DAILY-OMNI: TOWARDS AUDIO-VISUAL REASONING WITH TEMPORAL ALIGNMENT ACROSS MODALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent Multimodal Large Language Models (MLLMs) achieve promising performance on visual and audio benchmarks independently. However, the ability of these models to process cross-modal information synchronously remains largely unexplored. In this paper, we introduce: 1) **Daily-Omni**, an Audio-Visual Questioning and Answering benchmark comprising 684 videos of daily life scenarios from diverse sources, rich in both audio and visual information, and featuring 1197 multiple-choice QA pairs across 6 major tasks; 2) **Daily-Omni QA Generation Pipeline**, which includes automatic annotation, QA generation and QA optimization, significantly improves efficiency for human evaluation and scalability of the benchmark; 3) **Daily-Omni-Agent**, a training-free agent utilizing open-source Visual Language Model (VLM), Audio Language Model (ALM) and Automatic Speech Recognition (ASR) model to establish a baseline for this benchmark. The results show that current MLLMs still struggle significantly with tasks requiring audio-visual integration, but combining VLMs and ALMs with simple **temporal alignment techniques** can achieve substantially better performance.

1 INTRODUCTION

Non-textual modalities, such as vision and hearing, often convey richer and more nuanced information than text alone in daily life scenarios, playing a crucial role in our understanding of and interaction with the physical world. Therefore, advancements in Multimodal Large Language Models (MLLMs) capable of comprehensively understanding the multi-modal information are a crucial foundation for achieving artificial intelligence capable of interacting with the physical world and observing the outcome of its operations.

Recent MLLMs (Xu et al., 2025; Sun et al., 2024; Pichai et al., 2024; Zhang et al., 2024a; Fu et al., 2025b; Liu et al., 2025c; Cheng et al., 2024; Li et al., 2025; Xie & Wu, 2024; Guo et al., 2025; Team et al., 2024; Microsoft et al., 2025; Lu et al., 2024b; Liu et al., 2025a; Team et al., 2025) have demonstrated groundbreaking capabilities spanning audio (ASR, sound classification, captioning) and visual domains (OCR, VQA, video grounding), with significant accuracy improvements over previous benchmarks. However, existing methods still face several limitations. Firstly, many MLLMs predominantly focus on visual abilities, often neglecting the importance of other modalities like audio. This oversight may stem from the fact that current visual datasets are more abundant, of higher quality, and cover a broader range of tasks compared to audio datasets. Existing audio datasets (Panayotov et al., 2015; Wang et al., 2021; Poria et al., 2019; Chen et al., 2020; Gemmeke et al., 2017; Gong et al., 2022) tend to prioritize speech-related or music-related tasks and basic sound classification, often overlooking more complex yet crucial tasks such as reasoning over generic sounds. Consequently, many MLLMs incorporate only speech encoders as their primary auditory component or rely heavily on speech-related datasets for audio pre-training. This architectural limitation fundamentally restricts their ability to comprehend rich acoustic environments where non-speech sounds (e.g., environmental noises, mechanical failures, or emotional cues in non-verbal vocalizations) carry critical semantic information. Secondly, the current landscape lacks high-quality multimodal datasets that integrate temporally aligned auditory and visual information. Existing audio-visual datasets and benchmarks (Yun et al., 2021; Li et al., 2022; Yang et al., 2022; Li et al., 2024; Hong et al., 2025; Gong et al., 2024a; Sung-Bin et al., 2025; Geng et al., 2025) reveal three persistent limitations. First, several focus on specialized scenarios (Yun et al., 2021; Li et al., 2022) such as musical performances or panoramic environments, thereby introducing domain-specific biases. Second, many employ static image-

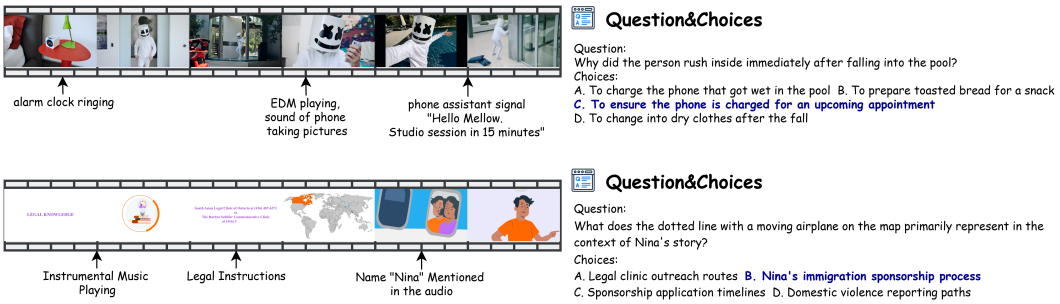


Figure 1: **Examples of Daily-Omni QAs.** The audio and visual information required for answering the questions are provided in the figure. The correct answer for the given questions are highlighted. More cases are presented in Appendix A.

audio pairs (Li et al., 2024; Gong et al., 2024a) that disregard crucial temporal dynamics inherent in real-world video contexts. Third, current tasks are often too narrow, with many benchmarks focusing on specific applications such as captioning or open-ended responses (Geng et al., 2025). The common absence of rigorous evaluation frameworks and standardized metrics makes it hard to reliably compare results. While WorldSense (Hong et al., 2025) established a valuable audio-visual multi-choice QA benchmark for daily scenarios through its meticulous dataset curation, two critical limitations persist: (1) it lacks a systematic framework for scalable QA generation, relying instead on manual annotation processes that hinder dataset expansion; and (2) the benchmark serves primarily as an evaluation tool, offering limited methodological guidance on enhancing model capabilities via explicit training protocols or architectural modifications. This dual limitation constrains both the benchmark's adaptability to emerging domains and its practical utility in driving model improvements.

In this paper, we introduce **Daily-Omni**, an Audio-Visual Questioning and Answering benchmark with 684 videos featuring daily life scenarios from various sources (Gemmeke et al., 2017; Fu et al., 2025a; Farré et al., 2024) with rich audio and visual information and 1197 multiple choice QA-pairs across 6 major tasks ranging from audio visual event aligning to complicated cross-modal reasoning. The videos are sampled from different datasets and segmented into 30-second or 60-second intervals to systematically evaluate model performance across different temporal contexts. We further introduce **Daily-Omni QA Generation Pipeline** which encompasses five automated modules: video annotation, annotation revision, audio-visual temporal alignment, QA generation, and QA optimization. This framework demonstrates remarkable scalability: a single annotator can complete quality filtering process for these 1197 high-quality QA pairs within 30 hours, achieving an approximate 30% acceptance rate from initially generated candidates. In addition, we propose **Daily-Omni Agent**, an audio-visual agent utilizing open-source Visual Language Model (VLM), Audio Language Model (ALM) and Automatic Speech Recognition (ASR) model without further finetuning to establish a baseline for this benchmark. We evaluated recent MLLMs and our agent on the Daily-Omni benchmark, where the Daily-Omni Agent achieved **state-of-the-art** performance among open-source methods. The experimental results show that current MLLMs still face significant challenges in tasks requiring deep audio-visual temporal integration. They also reveal that by combining existing visual and audio language models with simple temporal alignment techniques, as demonstrated by our Daily-Omni Agent, substantially improved performance can be achieved, underscoring a promising direction for enhancing multimodal reasoning.

2 Related Works

2.1 Multimodal Large Language Models

Recent advancements in Multimodal Large Language Models (MLLMs) primarily fall into three categories: Audio Language Models (ALMs) incorporating audio capabilities (Tang et al., 2024; Gong et al., 2024b; Chu et al., 2023; 2024; Ghosh et al., 2025; 2024; Goel et al., 2025), Visual Language Models (VLMs) adding visual understanding (Liu et al., 2025b; Bai et al., 2025; Lu et al., 2024a; Wang et al., 2024), and Omni-modal Language Models (OLMs) combining both audio and visual modalities (Pichai et al., 2024; Li et al., 2025; Zhang et al., 2024a; Team et al., 2024; Cheng et al., 2024; Liu et al., 2025c; Microsoft et al., 2025; Fu et al., 2025b; Guo et al., 2025; Xu et al., 2025;

Liu et al., 2025a; Team et al., 2025). These MLLMs often employ a modular architecture, utilizing separate encoders for audio and visual inputs. In the audio domain, Radford et al. (2023); Chen et al. (2023); LI et al. (2024) proposed audio encoders to extract sound representations. Some models Tang et al. (2024); Liu et al. (2025c); Zhang et al. (2024a) even integrate multiple audio encoders specialized for different sound types (e.g., speech, music). Similarly, visual encoders (Li et al., 2023; Dai et al., 2023; Liu et al., 2023; Dehghani et al., 2023) are used to process images and videos. However, this modular approach struggles to capture the crucial temporal correlations inherent in synchronized audio-visual streams. This limitation arises because audio and visual inputs are encoded independently. Although multi-modal positional embedding techniques like TMRoPE (Xu et al., 2025) enhance cross-modal temporal understanding to some extent, effective methods for temporally aligning multimodal data remain relatively scarce. Additionally, in several MLLMs (Microsoft et al., 2025; Fu et al., 2025b; Zhang et al., 2024a) , the audio encoder primarily serves to process user instructions–akin to how the text modality is used–rather than perceiving the environment.

2.2 Audio-Visual Understanding Datasets and Benchmarks

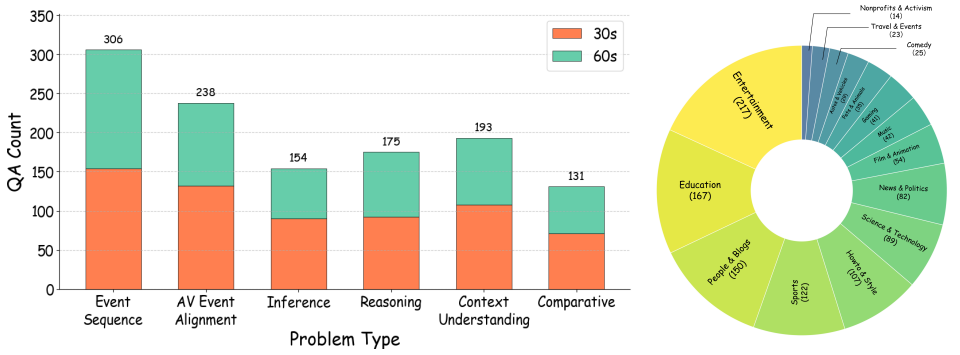
The development of uni-modal datasets and associated tasks for audio and vision has driven advancements in Audio Language Models (ALMs) and Visual Language Models (VLMs). Visual datasets and benchmarks (Li et al., 2024b; Fu et al., 2025a; Wu et al., 2021; Mangalam et al., 2023; Liu et al., 2023; Wu et al., 2024; Zhang et al., 2024b; Yue et al., 2024; 2025; Li et al., 2024a; Fu et al., 2024; Gao et al., 2017; Hu et al., 2025) primarily focus on tasks for static images (OCR, grounding, segmentation, classification, and question-answering) and dynamic videos (captioning, temporal grounding, and understanding), while audio datasets and benchmarks (Ghosh et al., 2025; Gemmeke et al., 2017; Chen et al., 2020; Panayotov et al., 2015; Yang et al., 2024) address speech-related tasks (ASR, emotion recognition, and entity recognition) and non-speech tasks (such as sound classification and audio grounding). However, while attempts at creating audio-visual datasets date back to at least 2021 (Yun et al., 2021), these early efforts often had significant limitations. For example, Music-AVQA (Li et al., 2022) focused specifically on music performance videos, while Pano-AVQA (Yun et al., 2021) centered on panoramic videos. Others, such as AVQA (Yang et al., 2022) and OmniBench (Li et al., 2024), were restricted to short, simple videos or static images. Furthermore, AV-Odyssey (Gong et al., 2024a) heavily emphasized specific audio tasks, such as recognizing timbre and loudness, rather than broader audio-visual understanding. While WorldSense (Hong et al., 2025), a concurrent work, also provides a valuable benchmark for real-world audio-visual question-answering, the efficient scalability of AVQA datasets and the enhancement of OLM abilities still require further exploration.

3 Daily-Omni

This section details the Daily-Omni framework, which comprises three core components: multi-modal data curation, hierarchical video annotation, and QA synthesis and evaluation. Furthermore, we delineate the evaluation paradigm implemented through the Daily-Omni Agent, which serves as our benchmarking baseline for systematic assessment. Daily-Omni aims to establish a benchmarking framework designed to systematically evaluate MLLMs' ability to perceive and reason in real-life audio-visual scenarios. As illustrated in Figure 2, the Daily-Omni benchmark comprises videos from all **11** YouTube categories. The QAs are meticulously crafted to assess a spectrum of multi-modal capabilities, spanning from basic cross-modal perception to intricate reasoning. All questions are designed to require the integration of audio, visual, and textual information for a correct answer. In total, the Daily-Omni benchmark comprises 684 videos and 1197 QA pairs. Of these QAs, 550 correspond to 60-second videos and 647 correspond to 30-second videos. We compared the features of Daily-Omni with various audio-visual benchmarks in Table 1.

3.1 Data Curation

Daily-life scenarios offer abundant information across both audio and visual modalities. Within the visual modality, our data curation focuses on selecting videos characterized by **significant temporal dynamics**. Conversely, static scenes such as a vlogger addressing the camera with limited motion offer minimal temporal information for questioning and are usually excluded. Regarding the audio modality, recognizing that prior datasets have concentrated heavily on speech, our benchmark is designed to encompass a **broader range of everyday sounds**, such as music, speech, and other sound



(a) Distribution of QA type (b) Distribution of video category

Figure 2: **Distribution of 1197 Daily-Omni QA pairs.**

events. These acoustic signals should be present within the videos, appearing either concurrently or consecutively. Furthermore, to maintain focus on audio-visual processing and avoid challenges related to multilingual text comprehension, videos with languages other than English are also excluded.

Table 1: **Comparison of audio-visual benchmarks.** We detail publication, modality (A: audio, V: video, I: image), size, and question type (MC: multiple choice, DF: defined word, BB: bounding boxes). 'Efficient Scalability' refers to automated expansion methods. 'Open-Domain' and 'General Sound' denote genre and sound diversity.

Benchmarks	Pub	Modality	#Video	#QA Pairs	Question Type	Efficient Scalability	Open-Domain	General Sound
AVQA	ACM MM'22	V+A	57,015	57,335	MC	✗	✓	✗
Music-AVQA	CVPR'22	V+A	9,288	45,867	DF	✗	✗	✓
Pano-AVQA	ICCV'21	V+A	5,400	51,700	DF & BB	✗	✗	✗
OmniBench	ARXIV'24	I+A	✗	1,142	MC	✗	✓	✓
AV-Odyssey	ARXIV'24	I+A	✗	4,555	MC	✗	✓	✓
WorldSense	ARXIV'25	V+A	1,662	3,172	MC	✗	✓	✓
Daily-Omni	–	V+A	684	1197	MC	✓	✓	✓

Our video data originates from AudioSet (Gemmeke et al., 2017), Video-MME (Fu et al., 2025a), and FineVideo (Farré et al., 2024). While AudioSet provides 10-second clips primarily for audio classification, we retrieved the original, full-length source videos. These were then processed into longer 30s or 60s segments which contains the original 10s segments to ensure they contain certain types of sound event. We employ Whisper-V3-Large (Radford et al., 2023) to ensure inclusion of spoken content. For Video-MME and FineVideo, our selection criteria prioritized videos containing substantial audio information alongside rich temporal visual dynamics. The inclusion of these datasets also increases diversity, as they feature more recent videos compared to Audio Set (primarily pre-2017 uploads), thus broadening the representation of genres and styles.

3.2 Data Annotation & QA Construction

We developed a pipeline that employs MLLMs to generate and revise visual and audio annotations. Concurrently, Reasoning Large Language Models (LLMs) are utilized to construct and optimize the associated questions, choices, and answers. To obtain detailed annotations while ensuring cost-effectiveness, we specifically used Gemini 2.0 Flash (Pichai et al., 2024) for the annotation task and Deepseek-R1 (DeepSeek-AI et al., 2025a) for QA construction and optimization. Figure 3 provides a outline of this process.

Segment Annotation Recognizing that even state-of-the-art Multimodal Large Language Models (MLLMs) can exhibit cross-modal hallucinations (Sung-Bin et al., 2025), we employed Gemini 2.0 Flash to annotate the audio and visual modalities independently. Additionally, since MLLM performance is known to degrade when processing long audio clips, we segmented the videos prior to annotation. Specifically, each clip was divided into three equal, shorter segments (e.g., 10s segments for 30s clips, 20s segments for 60s clips) to improve the MLLM's subsequent audio annotation quality.

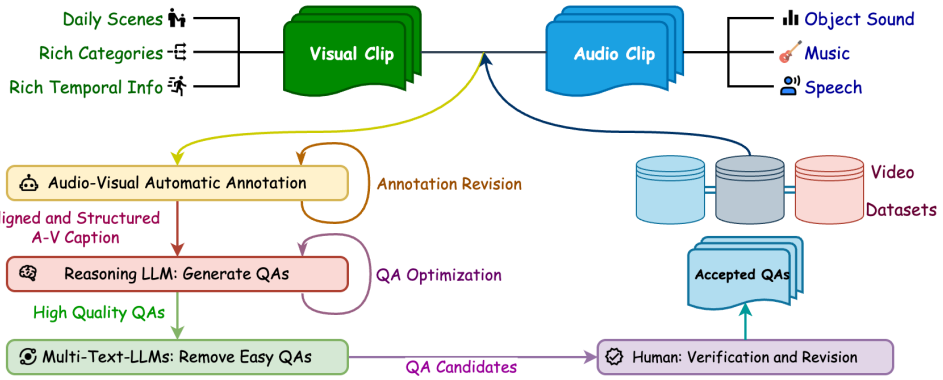


Figure 3: **The outline of Daily-Omni QA construction pipeline.** The arrows indicates the sequence of the processes.

Visual & Audio Revision After obtaining visual annotations from the video segments (processed without audio) and audio annotations from the corresponding audio segments, we use Gemini 2.0 Flash to perform a consistency check on the visual annotations. For this step, we prompt the model with the complete video clip, allowing it to review the segment annotations and ensure overall coherence. For example, referencing the full-length video, the model verifies whether a person described in an early segment is the same individual appearing in subsequent segments and generates a consistent revised annotation. Subsequently, audio annotations undergo refinement using a Reasoning LLM (Gemini 2.0 was employed in the initial phase of the project, with Deepseek-R1 used in later stages). This model leverages the consistent visual annotations to perform cross-modal correction, rectifying sound misidentifications and identifying sound sources. For example, if the audio is annotated as a 'generic impact sound' but the visual annotation for the same segment shows a 'door slamming shut', the Reasoning LLM uses this visual context to correct the audio description to 'door slamming sound' and attributes the sound's origin to the observed door.

Visual & Audio Event Alignment At this stage, we have sequences of visual and audio events annotated within consecutive 10s or 20s segments. However, these segment-level annotations do not explicitly specify the temporal alignment between individual visual and audio events – i.e., which specific events occurred simultaneously. To establish this precise cross-modal event concurrency, we proposed a novel **event aligning technique**. By prompting Gemini 2.0 Flash with the complete audio-visual clip, we instruct it to identify the visual event(s) occurring concurrently with each identified audio event. These aligned audio-visual event pairs provide sufficient information to infer the temporal relationships between any audio and visual event within the sequence. The details of annotation generation, revision and event alignment is shown in Figure 4.

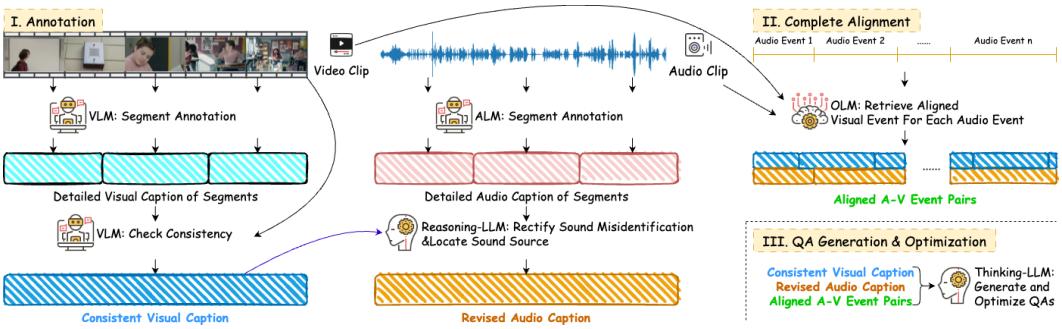


Figure 4: **Details of Daily-Omni annotation generation, revision and event alignment.** For cost-efficiency, we align all events with one query.

QA Construction Using the consistent visual annotations, revised audio annotations, and aligned event pairs derived from each video, we prompted Deepseek-R1 to generate multi-choice questions covering the following types: (1) AV Event Alignment: Questions to determine which audio and visual events occurred simultaneously with each other; (2) Event Sequence: Questions to determine the temporal sequence of visual and audio events in the video; (3) Reasoning: Questions to explain the

cause or reason behind the occurrence of a visual or audio event in the video; (4) Inference: Questions to speculate on information not explicitly presented in the video; (5) Comparative: Questions to compare the similarity or difference between the audio and visual information of two or more events in the video; (6) Context Understanding: Questions to determine the contextual information surrounding a specific event in the video. Since Daily-Omni aims to evaluate MLLM perception and reasoning within real-life audio-visual scenarios, we deliberately **exclude certain question types**. Specifically, counting and measuring questions are omitted, as they often rely on a single modality rather than integrated multi-modal understanding. Similarly, purely knowledge-based queries, such as celebrity identification, fall outside our intended scope.

QA Optimization and Quality Control By avoiding strict question templates, the generated QAs exhibit greater creativity, incorporate complicated logic, and sometimes feature obscure descriptions, making them more closely resemble questions asked in real-life scenarios. However, the generated questions and choices can sometimes contain excessive textual information, potentially allowing powerful models to infer the correct answer using text alone, without engaging with the audio-visual content. Therefore, Deepseek-R1 was employed again to remove superfluous textual information from the questions and choices. Additionally, it replaced obviously incorrect options with more challenging distractors, thereby increasing the difficulty and reducing the potential for correct answers based on guessing alone. Subsequently, we evaluated the optimized questions using two powerful LLMs, GPT-4o (OpenAI et al., 2024) and Deepseek-V3 (DeepSeek-AI et al., 2025b), providing them with only the textual questions and choices (no audio-visual context). Questions that could be answered correctly by both LLMs under this text-only condition were discarded, as they did not necessitate multimodal reasoning. This automated filtering step resulted in approximately 47% of the candidate QAs being discarded. Finally, the remaining QAs underwent manual evaluation for quality control. Human evaluators examined each QA, verifying: (1) that there was exactly one unambiguously correct answer among the choices, (2) that the proposed answer was indeed the correct one, and (3) that answering the question genuinely required comprehensive audio-visual capabilities. Based on this assessment, evaluators either accepted the QA for inclusion in the final benchmark or rejected it. Facilitated by the automated pipeline, the final human evaluation process was highly efficient. A **single annotator** used less than **30 hours** to review the candidates and establish the final set of 1197 QAs, corresponding to an acceptance rate of approximately **30%** during this manual review stage.

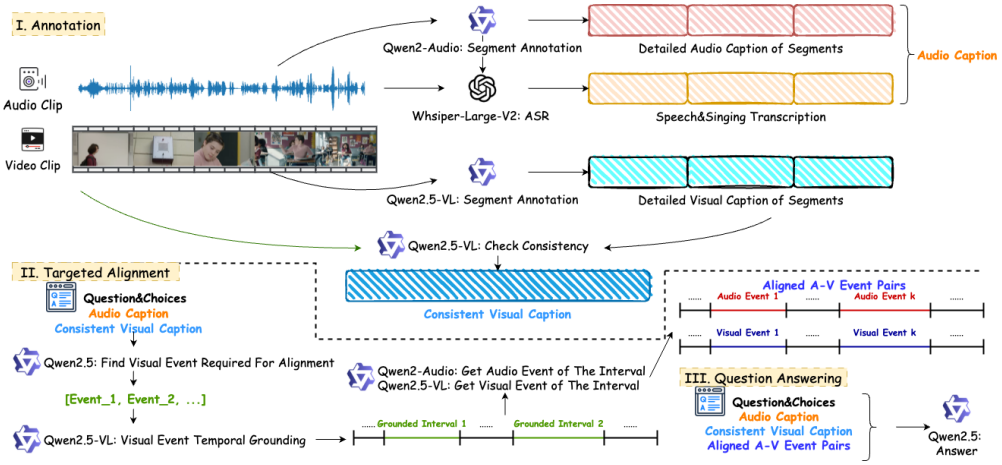


Figure 5: **The outline of Daily-Omni Agent workflow.**

3.3 DAILY-OMNI AGENT

To establish a baseline for MLLMs and demonstrate the importance of temporal awareness in audio-visual question answering, we constructed the **"Daily-Omni Agent"**. This agent, designed to understand audio-visual context and answer related questions, integrates several models: Qwen2-Audio (7B parameters) (Chu et al., 2024), Qwen2.5-VL-7B (Bai et al., 2025), Whisper-Large-V2 (Radford et al., 2023), and the text-based Qwen2.5-14B-Instruct (Qwen et al., 2025). As shown in Figure 5, when presented with a question, its choices, and the associated video context, the agent first divides both the video and audio streams into three segments of equal duration. Subsequently,

it independently generates annotations for these segments: visual annotations using Qwen2.5-VL and audio annotations using Qwen2-Audio. Additionally, we use Whisper-Large-V2 to provide a transciption of speech and singing of each segment, as Qwen2-Audio tends to omit this information in its annotations. Following the segment annotation, the agent utilizes Qwen2.5-VL to perform a **consistency check** on the visual annotations using the complete video, mirroring the revision process described previously (Section 3.2).

While generating aligned audio-visual event pairs would ideally provide richer temporal information for reasoning, implementing this step within the agent presents practical challenges. Unlike the data curation pipeline, the agent lacks access to a highly capable OLM like Gemini 2.0 Flash for precise event alignment. Furthermore, providing a large number of fine-grained event pairs risks overwhelming the context capacity or reasoning capabilities of the agent's LLM. Therefore, as an alternative to full event alignment, we adopt a innovative **targeted approach**: First, we prompt Qwen2.5-14B-Instruct with the visual and audio annotations, the question, and its choices. The model's task is to identify a list of specific **visual events** whose temporal localization is deemed necessary to answer the question correctly. Subsequently, we utilize Qwen2.5-VL-7B, functioning as a video temporal grounding model, to determine the start and end timestamps for each of these identified critical events. If the duration of a grounded event interval falls below a predefined threshold, the agent classifies this interval as critical. It then retrieves descriptions of both visual and audio events occurring within this specific, brief period using the previously mentioned approach, thereby establishing a localized alignment between concurrent events. Finally, we prompted Qwen2.5-14B-Instruct to determine the correct answer by providing it with the question, choices, the previously generated visual and audio annotations, and the extracted aligned event pairs.

4 EXPERIMENT

In this section, we present a comprehensive evaluation of recent Multimodal Large Language Models (MLLMs) on our benchmark to delineate their capability boundaries. Furthermore, we conduct ablation studies to investigate the key factors influencing model performance.

4.1 SETTINGS

Our evaluation encompassed three distinct types of models. Firstly, we examine **OLMs**, including several open-source contenders like VideoLLaMA 2 (Cheng et al., 2024), Unified-IO 2 (Lu et al., 2024b), Qwen2.5-Omni (Xu et al., 2025), Ola (Liu et al., 2025c) and our **Daily-Omni Agent**, as well as the proprietary Gemini models (Pichai et al., 2024). Secondly, we assess **VLMs** including Qwen2.5-VL (Bai et al., 2025) and GPT-4o (OpenAI et al., 2024) and **ALMs** including Audio Flamingo 3 (Goel et al., 2025) and Qwen2-Audio (Chu et al., 2024). We also evaluate the performance of some OLMs when provided with only visual inputs. Finally, we test some **LLMs** such as Deepseek-V3 (DeepSeek-AI et al., 2025b), GPT-4o (OpenAI et al., 2024) and Qwen2.5-14B(Qwen et al., 2025) without visual and audio inputs to check whether our benchmark contains too much information in questions and choices. All model evaluations are conducted strictly according to instructions from their respective official repositories and accompanying documentation (such as "cookbooks" or developer guides).

4.2 MAIN RESULTS

Table 2 presents comprehensive evaluation results, shedding light on the real-world audio-visual understanding capabilities of contemporary MLLMs and our proposed **Daily-Omni Agent**. Firstly, earlier Omni-modal Language Models (OLMs) such as Unified-IO 2 and VideoLLaMA 2 demonstrate limited performance on our benchmarks. Notably, their results are, in several instances, even surpassed by text-only LLMs. Furthermore, the Unified-IO 2 series displays a perplexing degradation in performance with increasing model size, an observation consistent with findings from OmniBench (Li et al., 2024). A plausible explanation is an inadequate capacity within these language models to effectively process and synthesize cross-modal information.

Secondly, while overall performance still presents challenges, more recent open-source OLMs such as the Qwen2.5-Omni series and Ola demonstrate reasonable proficiency in inference and reasoning tasks. This suggests a foundational understanding of general visual and audio contexts. However,

their efficacy significantly diminishes on temporally-sensitive tasks like audio-visual event alignment and context understanding, where results remain largely unsatisfactory. Despite incorporating cross-modal positional embeddings designed to provide temporal awareness, these models still struggle to address questions requiring precise temporal capabilities. Conversely, proprietary OLMs such as Gemini 2.0 Flash exhibit markedly superior cross-modal temporal capabilities, attaining the leading overall performance score of **67.84%**. The observation that even top-tier OLMs have not yet breached the 70% accuracy threshold underscores the demanding nature of this benchmark, positing it as a valuable and pertinent objective for the advancement of contemporary OLMs.

Table 2: **Performance comparison of MLLMs on Daily-Omni**. Boldface and underline indicate the top two performers. Subscripts on 'Avg' for visual-only OLMs show the performance drop from their audio-visual counterparts. Random guess accuracy is 25%. See Appendix B for evaluation details.

Methods	AV Event Alignment	Comparative	Context Understanding	Event Sequence	Inference	Reasoning	30s Subset	60s Subset	Avg
Omni-Modal Language Models (With Visual and Audio)									
Qwen2.5-Omni (7B)	44.12	51.15	38.86	40.52	57.79	61.71	46.68	48.36	47.45
Qwen2.5-Omni (3B)	38.66	48.09	33.68	33.99	54.55	44.00	42.35	38.36	40.52
Ola (7B)	40.34	61.07	40.41	43.46	63.64	69.71	51.47	49.82	50.71
Unified-IO-2 L (1B)	27.31	22.90	26.42	27.78	29.87	29.14	27.67	27.09	27.40
Unified-IO-2 XL (3B)	30.25	30.53	25.39	29.08	33.12	21.71	28.13	28.55	28.32
Unified-IO-2 XXL (8B)	25.63	31.30	26.42	25.82	35.06	29.71	26.74	30.00	28.24
VideoLLaMA2 (7B)	35.71	35.88	35.75	31.70	40.91	34.29	38.02	31.82	35.17
Gemini 2.0 Flash	**62.18**	**73.28**	**63.73**	**63.72**	<u>76.62</u>	**75.43**	**67.23**	**68.55**	**67.84**
Gemini 2.0 Flash Lite	<u>55.04</u>	64.89	58.03	<u>54.25</u>	74.03	<u>72.00</u>	62.44	<u>60.00</u>	61.32
Daily-Omni (ours)	51.68	<u>68.70</u>	<u>60.10</u>	53.92	**78.57**	71.43	<u>63.99</u>	59.27	<u>61.82</u>
Omni-Modal Language Models (Visual Only)									
Qwen2.5-Omni (7B)	38.24	48.85	34.72	36.27	51.95	45.71	40.80	41.64	$41.19_{-6.3}$
Qwen2.5-Omni (3B)	33.61	42.75	36.27	33.33	49.35	38.86	38.79	36.55	$37.76_{-2.8}$
Gemini 2.0 Flash	39.08	64.12	56.48	56.21	67.53	62.29	56.57	55.45	$56.06_{-11.8}$
Gemini 2.0 Flash Lite	43.70	58.02	53.89	45.10	64.29	60.57	53.01	51.64	$52.38_{-8.9}$
Visual Language Models (Visual Only)									
GPT-4o	47.90	62.60	52.33	52.61	66.23	66.29	55.64	57.45	56.47
Qwen2.5-VL (7B)	36.97	46.56	33.68	37.91	51.95	44.00	39.26	42.36	40.68
Qwen2.5-VL (3B)	35.71	43.51	34.72	33.66	43.51	39.43	37.71	37.09	37.43
Audio Language Models (Audio Only)									
Audio Flamingo 3 (7B)	40.76	55.73	43.01	40.52	65.58	68.00	50.23	49.45	49.87
Qwen2-Audio (7B)	28.99	35.88	27.46	32.03	33.77	33.14	31.22	31.82	31.50
Textual Language Models (Without Visual and Audio)									
GPT-4o	33.19	43.51	28.50	30.39	44.81	46.86	36.48	36.18	36.34
Deepseek-V3 (671B)	31.93	41.22	29.02	29.41	44.81	46.29	35.24	36.00	35.59
Qwen2.5-Instruct (14B)	30.25	39.69	27.98	28.43	42.21	42.86	32.15	35.82	33.83

Thirdly, upon ablating audio input and providing only visual data, all evaluated OLMs exhibit a substantial decline in performance. Notably, the magnitude of this performance loss tends to be greater for models that demonstrated superior initial performance when leveraging both modalities. The significant performance drop, especially in high-performing models, when deprived of audio, validates that the tasks within this benchmark genuinely require and benefit from the integration of both auditory and visual information, rather than being solvable by visual cues alone. This dependence on both modalities is further evidenced by the performance of leading Visual Language Models (VLMs). For instance, Qwen2.5-VL (7B) achieves only 40.68% accuracy, and even the highly capable GPT-4o (with its visual input) reaches just 56.47%. Given that these are state-of-the-art models within their respective size categories (sub-7B and larger proprietary models), their comparatively modest performance on our benchmark further underscores its value in demanding **genuine audio-visual integration**, a capability absent in previous multi-modal benchmarks.

These results reveal that: (1) Current OLMs demonstrate reasonable general audio and visual understanding capabilities when processing videos under 60 seconds in duration. However, they continue to struggle with tasks demanding sophisticated cross-modal temporal awareness or the integration of information across extended time intervals. (2) Our proposed question-answer generation pipeline proves effective in creating challenging evaluation instances that accurately probe the audio-visual understanding capabilities of MLLMs.

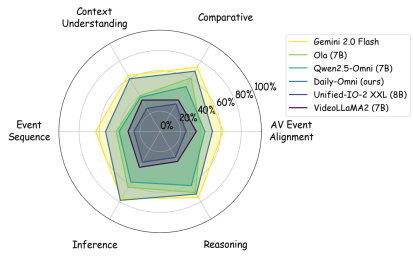


Figure 6: **MLLMs' accuracy over different question categories**

Align Method	Average Accuracy	Aligned Event Pairs per Question
No Alignment	60.65	0
Naive Alignment	59.65	0.68
Smart Alignment	**61.82**	**1.11**

Table 3: **Impact of different aligning methods on average accuracy and aligned event pairs per question.**

4.3 Results of Daily-Omni Agent

Our Daily-Omni Agent, leveraging Qwen2.5-VL-7B, Qwen2-Audio (7B), and Qwen2.5-Instruct-14B, achieves **61.82%** overall accuracy. This performance is **state-of-the-art** among open-source methods and surpasses smaller proprietary MLLMs. The success of our agent suggests that by simply leveraging VLM and ALM with time-period-level alignment coupled with event-level alignment for some key events, we are able to create a very strong OLM.

To further study the effect of event alignment, we conducted an ablation study evaluating the Daily-Omni Agent's performance under three alignment scenarios: (1) **No Alignment**: Generating comprehensive visual and audio captions as a base, but omitting any form of event alignment. (2) **Naive Alignment**: Generating comprehensive visual and audio captions, followed by a step where the VLM is prompted to extract question-relevant events and their temporal boundaries (begin/end times). The ALM is then utilized to produce audio captions specifically for those extracted segments and attaining aligned audio-visual event pairs in the process. (3) **Smart Alignment**: Generating comprehensive visual and audio captions and aligned events as stated in Section **3.3**. Note that for (2) and (3), we only consider generating an align event pair when the provided segment has a duration below a certain threshold to prevent confusion. The overall accuracy and the average number of aligned event pairs identified per question for each method are presented in Table 3. As expected, the Smart Alignment method achieves the highest average accuracy. This demonstrates that explicitly identifying and aligning relevant events significantly boosts the model's overall performance compared to simply generating global captions. Conversely, the Naive Alignment method exhibited a slight decline in accuracy relative to the No Alignment baseline. Careful analysis of the generated events and their temporal boundaries suggests that this outcome is likely attributable to the limitations of the Qwen2.5-VL-7B model. It appears that this model is not sufficiently powerful to reliably identify and retrieve the target event through a single query. Moreover, the temporal grounding process itself frequently yields imprecise results, leading to erroneous alignment. This issue with temporal grounding affects even the Smart Alignment method, occasionally producing incorrect or confusing aligned event pairs. Consequently, we hypothesize that equipping the agent with a more powerful open-vocabulary video temporal grounding model would unlock further significant improvements in its performance, mitigating the impact of imprecise temporal grounding.

5 Conclusion

This paper introduced Daily-Omni, a novel Audio-Visual Question Answering benchmark designed to evaluate MLLMs on temporally-aligned multimodal reasoning in daily life scenarios. We also proposed an efficient Daily-Omni QA Generation Pipeline and the training-free Daily-Omni Agent, a strong open-source baseline. Our evaluation revealed that while recent MLLMs show general audio-visual understanding, they significantly struggle with precise cross-modal temporal awareness The results underscore our pipeline's value and effectiveness in creating such challenging questions. The Daily-Omni Agent, with its targeted temporal alignment, achieved competitive results, and our ablation study confirmed that a **Smart Alignment** strategy improves performance, unlike naive approaches. In summary, Daily-Omni highlights the current MLLM frontiers in complex audio-visual temporal reasoning. Future efforts should prioritize developing more accurate and robust multimodal temporal grounding techniques for truly sophisticated audio-visual understanding.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, February 2025. URL http://arxiv.org/abs/2502.13923. arXiv:2502.13923 [cs].

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020. doi: 10.1109/ICASSP40776.2020.9053174.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 5178–5193. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/chen23ag.html.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs, October 2024. URL http://arxiv.org/abs/2406.07476. arXiv:2406.07476.

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, December 2023. URL http://arxiv.org/abs/2311.07919. arXiv:2311.07919 [eess].

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, 2024.

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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha,

Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025a.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025b.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Peter Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey A. Gritsenko, Mario Lucic, and Neil Houlsby. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. `https://huggingface.co/datasets/HuggingFaceFV/finevideo`, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, June 2025a.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. VITA-1.5: Towards GPT-4o Level Real-Time Vision and Speech Interaction, January 2025b. arXiv:2501.01957 [cs].

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, 2024.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5277–5285, 2017. doi: 10.1109/ICCV.2017.563.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18959–18969, 2025.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6288–6313, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Forty-second International Conference on Machine Learning*, 2025.

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

Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han, Benyou Wang, Yutong Bai, Zhuoran Yang, and Xiangyu Yue. AV-Odyssey Bench: Can Your Multimodal LLMs Really Understand Audio-Visual Information?, December 2024a. arXiv:2412.02611 [cs].

Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2022. doi: 10.1109/icassp43922.2022.9746828.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. Listen, think, and understand. In *The Twelfth International Conference on Learning Representations*, 2024b.

Qingpei Guo, Kaiyou Song, Zipeng Feng, Ziping Ma, Qinglong Zhang, Sirui Gao, Xuzheng Yu, Yunxiao Sun, Tai-WeiChang, Jingdong Chen, Ming Yang, and Jun Zhou. M2-omni: Advancing Omni-MLLM for Comprehensive Modality Support with Competitive Performance, February 2025. arXiv:2502.18778 [cs].

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. WorldSense: Evaluating Real-world Omnimodal Understanding for Multimodal LLMs, February 2025. arXiv:2502.04326 [cs].

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos, 2025. URL https://arxiv.org/abs/2501.13826.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking Multimodal Large Language Models . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, Los Alamitos, CA, USA, June 2024a. IEEE Computer Society. doi: 10.1109/CVPR52733.2024.01263.

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

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19086–19096, 2022. doi: 10.1109/CVPR52688.2022.01852.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li, Yuqi Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Linchu Xiong, Yozhen Wu, Jiahui Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang, Fuzhong Chen, Xuezhen Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yijia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi Ma, Kegeng Wu, Jia li, Aiyuan Yang, Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin Huang, Kun Li, Wenjing Luo, Yifei Duan, Lingling Zhu, Ran Xiao, Zhe Su, Jiani Pu, Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yujing Qiao, Lei Zhang, Yanjun Shen, Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao, Hao Liang, Youquan Li, Yanzhao Qin, Linzhuang Sun, Jianhua Xu, Haoze Sun, Mingan Lin, Zenan Zhou, and Weipeng Chen. Baichuan-Omni-1.5 Technical Report, January 2025. arXiv:2501.15368 [cs].

Yizhi LI, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. MERT: Acoustic music understanding model with large-scale self-supervised training. In *The Twelfth International Conference on Learning Representations*, 2024.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. OmniBench: Towards The Future of Universal Omni-Language Models, October 2024. arXiv:2409.15272 [cs].

Che Liu, Yingji Zhang, Dong Zhang, Weijie Zhang, Chenggong Gong, Haohan Li, Yu Lu, Shilin Zhou, Yue Lu, Ziliang Gan, Ziao Wang, Junwei Liao, Haipang Wu, Ji Liu, André Freitas, Qifan Wang, Zenglin Xu, Rongjuncheng Zhang, and Yong Dai. Nexus-O: An Omni-Perceptive And -Interactive Model for Language, Audio, And Vision, March 2025a. arXiv:2503.01879 [cs].

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, Ranjay Krishna, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. Nvila: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4122–4134, June 2025b.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the Frontiers of Omni-Modal Language Model with Progressive Modality Alignment, February 2025c. arXiv:2502.04328 [cs].

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024a.

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 (CVPR)*, pp. 26439–26455, June 2024b.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, Mei Gao, Min Gao, Amit Garg, Abhishek Goswami, Junheng Hao, Amr Hendy, Yuxuan Hu, Xin Jin, Mahmoud Khademi, Dongwoo Kim, Young Jin Kim, Gina Lee, Jinyu Li, Yunsheng Li, Chen Liang, Xihui Lin, Zeqi Lin, Mengchen Liu, Yang Liu, Gilsinia Lopez, Chong Luo, Piyush Madan, Vadim Mazalov, Arindam Mitra, Ali Mousavi, Anh Nguyen, Jing Pan, Daniel Perez-Becker, Jacob Platin, Thomas Portet, Kai Qiu, Bo Ren, Liliang Ren, Sambuddha Roy, Ning Shang, Yelong Shen, Saksham Singhal, Subhojit Som, Xia Song, Tetyana Sych, Praneetha Vaddamanu, Shuohang Wang, Yiming Wang, Zhenghao Wang, Haibin Wu, Haoran Xu, Weijian Xu, Yifan Yang, Ziyi Yang, Donghan Yu, Ishmam Zabir, Jianwen Zhang, Li Lyna Zhang, Yunan Zhang, and Xiren Zhou. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras, 2025.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A. J. Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz,

Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. GPT-4o System Card, October 2024. arXiv:2410.21276 [cs].

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. Introducing Gemini 2.0: our new AI model for the agentic era, December 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1050.

Qwen, :, 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, Tianyi Tang, 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, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

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. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models, 2025.

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*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Reka Team, Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, Kaloyan Aleksiev, Lei Li, Matthew Henderson, Max Bain, Mikel Artetxe, Nishant Relan, Piotr Padlewski, Qi Liu, Ren Chen, Samuel Phua, Yazheng Yang, Yi Tay, Yuqi Wang, Zhongkai Zhu, and Zhihui Xie. Reka Core, Flash, and Edge: A Series of Powerful Multimodal Language Models, April 2024. arXiv:2404.12387 [cs].

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. Covost 2 and massively multilingual speech translation. In *Interspeech 2021*, pp. 2247–2251, 2021. doi: 10.21437/Interspeech.2021-2027.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need, 2024.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Zhifei Xie and Changqiao Wu. Mini-Omni2: Towards Open-source GPT-4o with Vision, Speech and Duplex Capabilities, November 2024. arXiv:2410.11190.

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, and Junyang Lin. Qwen2.5-Omni Technical Report, March 2025. arXiv:2503.20215 [cs].

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 3480–3491, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548291.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. AIR-bench: Benchmarking large audio-language models via generative comprehension. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.109. URL https://aclanthology.org/2024.acl-long.109/.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. doi: 10.1109/CVPR52733.2024.00913.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.736. URL https://aclanthology.org/2025.acl-long.736/.

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360° videos. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2011–2021, 2021. doi: 10.1109/ICCV48922.2021.00204.

Pan Zhang, Xiaoyi Dong, Yuhang Cao, Yuhang Zang, Rui Qian, Xilin Wei, Lin Chen, Yifei Li, Junbo Niu, Shuangrui Ding, Qipeng Guo, Haodong Duan, Xin Chen, Han Lv, Zheng Nie, Min Zhang, Bin Wang, Wenwei Zhang, Xinyue Zhang, Jiaye Ge, Wei Li, Jingwen Li, Zhongying Tu, Conghui He, Xingcheng Zhang, Kai Chen, Yu Qiao, Dahua Lin, and Jiaqi Wang. InternLM-XComposer2.5-OmniLive: A Comprehensive Multimodal System for Long-term Streaming Video and Audio Interactions, December 2024a. arXiv:2412.09596 [cs].

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data, 2024b.

A Case Study

This section delves into selected case studies of model performance on Daily-Omni. Such an analysis of specific examples offers deeper insights into both the nature of the Daily-Omni benchmark and the current capabilities and limitations of the models. Figure 7 presents three illustrative questions from the Daily-Omni benchmark. It displays the responses of the Gemini 2.0 Flash and Qwen2.5-Omni models under two conditions: processing full audio-visual inputs versus visual-only inputs.

Case 1: Chronological Event Ordering This question requires determining the first event to occur in the video from a given list of audio-visual events. To correctly ascertain the existence and chronological position of choices B (Player discussing a personal milestone) and D (Female presenter

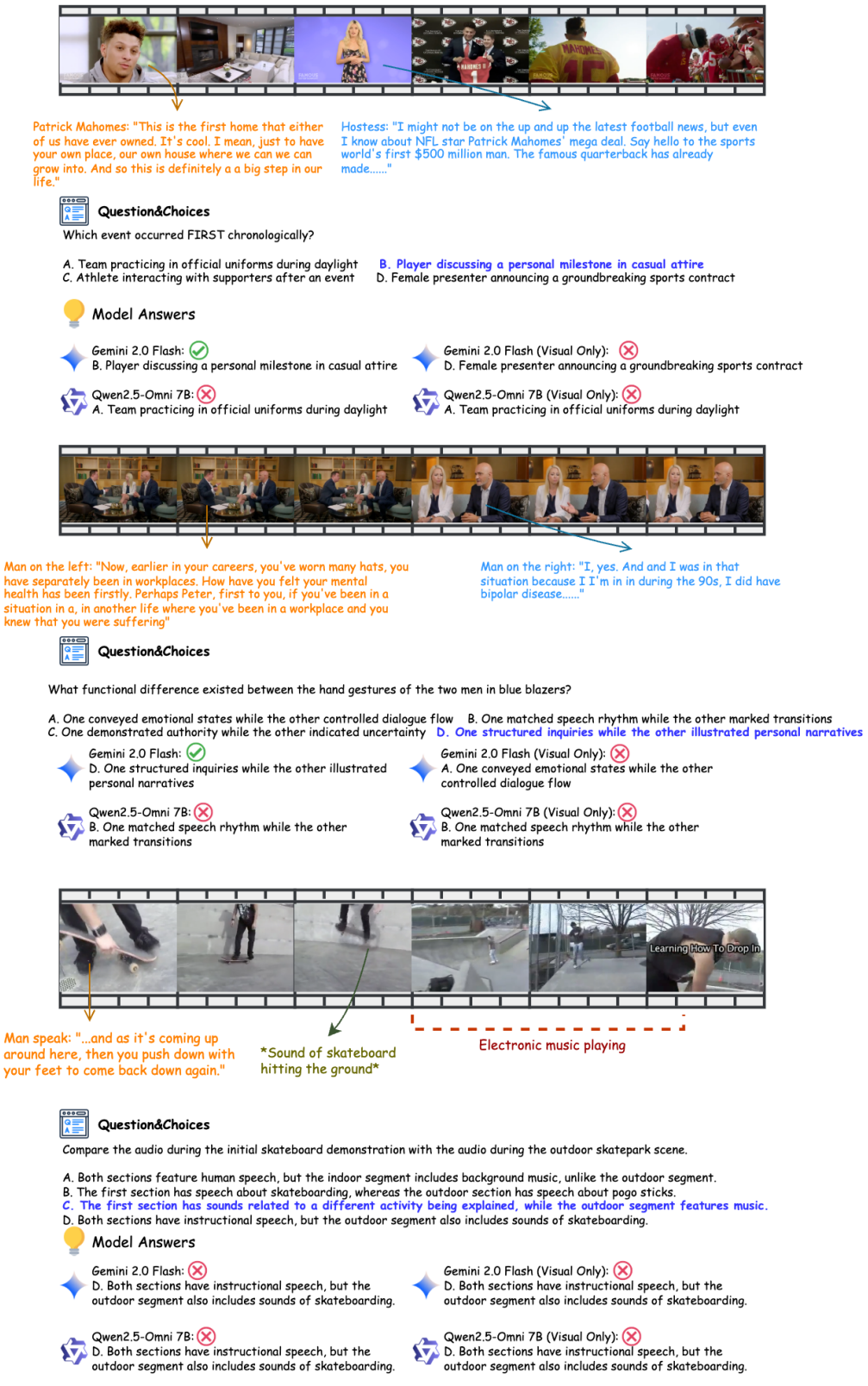


Figure 7: **OLM responses to Daily-Omni questions: A comparison of audio-visual versus visual-only inputs.** The figure presents examples of OLM performance when using full audio-visual modalities compared to visual-only input. For each case, it shows the audio-visual content, the question with choices (correct answer highlighted), and the models' answers, indicating correctness for each input condition.

announcing a groundbreaking sports contract), the model must leverage both audio and visual sensory inputs. Gemini 2.0 Flash successfully identifies choice B as the initial event, whereas Qwen2.5-Omni 7B fails to do so. Both model failed to answer correctly when only taking visual input. This disparity suggests that merely possessing multi-modal sensory capabilities is insufficient; models also require robust cross-modal reasoning to effectively integrate audio-visual information and achieve correct temporal ordering.

Case 2: Audio-Visual Inference This case assesses the ability to discern the functional differences between the hand gestures of two individuals engaged in a discussion. While visual input captures the gestures themselves, audio input is also critical for interpreting their communicative purpose as it reveals what is being said and who said it. Gemini 2.0 Flash correctly identified the functional difference (D) with audio-visual input but failed with visual-only input, selecting a more generic but incorrect interpretation (A). Qwen2.5-Omni 7B, failed to make the correct distinction in both audio-visual and visual-only conditions, selecting an unreasonable option (B). This question underscores that merely understanding the speech content is insufficient; accurately identifying the speaker for each utterance is also crucial. Successfully attributing speech requires robust audio-visual temporal integration—the ability to precisely synchronize the spoken audio with the visual depiction of the active speaker. Following this, strong reasoning ability is needed to infer the distinct roles and communicative intents (e.g., inquirer vs. narrator) based on the combined audio-visual information, and thereby understand the different functions of their gestures. The failure of Qwen2.5-Omni 7B, even with audio-visual input, suggests potential limitations in one or both of these sophisticated cross-modal capabilities.

Case 3: Sound Understanding This question directly tests the model's ability to analyze and compare audio characteristics across two different segments of a video-an skateboard demonstration and an outdoor skatepark scene. The task involves identifying distinct audio elements such as human speech, specific sound effects (e.g., skateboard impact), and background music. This is fundamentally an audio-centric task where visual context primarily helps situate the sounds. Strikingly, both Gemini 2.0 Flash and Qwen2.5-Omni failed to correctly compare the audio content, selecting the same incorrect option (D) irrespective of whether they received audio-visual or visual-only input. Their failure in the audio-visual condition suggests a significant challenge in fine-grained auditory scene analysis, such as distinguishing non-verbal sound event and music, or accurately mapping these perceived audio features to the descriptive choices provided.

B Evaluation Details

The models listed in Table 2 were evaluated as follows: Qwen2.5-Omni (7B and 3B), Ola, Unified-IO-2 (L, XL, and XXL), VideoLLaMA2, and Qwen2.5-VL (7B and 3B) were deployed and tested on a local server. In contrast, Gemini 2.0 Flash, Gemini 2.0 Flash Lite, GPT-4o, and Qwen2.5-Instruct (14B) were evaluated via their respective APIs.

For the Daily-Omni Agent, Qwen2.5-VL-7B-Instruct and Qwen2.5-14B-Instruct were accessed via the API provided by Alibaba Cloud's Bailian platform (accessible at `https://bailian.console.aliyun.com/`) to ensure operational efficiency. Similarly, Whisper-Large-V2 was utilized via the API provided by OpenAI. The Qwen2-Audio component, on the other hand, was deployed and tested locally. It is important to note that all component models of Daily-Omni are capable of local deployment. According to official documentation from Aliyun (https://help.aliyun.com/zh/model-studio/model-user-guide/) and OpenAI (https://openai.com/index/introducing-chatgpt-and-whisper-apis/), the Qwen models offered through Bailian and the Whisper-Large-V2 model accessed via the OpenAI API are identical to their respective open-source or standard versions. Therefore, no performance differences are anticipated.

Our code implements direct passing of local video path to the Qwen2.5-VL API. However, this functionality might require you to contact Aliyun customer service to enable. If direct path input is not activated, you can alternatively pass a list of video frames, though this may result in suboptimal performance.