Qwen2.5-Omni Technical Report

Qwen Team

https://huggingface.co/Qwen
https://modelscope.cn/organization/qwen
https://github.com/QwenLM/Qwen2.5-Omni

Abstract

In this report, we present Qwen2.5-Omni, an end-to-end multimodal model designed to perceive diverse modalities, including text, images, audio, and video, while simultaneously generating text and natural speech responses in a streaming manner. To enable the streaming of multimodal information inputs, both audio and visual encoders utilize a block-wise processing approach. This strategy effectively decouples the handling of long sequences of multimodal data, assigning the perceptual responsibilities to the multimodal encoder and entrusting the modeling of extended sequences to a large language model. Such a division of labor enhances the fusion of different modalities via the shared attention mechanism. To synchronize the timestamps of video inputs with audio, we organize the audio and video sequentially in an interleaved manner and propose a novel position embedding approach, named **TMRoPE** (**T**ime-aligned **M**ultimodal **RoPE**). To concurrently generate text and speech while avoiding interference between the two modalities, we propose **Thinker-Talker** architecture. In this framework, Thinker functions as a large language model tasked with text generation, while Talker is a dual-track autoregressive model that directly utilizes the hidden representations from the Thinker to produce audio tokens as output. Both the Thinker and Talker models are designed to be trained and inferred in an end-to-end manner. For decoding audio tokens in a streaming manner, we introduce a sliding-window DiT that restricts the receptive field, aiming to reduce the initial package delay. Qwen2.5-Omni is comparable with similarly sized Qwen2.5-VL and outperforms Qwen2-Audio. Furthermore, Qwen2.5-Omni achieves state-of-the-art performance on multimodal benchmarks like Omni-Bench. Notably, Qwen2.5-Omni 's performance in end-to-end speech instruction following is comparable to its capabilities with text inputs, as evidenced by benchmarks such as MMLU and GSM8K. As for speech generation, Qwen2.5-Omni's streaming Talker outperforms most existing streaming and non-streaming alternatives in robustness and naturalness.
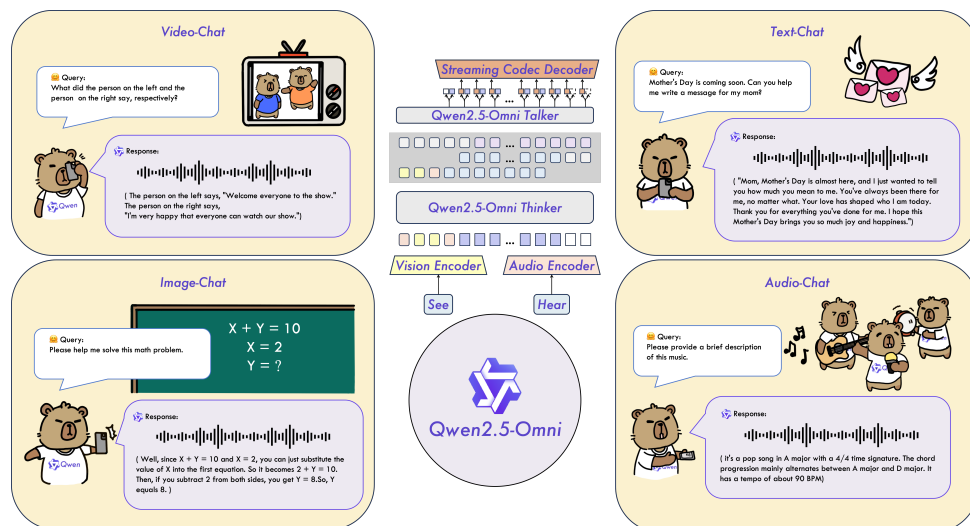
Figure 1: Qwen2.5-Omni is a unified end-to-end model capable of processing multiple modalities, such as text, audio, image and video, and generating real-time text or speech response. Based on these features, Qwen2.5-Omni supports a wide range of tasks, including but not limited to voice dialogue, video dialogue, and video reasoning.

1 Introduction

In daily life, humans are capable of simultaneously perceiving the visual and auditory information around them. After processing this information through the brain, they express feedback through writing, vocalization, or using tools (and physical actions), thereby engaging in information exchange with various organisms in the world and exhibiting intelligence. In recent years, general artificial intelligence has become increasingly visible, largely due to advancements in Large Language Models (LLMs) (Brown et al., 2020; OpenAI, 2023; 2024; Gemini Team, 2024; Anthropic, 2023a;b; 2024; Bai et al., 2023a; Yang et al., 2024a; Touvron et al., 2023a;b; Dubey et al., 2024a). These models, trained on vast amounts of textual data, represent high-level discrete representation created by humans, showcasing the ability to solve complex problems and learn rapidly. Furthermore, in the realm of understanding, Language-Audio-Language Models (LALMs) (OpenAI, 2024; Tang et al., 2024; Chu et al., 2023b; 2024b) and Language-Visual-Language Models (LVLMs) (Li et al., 2023; Liu et al., 2023b; Dai et al., 2023; Zhu et al., 2023; Huang et al., 2023; Bai et al., 2023b; Liu et al., 2023a; Wang et al., 2023b; OpenAI., 2023; Gemini Team, 2024) have helped LLMs to further extend auditory and visual capabilities in an end-to-end manner. However, efficiently unifying all these different understanding modalities in an end-to-end fashion, utilizing as much data as possible, and providing responses in both text and speech streams akin to human communication still presents a significant challenge.

The development of a unified and intelligent omni-model requires careful consideration of several key factors. First, it is crucial to implement a systematic method for the joint training of various modalities, including text, images, videos, and audio, to foster mutual enhancement among them. This alignment is particularly important for video content, where synchronization of the temporal aspects of audio and visual signals is necessary. Second, it is essential to manage potential interference among outputs from different modalities, ensuring that the training processes for outputs such as text and voice tokens do not disrupt each other. Finally, there is a need to explore architectural designs that enable real-time understanding of multimodal information and allow for efficient audio output streaming, thereby reducing initial latency.

In this report, we introduce Qwen2.5-Omni, a unified single model capable of processing multiple modalities and generating text and natural speech responses simultaneously in a streaming format. To tackle the first challenge, we propose a novel position embedding approach, named **TMRoPE** (**T**ime-aligned **M**ultimodal **RoPE**). We organize these audio and video frames in an interleaved structure to represent video sequences in time order. For the second challenge, we present Thinker-Talker architecture, wherein Thinker is tasked with text generation while the Talker focuses on generating streaming speech tokens. Talker receives high-level representations directly from Thinker. This design is inspired by the way humans utilize different organs to produce various signals, which are simultaneously coordinated through the same neural networks. As a result, Thinker-Talker architecture is end-to-end jointly trained, with each component dedicated to generating distinct signals. To address the challenges associated with streaming and to facilitate the pre-filling necessary for real-time comprehension of multimodal signals, we propose modifications to all multimodal encoders by adopting a block-wise streaming processing approach. In order to support streaming speech generation, we implement a dual-track autoregressive model that generates speech tokens, alongside a DiT model which converts these tokens into waveforms, thereby enabling streaming audio generation and minimizing initial latency. This design aims to enable the model to process multimodal information in real-time and effectively perform pre-filling, thereby enabling the concurrent generation of text and speech signals.

Qwen2.5-Omni is comparable with the similarly sized Qwen2.5-VL (Wang et al., 2024c) and outperforms Qwen2-Audio (Chu et al., 2024b) in image and audio capabilities respectively. Furthermore, Qwen2.5-Omni achieves state-of-the-art performance on multimodal benchmarks such as OmniBench (Li et al., 2024b) and AV-Odyssey Bench (Gong et al., 2024). Notably, Qwen2.5-Omni's performance in end-to-end speech instruction following is comparable to its capabilities with text inputs, as evidenced by benchmarks such as MMLU (Hendrycks et al., 2021a) and GSM8K (Cobbe et al., 2021). As for speech generation, Qwen2.5-Omni achieves 1.42%, 2.33% and 6.54% WER on seed-tts-eval (Anastassiou et al., 2024) test-zh, test-en and test-hard set respectively, outperforming MaskGCT (Wang et al., 2024e) and CosyVoice 2 (Du et al., 2024).

The key features of Qwen2.5-Omni can be summarized as:

- We introduce Qwen2.5-Omni, a unified model that can perceive all modalities and simultaneously generate text and natural speech responses in a streaming fashion.
- We present a novel positional embedding algorithm, termed TMRoPE, which explicitly incorporates temporal information for synchronizing audio and video.
- We propose the Thinker-Talker Architecture to facilitate real-time comprehension and speech generation.

- Qwen2.5-Omni demonstrates strong performance across all modalities when benchmarked against similarly sized single-modality models. It significantly enhances the capability of following voice commands, achieving performance levels comparable to pure text input. For tasks that involve integrating multiple modalities, such as those evaluated in OmniBench (Li et al., 2024b), Qwen2.5-Omni achieves state-of-the-art performance. Notably, Qwen2.5-Omni achieves strong performance on seed-tts-eval (Anastassiou et al., 2024), demonstrating robust speech generation abilities.
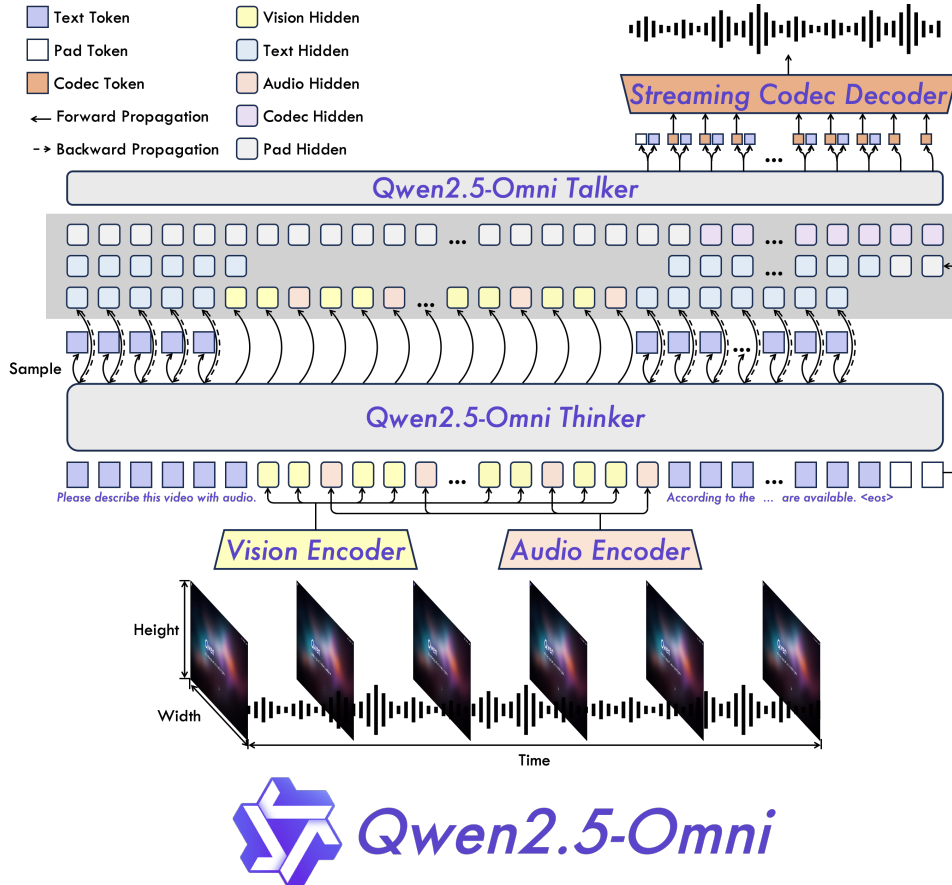
2 Architecture

Figure 2: The overview of Qwen2.5-Omni. Qwen2.5-Omni adpots 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.

2.1 Overview

As shown in Figure 2, Qwen2.5-Omni employs Thinker-Talker architecture. Thinker functions like a brain, responsible for processing and understanding inputs from text, audio and video modalities, generating high-level representations and corresponding text. Talker operates like a human mouth, taking in the high-level representations and text produced by the Thinker in a streaming manner, and outputting discrete tokens of speech fluidly. Thinker is a Transformer decoder, accompanied by encoders for audio and image that facilitate information extraction. In contrast, Talker is designed as a dual-track autoregressive Transformer Decoder architecture, motivated by Mini-Omni (Xie & Wu, 2024). During both training and inference, Talker directly receives high-dimensional representations from Thinker and shares all of Thinker's historical context information. Consequently, the entire architecture operates as a cohesive single model, enabling end-to-end training and inference.

In the following sections, we first introduce how Qwen2.5-Omni perceives various input signals and present our proposed novel positional encoding algorithm, TMRoPE. Subsequently, the details of text and speech generation are presented. Finally, we highlight the improvements made in the understanding and generation modules to facilitate efficient streaming inference.
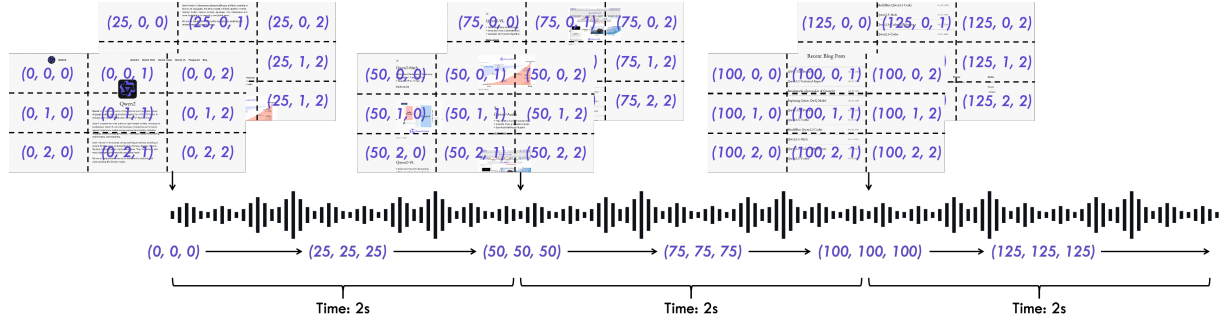
2.2 Perceivation

Figure 3: An illustration of Time-aligned Multimodal RoPE (TMRoPE).

Text, Audio, Image and Video (w/o Audio). Thinker processes text, audio, images, and video (without the audio track) by converting them into a series of hidden representations for input. For tokenizing text, we use Qwen's tokenizer (Yang et al., 2024a), which applies byte-level byte-pair encoding with a vocabulary comprising 151,643 regular tokens. Regarding audio input and audio from videos, we resample it to a frequency of 16kHz and transform the raw waveform into a 128-channel mel-spectrogram with a window size of 25ms and a hop size of 10ms. We adopt the audio encoder from Qwen2-Audio (Chu et al., 2024b), to make each frame of audio representation roughly corresponds to a 40ms segment of the original audio signal. Furthermore, we employ the vision encoder from Qwen2.5-VL (Bai et al., 2025), which is based on the Vision Transformer (ViT) model with approximately 675 million parameters, enabling it to effectively handle both image and video inputs. The vision encoder employs a mixed training regimen incorporating both image and video data, ensuring proficiency in image understanding and video comprehension. To preserve video information as completely as possible while adapting to the audio sampling rate, we sample the video using a dynamic frame rate. Additionally, for consistency, each image is treated as two identical frames.

Video and TMRoPE. We propose a time-interleaving algorithm for audio and video, along with a novel position encoding approach. As shown in Figure 3, TMRoPE encodes the 3-D positional information of multimodal inputs, which is Multimodal Rotary Position Embedding (M-RoPE) (Bai et al., 2023b) with absolute temporal positions. This is achieved by deconstructing the original rotary embedding into three components: temporal, height, and width. For text inputs, these components utilize identical position IDs, making M-RoPE functionally equivalent to 1D-RoPE. Similarly, for audio inputs, we also use identical position IDs and introduce absolute temporal position encoding, with one temporal ID corresponding to 40ms.

When processing images, the temporal IDs of each visual token remain constant, while distinct IDs are assigned to the height and width components based on the token's position in the image. When the input is video with audio, the audio is still encoded with identical position IDs for every 40ms per frame, and the video is treated as a series of images with temporal ID increments for each frame, while the height and width components follow the same ID assignment pattern as images. Since the frame rate in video is not fixed, we dynamically adjust the temporal IDs between frames based on the actual time corresponding to each frame to ensure that one temporal ID corresponds to 40ms. In scenarios where the model's input encompasses multiple modalities, position numbering for each modality is initialized by incrementing the maximum position ID of the preceding modality by one. TMRoPE enhances positional information modeling, maximizing the integration of various modalities, enabling Qwen2.5-Omni to simultaneously understand and analyze information from multiple modalities.

After incorporating positional information into each modality, we arrange the representations in order. To enable the model to receive both visual and auditory information simultaneously, as shown in Figure 3, we have a special design for video with audio called the time-interleaving method, which segments the representation in the video with audio into chunks every 2 seconds according to the actual time. We then arrange the visual representation at the front and the audio representation at the back within the 2 seconds, interleaving the representations of the video with audio.

2.3 Generation

Text. Text is generated directly by Thinker. The logic of text generation is fundamentally the same as that employed by widely used LLMs, which generate text through autoregressive sampling based on the

probability distribution over the vocabulary. The generation process may incorporate techniques such as repetition penalty and top-p sampling to enhance its diversity.

Speech. Talker receives both high-level representations and embeddings of the text tokens sampled by Thinker. The integration of high-dimensional representations and discrete sampling tokens is essential in this context. As a streaming algorithm, voice generation must anticipate the content's tone and attitude before the entire text is fully generated. The high-dimensional representations provided by Thinker implicitly convey this information, enabling a more natural streaming generation process. Furthermore, Thinker's representations primarily express semantic similarity in the representational space rather than phonetic similarity. Consequently, even phonetically distinct words may have very similar high-level representations, necessitating the input of sampled discrete tokens to eliminate such uncertainty.

We designed an efficient speech codec named *qwen-tts-tokenizer*. *qwen-tts-tokenizer* efficiently represents key information of speech and can be decoded to speech streamingly through a causal audio decoder. After receiving the information, Talker starts to autoregressively generate audio tokens and text tokens. The generation of speech does not require word-level and timestamp-level alignment with the text. This significantly simplifies the requirements for training data and the inference process.

2.4 Designs for Streaming

In the context of streaming audio and video interactions, the initial packet latency is a critical indicator of the system's streaming performance. This latency is influenced by several factors: 1) the delay caused by the processing of multimodal information inputs; 2) the latency from the moment the first text input is received until the first voice token is output; 3) the delay in converting the first segment of speech into audio; and 4) the inherent latency of the architecture itself, which is related to model size, computational FLOPs, and other factors. This paper will subsequently discuss the algorithmic and architectural improvements made to reduce these latencies across these four dimensions.

Support Prefilling. Chunked-prefills is a mechanism widely used in modern inference framework. To support it in modalities interation, we modified the audio and visual encoders to support block-wise attention along the temporal dimension. Specifically, the audio encoder is changed from full attention over the entire audio to performing attention in blocks of 2 seconds each. The vision encoder utilizes flash attention for efficient training and inference with a simple MLP layer that merges adjacent 2×2 tokens into a single token. The patch size is set to 14, which allows images of different resolutions to be packed into a sequence.

Streaming Codec Generation. To facilitate the streaming of audio, especially for extended sequences, we propose a sliding window block attention mechanism that restricts the current token's access to a limited context. Specifically, we utilize a Flow-Matching (Lipman et al.) DiT model. The input code is transformed into a mel-spectrogram using Flow-Matching, followed by a modified BigVGAN (Lee et al.) to reconstruct the generated mel-spectrogram back into the waveform.
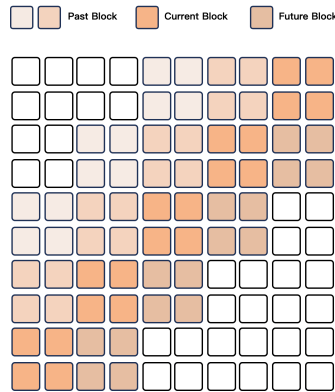
Figure 4: An illustration of sliding window block attention mechanism in DiT for codec to wav generation.

As shown in Figure 4, to generate waveforms from code, we group adjacent codes into blocks and use these for our attention mask. We limit the DiT's receptive field to 4 blocks, including a lookback of 2 blocks and a lookahead of 1 block. During decoding, we generate the mel-spectrum in chunks using Flow Matching, ensuring that each code chunk has access to the necessary contextual blocks. This approach enhances the quality of streaming outputs by maintaining contextual information. We also use this chunk-by-chunk method for BigVGAN's fixed receptive field to facilitate streaming waveform generation

3 Pre-training

Qwen2.5-Omni consists of three training stages. In the first stage, we lock the LLM parameters and focus exclusively on training the vision encoder and audio encoder, 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 32k to enhance the model's ability to understand complex long-sequence data.

The model is pre-trained on a diverse dataset that includes various types such as image-text, video-text, video-audio, audio-text and text corpus. We replace the hierarchical tags with the natural language prompts following Qwen2-Audio (Chu et al., 2024a), which can improve better generalization ability and better instruction following ability.

During the initial pre-training phase, the LLM component of Qwen2.5-Omni is initialized using the parameters from Qwen2.5 (Yang et al., 2024b), while the vision encoder is the same as Qwen2.5-VL, and the audio encoder is initialized with the Whisper-large-v3 (Radford et al., 2023). The two encoders are trained separately on the fixed LLM, with both initially focusing on training their respective adapters before training the encoders. This foundational training is crucial in equipping the model with a robust understanding of core visual-textual and audio-textual correlations and alignments.

The second phase of pre-training marks a significant advancement by incorporating an additional 800 billion tokens of image and video related data, 300 billion tokens of audio related data, and 100 billion tokens of video with audio related data. This phase introduces a larger volume of mixed multimodal data and a wider variety of tasks, which enhances the interaction and deepens the understanding between auditory, visual, and textual information. The inclusion of multimodal, multitask datasets is crucial for developing the model's ability to handle multiple tasks and modalities simultaneously, a vital capability for managing complex real-world datasets. Moreover, pure text data plays an essential role in maintaining and improving language proficiency.

To improve training efficiency, we limited the maximum token length to 8192 tokens in the previous stages. Then, we incorporate long audio and long video data and extend the original text, audio, image, and video data to 32,768 tokens for training. Experimental results indicate that our data shows significant improvement in supporting long sequence data

4 Post-training

4.1 Data Format

> **The Dataset Format Example of ChatML**
>
> <|im_start|>user
> <|vision_start|>Video.mp4 [Two people are talking in the video]<|vision_end|>What are the people in the video saying?<|im_end|>
> <|im_start|>assistant
> Both pictures are of SpongeBob SquarePants.The person in the red clothes says, "Hello, how's the weather today?" The person in the black clothes responds, "Hello, the weather is quite nice today."<|im_end|>
> <|im_start|>user
> <|vision_start|>Video.mp4 [A person in the video is saying, "Please describe the person in front of you."]<|vision_end|><|im_end|>
> <|im_start|>assistant
> The person in front of you is wearing glasses and a brown jacket over a blue shirt. They appear to be speaking or reacting to something, as their mouth is open and they seem engaged. The background shows a room with a wall-mounted air conditioner, a clothing rack with various garments hanging on it, and a large screen displaying an image of a cityscape at night. The lighting in the room is warm and cozy.<|im_end|>

4.2 Thinker

During the post-training phase, we employ instruction-following data with ChatML (OpenAI, 2022) format for instruction-finetuning. Our dataset incorporates pure text-based dialogue data, visual-modality conversation data, audio-modality conversation data and mix-modality conversation data.

4.3 Talker

We introduced a three-stage training process for Talker, allowing Qwen2.5-Omni to generate text and speech responses simultaneously. In the first stage, we train Talker to learn context continuation. The second stage utilized DPO (Rafailov et al., 2023) to enhance the stability of speech generation. In the third stage, we applied multi-speaker instruction fine-tuning to improve the naturalness and controllability of the speech responses.

During the In-Context Learning (ICL) training phase, in addition to utilizing text supervision similar to that of Thinker, we perform a speech continuation task through next-token prediction, leveraging an extensive dataset of dialogues that incorporate multimodal contexts and spoken responses. Talker learns to establish a monotonic mapping from semantic representation to speech, while also acquiring the ability to express speech with diverse attributes that are contextually appropriate, such as prosody, emotion, and accent. Additionally, we implement timbre disentanglement techniques to prevent the model from associating specific voices with infrequent textual patterns.

$$\mathcal{L}_{\text{DPO}}(\mathcal{P}_\theta; \mathcal{P}_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\mathcal{P}_\theta(y_w \mid x)}{\mathcal{P}_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\mathcal{P}_\theta(y_l \mid x)}{\mathcal{P}_{\text{ref}}(y_l \mid x)} \right) \right]. \tag{1}$$

To broaden the coverage of speakers and scenarios, the pretraining data inevitably contains label noise and pronunciation errors, leading to model hallucinations. To mitigate this issue, we introduce a reinforcement learning phase to improve the stability of speech generation. Specifically, for each request and response text paired with the reference speech, we build a dataset $\mathcal{D}$ with the triplet data (x, y_w, y_l), where x is the input sequence with input text, and y_w and y_l are the good and bad generated speech sequences respectively. We rank these samples based on their reward scores associated with word error rate (WER) and the punctuation pause error rate.

Lastly, we performed speaker fine-tuning on the aforementioned base model, enabling Talker to adopt specific voices and improve its naturalness.

5 Evaluation

We conduct comprehensive evaluation of Qwen2.5-Omni. The model is divided into two main categories: understanding (X→Text) and speech generation (X→Speech).

5.1 Evaluation of X→Text

In this section, we evaluate Qwen2.5-Omni's ability to comprehend various multimodal inputs (text, audio, image, and video) and generate textual responses.

Text→Text   Our evaluation of Qwen2.5-Omni on text → text primarily focuses on general evaluation, mathematics & science ability and coding ability. Specifically, we utilize MMLU-Pro (Wang et al., 2024f), MMLU-redux (Gema et al., 2024) and Livebench0803 (White et al., 2024) for general evaluation, GPQA (Rein et al., 2023), GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021b) for mathematics & science, HumanEval (Chen et al., 2021), MBPP (Austin et al., 2021), MultiPL-E (Cassano et al., 2023) and LiveCodeBench 2305-2409 (Jain et al., 2024) for coding.

Audio→Text   The evaluation of Qwen2.5-Omni for audio → text includes audio understanding, audio reasoning, and voice-chatting. Specifically, we perform a comprehensive evaluation on Automatic Speech Recognition (ASR), Speech-to-Text Translation (S2TT), Speech Entity Recognition (SER), Vocal Sound classification (VSC) and Music, which assesses the performance of Qwen2.5-Omni on a broad range of audio understanding tasks. We utilize MMAU (Sakshi et al., 2024) for audio reasoning tasks, VoiceBench (Chen et al., 2024b) and a self-curated speech-instruction benchmark for voice-chatting tasks.

Image→Text   The evaluation of Qwen2.5-Omni for image → text primarily emphasizes the performance in college-level problems, math, general visual question answering and OCR-related tasks. Specifically, we utilize MMMU (Yue et al., 2023) and MMMU-Pro (Yue et al., 2024) for college-level problems evaluation, MathVista (Lu et al., 2024b) and MathVision (Wang et al., 2024b) for math. For general visual question answering, we evaluate the performance on benchmark datasets such as MMBench-V1.1 (Liu et al., 2023c), MMVet (Yu et al., 2024), MMStar (Chen et al., 2024a), MME (Fu et al., 2023), MuirBench (Wang et al., 2024a), CRPE (Wang et al., 2024d), RealWorldQA (X.AI, 2024), MMERealWorld (Zhang et al., 2024), and MM-MT-Bench (Agrawal et al., 2024). Additionally, we evaluate Qwen2.5-Omni on various OCR benchmarks, such as AI2D (Kembhavi et al., 2016), TextVQA (Singh et al., 2019), DocVQA (Mathew

et al., 2021), ChartQA (Masry et al., 2022), and OCRBench_v2 (Fu et al., 2024b). Furthermore, we also evaluate the visual grounding capability of our model on the referring expression comprehension benchmarks (Kazemzadeh et al., 2014; Mao et al., 2016), object detection in the wild (Li et al., 2022) and a self-curated point grounding benchmark.

Video (w/o Audio)→Text   We assess our model on several representative video understanding tasks like Video-MME (Fu et al., 2024a), MVBench (Li et al., 2024a), and EgoSchema (Mangalam et al., 2023).

Multimodality→Text   We demonstrate the ability of our model for mixed-modality (image, audio and text) prompts on OmniBench (Li et al., 2024b).

5.1.1 Performance of Text→Text

We compare Qwen2.5-Omni with other leading large language model of similar size (7B). As shown in Table 1, the performance of Qwen2.5-Omni generally falls between Qwen2-7B and Qwen2.5-7B. Our model outperforms Qwen2-7B on most benchmarks, such as MMLU-Pro, MMLU-redux, MATH, GSM8K, MBPP, MultiPL-E and LiveCodeBench, which demonstrates the exceptional capabilities of our model for Text→Text.

Table 1: **Text → Text performance of 7B+ pure text models and Qwen2.5-Omni**

Datasets	Gemma2-9B	Llama3.1-8B	Qwen2-7B	Qwen2.5-7B	Qwen2.5-Omni-7B
General Tasks					
MMLU-Pro	52.1	48.3	44.1	**56.3**	47.0
MMLU-redux	72.8	67.2	67.3	**75.4**	71.0
LiveBench$_{0831}$	30.6	26.7	29.2	**35.9**	29.6
Mathematics & Science Tasks					
GPQA	32.8	32.8	34.3	**36.4**	30.8
MATH	44.3	51.9	52.9	**75.5**	71.5
GSM8K	76.7	84.5	85.7	**91.6**	88.7
Coding Tasks					
HumanEval	68.9	72.6	79.9	**84.8**	78.7
MBPP	74.9	69.6	67.2	**79.2**	73.2
MultiPL-E	53.4	50.7	59.1	**70.4**	65.8
LiveCodeBench$_{2305-2409}$	18.9	8.3	23.9	**28.7**	24.6

5.1.2 Performance of Audio→Text

We compare Qwen2.5-Omni with other leading specialist or generalist models on diverse audio understanding, audio reasoning, and voice-chatting benchmarks. As shown in Table 2 and 3, Qwen2.5-Omni delivers better or comparable performance with other state-of-the-art methods on audio understanding. For instance, it achieves superior ASR and S2TT performance on Fleurs_zh, CommonVoice_en, CommonVoice_zh, CoVoST2_en-de and CoVoST2_zh-en test sets, surpassing previous state-of-the-art models like Whisper-large-v3, Qwen2Audio, MinMo and other Omni models. Qwen2.5-Omni also achieves state-of-the-art performance on general audio understanding tasks like music and VSC. Additionally, Qwen2.5-Omni achieves state-of-the-art results on audio reasoning with superior performance on sound, music and speech subsets of MMAU benchmark. These results demonstrate the powerful capabilities of Qwen2.5-Omni in general audio understanding and reasoning.

Additionally, on VoiceBench, Qwen2.5-Omni achieves an impressive average score of 74.12, surpassing other audio language models and omni models of similar size. This showcases our model's strong capabilities in speech interaction. To further explore the performance of diverse speech interaction, we convert text instructions from several pure-text benchmarks into speech and evaluate Qwen2.5-Omni, Qwen2-Audio and Qwen2-7B on the in-house voice-chat benchmark. About 90% of text-instructions are utilized. We use speech instruction for Qwen2.5-Omni and Qwen2-Audio, and text instruction for Qwen2-7B. As shown in Table 4, compared to Qwen2-Audio, Qwen2.5-Omni significantly narrowes the gap with Qwen2-7B, which uses text instructions. This reflects our model's substantial progress in diversified end-to-end speech interaction.

Table 2: **Audio → text performance of State-of-the-art and Qwen2.5-Omni**

Datasets	Model	Performance
	ASR	
Librispeech *dev-clean* ∣ *dev-other* ∣ *test-clean* ∣ *test-other*	SALMONN (Tang et al., 2024)	- ∣ - ∣ 2.1 ∣ 4.9
	SpeechVerse (Das et al., 2024)	- ∣ - ∣ 2.1 ∣ 4.4
	Whisper-large-v3 (Radford et al., 2023)	- ∣ - ∣ 1.8 ∣ 3.6
	Llama-3-8B (Dubey et al., 2024b)	- ∣ - ∣ - ∣ 3.4
	Llama-3-70B (Dubey et al., 2024b)	- ∣ - ∣ - ∣ 3.1
	Seed-ASR-Multilingual (Bai et al., 2024)	- ∣ - ∣ **1.6** ∣ **2.8**
	MiniCPM-o (Yao et al., 2024)	- ∣ - ∣ 1.7 ∣ -
	MinMo (Chen et al., 2025)	- ∣ - ∣ 1.7 ∣ 3.9
	Qwen-Audio (Chu et al., 2023a)	1.8 ∣ 4.0 ∣ 2.0 ∣ 4.2
	Qwen2-Audio (Chu et al., 2024a)	**1.3** ∣ **3.4** ∣ **1.6** ∣ 3.6
	Qwen2.5-Omni-7B	1.6 ∣ 3.5 ∣ 1.8 ∣ 3.4
Common Voice 15 *en* ∣ *zh* ∣ *yue* ∣ *fr*	Whisper-large-v3 (Radford et al., 2023)	9.3 ∣ 12.8 ∣ 10.9 ∣ 10.8
	MinMo (Chen et al., 2025)	7.9 ∣ 6.3 ∣ 6.4 ∣ 8.5
	Qwen2-Audio (Chu et al., 2024a)	8.6 ∣ 6.9 ∣ **5.9** ∣ 9.6
	Qwen2.5-Omni-7B	**7.6** ∣ **5.2** ∣ 7.3 ∣ **7.5**
Fleurs *zh* ∣ *en*	Whisper-large-v3 (Radford et al., 2023)	7.7 ∣ 4.1
	Seed-ASR-Multilingual (Bai et al., 2024)	- ∣ **3.4**
	Megrez-3B-Omni (Infinigence)	10.8 ∣ -
	MiniCPM-o (Yao et al., 2024)	4.4 ∣ -
	MinMo (Chen et al., 2025)	**3.0** ∣ 3.8
	Qwen2-Audio (Chu et al., 2024a)	7.5 ∣ -
	Qwen2.5-Omni-7B	**3.0** ∣ 4.1
Wenetspeech *test-net* ∣ *test-meeting*	Seed-ASR-Chinese (Bai et al., 2024)	**4.7** ∣ **5.7**
	Megrez-3B-Omni (Infinigence)	- ∣ 16.4
	MiniCPM-o (Yao et al., 2024)	6.9 ∣ -
	MinMo (Chen et al., 2025)	6.8 ∣ 7.4
	Qwen2.5-Omni-7B	5.9 ∣ 7.7
Voxpopuli-V1.0-en	Llama-3-8B (Dubey et al., 2024b)	6.2
	Llama-3-70B (Dubey et al., 2024b)	**5.7**
	Qwen2.5-Omni-7B	5.8
	S2TT	
CoVoST2 *en-de* ∣ *de-en* ∣ *en-zh* ∣ *zh-en*	SALMONN (Tang et al., 2024)	18.6 ∣ - ∣ 33.1 ∣ -
	SpeechLLaMA (Wu et al., 2023)	- ∣ 27.1 ∣ - ∣ 12.3
	BLSP (Wang et al., 2023a)	14.1 ∣ - ∣ - ∣ -
	MiniCPM-o (Yao et al., 2024)	- ∣ - ∣ **48.2** ∣ 27.2
	MinMo (Chen et al., 2025)	- ∣ **39.9** ∣ 46.7 ∣ 26.0
	Qwen-Audio (Chu et al., 2023a)	25.1 ∣ 33.9 ∣ 41.5 ∣ 15.7
	Qwen2-Audio (Chu et al., 2024a)	29.9 ∣ 35.2 ∣ 45.2 ∣ 24.4
	Qwen2.5-Omni-7B	**30.2** ∣ 37.7 ∣ 41.4 ∣ **29.4**

Table 3: **Audio → text performance of State-of-the-art and Qwen2.5-Omni**

Datasets	Model	Performance
	SER	
Meld	WavLM-large (Chen et al., 2022)	0.542
	MiniCPM-o (Yao et al., 2024)	0.524
	Qwen-Audio (Chu et al., 2023a)	0.557
	Qwen2-Audio (Chu et al., 2024a)	0.553
	Qwen2.5-Omni-7B	**0.570**
	VSC	
VocalSound	CLAP (Elizalde et al., 2022)	0.495
	Pengi (Deshmukh et al., 2023)	0.604
	Qwen-Audio (Chu et al., 2023a)	0.929
	Qwen2-Audio (Chu et al., 2024a)	**0.939**
	Qwen2.5-Omni-7B	**0.939**
	Music	
GiantSteps	LLark-7B (Gardner et al., 2023)	0.86
Tempo	Qwen2.5-Omni-7B	**0.88**
MusicCaps	LP-MusicCaps (Doh et al., 2023)	0.291 | 0.149 | 0.089 | **0.061** | **0.129** | 0.130
	Qwen2.5-Omni-7B	**0.328** | **0.162** | **0.090** | 0.055 | 0.127 | **0.225**
	Audio Reasoning	
MMAU	Gemini-Pro-V1.5 (Team et al., 2024)	56.75 | 49.40 | 58.55 | 54.90
Sound | Music |	Qwen2-Audio (Chu et al., 2024a)	54.95 | 50.98 | 42.04 | 49.20
Speech | Avg	Qwen2.5-Omni-7B	**67.87** | **69.16** | **59.76** | **65.60**
	Voice Chatting	
VoiceBench	Ultravox-v0.4.1-LLaMA-3.1-8B	**4.55** | 3.90 | 53.35 | 47.17
AlpacaEval | CommonEval |	MERaLiON (He et al., 2024)	4.50 | 3.77 | 55.06 | 34.95
SD-QA | MMSU	Megrez-3B-Omni (Infinigence)	3.50 | 2.95 | 25.95 | 27.03
	Lyra-Base (Zhong et al., 2024)	3.85 | 3.50 | 38.25 | 49.74
	MiniCPM-o (Yao et al., 2024)	4.42 | **4.15** | 50.72 | 54.78
	Baichuan-Omni-1.5 (Li et al., 2025)	4.50 | 4.05 | 43.40 | 57.25
	Qwen2-Audio (Chu et al., 2024a)	3.74 | 3.43 | 35.71 | 35.72
	Qwen2.5-Omni-7B	4.49 | 3.93 | **55.71** | **61.32**
VoiceBench	Ultravox-v0.4.1-LLaMA-3.1-8B	65.27 | **66.88** | 98.46 | 71.45
OpenBookQA | IFEval |	MERaLiON (He et al., 2024)	27.23 | 62.93 | 94.81 | 62.91
AdvBench | Avg	Megrez-3B-Omni (Infinigence)	28.35 | 25.71 | 87.69 | 46.25
	Lyra-Base (Zhong et al., 2024)	72.75 | 36.28 | 59.62 | 57.66
	MiniCPM-o (Yao et al., 2024)	78.02 | 49.25 | 97.69 | 71.69
	Baichuan-Omni-1.5 (Li et al., 2025)	74.51 | 54.54 | 97.31 | 71.14
	Qwen2-Audio (Chu et al., 2024a)	49.45 | 26.33 | 96.73 | 55.35
	Qwen2.5-Omni-7B	**81.10** | 52.87 | **99.42** | **74.12**

Table 4: Performance of Qwen2.5-Omni and other models for Chatting, * means that approximately 90% of text instructions suitable for speech are used.

Datasets	Qwen2-7B (text)	Qwen2-Audio	Qwen2.5-Omni-7B
MMLU*	69.3	33.2	65.6
CEval*	78.4	38.6	61.1
IFEval*	53.3	15.6	41.7
GSM8K*	82.3	18.4	85.4
Math23K*	92.3	23.0	87.1
Math401*	75.5	20.4	62.2

5.1.3 Performance of Image → Text

To comprehensively evaluate the capabilities on Image → Text, we compare Qwen2.5-Omni with the recent state-of-the-art large vision language model Qwen2.5-VL-7B and other best-performing omni models. As illustrated in Table 5, Qwen2.5-Omni demonstrates comparable performance to Qwen2.5-VL-7B, and attains better results on MMMU, MathVision, MMBench-V1.1-EN, TextVQA, DocVQA and ChartQA than any other open-sourced omni models. Additionally, Qwen2.5-Omni also surpasses GPT-4o-mini on most benchmarks. These results reveal the excellent capability of our model on image understanding.

Table 5: **Image → Text performance of 7B+ models and Qwen2.5-Omni**

Datasets	GPT-4o-mini	Qwen2.5-VL-7B	Other Best	Qwen2.5-Omni-7B
College-level Problems				
MMMU_{val}	**60.0**	58.6	53.9 (Li et al., 2025)	59.2
$\text{MMMU-Pro}_{overall}$	37.6	**38.3**	-	36.6
Mathematical				
$\text{MathVista}_{testmini}$	52.5	68.2	**71.9** (Yao et al., 2024)	67.9
MathVision_{full}	-	**25.1**	23.1 (Yao et al., 2024)	25.0
General Visual Question Answering				
$\text{MMBench-V1.1-EN}_{test}$	76.0	**82.6**	80.5 (Yao et al., 2024)	81.8
MMVet_{turbo}	66.9	67.1	**67.5** (Yao et al., 2024)	66.8
MMStar	54.8	63.9	**64.0** (Yao et al., 2024)	**64.0**
MME_{sum}	2003	2347	**2372** (Yao et al., 2024)	2340
MuirBench	-	**59.6**	-	59.2
$\text{CRPE}_{relation}$	-	76.4	-	**76.5**
RealWorldQA_{avg}	-	68.5	**71.9** (Infinigence)	70.3
$\text{MME-RealWorld}_{en}$	-	57.4	-	**61.6**
MM-MT-Bench	-	**6.3**	-	6.0
OCR-related Tasks				
AI2D	-	83.9	**85.8** (Yao et al., 2024)	83.2
TextVQA_{val}	-	**84.9**	83.2 (Li et al., 2025)	84.4
DocVQA_{test}	-	**95.7**	93.5 (Yao et al., 2024)	95.2
$\text{ChartQA}_{test\ Avg}$	-	**87.3**	84.9 (Li et al., 2025)	85.3
OCRBench_V2_{en}	-	56.3	-	**57.8**

Table 6: **Grounding performance of Qwen2.5-Omni and other models**

Datasets	Gemini 1.5 Pro	Grounding DINO	Qwen2.5-VL-7B	Qwen2.5-Omni-7B
Refcoco_{val}	73.2	**90.6**	90.0	90.5
Refcoco_{textA}	72.9	93.2	92.5	**93.5**
Refcoco_{textB}	74.6	**88.2**	85.4	86.6
Refcoco+_{val}	62.5	**88.2**	84.2	85.4
Refcoco+_{textA}	63.9	89.0	89.1	**91.0**
Refcoco+_{textB}	65.0	75.9	76.9	**79.3**
Refcocog_{val}	75.2	86.1	87.2	**87.4**
Refcocog_{test}	76.2	87.0	87.2	**87.9**
ODinW	36.7	**55.0**	37.3	42.2
PointGrounding	-	-	**67.3**	66.5

For visual grounding, we compare Qwen2.5-Omni with Qwen2.5-VL-7B and other leading LVLMs including Gemini and Grounding-DINO (Liu et al., 2024). As illustrated in Table 6, our model outperforms other models across most benchmarks from box-grounding to point-grounding and achieves a good performance of 42.2mAP on open-vocabulary object detection, which reveals the strong visual grounding capability of our model.

5.1.4 Performance of Video → Text

Similar to Image→Text, we compare Qwen2.5-Omni with Qwen2.5-VL-7B and other omni models. As shown in Table 7, Qwen2.5-Omni outperforms all other state-of-the-art open-sourced omni models and GPT-4o-Mini, and attains better or competitive results compared to Qwen2.5-VL-7B, which demonstrates the superior performance on video understanding.

5.1.5 Performance of Multimodality→Text

As shown in Table 8, Qwen2.5-Omni achieves state-of-the-art performance on OmniBench, surpassing other Omni models by a large margin, which demonstrates the superiority of our model in multimodality understanding.

Table 7: **Video → text performance of 7B+ models and Qwen2.5-Omni**

Datasets	GPT-4o-mini	Qwen2.5-VL-7B	Other Best	Qwen2.5-Omni-7B
		Video Understanding		
Video-MME$_{w/o\ sub}$	64.8	**65.1**	63.9 (Yao et al., 2024)	64.3
Video-MME$_{w\ sub}$	-	71.6	67.9 (Yao et al., 2024)	**72.4**
MVBench	-	69.6	67.2 (Zhong et al., 2024)	**70.3**
EgoSchema$_{test}$	-	65.0	63.2 (Zhong et al., 2024)	**68.6**

Table 8: Multimodality → Text performance of State-of-the-art and Qwen2.5-Omni

Datasets	Model	Performance
	Multimodal Understanding	
OmniBench *Speech ∣ Sound Event ∣ Music ∣ Avg*	Gemini-1.5-Pro (Team et al., 2024)	42.67% ∣ 42.26% ∣ 46.23% ∣ 42.91%
	MIO-Instruct (Wang et al., 2024g) (7B)	36.96% ∣ 33.58% ∣ 11.32% ∣ 33.80%
	AnyGPT (7B) (Zhan et al., 2024)	17.77% ∣ 20.75% ∣ 13.21% ∣ 18.04%
	video-SALMONN (13B) (Sun et al., 2024)	34.11% ∣ 31.70% ∣ **56.60**% ∣ 35.64%
	UnifiedIO2-xlarge (3.2B) (Lu et al., 2024a)	39.56% ∣ 36.98% ∣ 29.25% ∣ 38.00%
	UnifiedIO2-xxlarge (6.8B) (Lu et al., 2024a)	34.24% ∣ 36.98% ∣ 24.53% ∣ 33.98%
	MiniCPM-o (Yao et al., 2024)	- ∣ - ∣ - ∣ 40.5%
	Baichuan-Omni-1.5 (Li et al., 2025)	- ∣ - ∣ - ∣ 42.9%
	Qwen2.5-Omni-7B	**55.25**% ∣ **60.00**% ∣ 52.83% ∣ **56.13**%

5.2 Evaluation of X→Speech

In this section, we evaluate the speech generation capabilities of Qwen2.5-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 two aspects: Zero-shot and Single-Speaker speech generation capabilities.

- **Zero-Shot Speech Generation** We assessed the content consistency (WER) and speaker similarity (SIM) of our model in zero-shot speech generation on SEED (Anastassiou et al., 2024).
- **Single-Speaker Speech Generation** We assessed the stability of our speaker fine-tuned model on the SEED (Anastassiou et al., 2024), and evaluated the subjective naturalness (NMOS) of the generated speech on a self-created dataset.

5.2.1 Evaluation of Zero-Shot Speech Generation.

We compared the Qwen2.5-Omni with state-of-the-art zero-shot TTS systems. As shown in Table 9, Qwen2.5-Omni demonstrates highly competitive performance, highlighting its robust speech understanding and generation capabilities developed through in-context learning (ICL). Additionally, after reinforcement learning (RL) optimization, Qwen2.5-Omni showed significant improvements in generation stability, with marked reductions in attention misalignment, pronunciation errors, and inappropriate pauses on the challenging test-hard dataset.

5.2.2 Evaluation of Single-Speaker Speech Generation.

We compared the Qwen2.5-Omni model before and after speaker fine-tuning, as well as with human recordings. As shown in Table 10, the speaker-finetuned Qwen2.5-Omni more precisely captured the nuanced prosodic styles of the target speakers while preserving the foundational stability provided by the base model, achieving performance that approaches human-level quality across both subjective and objective metrics.

6 Conclusion

Qwen2.5-Omni is a unified model designed to understand and generate multiple modalities, including text and real-time speech. To enhance video integration, we've introduced a new positional embedding method called TMRoPE, which aligns audio and video timing. Our Thinker-Talker framework supports real-time speech generation while minimizing interference across different modalities. Additionally, we employ techniques such as block-wise audio/vision encoding and a sliding window mechanism for code-to-wav generation. This innovative model excels in complex audio-visual interactions and emotional

Table 9: Zero-Shot Speech Generation

Datasets	Model	Performance
	Content Consistency	
SEED *test-zh ∣ test-en ∣ test-hard*	Seed-TTS$_{ICL}$ (Anastassiou et al., 2024)	1.11 ∣ 2.24 ∣ 7.58
	Seed-TTS$_{RL}$ (Anastassiou et al., 2024)	**1.00** ∣ 1.94 ∣ **6.42**
	MaskGCT (Wang et al., 2024e)	2.27 ∣ 2.62 ∣ 10.27
	E2 TTS (Eskimez et al., 2024)	1.97 ∣ 2.19 ∣ -
	F5-TTS (Chen et al., 2024c)	1.56 ∣ **1.83** ∣ 8.67
	CosyVoice 2 (Du et al., 2024)	1.45 ∣ 2.57 ∣ 6.83
	CosyVoice 2-S (Du et al., 2024)	1.45 ∣ 2.38 ∣ 8.08
	Qwen2.5-Omni-7B$_{ICL}$	1.70 ∣ 2.72 ∣ 7.97
	Qwen2.5-Omni-7B$_{RL}$	1.42 ∣ 2.33 ∣ 6.54
	Speaker Similarity	
SEED *test-zh ∣ test-en ∣ test-hard*	Seed-TTS$_{ICL}$ (Anastassiou et al., 2024)	0.796 ∣ 0.762 ∣ 0.776
	Seed-TTS$_{RL}$ (Anastassiou et al., 2024)	**0.801** ∣ **0.766** ∣ **0.782**
	MaskGCT (Wang et al., 2024e)	0.774 ∣ 0.714 ∣ 0.748
	E2 TTS (Eskimez et al., 2024)	0.730 ∣ 0.710 ∣ -
	F5-TTS (Chen et al., 2024c)	0.741 ∣ 0.647 ∣ 0.713
	CosyVoice 2 (Du et al., 2024)	0.748 ∣ 0.652 ∣ 0.724
	CosyVoice 2-S (Du et al., 2024)	0.753 ∣ 0.654 ∣ 0.732
	Qwen2.5-Omni-7B$_{ICL}$	0.752 ∣ 0.632 ∣ 0.747
	Qwen2.5-Omni-7B$_{RL}$	0.754 ∣ 0.641 ∣ 0.752

Table 10: Single-Speaker Speech Generation

Datasets	Model	Performance
	Content Consistency	
SEED *test-zh ∣ test-en ∣ test-hard*	Human	**1.25** ∣ 2.14 ∣ -
	Qwen2.5-Omni $_{RL}$	1.30 ∣ 2.33 ∣ 6.54
	Qwen2.5-Omni $_{Speaker A}$	1.29 ∣ 1.86 ∣ 6.59
	Qwen2.5-Omni $_{Speaker B}$	1.37 ∣ 1.89 ∣ 7.25
	Qwen2.5-Omni $_{Speaker C}$	1.30 ∣ 2.13 ∣ **6.43**
	Qwen2.5-Omni $_{Speaker D}$	1.28 ∣ **1.83** ∣ 7.16
	Naturalness	
NMOS *zh ∣ en*	Human	**4.51** ∣ -
	Qwen2.5-Omni $_{Speaker A}$	4.46 ∣ 4.51
	Qwen2.5-Omni $_{Speaker B}$	**4.51** ∣ **4.62**
	Qwen2.5-Omni $_{Speaker C}$	4.50 ∣ 4.60
	Qwen2.5-Omni $_{Speaker D}$	4.48 ∣ 4.58

context in speech dialogues. Comprehensive evaluations show that Qwen2.5-Omni outperforms similarly sized single-modality models, particularly in following voice commands, and achieves state-of-the-art performance in multi-modal tasks.

In the development of the model, we have identified several critical issues that have often been overlooked by researchers in previous academic studies, such as video OCR and audio-video collaborative understanding. Addressing these challenges necessitates collaboration between the academic and industrial sectors, particularly in building comprehensive evaluation benchmarks and research datasets. We believe Qwen2.5-Omni represents a significant advancement toward artificial general intelligence (AGI). Our future goals include developing a more robust and faster model with expanded output capabilities across various modalities like images, videos, and music.

7 Authors

Core Contributors: Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, Junyang Lin

Contributors[1]: An Yang, Anfeng Li, Baosong Yang, Bei Chen, Bin Lin, Binyuan Hui, Bo Zheng, Bowen Yu, Cheng Chen, Chengen Huang, Chenhan Yuan, Chengyuan Li, Daren Chen, Dayiheng Liu, Dake Guo, Fan Zhou, Fei Huang, Guangdong Zhou, Hang Zhang, Haoran Lian, Haoyang Zhang, He Wang, Humen Zhong, Jian Yang, Jiandong Jiang, Jianhong Tu, Jianqiang Wan, Jianyuan Zeng, Jun Tang, Jianwei Zhang, Jianxin Yang, Jianyuan Zeng, Jing Zhou, Jingren Zhou, Kexin Yang, Lei Xie, Linhan Ma, Lingchen Meng, Le Yu, Mei Li, Miao Hong, Mingfeng Xue, Mingkun Yang, Mingze Li, Na Ni, Pei Zhang, Peiyang Zhang, Peng Liu, Peng Wang, Peng Zhang, Pengfei Wang, Rui Hu, Rui Men, Qiuyue Wang, Qing Fu, Shixuan Liu, Sibo Song, Siqi Zhang, Song Chen, Tianyi Tang, Tao He, Ting He, Wenbin Ge, Wei Ding, Xiaodong Deng, Xinyao Niu, Xipin Wei, Xue Bin, Xuejing Liu, Xingzhang Ren, Xuancheng Ren, Yang Liu, Yanpeng Li, Yang Liu, Yang Su, Yichang Zhang, Yuqiong Liu, Yuanjun Lv, Yuanzhi Zhu, Yuxuan Cai, Zeyu Cui, Zheng Li, Zhenru Zhang, Zihan Qiu, Zhaohai Li, Zhibo Yang, Zhipeng Zhou, Zhiyuan Zhu

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.

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.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. *CoRR*, abs/2108.07732, 2021.

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.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chuang Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, et al. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*, 2024.

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.

---

[1]Alphabetical order.

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.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Joshua Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *CoRR*, abs/2107.03374, 2021.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, Yabin Li, Xiang Lv, Jiaqing Liu, Haoneng Luo, Bin Ma, Chongjia Ni, Xian Shi, Jialong Tang, Hui Wang, Hao Wang, Wen Wang, Yuxuan Wang, Yunlan Xu, Fan Yu, Zhijie Yan, Yexin Yang, Baosong Yang, Xian Yang, Guanrou Yang, Tianyu Zhao, Qinglin Zhang, Shiliang Zhang, Nan Zhao, Pei Zhang, Chong Zhang, and Jinren Zhou. Minmo: A multimodal large language model for seamless voice interaction, 2025. URL https://arxiv.org/abs/2501.06282.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*, 2022.

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.

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

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, 2023b.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *CoRR*, abs/2407.10759, 2024a.

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*, 2024b.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168, 2021.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv:2305.06500*, 2023.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, David Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, et al. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*, 2024.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. *CoRR*, 2023.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*, 2023.

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.

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, 2024a.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv:2407.21783*, 2024b.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: learning audio concepts from natural language supervision. abs/2206.04769, 2022.

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, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv:2306.13394*, 2023.

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*, 2024a.

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

Joshua P Gardner, Simon Durand, Daniel Stoller, and Rachel M Bittner. Llark: A multimodal instruction-following language model for music. In *Forty-first International Conference on Machine Learning*, 2023.

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.

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, et al. Av-odyssey bench: Can your multimodal llms really understand audio-visual information? *arXiv preprint arXiv:2412.02611*, 2024.

Yingxu He, Zhuohan Liu, Shuo Sun, Bin Wang, Wenyu Zhang, Xunlong Zou, Nancy F Chen, and Ai Ti Aw. Meralion-audiollm: Technical report. *arXiv preprint arXiv:2412.09818*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*, 2021b.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv:2302.14045*, 2023.

Infinigence. Infini-megrez-omni. URL https://github.com/infinigence/Infini-Megrez-Omni.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. *CoRR*, abs/2403.07974, 2024.

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

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

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *ICLR 2023*.

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.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024a.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training, 2022. URL https://arxiv.org/abs/2112.03857.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR 2023*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv:2310.03744*, 2023a.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL https://arxiv.org/abs/2303.05499.

Yuan Liu, Haodong Duan, Bo Li Yuanhan Zhang, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*, 2023c.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26439–26455, 2024a.

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*, 2024b.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *NeurIPS*, 2023.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions, 2016. URL https://arxiv.org/abs/1511.02283.

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.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *WACV*, 2021.

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

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

OpenAI. Gpt-4v(ision) system card, 2023. URL https://openai.com/research/gpt-4v-system-card.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, 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.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. SALMONN: towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv:2302.13971*, 2023a.

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*, 2023b.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv:2309.00916*, 2023a.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding, 2024a. URL https://arxiv.org/abs/2406.09411.

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

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

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv:2311.03079*, 2023b.

Weiyun Wang, Yiming Ren, Haowen Luo, Tiantong Li, Chenxiang Yan, Zhe Chen, Wenhai Wang, Qingyun Li, Lewei Lu, Xizhou Zhu, Yu Qiao, and Jifeng Dai. The all-seeing project v2: Towards general relation comprehension of the open world, 2024d. URL https://arxiv.org/abs/2402.19474.

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*, 2024e.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *CoRR*, abs/2406.01574, 2024f.

Zekun Wang, King Zhu, Chunpu Xu, Wangchunshu Zhou, Jiaheng Liu, Yibo Zhang, Jiashuo Wang, Ning Shi, Siyu Li, Yizhi Li, et al. Mio: A foundation model on multimodal tokens. *arXiv preprint arXiv:2409.17692*, 2024g.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Siddartha Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. LiveBench: A challenging, contamination-free LLM benchmark. *CoRR*, abs/2406.19314, 2024.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, and Yu Wu. On decoder-only architecture for speech-to-text and large language model integration. abs/2307.03917, 2023.

X.AI. Grok-1.5 vision preview., 2024. URL https://x.ai/blog/grok-1.5v.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024b.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*, 2024.

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.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*, 2024.

Zhisheng Zhong, Chengyao Wang, Yuqi Liu, Senqiao Yang, Longxiang Tang, Yuechen Zhang, Jingyao Li, Tianyuan Qu, Yanwei Li, Yukang Chen, et al. Lyra: An efficient and speech-centric framework for omni-cognition. *arXiv preprint arXiv:2412.09501*, 2024.

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.