本文由 AINLP 公众号整理翻译，更多 LLM 资源请扫码关注!

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.

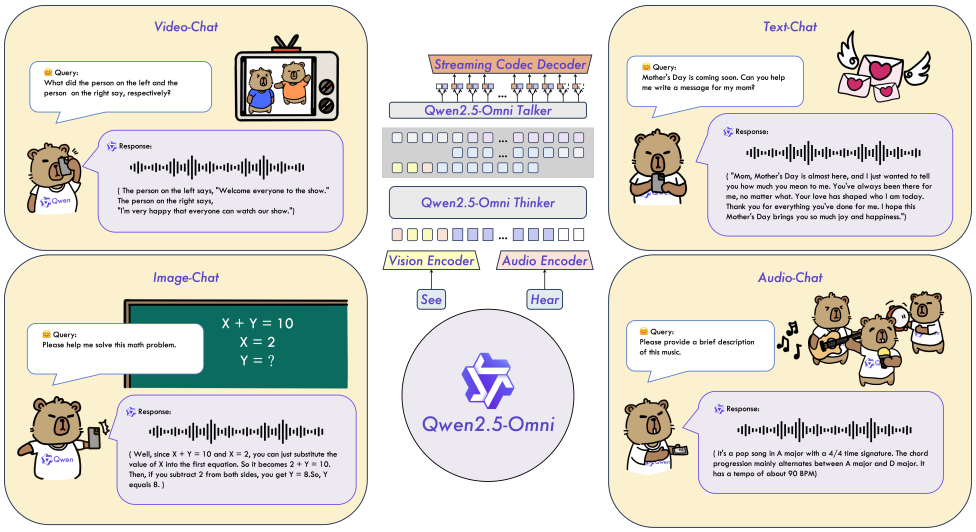

Figure 1: Qwen2.5-Omni is a unified end-to-end model capable of processing multiple modalities, such as text, audio, image and video, and generating real-time text or speech response. Based on these features, Qwen2.5-Omni supports a wide range of tasks, including but not limited to voice dialogue, video dialogue, and video reasoning.

Qwen2.5-Omni技术报告

Qwen团队

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

摘要

在本报告中，我们介绍了Qwen2.5-Omni，这是一种端到端的多模态模型，旨在感知多种模态，包括文本、图像、音频和视频，同时以流式方式生成文本和自然语音响应。为了实现多模态信息输入的流式处理，音频和视觉编码器都采用了块处理方法。这一策略有效地解耦了对长序列多模态数据的处理，将感知责任分配给多模态编码器，并将扩展序列的建模任务委托给大型语言模型。这种分工通过共享注意力机制增强了不同模态的融合。为了将视频输入的时间戳与音频同步，我们以交错的方式顺序组织音频和视频，并提出了一种新颖的位置嵌入方法，称为TMRoPE（时间对齐多模态RoPE）。为了同时生成文本和语音，同时避免两种模态之间的干扰，我们提出了Thinker-Talker架构。在该框架中，Thinker作为一个大型语言模型，负责文本生成，而Talker是一个双轨自回归模型，直接利用Thinker的隐藏表示生成音频标记作为输出。Thinker和Talker模型均设计为以端到端的方式进行训练和推理。为了以流式方式解码音频标记，我们引入了一种滑动窗口DiT，限制感受野，旨在减少初始包延迟。Qwen2.5-Omni与同样规模的Qwen2.5-VL相当，并且在性能上优于Qwen2-Audio。此外，Qwen2.5-Omni在Omni-Bench等多模态基准测试中达到了最先进的性能。值得注意的是，Qwen2.5-Omni在端到端语音指令跟随方面的表现与其在文本输入方面的能力相当，这一点在MMLU和GSM8K等基准测试中得到了证明。至于语音生成，Qwen2.5-Omni的流式Talker在鲁棒性和自然性方面优于大多数现有的流式和非流式替代方案。

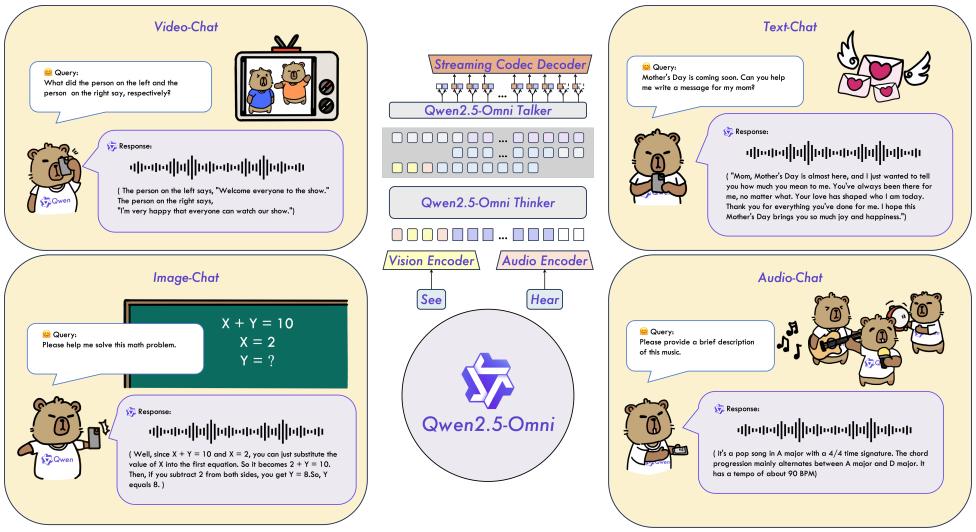

图1：Qwen2.5-Omni是一个统一的端到端模型，能够处理多种模态，如文本、音频、图像和视频，并生成实时文本或语音响应。基于这些特性，Qwen2.5-Omni支持广泛的任务，包括但不限于语音对话、视频对话和视频推理。

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.

1 引言

在日常生活中，人类能够同时感知周围的视觉和听觉信息。在通过大脑处理这些信息后，他们通过书写、发声或使用工具（以及身体动作）表达反馈，从而与世界上各种生物进行信息交流，展现出智能。近年来，通用人工智能变得越来越显著，这在很大程度上得益于大型语言模型（LLMs）的进步（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）。这些模型在大量文本数据上进行训练，代表了人类创造的高层次离散表示，展示了解决复杂问题和快速学习的能力。此外，在理解领域，语言-音频-语言模型（LALMs）（OpenAI, 2024; Tang et al., 2024; Chu et al., 2023b; 2024b）和语言-视觉-语言模型（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）帮助LLMs进一步以端到端的方式扩展听觉和视觉能力。然而，以端到端的方式有效统一所有这些不同的理解模态，尽可能利用大量数据，并提供类似于人类交流的文本和语音流的响应，仍然是一个重大挑战。

开发统一且智能的全模型需要仔细考虑几个关键因素。首先，实施一种系统的方法来联合训练各种模态，包括文本、图像、视频和音频，以促进它们之间的相互增强是至关重要的。这种对齐对于视频内容尤为重要，因为需要同步音频和视觉信号的时间方面。其次，必须管理不同模态输出之间的潜在干扰，确保文本和语音标记等输出的训练过程不会相互干扰。最后，需要探索能够实时理解多模态信息的架构设计，并允许高效的音频输出流，从而减少初始延迟。

在本报告中，我们介绍了Qwen2.5-Omni，这是一种统一的单一模型，能够同时以流式格式处理多种模态并生成文本和自然语音响应。为了解决第一个挑战，我们提出了一种新颖的位置嵌入方法，称为TMRoPE（时间对齐多模态RoPE）。我们将这些音频和视频帧组织成交错结构，以按时间顺序表示视频序列。针对第二个挑战，我们提出了Thinker-Talker架构，其中Thinker负责文本生成，而Talker专注于生成流式语音标记。Talker直接从Thinker接收高级表示。这一设计灵感来源于人类利用不同器官同时协调产生各种信号的方式，这些信号通过相同的神经网络进行协调。因此，Thinker-Talker架构是端到端联合训练的，每个组件专注于生成不同的信号。为了应对与流式处理相关的挑战，并促进实时理解多模态信号所需的预填充，我们通过采用块级流式处理方法对所有多模态编码器进行了修改。为了支持流式语音生成，我们实现了一个双轨自回归模型，该模型生成语音标记，同时还有一个DiT模型将这些标记转换为波形，从而实现流式音频生成并最小化初始延迟。该设计旨在使模型能够实时处理多模态信息并有效执行预填充，从而实现文本和语音信号的并发生成。

Qwen2.5-Omni与同样大小的Qwen2.5-VL（Wang et al., 2024c）相当，并在图像和音频能力上分别超越了Qwen2-Audio（Chu et al., 2024b）。此外，Qwen2.5-Omni在多模态基准测试中实现了最先进的性能，例如OmniBench（Li et al., 2024b）和AV-Odyssey Bench（Gong et al., 2024）。值得注意的是，Qwen2.5-Omni在端到端语音指令跟随中的表现与其文本输入的能力相当，这在MMLU（Hendrycks et al., 2021a）和GSM8K（Cobbe et al., 2021）等基准测试中得到了证明。至于语音生成，Qwen2.5-Omni在seed-tts-eval（Anastassiou et al., 2024）测试集zh、测试集en和测试集hard上分别达到了1.42%、2.33%和6.54%的WER，超越了MaskGCT（Wang et al., 2024e）和CosyVoice 2（Du et al., 2024）。

Qwen2.5-Omni的主要特点可以总结为：

- 我们介绍Qwen2.5-Omni，这是一个统一模型，可以感知所有模态，并以流式方式同时生成文本和自然语音响应。
- 我们提出了一种新颖的位置嵌入算法，称为 TMRoPE，它明确地结合了时间信息，以同步音频和视频。
- 我们提出了思考者-谈话者架构，以促进实时理解和语音生成。

- 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

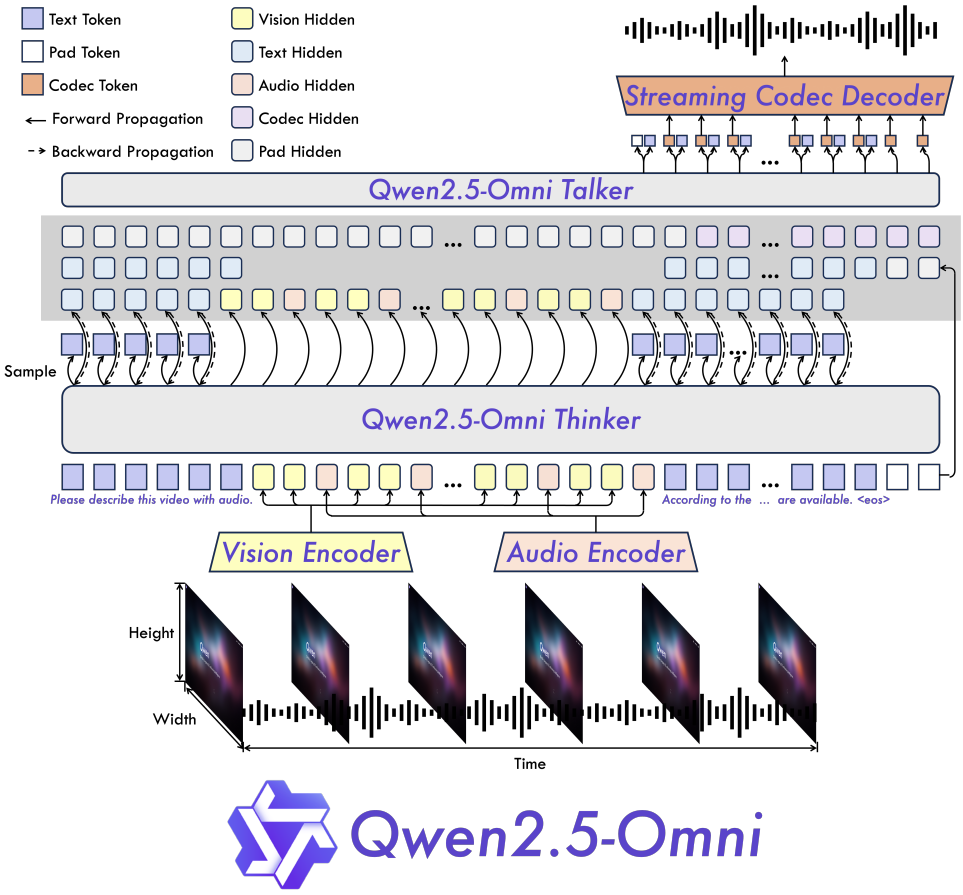

Figure 2: The overview of Qwen2.5-Omni. Qwen2.5-Omni adpots the Thinker-Talker architecture. Thinker is tasked with text generation while Talker focuses on generating streaming speech tokens by receives high-level representations directly from Thinker.

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.

- Qwen2.5-Omni在与同样规模的单模态模型进行基准测试时，展示了在所有模态上的强大性能。它显著增强了执行语音命令的能力，达到与纯文本输入相当的性能水平。对于涉及整合多种模态的任务，例如在OmniBench（Li et al., 2024b）中评估的任务，Qwen2.5-Omni达到了最先进的性能。值得注意的是，Qwen2.5-Omni在seed-tts-eval（Anastassiou et al., 2024）上表现出色，展示了强大的语音生成能力。

2 架构

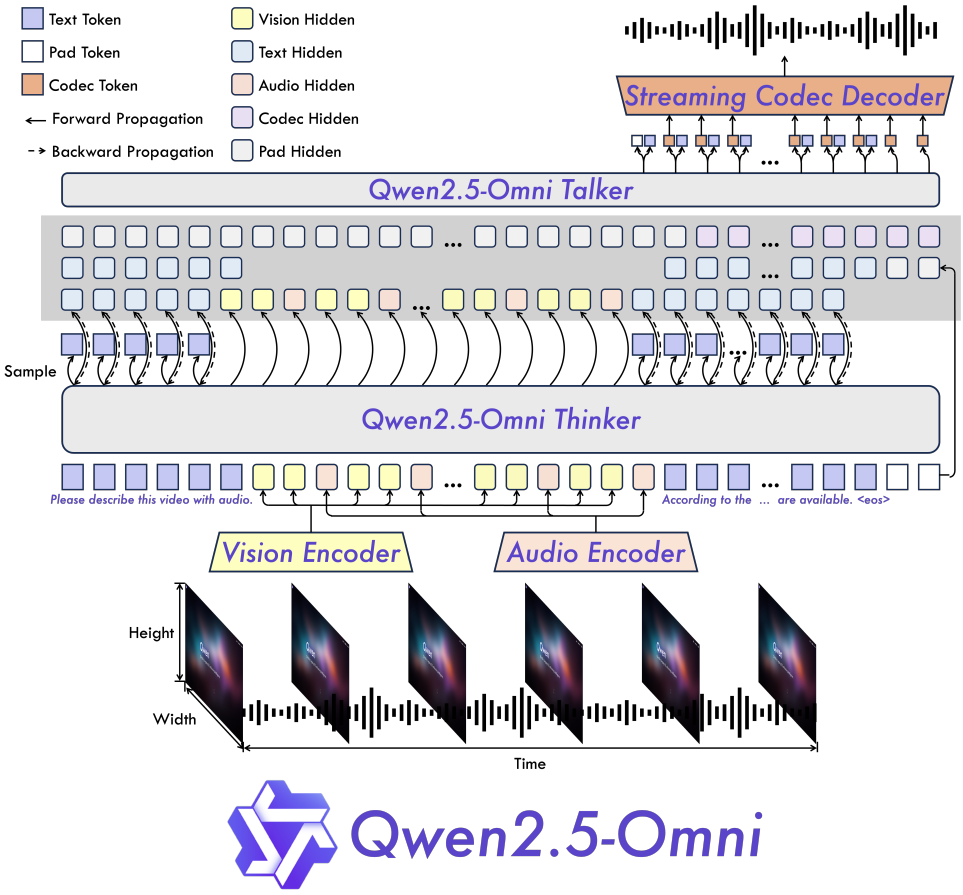

图2：Qwen2.5-Omni的概述。Qwen2.5-Omni采用了Thinker-Talker架构。Thinker负责文本生成，而Talker则专注于通过直接接收来自Thinker的高层表示来生成流式语音令牌。

2.1 概述

如图2所示，Qwen2.5-Omni采用了Thinker-Talker架构。Thinker像大脑一样，负责处理和理解来自文本、音频和视频模态的输入，生成高级表示和相应的文本。Talker则像人类的嘴，流式接收Thinker生成的高级表示和文本，并流畅地输出离散的语音标记。Thinker是一个Transformer解码器，配有音频和图像的编码器，以促进信息提取。相比之下，Talker被设计为双轨自回归Transformer解码器架构，受到Mini-Omni（Xie & Wu, 2024）的启发。在训练和推理过程中，Talker直接接收来自Thinker的高维表示，并共享Thinker的所有历史上下文信息。因此，整个架构作为一个统一的单一模型运行，实现端到端的训练和推理。

在接下来的章节中，我们首先介绍 Qwen2.5-Omni 如何感知各种输入信号，并提出我们新提出的位置编码算法 TMRoPE。随后，文本和语音生成的细节将被呈现。最后，我们强调在理解和生成模块中所做的改进，以促进高效的流式推理。

2.2 Perceivation

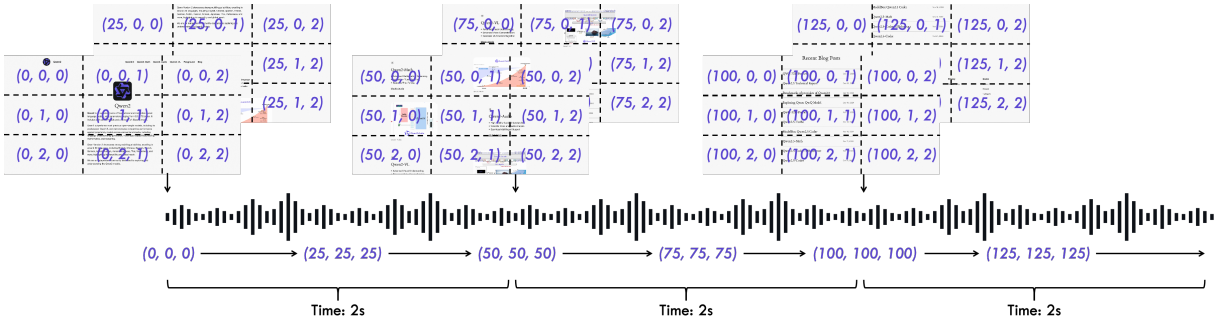

Figure 3: An illustration of Time-aligned Multimodal RoPE (TMRoPE).

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

2.2 感知

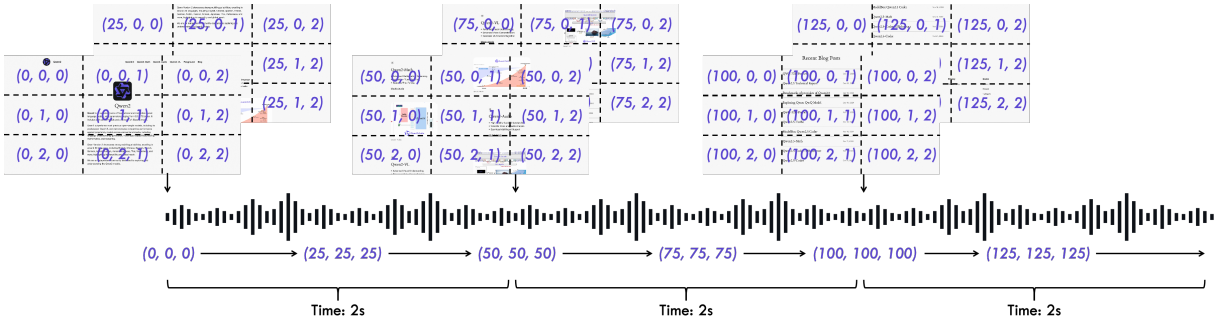

图3：时间对齐多模态RoPE（TMRoPE）的示意图。

文本、音频、图像和视频（不含音频）。思考者通过将文本、音频、图像和视频（不含音轨）转换为一系列隐藏表示进行处理。对于文本的标记化，我们使用Qwen的标记器（Yang et al., 2024a），该标记器应用字节级字节对编码，词汇表包含151,643个常规标记。关于音频输入和视频中的音频，我们将其重采样至16kHz的频率，并将原始波形转换为128通道的梅尔谱图，窗口大小为25ms，跳跃大小为10ms。我们采用Qwen2-Audio（Chu et al., 2024b）中的音频编码器，使每帧音频表示大致对应于原始音频信号的40ms片段。此外，我们使用Qwen2.5-VL（Bai et al., 2025）中的视觉编码器，该编码器基于大约675百万参数的视觉变换器（ViT）模型，使其能够有效处理图像和视频输入。视觉编码器采用混合训练方案，结合图像和视频数据，确保在图像理解和视频理解方面的熟练度。为了尽可能完整地保留视频信息，同时适应音频采样率，我们使用动态帧率对视频进行采样。此外，为了保持一致性，每个图像被视为两个相同的帧。

视频和TMRoPE。我们提出了一种用于音频和视频的时间交错算法，以及一种新颖的位置编码方法。如图3所示，TMRoPE编码了多模态输入的三维位置信息，这就是多模态旋转位置嵌入（M-RoPE）（Bai等，2023b），具有绝对时间位置。这是通过将原始旋转嵌入分解为三个组成部分来实现的：时间、高度和宽度。对于文本输入，这些组件使用相同的位置ID，使得M-RoPE在功能上等同于1D-RoPE。同样，对于音频输入，我们也使用相同的位置ID，并引入绝对时间位置编码，其中一个时间ID对应于40毫秒。

在处理图像时，每个视觉标记的时间 ID 保持不变，而根据标记在图像中的位置，为高度和宽度组件分配不同的 ID。当输入为带音频的视频时，音频仍然以每帧 40 毫秒的相同位置 ID 进行编码，而视频被视为一系列图像，每帧的时间 ID 递增，同时高度和宽度组件遵循与图像相同的 ID 分配模式。由于视频的帧率不是固定的，我们根据每帧对应的实际时间动态调整帧之间的时间 ID，以确保一个时间 ID 对应 40 毫秒。在模型输入包含多种模态的场景中，每种模态的位置编号通过将前一种模态的最大位置 ID 增加一来初始化。TMRoPE 增强了位置编码信息建模，最大化了各种模态的整合，使 Qwen2.5-Omni 能够同时理解和分析来自多种模态的信息。

在将位置信息纳入每种模态后，我们按顺序排列表示。为了使模型能够同时接收视觉和听觉信息，如图3所示，我们为带音频的视频设计了一种特殊的方法，称为时间交错方法，该方法根据实际时间每2秒将带音频的视频表示分段。然后，我们在2秒内将视觉表示放在前面，将音频表示放在后面，交错带音频的视频表示。

2.3 生成

文本。文本是由Thinker直接生成的。文本生成的逻辑与广泛使用的LLM所采用的逻辑基本相同，这些LLM通过基于自回归采样生成文本。

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.

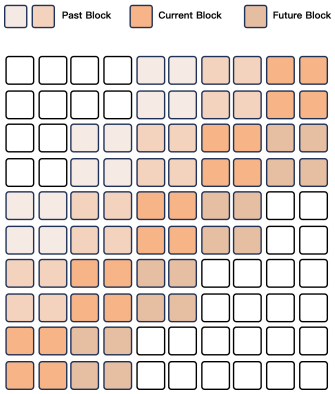

Figure 4: An illustration of sliding window block attention mechanism in DiT for codec to wav generation.

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

词汇表上的概率分布。生成过程可能会结合诸如重复惩罚和顶级采样等技术，以增强其多样性。

语音。说话者接收来自思考者的高层次表示和文本标记的嵌入。高维表示和离散采样标记的整合在此背景下至关重要。作为一种流式算法，语音生成必须在整个文本完全生成之前预测内容的语气和态度。思考者提供的高维表示隐含地传达了这些信息，从而使流式生成过程更加自然。此外，思考者的表示主要在表示空间中表达语义相似性，而不是语音相似性。因此，即使是语音上截然不同的词也可能具有非常相似的高层次表示，这就需要输入采样的离散标记以消除这种不确定性。

我们设计了一种高效的语音编解码器，名为 *qwen-tts-tokenizer*。*qwen-tts-tokenizer* 高效地表示语音的关键信息，并可以通过因果音频解码器流式解码为语音。在接收到信息后，发言者开始自回归地生成音频标记和文本标记。语音的生成不需要与文本进行词级和时间戳级的对齐。这大大简化了对训练数据和推理过程的要求。

2.4 流媒体设计

在音频和视频交互的流媒体环境中，初始数据包延迟是系统流媒体性能的一个关键指标。这个延迟受到几个因素的影响：1）多模态信息输入处理造成的延迟；2）从接收到第一个文本输入到输出第一个语音令牌之间的延迟；3）将第一段语音转换为音频的延迟；以及4）架构本身的固有延迟，这与模型大小、计算FLOPs和其他因素有关。本文将随后讨论在这四个维度上减少这些延迟所做的算法和架构改进。

支持预填充。分块预填充是一种在现代推理框架中广泛使用的机制。为了在模态交互中支持它，我们修改了音频和视觉编码器，以支持沿时间维度的块状注意力。具体而言，音频编码器从对整个音频的全注意力改为对每个2秒块进行注意力处理。视觉编码器利用闪存注意力进行高效的训练和推理，并使用一个简单的MLP层将相邻的2×2个标记合并为一个单一标记。补丁大小设置为14，这允许不同分辨率的图像被打包成一个序列。

流媒体编解码生成。为了促进音频的流媒体传输，特别是对于扩展序列，我们提出了一种滑动窗口块注意机制，该机制限制当前标记对有限上下文的访问。具体而言，我们利用了Flow-Matching（Lipman等）DiT模型。输入代码通过Flow-Matching转换为梅尔谱图，然后使用修改后的BigVGAN（Lee等）将生成的梅尔谱图重构回波形。

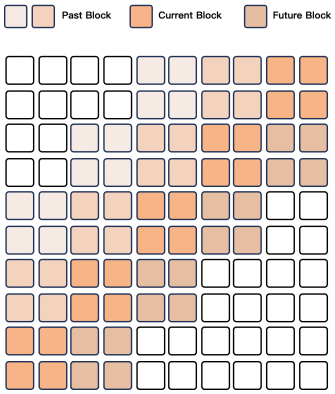

图4：DiT中滑动窗口块注意力机制的示意图，用于编解码器到wav的生成。

如图4所示，为了从代码生成波形，我们将相邻的代码分组为块，并将这些块用于我们的注意力掩码。我们将DiT的感受野限制为4个块，包括2个块的回顾和1个块的前瞻。在解码过程中，我们使用流匹配以块的形式生成梅尔谱，确保每个代码块可以访问必要的上下文块。这种方法通过保持上下文信息来增强流输出的质量。我们还使用这种逐块的方法来处理BigVGAN的固定感受野，以促进流波形生成。

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 incorporat 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.

3 预训练

Qwen2.5-Omni由三个训练阶段组成。在第一阶段，我们锁定LLM参数，专注于训练视觉编码器和音频编码器，利用大量音频-文本和图像-文本对来增强LLM的语义理解。在第二阶段，我们解冻所有参数，并使用更广泛的多模态数据进行更全面的学习。在最后阶段，我们使用序列长度为32k的数据来增强模型理解复杂长序列数据的能力。

该模型在一个多样化的数据集上进行了预训练，该数据集包括图像-文本、视频-文本、视频-音频、音频-文本和文本语料库等多种类型。我们用自然语言提示替换了层次标签，遵循Qwen2-Audio（Chu et al., 2024a），这可以提高更好的泛化能力和更好的指令遵循能力。

在初始的预训练阶段，Qwen2.5-Omni 的 LLM 组件使用 Qwen2.5（Yang et al., 2024b）中的参数进行初始化，而视觉编码器与 Qwen2.5-VL 相同，音频编码器则使用 Whisper-large-v3（Radford et al., 2023）进行初始化。这两个编码器在固定的 LLM 上分别进行训练，最初都专注于训练各自的适配器，然后再训练编码器。这一基础训练对于使模型具备对核心视觉-文本和音频-文本相关性及对齐的强大理解至关重要。

预训练的第二阶段标志着一个重要的进展，通过引入额外的8000亿个与图像和视频相关的数据令牌、3000亿个与音频相关的数据令牌，以及1000亿个与音频相关的视频数据令牌。这个阶段引入了更大规模的混合多模态数据和更广泛的任务种类，从而增强了听觉、视觉和文本信息之间的互动，并加深了理解。多模态、多任务数据集的纳入对于发展模型同时处理多个任务和模态的能力至关重要，这是一项管理复杂现实世界数据集的重要能力。此外，纯文本数据在维持和提高语言能力方面发挥着至关重要的作用。

为了提高训练效率，我们在之前的阶段将最大令牌长度限制为8192个令牌。然后，我们结合了长音频和长视频数据，并将原始文本、音频、图像和视频数据扩展到32,768个令牌进行训练。实验结果表明，我们的数据在支持长序列数据方面显示出显著改善。

4 训练后

4.1 数据格式

<|im_start|>用户
<|vision_start|>视频.mp4 [视频中有两个人在交谈]<|vision_end|>视频中的人们在说什么？<|im_end|><|im_start|>两张图片都是海绵宝宝。穿红衣服的人说：“你好，今天天气怎么样？”穿黑衣服的人回答：“你好，今天天气很好。”<|im_end|><|im_start|>用户 <|vision_start|>视频.mp4 [视频中的一个人说：“请描述你面前的人。”]<|vision_end|><|im_end|><|im_start|>面前的人戴着眼镜，穿着棕色夹克，里面是一件蓝色衬衫。他们似乎在说话或对某事做出反应，因为他们的嘴巴是张开的，看起来很投入。背景显示一个房间，墙上挂着空调，衣架上挂着各种衣物，墙上有一个大屏幕，显示着夜晚的城市风景。房间的灯光温暖而舒适。<|im_end|>

4.2 思考者

在后训练阶段，我们使用 ChatML (OpenAI, 2022) 格式的指令跟随数据进行指令微调。我们的数据集包含纯文本对话数据、视觉模态对话数据、音频模态对话数据和混合模态对话数据。

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.

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

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 $(\boldsymbol{x}, \boldsymbol{y}_w, \boldsymbol{y}_l)$, where $\boldsymbol{x}$ is the input sequence with input text, and $\boldsymbol{y}_w$ and $\boldsymbol{y}_l$ 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

4.3 说话者

我们为Talker引入了一个三阶段的训练过程，使Qwen2.5-Omni能够同时生成文本和语音响应。在第一阶段，我们训练Talker学习上下文延续。在第二阶段，利用DPO（Rafailov等，2023）来增强语音生成的稳定性。在第三阶段，我们应用了多说话人指令微调，以提高语音响应的自然性和可控性。

在上下文学习（ICL）训练阶段，除了利用类似于Thinker的文本监督外，我们还通过下一个标记预测执行语音续写任务，利用包含多模态上下文和口语响应的大量对话数据集。Talker学习建立从语义表示到语音的单调映射，同时还获得了以上下文适当的多样属性（如韵律、情感和口音）表达语音的能力。此外，我们实施音色解耦技术，以防止模型将特定声音与不常见的文本模式关联。

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

为了扩大说话者和场景的覆盖范围，预训练数据不可避免地包含标签噪声和发音错误，从而导致模型幻觉。为了解决这个问题，我们引入了一个强化学习阶段，以提高语音生成的稳定性。具体来说，对于每个请求和与参考语音配对的响应文本，我们构建了一个数据集 $\mathcal{D}$，其中包含三元组数据 (x, y_w, y_l)，其中 x 是输入文本的输入序列，而 y_w 和 y_l 分别是生成的良好和不良语音序列。我们根据与词错误率 (WER) 和标点暂停错误率相关的奖励分数对这些样本进行排名。

最后，我们对上述基础模型进行了说话者微调，使Talker能够采用特定的声音并提高其自然性。

5 评估

我们对Qwen2.5-Omni进行了全面评估。该模型分为两个主要类别：理解（X→Text）和语音生成（X→Speech）。

5.1 对X→文本的评估

在本节中，我们评估Qwen2.5-Omni理解各种多模态输入（文本、音频、图像和视频）并生成文本响应的能力。

文本→文本 我们对Qwen2.5-Omni在文本→文本上的评估主要集中在一般评估、数学与科学能力以及编码能力上。具体而言，我们利用MMLU-Pro（Wang et al., 2024f）、MMLU-redux（Gema et al., 2024）和Livebench0803（White et al., 2024）进行一般评估，GPQA（Rein et al., 2023）、GSM8K（Cobbe et al., 2021）和MATH（Hendrycks et al., 2021b）进行数学与科学评估，HumanEval（Chen et al., 2021）、MBPP（Austin et al., 2021）、MultiPL-E（Cassano et al., 2023）和LiveCodeBench 2305-2409（Jain et al., 2024）进行编码评估。

音频→文本 对Qwen2.5-Omni在音频→文本方面的评估包括音频理解、音频推理和语音聊天。具体而言，我们对自动语音识别（ASR）、语音转文本翻译（S2TT）、语音实体识别（SER）、声音分类（VSC）和音乐进行全面评估，以评估Qwen2.5-Omni在广泛的音频理解任务上的表现。我们利用MMAU（Sakshi等，2024）进行音频推理任务，使用VoiceBench（Chen等，2024b）和自我策划的语音指令基准进行语音聊天任务。

图像→文本 对于图像→文本的Qwen2.5-Omni评估主要强调在大学水平问题、数学、一般视觉问答和与OCR相关任务中的表现。具体而言，我们利用MMMU（Yue et al., 2023）和MMMU-Pro（Yue et al., 2024）进行大学水平问题的评估，使用MathVista（Lu et al., 2024b）和MathVision（Wang et al., 2024b）进行数学评估。对于一般视觉问答，我们在基准数据集上评估表现，如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）和MM-MT-Bench（Agrawal et al., 2024）。此外，我们还在各种OCR基准上评估Qwen2.5-Omni，如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
$\text{LiveCodeBench}_{2305\text{-}2409}$	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.

et al., 2021），ChartQA（Masry et al., 2022），以及OCRBench_v2（Fu et al., 2024b）。此外，我们还在指称表达理解基准（Kazemzadeh et al., 2014；Mao et al., 2016）、野外物体检测（Li et al., 2022）以及自我策划的点定位基准上评估了我们模型的视觉定位能力。

视频（无音频）→文本 我们在几个具有代表性的视频理解任务上评估我们的模型，如 Video-MME (Fu et al., 2024a)、MVBench (Li et al., 2024a) 和 EgoSchema (Mangalam et al., 2023)。

多模态→文本 我们展示了我们的模型在OmniBench（Li et al., 2024b）上处理混合模态（图像、音频和文本）提示的能力。

5.1.1 文本→文本的性能

我们将Qwen2.5-Omni与其他同类大型语言模型（7B）进行比较。如表1所示，Qwen2.5-Omni的性能通常介于Qwen2-7B和Qwen2.5-7B之间。我们的模型在大多数基准测试中优于Qwen2-7B，例如MMLU-Pro、MMLU-redux、MATH、GSM8K、MBPP、MultiPL-E和LiveCodeBench，这证明了我们模型在Text→Text方面的卓越能力。

表1：文本 → 7B+ 纯文本模型和 Qwen2.5-Omni 的文本性能

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
$\text{LiveCodeBench}_{2305\text{-}2409}$	18.9	8.3	23.9	**28.7**	24.6

5.1.2 音频→文本的性能

我们将Qwen2.5-Omni与其他领先的专业或通用模型在多样的音频理解、音频推理和语音聊天基准上进行了比较。如表2和表3所示，Qwen2.5-Omni在音频理解方面提供了比其他最先进的方法更好或可比的性能。例如，它在Fleurs_zh、CommonVoice_en、CommonVoice_zh、CoVoST2_en-de和CoVoST2_zh-en测试集上实现了优越的ASR和S2TT性能，超越了之前的最先进模型，如Whisper-large-v3、Qwen2Audio、MinMo和其他Omni模型。Qwen2.5-Omni在音乐和VSC等通用音频理解任务上也达到了最先进的性能。此外，Qwen2.5-Omni在音频推理方面取得了最先进的结果，在MMAU基准的声音、音乐和语音子集上表现优越。这些结果展示了Qwen2.5-Omni在通用音频理解和推理方面的强大能力。

此外，在VoiceBench上，Qwen2.5-Omni取得了令人印象深刻的平均分74.12，超越了其他相似规模的音频语言模型和全能模型。这展示了我们模型在语音交互方面的强大能力。为了进一步探索多样化语音交互的性能，我们将几个纯文本基准的文本指令转换为语音，并在内部语音聊天基准上评估Qwen2.5-Omni、Qwen2-Audio和Qwen2-7B。大约90%的文本指令被利用。我们对Qwen2.5-Omni和Qwen2-Audio使用语音指令，对Qwen2-7B使用文本指令。如表4所示，与Qwen2-Audio相比，Qwen2.5-Omni显著缩小了与使用文本指令的Qwen2-7B之间的差距。这反映了我们模型在多样化端到端语音交互方面的重大进展。

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**

表2：最先进技术和Qwen2.5-Omni的音频 → 文本性能

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 *Tempo*	LLark-7B (Gardner et al., 2023)	0.86
	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 *Sound* | *Music* | *Speech* | *Avg*	Gemini-Pro-V1.5 (Team et al., 2024)	56.75 | 49.40 | 58.55 | 54.90
	Qwen2-Audio (Chu et al., 2024a)	54.95 | 50.98 | 42.04 | 49.20
	Qwen2.5-Omni-7B	**67.87** | **69.16** | **59.76** | **65.60**
	Voice Chatting	
VoiceBench *AlpacaEval* | *CommonEval* | *SD-QA* | *MMSU*	Ultravox-v0.4.1-LLaMA-3.1-8B	**4.55** | 3.90 | 53.35 | 47.17
	MERaLiON (He et al., 2024)	4.50 | 3.77 | 55.06 | 34.95
	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 *OpenBookQA* | *IFEval* | *AdvBench* | *Avg*	Ultravox-v0.4.1-LLaMA-3.1-8B	65.27 | **66.88** | 98.46 | 71.45
	MERaLiON (He et al., 2024)	27.23 | 62.93 | 94.81 | 62.91
	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.

表3：最先进技术和Qwen2.5-Omni的音频 → 文本表现

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 *Tempo*	LLark-7B (Gardner et al., 2023)	0.86
	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 *Sound* | *Music* | *Speech* | *Avg*	Gemini-Pro-V1.5 (Team et al., 2024)	56.75 | 49.40 | 58.55 | 54.90
	Qwen2-Audio (Chu et al., 2024a)	54.95 | 50.98 | 42.04 | 49.20
	Qwen2.5-Omni-7B	**67.87** | **69.16** | **59.76** | **65.60**
	Voice Chatting	
VoiceBench *AlpacaEval* | *CommonEval* | *SD-QA* | *MMSU*	Ultravox-v0.4.1-LLaMA-3.1-8B	**4.55** | 3.90 | 53.35 | 47.17
	MERaLiON (He et al., 2024)	4.50 | 3.77 | 55.06 | 34.95
	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 *OpenBookQA* | *IFEval* | *AdvBench* | *Avg*	Ultravox-v0.4.1-LLaMA-3.1-8B	65.27 | **66.88** | 98.46 | 71.45
	MERaLiON (He et al., 2024)	27.23 | 62.93 | 94.81 | 62.91
	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**

表4：Qwen2.5-Omni及其他模型在聊天中的表现，*意味着大约90%的适合语音的文本指令被使用。

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 图像 → 文本的性能

为了全面评估图像 → 文本的能力，我们将 Qwen2.5-Omni 与最近的最先进的大型视觉语言模型 Qwen2.5-VL-7B 以及其他表现最佳的全能模型进行比较。如表 5 所示，Qwen2.5-Omni 的性能与 Qwen2.5-VL-7B 相当，并且在 MMMU、MathVision、MMBench-V1.1-EN、TextVQA、DocVQA 和 ChartQA 上的结果优于任何其他开源全能模型。此外，Qwen2.5-Omni 在大多数基准测试中也超越了 GPT-4o-mini。这些结果揭示了我们模型在图像理解方面的出色能力。

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
$MMMU\text{-}Pro_{overall}$	37.6	**38.3**	-	36.6
Mathematical				
$MathVista_{testmini}$	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				
$MMBench\text{-}V1.1\text{-}EN_{test}$	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
$CRPE_{relation}$	-	76.4	-	**76.5**
$RealWorldQA_{avg}$	-	68.5	**71.9** (Infinigence)	70.3
$MME\text{-}RealWorld_{en}$	-	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
$ChartQA_{test\ Avg}$	-	**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.

表5：图像 → 文本性能的 7B+ 模型和 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
$MMMU\text{-}Pro_{overall}$	37.6	**38.3**	-	36.6
Mathematical				
$MathVista_{testmini}$	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				
$MMBench\text{-}V1.1\text{-}EN_{test}$	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
$CRPE_{relation}$	-	76.4	-	**76.5**
$RealWorldQA_{avg}$	-	68.5	**71.9** (Infinigence)	70.3
$MME\text{-}RealWorld_{en}$	-	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
$ChartQA_{test\ Avg}$	-	**87.3**	84.9 (Li et al., 2025)	85.3
$OCRBench\_V2_{en}$	-	56.3	-	**57.8**

表6：Qwen2.5-Omni及其他模型的基础性能

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

对于视觉定位，我们将Qwen2.5-Omni与Qwen2.5-VL-7B及其他领先的LVLM进行比较，包括Gemini和Grounding-DINO（Liu et al., 2024）。如表6所示，我们的模型在从框定位到点定位的大多数基准测试中优于其他模型，并在开放词汇物体检测中取得了42.2mAP的良好表现，这揭示了我们模型强大的视觉定位能力。

5.1.4 视频 → 文本的性能

类似于图像→文本，我们将Qwen2.5-Omni与Qwen2.5-VL-7B及其他全能模型进行比较。如表7所示，Qwen2.5-Omni在所有其他最先进的开源全能模型和GPT-4o-Mini中表现优越，并且与Qwen2.5-VL-7B相比，取得了更好或具有竞争力的结果，这证明了其在视频理解方面的卓越性能。

5.1.5 多模态性能→文本

如表8所示，Qwen2.5-Omni在OmniBench上实现了最先进的性能，远超其他Omni模型，这证明了我们模型在多模态理解方面的优越性。

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

表7：7B+模型和Qwen2.5-Omni的视频→文本性能

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**

表8：多模态 → 先进技术与 Qwen2.5-Omni 的文本表现

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 X→语音的评估

在本节中，我们评估Qwen2.5-Omni的语音生成能力。由于缺乏相关评估，语音生成的评估主要集中在给定文本的语音生成上，类似于文本到语音（TTS），从两个方面进行：零样本和单一说话者的语音生成能力。

- 零样本语音生成 我们评估了我们模型在SEED（Anastassiou等，2024）上的内容一致性（WER）和说话人相似性（SIM）。
- 单声道语音生成 我们评估了在SEED（Anastassiou等，2024）上我们经过微调的说话者模型的稳定性，并在自创的数据集上评估了生成语音的主观自然性（NMOS）。

5.2.1 零样本语音生成的评估。

我们将Qwen2.5-Omni与最先进的零-shot TTS系统进行了比较。如表9所示，Qwen2.5-Omni展现出高度竞争的性能，突显了其通过上下文学习（ICL）开发的强大语音理解和生成能力。此外，在强化学习（RL）优化后，Qwen2.5-Omni在生成稳定性方面显示出显著改善，注意力错位、发音错误和不当停顿在具有挑战性的test-hard数据集上明显减少。

5.2.2 单一发言人语音生成的评估。

我们比较了Qwen2.5-Omni模型在扬声器微调前后的表现，以及与人类录音的对比。如表10所示，经过扬声器微调的Qwen2.5-Omni更准确地捕捉了目标扬声器的细微韵律风格，同时保持了基础模型提供的基础稳定性，在主观和客观指标上达到了接近人类水平的表现。

6 结论

Qwen2.5-Omni 是一个统一模型，旨在理解和生成多种模态，包括文本和实时语音。为了增强视频集成，我们引入了一种新的位置嵌入方法，称为 TMRoPE，它对齐了音频和视频的时序。我们的 Thinker-Talker 框架支持实时语音生成，同时最小化不同模态之间的干扰。此外，我们采用了块级音频/视觉编码和滑动窗口机制等技术，用于代码到 wav 的生成。这个创新模型在复杂的音频-视觉交互和情感方面表现出色。

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

表9：零样本语音生成

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

表10：单一发言者语音生成

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

在语音对话中的上下文。综合评估显示，Qwen2.5-Omni在执行语音命令方面优于同等规模的单一模态模型，并在多模态任务中达到了最先进的性能。

在模型的开发过程中，我们识别出几个关键问题，这些问题在以往的学术研究中常常被研究人员忽视，例如视频OCR和音视频协同理解。解决这些挑战需要学术界和工业界之间的合作，特别是在建立全面的评估基准和研究数据集方面。我们相信Qwen2.5-Omni代表了朝着人工通用智能（AGI）迈出的重要一步。我们未来的目标包括开发一个更强大、更快速的模型，具备在图像、视频和音乐等多种模态下扩展输出能力。

7 位作者

核心贡献者：徐金，郭志芳，何金正，胡航瑞，何婷，白帅，陈克勤，王佳林，范扬，邓凯，张斌，王雄，楚云飞，林俊阳

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.

贡献者[1]: 安杨, 李安丰, 杨宝松, 陈贝, 林彬, 惠彬源, 郑博, 余博文, 陈成, 黄承根, 袁晨汉, 李承源, 陈达仁, 刘大义恒, 郭大可, 周凡, 黄飞, 周广东, 张航, 连浩然, 张浩阳, 王赫, 钟虎门, 杨剑, 姜建东, 涂建宏, 万建强, 曾建元, 唐俊, 张建伟, 杨建新, 曾建元, 周静, 周静仁, 杨可欣, 谢雷, 马林汉, 孟灵辰, 余乐, 李梅, 洪苗, 薛明峰, 杨明坤, 李明泽, 倪娜, 张佩, 张佩扬, 刘鹏, 王鹏, 张鹏, 王鹏飞, 胡瑞, 门瑞, 王秋月, 傅青, 刘世轩, 宋思博, 张思琪, 陈松, 唐天逸, 何涛, 何婷, 葛文彬, 丁伟, 邓晓东, 牛新尧, 魏西平, 邢斌, 刘雪晶, 任兴章, 任宣诚, 刘杨, 李彦鹏, 刘杨, 苏杨, 张宜昌, 刘玉琼, 吕元俊, 朱元志, 蔡宇轩, 崔泽宇, 李正, 张振如, 邱子涵, 李兆海, 杨志博, 周志鹏, 朱志远

参考文献

P拉维什·阿格拉瓦尔，西蒙·安东尼亚克，艾玛·布哈娜，巴蒂斯特·布特，德文德拉·查普洛特，杰西卡·楚德诺夫斯基，迪奥戈·科斯塔，博多温·德·莫尼考，索拉布·加尔格，西奥菲尔·热尔维，索哈姆·戈什，阿梅莉·海柳，保罗·雅各布，阿尔伯特·Q·姜，卡尔蒂克·坎德尔瓦尔，蒂莫西·拉克鲁瓦，纪尧姆·兰普尔，迭戈·拉斯·卡萨斯，蒂博·拉夫里尔，特文·勒·斯卡奥，安迪·洛，威廉·马歇尔，路易斯·马丁，阿图尔·门施，帕万库马尔·穆迪雷迪，瓦莱拉·涅米奇科娃，玛丽·佩拉特，帕特里克·冯·普拉滕，尼基尔·拉古拉曼，巴蒂斯特·罗齐埃，亚历山大·萨布莱罗尔，露西尔·索尔尼耶，罗曼·索维斯特，温迪·尚，罗曼·索列茨基，劳伦斯·斯图尔特，皮埃尔·斯托克，约阿希姆·斯图德尼亚，桑迪普·苏布拉马尼安，萨加尔·瓦泽，托马斯·王，和索非亚·杨。Pixtral 12b，2024。网址 https://arxiv.org/abs/2410.0 7073。

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao 等. Seed-tts: 一系列高质量多功能语音生成模型. *arXiv preprint arXiv:2406.02430*, 2024.

Anthropic. 介绍Claude，2023a。网址 https://www.anthropic.com/index/introducing-claude。

Anthropic. Claude 2. 技术报告，Anthropic，2023b。网址 https://www-files.anthropic.com/pro duction/images/Model-Card-Claude-2.pdf。

Anthropic。Claude 3 模型系列：Opus、Sonnet、Haiku。技术报告，Anthropic，人工智能，2024。网址 https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf。

雅各布·奥斯丁，奥古斯都·奥德纳，麦克斯韦·I·奈，马滕·博斯马，亨利克·米哈维尔斯基，大卫·多汉，艾伦·姜，凯莉·J·蔡，迈克尔·特里，阮·V·黎，查尔斯·萨顿。使用大型语言模型进行程序合成。*CoRR*，abs/2108.07732，2021。

J在泽白、帅白、云飞楚、泽宇崔、凯当、晓东邓、杨帆、文彬葛、余汉、飞黄、宾源惠、罗吉、梅丽、俊阳林、润基林、大义恒刘、高刘、成强卢、克明卢、建新马、瑞门、兴章任、宣诚任、传奇谭、思南谭、建宏涂、彭王、世杰王、伟王、盛光吴、本峰徐、金徐、安杨、浩杨、建杨、树生杨、杨瑶、博文余、鸿毅袁、正元、建伟张、兴轩张、怡畅张、振如张、常周、景仁周、晓欢周和天航朱。Qwen技术报告。*CoRR*，abs/2309.16609，2023a。

金泽白、帅白、舒生杨、世杰王、思南谭、鹏王、俊阳林、常周和景仁周。Qwen-VL：一个具有多种能力的前沿大型视觉语言模型。*CoRR*，abs/2308.12966，2023b。

帅白，柯勤陈，雪晶刘，佳林王，文彬葛，思博宋，凯当，彭王，世杰王，俊唐等。Qwen2.5-vl技术报告。*arXiv preprint arXiv:2502.13923*，2025。

叶白，陈京平，陈继彤，陈伟，陈卓，丁创，董林浩，董倩倩，杜宇娇，高可攀，等。Seed-asr：利用基于大语言模型的语音识别理解多样的语音和上下文。*arXiv preprint arXiv:2407.04675*，2024。

汤姆·布朗，本杰明·曼，尼克·赖德，梅拉妮·苏比亚，贾里德·D·卡普兰，普拉富拉·达里瓦尔，阿尔文·尼拉坎坦，普拉纳夫·夏姆，吉里什·萨斯特里，阿曼达·阿斯克尔等。语言模型是少量学习者。在 *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.

费德里科·卡萨诺，约翰·古瓦尔，丹尼尔·阮，悉尼·阮，露娜·菲普斯-科斯廷，唐纳德·平克尼，明浩·易，杨天子，卡罗琳·简·安德森，莫莉·Q·费尔德曼，阿尔君·古哈，迈克尔·格林伯格，和阿比纳夫·江达。MultiPL-E：一种可扩展的多语言神经代码生成基准测试方法。*IEEE Trans. Software Eng.*，49(7)：3675–3691，2023。

林晨，李金松，董晓怡，张潘，臧宇航，陈泽辉，段浩东，王佳琪，乔宇，林大华，等。我们在评估大型视觉-语言模型的正确道路上吗？*arXiv:2403.20330*，2024a。

M陈克，杰瑞·特沃雷克，金熙宇，袁启明，亨里克·庞德·德·奥利维拉·平托，贾里德·卡普兰，哈里森·爱德华兹，尤里·布尔达，尼古拉斯·约瑟夫，格雷格·布罗克曼，亚历克斯·雷，劳尔·普里，格雷琴·克鲁格，迈克尔·彼得罗夫，海迪·克拉夫，吉里什·萨斯特里，帕梅拉·米什金，布鲁克·陈，斯科特·格雷，尼克·赖德，米哈伊尔·帕夫洛夫，阿莱西娅·鲍尔，卢卡斯·凯泽，穆罕默德·巴瓦里安，克莱门斯·温特，菲利普·蒂莱，费利佩·佩特罗斯基·苏奇，戴夫·卡明斯，马蒂亚斯·普拉普特，福提奥斯·钱齐斯，伊丽莎白·巴恩斯，阿里尔·赫伯特-沃斯，威廉·赫布根·古斯，亚历克斯·尼科尔，亚历克斯·佩诺，尼古拉斯·特扎克，唐杰，伊戈尔·巴布什金，苏奇尔·巴拉吉，尚坦努·贾因，威廉·桑德斯，克里斯托弗·赫斯，安德鲁·N·卡尔，扬·莱克，约书亚·阿基亚姆，维丹特·米斯拉，埃文·森川，亚历克·拉德福德，马修·奈特，迈尔斯·布伦达奇，米拉·穆拉蒂，凯蒂·梅耶，彼得·维林德，鲍勃·麦克格鲁，达里奥·阿莫代，萨姆·麦肯德利什，伊利亚·苏茨克弗，以及沃伊切赫·扎伦巴。评估在代码上训练的大型语言模型。*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, 和 Jinren Zhou. Minmo: 一种用于无缝语音交互的多模态大型语言模型, 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 和 Furu Wei. Wavlm：用于全栈语音处理的大规模自监督预训练。*IEEE J. Sel. Top. Signal Process.*，2022。

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, 和 Haizhou Li. Voicebench: 基于大型语言模型的语音助手基准测试. *arXiv preprint arXiv:2410.17196*, 2024b.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, 和 Xie Chen. F5- tts: 一个通过流匹配伪造流畅和忠实语音的讲故事者。*arXiv preprint arXiv:2410.06885*, 2024c.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, 和 Jingren Zhou. Qwen-audio: 通过统一的大规模音频-语言模型推进通用音频理解. *arXiv preprint arXiv:2311.07919*, 2023a.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, 和 Jingren Zhou. Qwen-Audio: 通过统一的大规模音频-语言模型推进通用音频理解. *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, 和 Jingren Zhou. Qwen2-audio 技术报告. *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 等. Qwen2-audio 技术报告. *arXiv preprint arXiv:2407.10759*, 2024b.

卡尔·科布，维尼特·科萨拉朱，穆罕默德·巴瓦里安，马克·陈，许宇俊，卢卡斯·凯泽，马蒂亚斯·普拉普特，杰瑞·特沃雷克，雅各布·希尔顿，中野礼一郎，克里斯托弗·赫塞，约翰·舒尔曼。训练验证器以解决数学文字问题。*CoRR*，abs/2110.14168，2021。

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, 和 Steven Hoi. Instructblip: 朝着具有指令调优的通用视觉-语言模型迈进。*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 等人。Speechverse：一个大规模可泛化的音频语言模型。*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.

Soham Deshmukh, Benjamin Elizalde, Rita Singh 和 Huaming Wang. Pengi：一个用于音频任务的音频语言模型。*CoRR*，2023。

SeungHeon Doh, Keunwoo Choi, Jongpil Lee 和 Juhan Nam. Lp-musiccaps: 基于 LLM 的伪音乐字幕生成. *arXiv preprint arXiv:2307.16372*, 2023.

杜志豪，王宇轩，陈倩，施贤，吕翔，赵天宇，高志福，杨叶欣，高长风，王辉，等。Cosyvoice 2：基于大语言模型的可扩展流式语音合成。*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, 等等。Llama 3 模型群。*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 等人。Llama 3 模型群。*arXiv:2407.21783*，2024b。

本杰明·埃利扎尔德，索哈姆·德什穆克，马哈茂德·阿尔·伊斯梅尔，和华明·王。CLAP：从自然语言监督中学习音频概念。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 等人。E2 tts：令人尴尬的简单完全非自回归零-shot tts。在 *2024 IEEE Spoken Language Technology Workshop (SLT)*，第 682–689 页。IEEE，2024。

超优傅，裴贤陈，云航沈，玉雷秦，梦丹张，徐林，振宇邱，伟林，金瑞杨，夏武郑，等。Mme：多模态大语言模型的综合评估基准。*arXiv:2306.13394*，2023。

超优傅，余翰戴，永东罗，雷李，书怀任，仁瑞张，子涵王，晨宇周，云航沈，梦丹张，等。视频-mme：首个多模态大语言模型在视频分析中的综合评估基准。*arXiv:2405.21075*，2024a。

凌傅，杨彪，邝哲彬，宋佳俊，李宇哲，朱凌浩，罗启迪，王新宇，卢浩，黄铭鑫，李章，唐国志，单斌，林春辉，刘琦，吴冰洪，冯浩，刘浩，黄灿，唐景群，陈伟，金连文，刘玉良，白翔。Ocrbench v2：用于评估大型多模态模型在视觉文本定位和推理方面的改进基准，2024b。网址 https://arxiv.org/abs/2501.00321。

约书亚·P·加德纳，西蒙·杜兰，丹尼尔·斯托勒和瑞秋·M·比特纳。Llark：一种用于音乐的多模态指令跟随语言模型。在 *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 等人。我们完成 mmlu 了吗？*CoRR*，abs/2406.04127，2024。

双子团队。双子 1.5：在数百万个上下文标记中解锁多模态理解。技术报告，谷歌，2024。网址 https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf。

高雄龚，凯拓冯，博豪李，怡冰王，莫凡程，世佳杨，嘉铭韩，本友王，宇彤白，卓然杨，等。Av-odyssey bench: 你的多模态大语言模型真的能理解音视频信息吗？*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.

何英旭, 刘卓涵, 孙硕, 王斌, 张文宇, 邹勋龙, 南希·F·陈, 和艾·蒂·阿沃。Meralion-audiollm：技术报告。*arXiv preprint arXiv:2412.09818*，2024。

丹·亨德里克斯、科林·伯恩斯、史蒂文·巴萨特、安迪·邹、曼塔斯·梅泽卡、道恩·宋和雅各布·斯坦哈特。测量大规模多任务语言理解。在 *ICLR*。OpenReview.net，2021a。

丹·亨德里克斯、科林·伯恩斯、索拉夫·卡达瓦斯、阿库尔·阿罗拉、史蒂文·巴萨特、埃里克·唐、道恩·宋和雅各布·斯坦哈特。使用 MATH 数据集测量数学问题解决能力。在 *NeurIPS Datasets and Benchmarks*，2021b。

黄少寒，董立，王文辉，郝雅如，萨克尚·辛格哈尔，马树明，吕腾超，崔雷，欧维斯·汗·穆罕默德，刘强，等。语言并不是你所需要的一切：将感知与语言模型对齐。*arXiv:2302.14045*，2023。

无限智能。无限-梅格雷兹-全能。网址 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, 和 Ion Stoica. LiveCodeBench: 大型语言模型代码的整体和无污染评估. *CoRR*, abs/2403.07974, 2024. Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, 和 Tamara Berg. ReferItGame: 在自然场景照片中指代物体. 在 Alessandro Moschitti, Bo Pang, 和 Walter Daelemans (编), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 第 787–798 页, 卡塔尔多哈, 2014 年 10 月. 计算语言学协会. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086/. Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, 和 Ali Farhadi. 一张图胜过十张图像. 在 *ECCV*, 2016. Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, 和 Sungroh Yoon. Bigvgan: 一种具有大规模训练的通用神经声码器. 在 *ICLR 2023*. Junnan Li, Dongxu Li, Silvio Savarese, 和 Steven Hoi. Blip-2: 使用冻结图像编码器和大型语言模型进行语言-图像预训练. *arXiv:2301.12597*, 2023. Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, 等. Mvbench: 一个全面的多模态视频理解基准. 在 *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, 和 Jianfeng Gao. 基于基础的语言-图像预训练, 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, 等. Baichuan-omni-1.5 技术报告. *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, 等. Omnibench: 朝着通用全语言模型的未来. *arXiv preprint arXiv:2409.15272*, 2024b. Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, 和 Matthew Le. 生成建模的流匹配. 在 *ICLR 2023*. Haotian Liu, Chunyuan Li, Yuheng Li, 和 Yong Jae Lee. 通过视觉指令调优改进基线. *arXiv:2310.03744*, 2023a. Haotian Liu, Chunyuan Li, Qingyang Wu, 和 Yong Jae Lee. 视觉指令调优. *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, 和 Lei Zhang. Grounding dino: 将 dino 与基础预训练结合用于开放集物体检测, 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, 和 Dahua Lin. Mmbench: 你的多模态模型是全能选手吗? *arXiv:2307.06281*, 2023c. Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, 和 Aniruddha Kembhavi. Unified-io 2: 通过视觉语言音频和动作扩展自回归多模态模型. 在 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 第 26439–26455 页, 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.

潘璐，Hritik Bansal，Tony Xia，刘嘉诚，李春元，Hannaneh Hajishirzi，Hao Cheng，Kai-Wei Chang，Michel Galley，和高剑锋。Mathvista：在视觉上下文中评估基础模型的数学推理。在 *ICLR*，2024b。Karttikeya Mangalam，Raiymbek Akshulakov，和Jitendra Malik。Egoschema：一个用于非常长视频语言理解的诊断基准。在 *NeurIPS*，2023。毛俊华，黄Jonathan，亚历山大·托谢夫，Oana Camburu，Alan Yuille，和Kevin Murphy。生成和理解明确的物体描述，2016。网址 https://arxiv.org/abs/1511.02283。Ahmed Masry，Do Xuan Long，Jia Qing Tan，Shafiq Joty，和Enamul Hoque。Chartqa：一个关于图表的视觉和逻辑推理问答基准。*arXiv:2203.10244*，2022。Minesh Mathew，Dimosthenis Karatzas，和CV Jawahar。Docvqa：一个用于文档图像的vqa数据集。在 *WACV*，2021。OpenAI。ChatML，2022。网址 https://github.com/openai/openai-python/blob/e389823ba013a24b4c32ce38fa0bd87e6bccae94/chatml.md。OpenAI。GPT4技术报告。*CoRR*，abs/2303.08774，2023。OpenAI。Gpt-4v(ision)系统卡，2023。网址 https://openai.com/research/gpt-4v-system-card。OpenAI。Hello GPT-4o，2024。网址 https://openai.com/index/hello-gpt-4o/。Alec Radford，金钟旭，Tao Xu，Greg Brockman，Christine McLeavey，和Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别。在 *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，和Chelsea Finn。直接偏好优化：你的语言模型实际上是一个奖励模型。在 *NeurIPS*，2023。David Rein，Betty Li Hou，Asa Cooper Stickland，Jackson Petty，Richard Yuanzhe Pang，Julien Dirani，Julian Michael，和Samuel R. Bowman。GPQA：一个研究生级别的Google-proof问答基准。*CoRR*，abs/2311.12022，2023。S Sakshi，Utkarsh Tyagi，Sonal Kumar，Ashish Seth，Ramaneswaran Selvakumar，Oriol Nieto，Ramani Duraiswami，Sreyan Ghosh，和Dinesh Manocha。Mmau：一个大规模多任务音频理解和推理基准，2024。网址 https://arxiv.org/abs/2410.19168。Amanpreet Singh，Vivek Natarajan，Meet Shah，Yu Jiang，Xinlei Chen，Dhruv Batra，Devi Parikh，和Marcus Rohrbach。朝着能够阅读的vqa模型迈进。在 *CVPR*，2019。光志孙，文怡余，常立唐，贤钊陈，天坦，韦李，卢璐，泽军马，宇轩王，和超张。video-salmonn：语音增强的视听大型语言模型。*arXiv preprint arXiv:2406.15704*，2024。常立唐，文怡余，光志孙，贤钊陈，天坦，韦李，卢璐，泽军马，和超张。SALMONN：朝着大型语言模型的通用听觉能力迈进。在 *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*，2024。Gemini团队，Petko Georgiev，Ving Ian Lei，Ryan Burnell，Libin Bai，Anmol Gulati，Garrett Tanzer，Damien Vincent，Zhufeng Pan，Shibo Wang，等。Gemini 1.5：解锁数百万上下文标记的多模态理解。*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，等。Llama：开放和高效的基础语言模型。*arXiv:2302.13971*，2023a。Hugo Touvron，Louis Martin，Kevin Stone，Peter Albert，Amjad Almahairi，Yasmine Babaei，Nikolay Bashlykov，Soumya Batra，Prajjwal Bhargava，Shruti Bhosale，等。Llama 2：开放的基础和微调聊天模型。*arXiv:2307.09288*，2023b。陈旺，敏鹏廖，钟强黄，金亮卢，俊宏吴，宇辰刘，承庆宗，和佳俊张。Blsp：通过续写行为对齐进行语言-语音预训练的引导。*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.

费王，邢宇傅，詹姆斯·Y·黄，泽坤·李，秦·刘，晓耕·刘，明宇·德里克·马，南·徐，文轩·周，凯·张，天怡·洛雷娜·严，文杰·杰基·莫，向辉·刘，潘·卢，春元·李，超伟·肖，凯伟·张，丹·罗斯，盛·张，霍丰·潘，和慕豪·陈。Muirbench：一个全面的强健多图像理解基准，2024a。网址 https://arxiv.org/abs/2406.09411。柯·王，俊廷·潘，伟康·石，子木·卢，明杰·詹，和洪生·李。用数学视觉数据集测量多模态数学推理。*arXiv:2402.14804*，2024b。彭·王，帅·白，思南·谭，世杰·王，志豪·范，金泽·白，克勤·陈，雪晶·刘，佳林·王，文彬·葛，杨·范，凯·邓，梦飞·杜，轩成·任，瑞·门，大义恒·刘，常·周，景仁·周，和俊阳·林。Qwen2-vl：增强视觉-语言模型对任何分辨率世界的感知。*CoRR*，abs/2409.12191，2024c。韦汉·王，青松·吕，文梦·余，文怡·洪，季·齐，燕·王，俊辉·季，卓毅·杨，雷·赵，希轩·宋，等。Cogvlm：预训练语言模型的视觉专家。*arXiv:2311.03079*，2023b。韦云·王，怡铭·任，浩文·罗，天通·李，晨翔·燕，哲·陈，文海·王，青云·李，乐伟·卢，希洲·朱，宇·乔，和季峰·戴。全视项目v2：朝着开放世界的一般关系理解，2024d。网址 https://arxiv.org/abs/2402.19474。元城·王，浩月·詹，李伟·刘，瑞红·曾，浩天·郭，嘉辰·郑，强·张，雪瑶·张，顺思·张，和志正·吴。Maskgct：使用掩码生成编码器变换器的零样本文本到语音。*arXiv preprint arXiv:2409.00750*，2024e。宇博·王，雪光·马，格·张，元生·倪，阿布拉尼尔·钱德拉，时光·郭，伟铭·任，阿兰·阿鲁尔拉杰，轩·何，子妍·姜，天乐·李，马克斯·库，凯·王，亚历克斯·庄，荣琦·范，向·岳，和文虎·陈。MMLU-Pro：一个更强健和具有挑战性的多任务语言理解基准。*CoRR*，abs/2406.01574，2024f。泽坤·王，金·朱，春普·徐，王春树·周，嘉恒·刘，怡博·张，佳硕·王，宁·石，思宇·李，怡智·李，等。Mio：一个基于多模态令牌的基础模型。*arXiv preprint arXiv:2409.17692*，2024g。科林·怀特，塞缪尔·杜利，曼利·罗伯茨，阿卡·帕尔，本杰明·费尔，西达尔塔·贾因，拉维德·施瓦茨-齐夫，尼尔·贾因，哈立德·赛义夫拉，西达尔塔·奈杜，钦梅·赫格，扬·勒昆，汤姆·戈德斯坦，威利·奈斯旺格，和米卡·戈德布鲁姆。LiveBench：一个具有挑战性、无污染的LLM基准。*CoRR*，abs/2406.19314，2024。简·吴，雅舍·高，卓·陈，龙·周，怡萌·朱，天瑞·王，金宇·李，淑杰·刘，博·任，林泉·刘，和宇·吴。关于仅解码器架构的语音到文本和大型语言模型集成。abs/2307.03917，2023。X.AI。Grok-1.5视觉预览，2024。网址 https://x.ai/blog/grok-1.5v。志飞·谢和长桥·吴。Mini-omni：语言模型可以在思考时听、说并进行流式处理。*arXiv preprint arXiv:2408.16725*，2024。安·杨，宝松·杨，宾源·惠，博·郑，博文·余，常·周，承鹏·李，承元·李，大义恒·刘，费·黄，等。Qwen2技术报告。*arXiv:2407.10671*，2024a。安·杨，宝松·杨，北辰·张，宾源·惠，博·郑，博文·余，承元·李，大义恒·刘，费·黄，浩然·魏，欢·林，简·杨，建宏·涂，建伟·张，建新·杨，佳希·杨，景仁·周，俊阳·林，凯·邓，克明·卢，克勤·包，克欣·杨，乐·余，美·李，明锋·薛，佩·张，秦·朱，瑞·门，润基·林，天浩·李，婷宇·夏，兴章·任，轩成·任，杨·范，杨·苏，怡畅·张，余·万，玉琼·刘，泽宇·崔，振如·张，和子涵·邱。Qwen2.5技术报告。*CoRR*，abs/2412.15115，2024b。袁·姚，天宇·余，傲·张，崇义·王，俊博·崔，洪基·朱，天池·蔡，浩宇·李，伟林·赵，志辉·何，等。Minicpm-v：一个在你手机上的gpt-4v级mllm。*arXiv preprint arXiv:2408.01800*，2024。伟豪·余，正元·杨，林杰·李，建峰·王，凯文·林，子诚·刘，欣超·王，和丽娟·王。Mm-vet：评估大型多模态模型的综合能力。在*ICML*，2024。向·岳，元生·倪，凯·张，天宇·郑，若琦·刘，格·张，塞缪尔·史蒂文斯，东福·姜，伟铭·任，宇轩·孙，等。Mmmu：一个针对专家AGI的大规模多学科多模态理解和推理基准。*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.

向月，郑天宇，倪元生，王宇博，张凯，童胜邦，孙宇轩，尹铭，余博涛，张戈，等。Mmmu-pro：一个更强大的多学科多模态理解基准。*arXiv preprint arXiv:2409.02813*，2024。Jun Zhan，戴俊奇，叶家生，周云华，张东，刘志耕，张鑫，袁瑞彬，张戈，李林扬，等。Anygpt：统一的多模态大语言模型，具有离散序列建模。*arXiv preprint arXiv:2402.12226*，2024。张怡凡，张焕宇，田浩辰，傅超友，张双青，吴俊飞，李峰，王琨，温青松，张张，等。Mme-realworld：你的多模态大语言模型能否挑战人类难以应对的高分辨率真实场景？*arXiv preprint arXiv:2408.13257*，2024。钟志胜，王承耀，刘宇琪，杨森桥，唐龙翔，张悦辰，李静瑶，曲天源，李彦伟，陈宇康，等。Lyra：一个高效且以语音为中心的全认知框架。*arXiv preprint arXiv:2412.09501*，2024。朱德尧，陈俊，沈晓倩，李向，和Mohamed Elhoseiny。Minigpt-4：通过先进的大型语言模型增强视觉-语言理解。*arXiv:2304.10592*，2023。