown in Table 11, the experimental results validate the efficacy of Qwen3-Omni across diverse audiovisual tasks. For general understanding, Qwen3-Omni-Instruct achieves state-of-the-art performance on the WorldSense benchmark, surpassing other Omni models by a substantial margin. This outcome demonstrates its effectiveness in foundational multimodal integration. Moreover, the model exhibits enhanced performance on complex reasoning tasks, as illustrated in Table 12, particularly on benchmarks that necessitate reasoning over interconnected audio and visual information. These findings collectively suggest that Qwen3-Omni possesses considerable potential for advanced perception and reasoning in real-world contexts.

Table 11: **AudioVisual → Text performance of Qwen3-Omni-Instruct and other non-reasoning baselines. The highest scores are shown in bold.**

Datasets	Previous Open-source SoTA	Gemini-2.5-Flash	Qwen2.5-Omni	Qwen3-Omni-30B-A3B -Instruct	Qwen3-Omni-Flash -Instruct
WorldSense	47.1(Yang et al., 2025b)	50.9	45.4	54.0	**54.1**

Table 12: **AudioVisual → Text performance of Qwen3-Omni-30B-A3B-Thinking and other reasoning baselines. The highest scores are shown in bold.**

Datasets	Previous Open-source SoTA	Gemini-2.5-Flash -Thinking	Qwen3-Omni-30B-A3B -Thinking	Qwen3-Omni-Flash -Thinking
DailyOmni	69.8(Tang et al., 2025)	72.7	75.8	**76.2**
VideoHolmes	55.6(Tang et al., 2025)	49.5	**57.3**	**57.3**

5.2 Evaluation of X→Speech

In this section, we evaluate the speech generation capabilities of Qwen3-Omni. Due to the lack of relevant assessments, the evaluation of speech generation focuses primarily speech generation given texts, similarity to text-to-speech (TTS), on following three aspects:

- **Zero-Shot Speech Generation**: We assess the content consistency (WER) and speaker similarity (SIM) of our model in zero-shot speech generation on SEED (Anastassiou et al., 2024).
- **Multilingual Speech Generation**: We assess the content consistency and speaker similarity of our model in zero-shot multilingual speech generation on MiniMax multilingual test set (Zhang et al., 2025).
- **Cross-Lingual Speech Generation**: We assess the content consistency of our model in zero-shot cross-lingual speech generation on CV3-Eval (Du et al., 2025).

5.2.1 Evaluation of Zero-Shot Speech Generation

We compare the Qwen3-Omni with state-of-the-art zero-shot TTS systems. As shown in Table 13, Qwen3-Omni demonstrates highly competitive performance, highlighting its robust speech understanding and generation capabilities developed through pretraining and continual pretraining. Additionally, with reinforcement learning (RL) optimization, Qwen3-Omni yields significant improvements in generation stability, which achieves the best performance in the test-en set.

Table 13: **Zero-Shot Speech Generation on Seed-TTS Test Set. The highest scores are shown in bold.**

Datasets	Model	Performance
	Content Consistency	
SEED *test-zh* | *test-en*	Seed-TTS$_{ICL}$ (Anastassiou et al., 2024)	1.11 | 2.24
	Seed-TTS$_{RL}$ (Anastassiou et al., 2024)	1.00 | 1.94
	MaskGCT (Wang et al., 2024c)	2.27 | 2.62
	E2 TTS (Eskimez et al., 2024)	1.97 | 2.19
	F5-TTS (Chen et al., 2024c)	1.56 | 1.83
	Spark TTS (Wang et al., 2025b)	1.20 | 1.98
	CosyVoice 2 (Du et al., 2024)	1.45 | 2.57
	CosyVoice 3 (Du et al., 2025)	**0.71** | 1.45
	Qwen2.5-Omni-7B (Xu et al., 2025)	1.42 | 2.33
	Qwen3-Omni-30B-A3B	1.07 | **1.39**

5.2.2 Evaluation of Multilingual Speech Generation

Qwen3-Omni supports speech generation across 10 languages. We evaluate its performance against both the MiniMax-Speech and ElevenLabs Multilingual v2 models for multilingual speech generation. As shown in Table 14, Qwen3-Omni surpasses these models by a significant margin for languages such as Chinese, English, and French, while delivering competitive results in the remaining languages. These findings indicate that Qwen3-Omni generates cloned speech with consistent stability and human-like voice across all evaluated languages.

Table 14: **Multilingual Speech Generation on MiniMax Multilingual Test Set. The highest scores are shown in bold.**

Language	Content Consistency			Speaker Similarity		
	Qwen3-Omni -30B-A3B	MiniMax	ElevenLabs	Qwen3-Omni -30B-A3B	MiniMax	ElevenLabs
Chinese	**0.716**	2.252	16.026	0.772	**0.780**	0.677
English	**1.069**	2.164	2.339	**0.773**	0.756	0.613
German	0.777	1.906	**0.572**	**0.738**	0.733	0.614
Italian	**1.067**	1.543	1.743	**0.742**	0.699	0.579
Portuguese	1.872	1.877	**1.331**	0.770	**0.805**	0.711
Spanish	1.765	**1.029**	1.084	0.744	**0.762**	0.615
Japanese	3.631	**3.519**	10.646	0.763	**0.776**	0.738
Korean	**1.670**	1.747	1.865	**0.778**	0.776	0.700
French	**2.505**	4.099	5.216	**0.689**	0.628	0.535
Russian	3.986	4.281	**3.878**	0.759	**0.761**	0.676

5.2.3 Evaluation of Cross-Lingual Speech Generation

Qwen3-Omni supports not only multilingual voice cloning but also cross-lingual voice cloning. We evaluate its performance against CosyVoice2 and CosyVoice3 for cross-lingual speech generation. As shown in Table 15, Qwen3-Omni outperforms CosyVoice3 in any-to-en (any language to English) and any-to-ko (any language to Korean) voice cloning. Notably, in any-to-ja (any language to Japanese) tasks, Qwen3-Omni achieves comparable performance to CosyVoice3 even without text normalization, despite CosyVoice3 converting all Japanese characters into phonetic kana. These results highlight Qwen3-Omni's superiority in cross-lingual speech generation, demonstrating its adaptability across diverse linguistic contexts.

Table 15: **Cross-Lingual Speech Generation on CosyVoice3 Cross-Lingual Test Set. The highest scores are shown in bold.**

Language	Qwen3-Omni-30B-A3B	CosyVoice3	CosyVoice2
en-to-zh	5.37	**5.09**	13.5
ja-to-zh	3.32	**3.05**	48.1
ko-to-zh	**0.99**	1.06	7.70
zh-to-en	**2.76**	2.98	6.47
ja-to-en	**3.31**	4.20	17.1
ko-to-en	**3.34**	4.19	11.2
zh-to-ja	8.29	**7.08**	13.1
en-to-ja	7.53	**6.80**	14.9
ko-to-ja	4.24	**3.93**	5.86
zh-to-ko	**5.13**	14.4	24.8
en-to-ko	**4.96**	5.87	21.9
ja-to-ko	**6.23**	7.92	21.5

6 Evaluating Non-Degradation Across Modalities

A standardized data integration methodology is rendered impractical by the heterogeneous nature of different modalities, each requiring distinct pre-training objectives and optimization techniques. To ensure a fair and rigorous evaluation, we therefore designed a controlled comparative study. Our approach involved pre-training three models with matched parameter counts: a text-only baseline, a vision-only baseline, and a multimodal “Omni” model. To isolate the effects of multimodality, all confounding variables were meticulously controlled. Specifically, the Omni model was trained on the identical text and vision corpora as the unimodal baselines. Moreover, we aligned critical training parameters across all models, including learning rate schedules, batch sizes, and the effective number of training epochs for each modality, which was normalized by adjusting data sampling ratios. Consequently, the sole differentiating factor in our experiment was the Omni model’s inclusion of supplementary audio and audio-visual data during its pre-training phase.

The results are shown in Table 16, we evaluate comprehensive benchmarks covering a variety of modalities, including the text modality (general tasks, math & STEM tasks, coding tasks, multilingual tasks), the visual modality (college-level problems, OCR-related tasks), and the video modality (video understanding tasks). The experimental results not only demonstrate that mixing unimodal and cross-modal data during the early stage of text pretraining can achieve better performance across all modalities, but also indicate that joint multimodal training enables mutual enhancement between different modalities, leading to improved performance in single modalities as well. This fully showcases the versatility and robustness of Qwen3-Omni across diverse evaluation criteria.

Due to the prohibitive experimental cost, we could not conduct a comprehensive sweep across all model scales. Based on Table 16 and our internal experiments, we observe: (1) early multimodal integration during pretraining allows language models to be co-trained with vision or audio without any degradation in language capability; (2) the inclusion of the text modality substantially improves performance in the vision and audio. In constrast, we do not observe measurable gains in language ability from adding visual or audio signals; (3) empirically, adding audio data consistently improves vision performance on the MMMU benchmark and OCR-related tasks

Table 16: **We compare the performance of 30A3 models that are contemporaneous and identical in size in Qwen series. To ensure experimental rigor, all models were trained under the same schedule, using identical datasets for their respective modalities and exactly matched training compute (FLOPs).**

	Datasets	Qwen3-30B-A3B -Base-202507	Qwen3-VL-30B-A3B -Base-202507	Qwen3-Omni-30B-A3B -Base-202507
General Tasks	MMLU	81.24	-	**81.69**
	MMLU-Redux	80.17	-	**80.60**
	MMLU-Pro	**61.81**	-	61.57
	SuperGPQA	38.24	-	**40.14**
	BBH	**83.79**	-	83.53
Math & STEAM Tasks	GSM8K	90.83	-	**91.36**
	MATH	**60.84**	-	60.42
Coding Tasks	EvalPlus	69.70	-	**73.96**
	MultiPL-E	**65.75**	-	64.79
	MBPP	72.60	-	**72.60**
	CRUX-O	66.94	-	**69.06**
Multilingual Tasks	MGSM	78.75	-	**79.93**
	INCLUDE	**65.17**	-	64.73
College-level Problems	MMMU$_{\text{val}}$	-	57.22	**59.33**
General Visual Question Answering	MMStar	-	67.2	**69.6**
	RealWorldQA$_{\text{avg}}$	-	**73.98**	71.89
OCR-related Tasks	AI2D	-	85.88	**86.62**
	TextVQA$_{\text{val}}$	-	**81.67**	81.65
	DocVQA$_{\text{test}}$	-	95.19	**95.27**
	InfoVQA$_{\text{test}}$	-	81.17	**83.31**
	ChartQA$_{\text{test Avg}}$	-	87.12	**87.52**
	OCRBench	-	85.8	**86.0**
Video Understanding Tasks	Video-MME$_{\text{w/o sub}}$	-	69.22	**69.25**
	MVBench	-	**71.87**	69.50
	LVBench	-	48.61	**51.07**

7 Conclusion

In this paper, we introduce Qwen3-Omni-30B-A3B, Qwen3-Omni-30B-A3B-Thinking, Qwen3-Omni-Flash-Instruct, and Qwen3-Omni-Flash-Thinking models. Qwen3-Omni-30B-A3B matches or surpasses the latest same-size unimodal Qwen models on text and vision benchmarks. Notably, on audio processing and dialogue benchmarks, it attains state-of-the-art performance among open-source systems on 32 benchmarks and is comparable to, or better than, the strong proprietary counterpart Gemini-2.5-Pro. The Qwen3-Omni-30B-A3B Thinking variant achieves further gains on complex tasks spanning text, vision, and audio-visual reasoning. Beyond accuracy, the model supports 119 text languages, 19 languages for speech recognition and 10 languages for speech synthesis, and enables audio understanding and interactive sessions up to 40 minutes. Thanks to its streaming architecture and multi-codebook design, Qwen3-Omni at the 30B-A3B scale still delivers an end-to-end first-packet latency of 234 ms.

Research fields often cycle between specialization and integration. In this context, we believe Qwen3-Omni represents a milestone: to our knowledge, it provides the first evidence that fully integrated, end-to-end multimodal training can be achieved without degrading core language capability and other modalities. We are eager to share these findings with the community and hope they will stimulate further research.

For practical usage, Qwen3-Omni-30B-A3B offers strong text and vision capabilities, robust and reliable ASR, interactive speech support in over 20 languages, very low first-packet latency for interactive use, and stable, naturalistic speech synthesis. Crucially, it exhibits advantages over cascaded pipelines, including stronger cross-modal reasoning, lower end-to-end latency, and lower system complexity and cost. In future work, we will further advance the model along multiple axes, including multi-speaker ASR, video OCR, audiovisual proactive learning, and enhanced support for agent-based workflows and function calling.

8 Authors

Core Contributors: 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, Junyang Lin

Contributors[1]: An Yang, Anfeng Li, Bei Chen, Beichen Zhang, Bin Lin, Binyuan Hui, Bohan Wang, Buxiao Wu, Chenfei Wu, Cheng Chen, Chen Qiang, Chenhan Yuan, Chenhao Li, Chenxu Lv, Chujie Zheng, Daren Chen, Dayiheng Liu, Dake Guo, Fei Huang, Gezhengyang Zhu, Guangdong Zhou, Hang Zhang, Hongjian Tu, Humen Zhong, Jialong Zuo, Jianhong Tu, Jianwei Zhang, Jiayi Leng, Jing Zhou, Jingren Zhou, Kai Dang, Kexin Yang, Kun Yan, Laiwen Zheng, Lei Xie, Lianghao Deng, Lingchen Meng, Mei Li, Miao Hong, Mingfeng Xue, Minsheng Li, Mingze Li, Peiyang Zhang, Peng Liu, Pengfei Wang, Ruibin Yuan, Rui Hu, Ruiyang Xu, Qidong Huang, Qin Zhu, Que Shen, Shen Li, Shixuan Liu, Sibo Song, Siqi Zhang, Song Chen, Su Hao, Tianyi Tang, Wenbin Ge, Wentao Yao, Wei Ding, Wei Wang, Xiaodong Deng, Xiaotong Chen, Xiao Li, Xian Yang, Xinyao Niu, Xudong Guo, Xin Le, Xuechun Wang, Xutong Jin, Xuancheng Ren, Yang Fan, Yang Liu, Yang Su, Yantao Liu, Yi Wu, Yichang Zhang, Yilei Chen, Yiming Dong, Yinger Zhang, Yizhong Cao, Yuchong Sun, Yuezhang Wang, Yuhao Wang, Yuqiong Liu, Yuanzhi Zhu, Yuxiang Chen, Yuxuan Cai, Yuxuan Liu, Zeyu Cui, Zheng Li, Zhenghao Xing, Zhenru Zhang, Zihan Qiu, ZiYue Jiang, Zhaohai Li, Zhi Li, Zhibo Yang, Zhihai Wang, Zhipeng Zhou

[1] Alphabetical order.

9 Appendix

9.1 More Evaluation on Speech and Music Understanding

This section reports the performance of the Qwen3-Omni-thinking model on tasks pertaining to ASR/S2TT and Music. As shown in Table 17 and 18, in the domains of ASR/S2TT and Music understanding, the Qwen3-Omni-Thinking model is outperformed by its Instruct counterpart, which indicates that for these predominantly perception-based tasks, the engagement of sophisticated reasoning processes fails to yield performance gains. In fact, it may even introduce a higher propensity for hallucinations.

Table 17: **Transcription performance for Audio→Text tasks (ASR & S2TT), comparing Qwen3-Omni-Thinking with the baselines. The highest scores are shown in bold.**

	Seed	**Voxtral**	**Voxtral**	**GPT-4o**	**Gemini-2.5**	**Qwen2.5**	**Qwen3-Omni**	**Qwen3-Omni**
	-ASR	**-Mini**	**-Small**	**-Transcribe**	**-Pro**	**-Omni**	**-30B-A3B-Thinking**	**-Flash-Thinking**
EN & ZH ASR (wer)								
Wenetspeech *net* | *meeting*	4.66 | 5.69	24.30 | 31.53	20.33 | 26.08	15.30 | 32.27	14.43 | 13.47	5.91 | 7.65	6.16 | 8.17	6.85 | 8.42
Librispeech *clean* | *other*	1.58 | 2.84	1.88 | 4.12	1.56 | 3.30	1.39 | 3.75	2.89 | 3.56	1.74 | 3.45	2.22 | 4.38	1.82 | 4.01
CV15-en	-	9.47	7.79	10.01	9.89	7.61	10.44	10.52
CV15-zh	-	24.67	19.30	9.84	8.00	5.13	6.25	6.61
Fleurs-en	3.40	3.96	3.77	3.32	2.94	3.77	3.75	3.67
Fleurs-zh	2.69	12.22	7.98	2.44	2.71	2.54	2.73	2.57
Multilingual ASR (wer)								
Fleurs-avg (19 lang)[a]	-	15.67	8.09	4.48	5.55	14.04	8.63	8.88
Lyric ASR (wer)								
MIR-1K (vocal-only)[b]	6.45	23.33	18.73	11.87	9.85	8.15	11.15	10.47
Opencpop-test	2.98	31.01	16.06	7.93	6.49	2.84	6.11	4.52
S2TT (BLEU)								
Fleurs-en2xx[c]	-	30.35	37.85	-	**39.25**	29.22	36.24	36.04
Fleurs-xx2en	-	27.54	32.81	-	**35.41**	28.61	30.50	30.22
Fleurs-zh2xx	-	17.03	22.05	-	**26.63**	17.97	23.74	23.77
Fleurs-xx2zh	-	28.75	34.82	-	**37.50**	27.68	34.51	34.49

[a] These 19 languages include Arabic, Cantonese, Chinese, Dutch, English, French, German, Indonesian, Italian, Japanese, Korean, Malay, Portuguese, Russian, Spanish, Thai, Turkish, Urdu, Vietnamese.

[b] Transcription is converted into Simplified Chinese.

[c] The results encompass translations across 15 languages: Arabic, Cantonese, Chinese, English, French, German, Indonesian, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Thai, Vietnamese. For notation, "en2xx" denotes translation from English into each of the other 14 target languages, where "xx" ranges over the remaining language codes.

Table 18: **Music understanding performance for Audio→Text tasks, comparing Qwen3-Omni-Thinking with baselines. The highest scores are shown in bold.**

	Best Specialist	**GPT-4o**	**Gemini-2.5**	**Qwen2.5**	**Qwen3-Omni**	**Qwen3-Omni**
	Models	**-Audio**	**-Pro**	**-Omni**	**-30B-A3B-Thinking**	**-Flash-Thinking**
RUL-MuchoMusic	47.6 (Audio Flamingo 3) (Goel et al., 2025)	36.1	**49.4**	47.3	48.3	48.4
GTZAN *Acc.*	87.9 (CLaMP 3) (Wu et al., 2025a)	76.5	81.0	81.7	**89.0**	**89.0**
MTG Genre *Micro F1*	**35.8** (MuQ-MuLan) (Zhu et al., 2025)	25.3	32.6	32.5	32.5	33.0
MTG Mood/Theme *Micro F1*	10.9 (MuQ-MuLan) (Zhu et al., 2025)	11.3	14.1	8.9	14.3	**15.4**
MTG Instrument *Micro F1*	**39.8** (MuQ-MuLan) (Zhu et al., 2025)	34.2	33.0	22.6	36.4	36.4
MTG Top50 *Micro F1*	**33.2** (MuQ-MuLan) (Zhu et al., 2025)	25.0	26.1	21.6	29.1	29.3
MagnaTagATune *Micro F1*	**41.6** (MuQ) (Zhu et al., 2025)	29.2	28.1	30.1	32.2	32.6

9.2 Qualitative Results from Qwen3-Omni-30B-A3B-Captioner

In this section, we illustrate the performance of our finetuned Qwen3-Omni-30B-A3B-Captioner through three representative case studies. The selected scenarios are designed to test the model's proficiency in: (1) analyzing expressive speech, (2) interpreting complex auditory scenes and sound effects, and (3) captioning composite audio that includes speech, music, and ambient sounds. For reproducibility, the original audio samples can be accessed from our public GitHub repository.

Analysis of Expressive Speech

The audio clip opens in a studio setting, marked by a faint, persistent electronic hiss and a subtle low-frequency hum, indicating professional but not entirely noise-free recording conditions. The male speaker, whose voice is delivered in a clear, energetic, and highly theatrical manner, begins with an assertive “对” (“Right!”), delivered with a sharp, rising intonation that immediately establishes a performative, self-aware tone. He continues, “这就是我，万人敬仰的太乙真人” (“This is me, Taiyi Zhenren, admired by thousands”), using exaggerated emphasis and a pronounced rising inflection on “万人敬仰” (“admired by thousands”) to convey mock grandeur and comic bravado. His articulation is crisp, and his voice is rich in midrange frequencies, suggesting close-mic’d recording with a touch of digital reverb for dramatic effect.

After a brief, deliberate pause, the speaker shifts to a more self-deprecating yet still playful register: “虽然有点婴儿肥，但也掩不住我逼人的帅气” (“Although I’m a bit chubby, it still can’t hide my impressive handsomeness”). Here, the tone becomes lighter and more conversational, with a subtle downward inflection on “婴儿肥” (“chubby”) and a playful, rising emphasis on “逼人的帅气” (“impressive handsomeness”), underscoring the comedic contrast between modesty and self-praise. The delivery is marked by expressive inflection and clear articulation, maintaining a consistent presence in the mix.

Throughout the speech, the recording remains clean and undistorted, with no background noises, environmental cues, or audience sounds. The voice is centrally positioned in the stereo field, and the use of reverb and EQ creates a sense of theatricality and spatial depth, while the electronic hiss and hum are subtle but present, hinting at professional studio equipment and post-processing. The clip concludes with a sudden, complete cut to silence, with no trailing sounds or fade-out, indicating a deliberate and precise edit.

In summary, this audio features a single male speaker delivering a self-introductory monologue in Mandarin, portraying the legendary Taoist figure Taiyi Zhenren in a comedic, over-the-top manner. The performance blends exaggerated bravado with self-deprecating humor, employing dramatic inflection, expressive pacing, and professional studio production to create a memorable, entertaining character portrayal. The absence of background noise and the clean, intentional editing reinforce the clip’s crafted, theatrical nature, making it suitable for online content, podcasts, or digital entertainment aimed at a modern, culturally literate audience.

Analysis of Complex Scene Sound Effect

The audio clip is a highly produced, cinematic soundscape, lasting 25 seconds, designed to evoke intense drama and imminent danger. It begins with a deep, resonant musical drone and a sharp, metallic clank, immediately establishing a mood of tension and foreboding. A slow, rhythmic, and ominous beat joins, layered with swelling orchestral strings and synthetic textures that build suspense and a sense of scale.

Around the 9-second mark, the soundscape explodes with the thunderous, mechanical roar of a massive engine, its deep, percussive pulses dominating the mix. This is accompanied by a high-pitched, metallic screech, suggesting immense machinery in motion or under strain. The orchestral music intensifies, becoming more urgent and climactic, with the addition of a high-frequency electronic whine that adds to the feeling of technological threat.

At the 15-second point, a colossal, explosive impact occurs, characterized by a sharp, high-frequency crack followed by a deep, resonant boom. The sound of shattering and debris fills the space, with reverberation suggesting a vast, hard-walled environment—perhaps a hangar, tunnel, or industrial facility. The orchestral music reaches its peak, becoming chaotic and overwhelming, before abruptly ceasing at the 19-second mark.

Immediately after the impact, the focus shifts to the sounds of a single person: heavy, strained breathing, the rustle of heavy fabric, and a low groan. These sounds, recorded with close-mic clarity and natural reverberation, indicate the individual is recovering from exertion or injury, likely amidst the aftermath of the catastrophic event. There are no spoken words, but the vocalizations and physical sounds strongly suggest the presence of a male character.

Throughout the clip, there are no voices, dialogue, or environmental cues that would suggest a specific culture, language, or geographic setting. The orchestral and electronic musical elements are universal to modern action, sci-fi, and thriller genres, and the production values are consistent with contemporary Hollywood or AAA video game trailers.

In summary, the audio is a tightly crafted, non-verbal scene of a catastrophic event—likely a crash, explosion, or mechanical disaster—in a vast, industrial or futuristic setting, immediately followed by the physical and emotional aftermath of a single survivor. The absence of speech, combined with the orchestration, sound design, and spatial cues, positions the clip as a generic but highly effective trailer or scene cue for a dramatic, action-oriented media product.

Analysis of Mixed Speech, Audio, and Music

The audio begins with a deep, resonant metallic clang, immediately followed by a powerful, low-frequency rumble that evokes the sound of a massive engine or heavy machinery. The space is acoustically vast, with long reverberation tails and a faint, high-pitched whine, suggesting a colossal, metallic environment such as a spaceship hangar or a vast industrial chamber. A mechanical whirring sound emerges, layered with metallic clanking and grinding, creating the impression of a large vehicle or portal activating. These sounds are punctuated by a sharp, high-frequency crackle, reminiscent of electrical arcs or energy discharges, further reinforcing the technological setting. The mechanical noises intensify, with rapid, high-pitched metallic screeches and impacts that imply the movement or collision of massive metal components.

Suddenly, the mechanical sounds fade, and the ambient rumble persists, revealing a subtle, high-frequency hiss—possibly from an air filtration system or atmospheric turbulence—while the vastness of the space remains evident through lingering echoes. A female voice, distant and high-pitched, emerges with a plaintive, childlike tone, asking, ”Are we there yet?” Her speech is slightly muffled and reverberant, indicating she is physically separated from the microphone, likely inside the vehicle or machinery. This is followed by a deeper, gravelly male voice, close to the microphone, responding with a gruff, impatient tone: ”We get there when we get there.” His voice is clear and assertive, contrasting with the female’s, and the exchange is typical of familial banter.

The mechanical rumble swells again, joined by a whooshing sound as if air is rushing past, and a rapid metallic clatter signals the rapid movement of machinery or vehicles. The environment is further emphasized by a sharp, high-frequency crackle, suggesting an energy surge or system overload. A third male voice, energetic and friendly, calls out from a moderate distance: ”How you doing, honey?” His tone is warm and affectionate, with a slight echo, and the use of ”honey” implies a familial relationship. Immediately after, the female voice, now closer and more urgent, responds with a high-pitched, exasperated tone: ”Do I have to answer?” Her delivery is quick, sharp, and filled with playful annoyance, reflecting a familiar and comfortable dynamic among the group.

As the mechanical sounds subside, a low-frequency hum remains, and the audio transitions into a brief, synthesized musical sting. This consists of a single sustained note from a low-frequency synthesizer, likely a bass or synth pad, which is cut off abruptly, suggesting the end of the scene or a transition to another segment. Throughout, the audio is of high fidelity, with no distortion or noise, and each sound is distinct and well-defined. The spatial characteristics—distance, direction, and reverberation—contribute to a vivid sense of a large, metallic, and technological environment. The dialogue is clear and expressive, with emotional tones ranging from impatience and warmth to playful annoyance. The use of ”honey” and the familial banter reinforce the impression of a close-knit group, likely family members, engaged in a shared journey within a science fiction or fantasy context.

In summary, the audio presents a dynamic, high-fidelity soundscape of a massive, metallic environment—possibly a spaceship or futuristic vehicle—where a group of family members engage in playful banter as they travel together. Mechanical sounds, spatial cues, and expressive dialogue combine to create a vivid sense of place and character, culminating in a synthesized musical sting that signals a narrative transition. The scene is rich in emotional nuance and technological detail, firmly situating the listener within a science fiction or fantasy setting.

References

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073.

AIME. AIME problems and solutions, 2025. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

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*, 2024.

Anthropic. Introducing Claude, 2023a. URL https://www.anthropic.com/index/introducing-claude.

Anthropic. Claude 2. Technical report, Anthropic, 2023b. URL https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf.

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic, AI, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *CoRR*, abs/2309.16609, 2023a.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023b.

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*, 2025.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The mtg-jamendo dataset for automatic music tagging. ICML, 2019.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Federico Cassano, John Gouwar, Daniel Nguyen, Sydney Nguyen, Luna Phipps-Costin, Donald Pinckney, Ming-Ho Yee, Yangtian Zi, Carolyn Jane Anderson, Molly Q. Feldman, Arjun Guha, Michael Greenberg, and Abhinav Jangda. MultiPL-E: A scalable and polyglot approach to benchmarking neural code generation. *IEEE Trans. Software Eng.*, 49(7):3675–3691, 2023.

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:2403.20330*, 2024a.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024b.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024c.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can MLLM think like holmes for complex video reasoning? *CoRR*, abs/2505.21374, 2025.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *CoRR*, abs/2311.07919, 2023.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, 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.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *CoRR*, abs/2505.17589, 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The Llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 682–689. IEEE, 2024.

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:2405.21075*, 2024.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, et al. Are we done with mmlu? *CoRR*, abs/2406.04127, 2024.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google, 2024. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14375–14385. IEEE, 2024.

Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, Shruti Bhosale, Chenguang Zhu, Karthik Abinav Sankararaman, Eryk Helenowski, Melanie Kambadur, Aditya Tayade, Hao Ma, Han Fang, and Sinong Wang. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *CoRR*, abs/2410.15553, 2024. doi: 10.48550/ARXIV.2410.15553. URL https://doi.org/10.48550/arXiv.2410.15553.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *CoRR*, abs/2502.04326, 2025.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *ISMIR*, pp. 387–392, 2009.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv:2301.12597*, 2023.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. ZebraLogic: On the scaling limits of LLMs for logical reasoning. *CoRR*, abs/2502.01100, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv:2304.08485*, 2023.

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 *ICLR*, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv:2203.10244*, 2022.

OpenAI. ChatML, 2022. URL https://github.com/openai/openai-python/blob/e389823ba013a24b4c32ce38fa0bd87e6bccae94/chatml.md.

OpenAI. GPT4 technical report. *CoRR*, abs/2303.08774, 2023.

OpenAI. Hello GPT-4o, 2024. URL https://openai.com/index/hello-gpt-4o/.

Samuel J. Paech. Creative writing v3, 2024. URL https://eqbench.com/creative_writing.html.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to count to ten. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 3147–3157. IEEE, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *CoRR*, abs/2311.12022, 2023.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark, 2024. URL https://arxiv.org/abs/2410.19168.

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced Transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn 2: Captioning-enhanced audio-visual large language models. *CoRR*, abs/2506.15220, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *https://arxiv.org/abs/2502.14786*, 2025.

George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing*, 10(5):293–302, 2002.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. *CoRR*, abs/2506.04779, 2025a. doi: 10.48550/ARXIV.2506.04779. URL https://doi.org/10.48550/arXiv.2506.04779.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv:2402.14804*, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark. *CoRR*, abs/2406.08035, 2024b.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *CoRR*, abs/2503.01710, 2025b.

Yiming Wang, Pei Zhang, Jialong Tang, Haoran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. Polymath: Evaluating mathematical reasoning in multilingual contexts. *CoRR*, abs/2504.18428, 2025c. doi: 10.48550/ARXIV.2504.18428. URL https://doi.org/10.48550/arXiv.2504.18428.

Yuancheng Wang, Haoyue Zhan, Liwei Liu, Ruihong Zeng, Haotian Guo, Jiachen Zheng, Qiang Zhang, Xueyao Zhang, Shunsi Zhang, and Zhizheng Wu. Maskgct: Zero-shot text-to-speech with masked generative codec transformer. *arXiv preprint arXiv:2409.00750*, 2024c.

Shangda Wu, Zhancheng Guo, Ruibin Yuan, Junyan Jiang, Seungheon Doh, Gus Xia, Juhan Nam, Xiaobing Li, Feng Yu, and Maosong Sun. Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages. *arXiv preprint arXiv:2502.10362*, 2025a.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. WritingBench: A comprehensive benchmark for generative writing. *CoRR*, abs/2503.05244, 2025b.

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

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Berkeley function calling leaderboard. https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html, 2024.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv:2407.10671*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Qize Yang, Shimin Yao, Weixuan chen, Shenghao Fu, Detao Bai, Jiaxing Zhao, Boyuan Sun, Bowen Yin, Xihan Wei, and Jingren Zhou. Humanomniv2: From understanding to omni-modal reasoning with context. *CoRR*, abs/2506.21277, 2025b.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, et al. Marble: Music audio representation benchmark for universal evaluation. *Advances in Neural Information Processing Systems*, 36:39626–39647, 2023.

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

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian McAuley, and Zachary Novack. Are you really listening? boosting perceptual awareness in music-qa benchmarks. *arXiv preprint arXiv:2504.00369*, 2025.

Bowen Zhang, Congchao Guo, Geng Yang, Hang Yu, Haozhe Zhang, Heidi Lei, Jialong Mai, Junjie Yan, Kaiyue Yang, Mingqi Yang, Peikai Huang, Ruiyang Jin, Sitan Jiang, Weihua Cheng, Yawei Li, Yichen Xiao, Yiying Zhou, Yongmao Zhang, Yuan Lu, and Yucen He. Minimax-speech: Intrinsic zero-shot text-to-speech with a learnable speaker encoder. *CoRR*, abs/2505.07916, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911, 2023.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 13691–13701. Computer Vision Foundation / IEEE, 2025a.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *CoRR*, abs/2505.17862, 2025b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv:2304.10592*, 2023.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. Muq: Self-supervised music representation learning with mel residual vector quantization. *arXiv preprint arXiv:2501.01108*, 2025.