

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 OMNIVIDEOBENCH: TOWARDS AUDIO-VISUAL UNDERSTANDING EVALUATION FOR OMNI MLLMS

Anonymous authors

Paper under double-blind review

ABSTRACT

Recent advances in multimodal large language models (MLLMs) have demonstrated substantial potential in video understanding. However, existing benchmarks fail to comprehensively evaluate synergistic reasoning capabilities across audio and visual modalities, often neglecting either one of the modalities or integrating them in a logically inconsistent manner. To bridge this gap, we introduce **OmniVideoBench**, a large-scale and rigorously designed benchmark dedicated to assessing synergistic audio–visual understanding, with a strong emphasis on modality complementarity and logical consistency. Specifically, OmniVideoBench comprises **1000** high-quality question–answer(QA) pairs, each annotated with step-by-step reasoning traces, derived from 628 diverse videos ranging from **several seconds to 30 minutes**, and manually verified to guarantee complete correctness and uniqueness. Moreover, OmniVideoBench encompasses **13** carefully designed question types, covering temporal reasoning, spatial localization, counting, causal inference, summarization, and beyond, thereby capturing the essential challenges of video understanding. Evaluation of multiple MLLMs on OmniVideoBench reveals a pronounced gap between model performance and human reasoning, with open-source models lagging significantly behind their closed-source counterparts, underscoring the inherent difficulty of genuine audio–visual reasoning. We will release OmniVideoBench to foster the development of MLLMs with stronger and more generalizable reasoning capabilities.

1 INTRODUCTION

Multimodal large language models (MLLMs) have recently made impressive progress in bridging vision, language, and audio (Yin et al., 2024; Song et al., 2025; Cheng et al., 2025). While early benchmarks primarily focused on image–text alignment or visual reasoning (Xu et al., 2025c; Chen et al., 2024c; Yue et al., 2024a), the integration of video and audio presents a quite different challenge: models must jointly process long temporal sequences, dynamic scene transitions, and complementary acoustic cues. Despite rapid advances, evaluation of MLLMs on audio–visual reasoning remains underdeveloped. Existing benchmarks (Li et al., 2024a; Hong et al., 2025) often *(i) focus on short video clips that underrepresent long-term temporal dependencies, (ii) emphasize a single modality (e.g., vision) while treating audio as auxiliary or optional*. As a result, current evaluations fail to capture the challenges inherent to comprehensive video understanding, where audio and vision must be integrated consistently and logically to support robust inference.

To address these limitations, we introduce **OmniVideoBench**, a high-quality benchmark designed for evaluating audio–visual reasoning abilities in MLLMs. Specifically, first, we collect 628 diverse videos spanning up to 30 minutes across 8 major categories and 68 subcategories, covering realistic contexts such as news, sports, documentaries, vlogs, and ego-centric recordings. Then, we construct 1,000 high-quality question–answer pairs based on these videos, and each pair is annotated with step-by-step reasoning chains as shown in Figure 1, where these reasoning steps explicitly indicate modality and evidence information. This design not only strengthens the reliability of the evaluation but also provides a unique signal for analyzing how models reason, rather than just the final answers.

Based on our OmniVideoBench, we conduct extensive evaluations of both closed-source and open-source MLLMs, and several insightful findings are as follows:

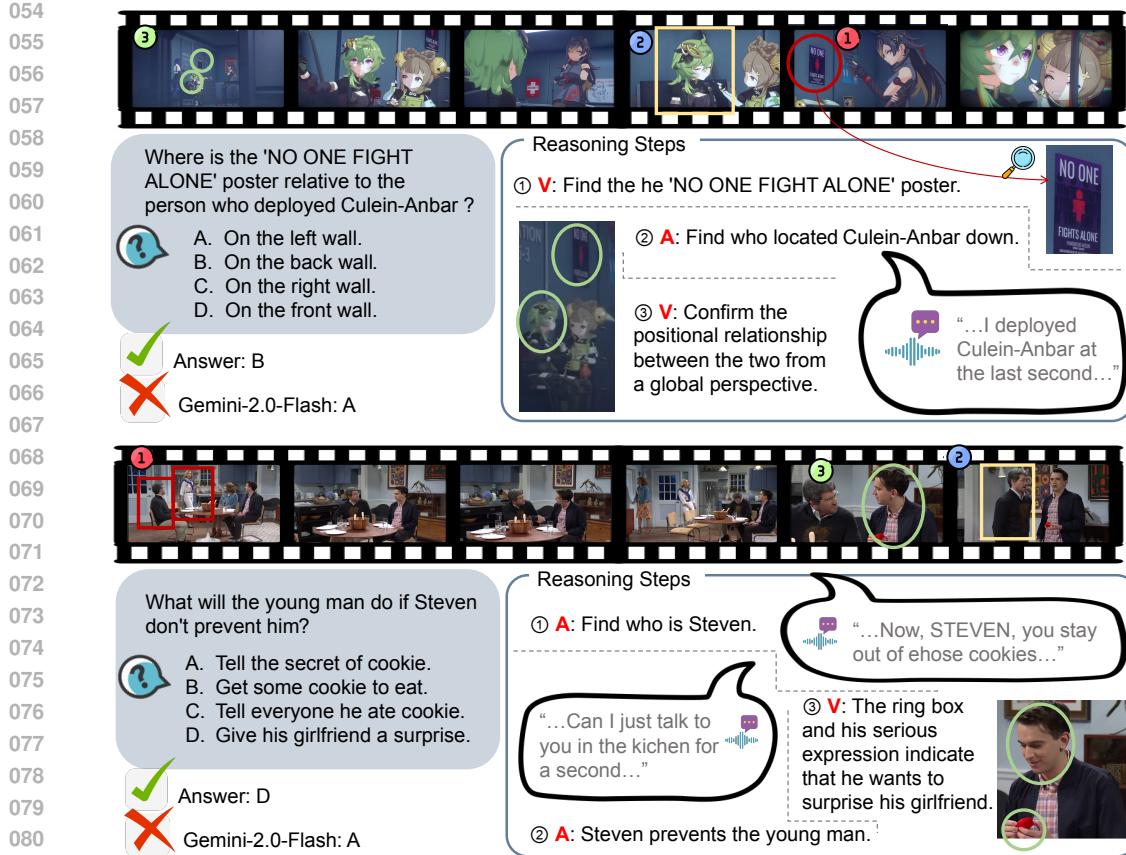


Figure 1: Examples in OmniVideoBench (“V” presents vision and “A” presents audio), and we present the atomic reasoning traces for these examples.

- **OmniVideoBench poses significant challenges for Omni-Modal Language Models.** Current MLLMs have not achieved a passing score (<60%) on OmniVideoBench. The best-performing model, Gemini-2.0-Pro, only achieves an accuracy of 58.90%. Except for the newly proposed Qwen3-Omni, the performance of open-source models is close to random.
- **Omni-understanding abilities on long videos have significant improvement room.** Although some leading models (such as Gemini-2.5-pro) demonstrate relatively robust performance on long videos, other models (e.g., Gemini-2.0-Flash, Qwen3-Omni-30B-A3B) still struggle on long video understanding.
- **Performance varies a lot for videos with different audio signals.** Gemini-2.5-Pro only achieves 38.46% accuracy on videos with music signal, while the results on sound and speech are 57.72% and 61.66%, respectively.
- **Performance on different task types differs a lot.** For example, Gemini-2.5-Pro achieves accuracy below 50% on the background and music understanding task, which requires low-semantic acoustic cues (e.g., musical style, tempo changes), and the accuracy results on the relationship reasoning and summarization tasks are more than 80%.

2 OMNIVIDEOBENCH

2.1 OVERVIEW

OmniVideoBench is a benchmark for evaluating the audio-visual collaborative reasoning of MLLMs. The main task in the evaluation requires a model to process a video, its audio, and associated text to generate a textual answer supported by explicit reasoning steps. This process assesses the model’s ability to synthesize information across modalities, from recognizing objects to

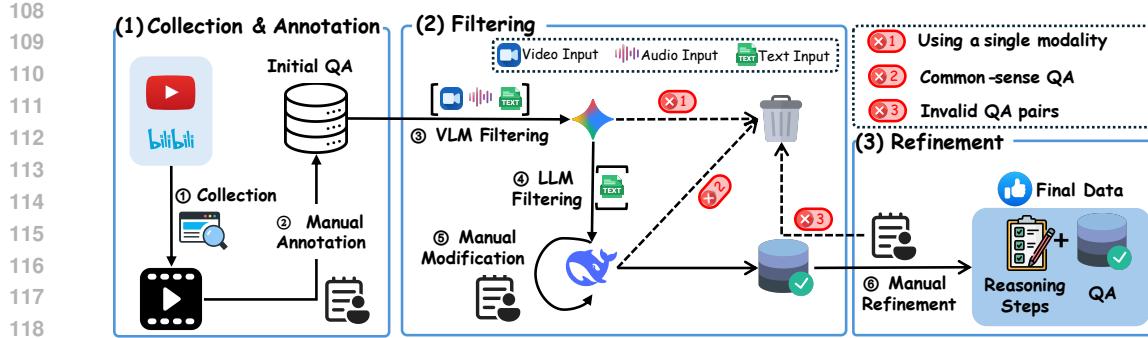


Figure 2: The complete pipeline of data collection, annotation, and refinement, where filtering and refinement serve as two key processes for quality assurance.

comprehending complex scene dynamics and context. This section details the benchmark’s design principles, annotation protocols, and dataset statistics.

2.2 VIDEO COLLECTION

OmniVideoBench is composed of real-world videos sourced from YouTube¹ and Bilibili². These videos feature rich audiovisual content; therefore, comprehensive understanding necessitates the accurate processing and integration of both audio and visual modalities for reasoning.

Regarding video richness, we primarily focus on two dimensions: type and duration. For type diversity, we categorize videos into eight broad classes: Vlog, News, Cartoon, Sports, Documentary, TV, Ego, and Others. Each class is further subdivided into nearly seventy fine-grained subcategories, which facilitates video retrieval and ensures broad coverage. Video categories are unevenly distributed. News and documentary videos have dense audio that nearly covers visual content, making them unsuitable for audio-visual reasoning tasks; thus, we manually controlled the video type distribution. For duration diversity, we restrict video lengths to the range of several seconds to 30 minutes, so as to evaluate reasoning across varying temporal scales.

Building upon this foundation, we established a set of rigorous video collection criteria that not only ensure the quality of the videos themselves, like resolution, but also guarantee the richness and diversity of their audio and visual content. To further avoid data overlap with existing training sets (e.g., popular TV shows), we restrict the selection to recent publications. The detailed collection principles are provided in Appendix B.

2.3 DATA ANNOTATION

After collecting high-quality videos, we carried out manual annotation. Compared with automated annotation, automated methods cap the evaluation ceiling by the capabilities of the annotating model, whereas manual annotation produces questions that are closer to real-world needs.

In Figure 2, we first designed multiple-choice questions consisting of the question stem, the correct answer, and several distractors, to facilitate convenient evaluation of model performance. At this stage, we obtained approximately 2,500 QA pairs. We categorize the tasks into 13 types: Fine-grained Perception, Spatial Reasoning, Attribute Comparison, Background & Music Understanding, Counting, Temporal Understanding, Summarization, Sentiment Analysis, Causal Reasoning, Relationship Reasoning, Reference Reasoning, Ego Reasoning, and Hypothetical Reasoning. In this design, each question is required to rely on audio-visual reasoning, and the answer must be both correct and unique with no alternative plausible interpretations in the video. Moreover, we require that questions should not depend on video resolution or frame rate. Cases where the target object is extremely small, blurred, and barely recognizable to the human eye, or where the relevant event occurs only within an instant, are excluded. In addition, we established the following rules to minimize the interference caused by extraneous textual information.

¹<https://www.youtube.com/>

²<https://www.bilibili.com/>

- **Questions should avoid redundant information.** We minimize unnecessary details in the question text, such as the gender, clothing, or exact speech of characters, as long as doing so does not affect the correctness or uniqueness of the answer. This serves two purposes: reducing textual cues the model could exploit and increasing question difficulty to better test its audio-visual understanding.
- **The length of answers is capped.** To prevent the answer text itself from providing excessive cues to the model, which could reduce the extent to which the evaluation reflects its understanding and reasoning over audio and visual modalities, we impose a limit on answer length. This constraint ensures that the results more faithfully capture the model’s multimodal comprehension and reasoning capabilities.
- **The format of options must be consistent.** Here, “format” refers to aspects such as length, tone, style, and variation patterns. If these features are inconsistent, they may provide the model with unintended cues for reasoning. For instance, when three options are considerably longer than the remaining one, when three options adopt a casual tone while the other is markedly formal, such discrepancies undermine the assumption that each option should have an equal probability of being chosen, thereby compromising the fairness of the evaluation.
- **Negative options must be relevant to the question.** We require that all distractors appear in the video and maintain relevance to the question. Without this constraint, the model could easily eliminate distractors, greatly reducing the need for reasoning.
- **Options should maintain a consistent semantic distance.** We formalize semantic distance as the number of differing semantic units between options. Let an option o_i be represented as a set of semantic units S_i . The semantic distance between two options o_i and o_j is defined as:

$$d(o_i, o_j) = |S_i \Delta S_j| \quad (1)$$

where Δ denotes the symmetric difference, capturing the distinct semantic units between two options. To prevent models from exploiting unbalanced textual cues rather than performing genuine audio-visual reasoning, we require that all distractors have consistent distances from one another and from the correct option.

2.4 QUALITY ASSURANCE

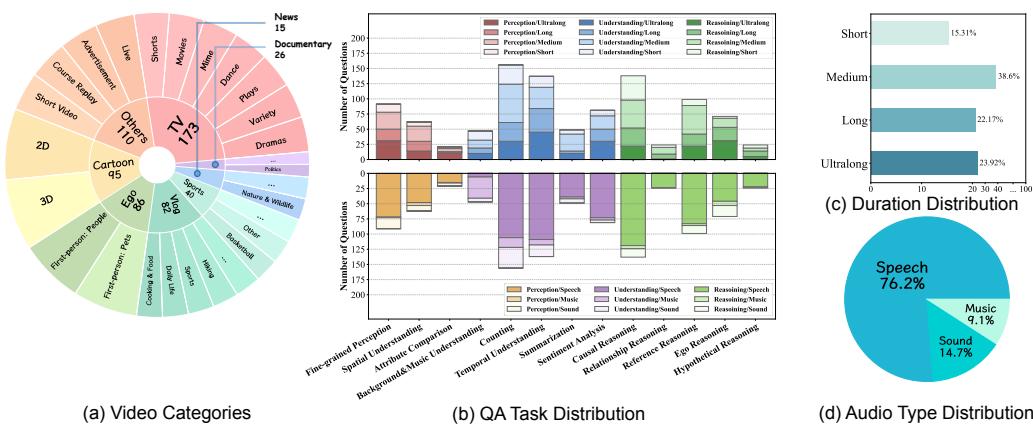
We employed an advanced MLLM (i.e., Gemini 2.0 Flash), with strong audiovisual perception and comprehension capabilities, as well as long-context processing ability, to filter out questions that could be resolved using only a single modality. If the model successfully selected the correct answer with a plausible explanation while relying solely on unimodal information, the corresponding question was removed. After this filtering stage, approximately 1,500 questions were retained.

Subsequently, we employed a large language model, DeepSeek-V3.1 (Liu et al., 2024a), with strong reasoning capabilities to filter out questions that could be answered solely based on textual information. Such cases primarily fall into two categories: first, questions that involve classical, well-known, or universally shared knowledge or objects, which can be answered without reference to the video content; and second, questions where the phrasing of the question, options, or answers provides unintended textual cues. For the former, we directly discarded the questions. For the latter, our annotators reviewed the reasoning process generated by the model and revised the textual formulations to eliminate such biases. After this stage of refinement, 1103 questions were retained.

Another group of annotators conducts the final refinement stage, thoroughly reviewing all questions to identify and remove those with incorrect, non-unique, or mismatched answers. After this validation, annotators enriched each question with step-by-step reasoning chains, where each step consists of three elements: modality, evidence, and inference. The modality specifies whether the step relies on audio or visual information; the evidence denotes the specific information extracted from the video; and the inference describes the reasoning derived from that information. We required each step to be atomic, meaning that it should involve only one modality and capture a minimal unit of evidence, such as a spoken sentence, an action, or the appearance of a character. This design ensures that the reasoning process is both detailed and comprehensive. Through this process, we obtained 1000 high-quality QA pairs with explicit step-by-step reasoning chains, forming a robust dataset for multimodal audio-visual reasoning.

216 Table 1: Dataset statistics divided into video-level and annotation-level information.
217

218 Video Statistics		219 Annotation Statistics	
#Major Categories	8	#Task Types	13
#Subcategories	68	Avg. Question Len.	14.68 words
Avg. Duration	384.24 s	Avg. Answer Len.	4.92 words
Min. Resolution	480p	Avg. Reasoning Steps	5.68
Max. Resolution	1080p	Audio Types (Sp:So:Mu)	762:147:91

239 Figure 3: (a) OmniVideoBench covers 8 major categories and 68 subcategories. (b) OmniVideoBench comprises 13 task types. The above part shows the video duration distribution across
240 different tasks, while the durations are categorized into four groups: “**Short**” for less than 1 minute,
241 “**Medium**” for 1–5 minutes, “**Long**” for 5–10 minutes, and “**Ultralong**” for more than 10 minutes.
242 The lower part illustrates the distribution of three types of audio (i.e., **Speech**, **Sound** and **Music**).
243 (c) Distribution of video durations across four time intervals. (d) Distribution of three audio types.
244

Benchmark	Modality	Qwen2.5-Omni	Multiple Domains	Video Type	Audio Type	Video Duration	Answer Type
AVQA (Yang et al., 2022)	V+A	/	✗	R	So	10	MC
Music-AVQA (Li et al., 2022)	V+A	/	✗	R+S	Mu	60	CLS
AVTRUSTBENCH (Chowdhury et al., 2025)	V+A	/	✓	R+S	Sp+So+Mu	10\60	MC
MMAU (Sakshi et al., 2024)	A	71.0	✓	/	Sp+So+Mu	/	MC
DAVE (Radevski et al., 2025)	V+A	31.0	✓	R+S	So	≤ 60	MC
AV-Odyssey (Gong et al., 2024)	I+A	/	✓	R	Sp+So+Mu	/	MC
AVH Bench (Judgement) (Sung-Bin et al., 2024)	V+A	74.7	✓	R+S	So	10	CLS
OmniBench (Li et al., 2024c)	I+A	56.1	✓	R	Sp+So+Mu	/	MC
Daily-Omni (Zhou et al., 2025)	V+A	47.5	✗	R	Sp+So+Mu	30\60	MC
WorldSense (Hong et al., 2025)	V+A	48.3	✓	R	Sp+So+Mu	15–656	MC
OmniVideoBench (Ours)	V+A	29.3	✓	R	Sp+So+Mu	4–1955	MC

255 Table 2: **Comparisons between different benchmarks and datasets.** V, I, A for modality represent
256 video, image and audio. **Qwen2.5-Omni** represents the performance of Qwen2.5-Omni-7B on these
257 benchmarks. **Multiple Domains** signifies whether the video includes diverse domains. R and S in
258 **Video Type** denote real-world and synthetic data. Sp, So, and Mu represent Speech, Sound, and
259 Music for **Audio Type**, respectively. **Video Duration** represents the duration in seconds. MC, CLS
260 for **Answer Type** indicate Multiple Choice and Classification from fixed vocabulary, respectively.
261262

2.5 DATASET STATISTICS

263264 As shown in Table 1, our OmniVideoBench dataset consists of 628 real-world videos with audio
265 tracks, spanning 8 major categories and 68 subcategories. The videos are of high quality and diverse
266 in content, with an average duration of 384.6 seconds, an average resolution of 480p, about 2k
267 ASR-transcribed tokens per video, and roughly three speakers per video. On the annotation side,
268 OmniVideoBench contains 1000 audio-visual reasoning QA pairs across 13 task types, with an
269 average question length of 14.68 words and an average answer length of 4.92 words. Each QA pair
is annotated with step-by-step reasoning chains averaging 5.68 steps. The reasoning process covers

270 both modalities, with 54% of steps grounded in vision and 46% in audio. There are 762, 147, 91
 271 QA pairs related to Speech, Sound and Music, respectively, highlighting the complementarity of
 272 modalities in multi-step reasoning. Moreover, we provide more detailed statistics in Figure 3.
 273

274 2.6 DATASET COMPARISON 275

276 As shown in Table 2, we compare OmniVideoBench
 277 with representative audio-video benchmarks. While AV-
 278 Odyssey (Gong et al., 2024) and OmniBench (Li et al.,
 279 2024c) operate on single images, OmniVideoBench tar-
 280 gets substantially more challenging videos with du-
 281 rations ranging from a few seconds to 30 minutes. Recent
 282 benchmarks such as AVTrustBench (Chowdhury et al.,
 283 2025), DAVE (Radevski et al., 2025), and MMAU (Sak-
 284 shi et al., 2024) begin emphasizing audio-video coor-
 285 dination, but typically focus on specific capabilities or
 286 short clips. AVHBM (Sung-Bin et al., 2024) also eval-
 287 uates audiovisual consistency, yet its tasks remain cen-
 288 tered on shorter videos and hallucination detection. Omni-
 289 VideoBench, by contrast, expands the scope to diverse
 290 video types, broader temporal spans, and fine-grained
 291 cross-modal reasoning, capturing richer dependencies be-
 292 tween audio and vision. Compared to Daily-Omni (Zhou
 293 et al., 2025) and WorldSense (Hong et al., 2025), which also utilize multi-domain videos, Omni-
 294 VideoBench places greater emphasis on explicit audiovisual collaboration. For instance, dis-
 295 abling audio causes Gemini-2.0-Flash’s performance to plummet to the near-random level, indi-
 296 cating that visual-only cues are insufficient. Furthermore, Figure 4 shows that widely used models
 297 such as Qwen2.5-Omni-7B perform closer to random guessing on our benchmark, indicating that
 298 OmniVideoBench presents significantly greater challenges than existing multimodal datasets.
 299

300 3 EXPERIMENTS

301 3.1 BASELINE MODELS

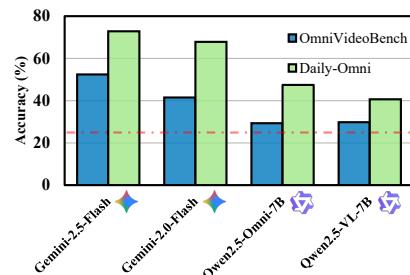
303 We evaluate open-source MLLMs (i.e., Qwen3-Omni series (Xu et al., 2025b), Qwen2.5-Omni
 304 series (Xu et al., 2025a), Baichuan-Omni-1.5 (Li et al., 2025), HumanOmni (Zhao et al., 2025),
 305 MiniCPM-o (Yao et al., 2024), VideoLLaMA2 (Cheng et al., 2024)), and various closed-source
 306 MLLMs (i.e., Gemini-2.5-Pro, Gemini-2.5-Flash (Comanici et al., 2025), and Gemini-2.0-Flash).
 307 We also evaluate the Qwen2.5-VL series (Bai et al., 2025) and DeepSeek-V3.1 (Liu et al., 2024a).
 308

3.2 HUMAN PERFORMANCE

310 It is worth noting that we also report the human test performance as human-level boundaries. We
 311 invited 10 qualified annotators, including 8 graduate students experienced in multimodal research
 312 and 2 experts trained in music-related analysis. Before the evaluation, 50 questions were randomly
 313 selected for 10 testers to answer simultaneously. The fact that the difference in results was no
 314 more than 5 questions indicates that the manual evaluation has minimal deviation and is viable. We
 315 consolidated questions requiring musical knowledge and assigned them to two music experts. The
 316 remaining questions were divided to ensure roughly equal distribution of speech and sound audio
 317 samples per set, which were then evenly assigned to high-level personnel such as graduate students
 318 to obtain persuasive results. Manual responses had no time constraints, yielding a final accuracy rate
 319 of 82.69%. This demonstrates that existing models still fall significantly short of human capabilities.
 320 We appreciate your suggestion and have included the human baseline results in the revised version
 321

3.3 MAIN RESULTS

322 In Table 3, we present evaluation results on OmniVideoBench and have the following observations:
 323



324 Figure 4: Performance comparison of
 325 selected models on OmniVideoBench
 326 and Daily-Omni. “Red line” denotes
 327 random guessing.

324 Table 3: Results of different models. The table reports accuracy on videos across three audio types
 325 and four duration ranges. Boldface highlights the best performance within each column.

327 Models	328 Audio Type			329 Video Duration				330 Avg.
	331 Music	332 Sound	333 Speech	334 < 1 min	335 1-5 mins	336 5-10 mins	337 > 10 mins	
<i>338 Omni-Modal Language Models (With Visual and Audio)</i>								
Gemini-2.5-Pro	38.46	57.72	61.66	57.83	64.43	55.02	55.94	58.90
Gemini-2.5-Flash	39.56	57.04	53.17	55.42	55.10	47.37	52.11	52.40
Gemini-2.0-Flash	29.67	40.27	43.21	49.40	43.15	41.05	34.87	41.50
Qwen3-Omni-30B-A3B	37.36	34.67	39.26	45.78	37.03	38.86	35.11	38.40
Baichuan-Omni-1.5	24.18	31.33	31.36	28.92	31.78	28.38	32.44	30.70
HumanOmni-7B	20.87	31.08	31.61	36.57	29.36	29.60	29.25	30.50
MiniCPM-o	27.47	28.57	30.24	31.43	28.49	34.53	26.15	29.70
Qwen2.5-Omni-7B	23.07	25.33	30.70	41.57	27.41	25.33	26.72	29.30
VideoLLaMA2-7B	26.37	30.67	29.25	32.00	28.20	29.60	28.29	29.20
<i>340 Omni-Modal Language Models (Visual Only)</i>								
Gemini-2.0-Flash	25.27	36.67	30.99	33.73	35.86	32.75	22.48	31.30
Qwen2.5-Omni-7B	27.47	26.67	26.22	28.31	27.11	24.45	25.95	26.40
<i>341 Visual Language Models (Visual Only)</i>								
Qwen2.5-VL-32B	32.97	32.00	31.49	38.55	31.20	29.26	30.53	31.80
Qwen2.5-VL-7B	29.67	31.33	29.51	25.90	30.03	31.88	30.15	29.80
Qwen2.5-VL-72B	26.37	29.33	29.91	33.13	30.03	31.88	24.43	29.50
<i>342 Baseline LLMs</i>								
DeepSeek-V3.1	28.57	26.17	27.28	30.91	27.57	25.00	26.44	27.60

- 348 • **Open-source models still lag significantly behind closed-source models.** Gemini-2.5-Pro
 349 achieves the best performance across most tasks. This underscores the urgent need for current
 350 open-source models to improve in multiple areas, including fine-grained perception, cross-
 351 modal reasoning, and speech awareness.
- 352 • **MLLMs show a performance degradation when dealing with music-related audio.** We ob-
 353 serve that models exhibit lower accuracy in responding to music-dominated videos compared
 354 to those containing human voices or ambient sounds, a phenomenon particularly pronounced
 355 in open-source models. Unlike human voices conveying explicit semantic content or ambi-
 356 ent sounds often corresponding to specific visual events, music primarily encodes abstract
 357 emotional and atmospheric information. Current MLLMs demonstrate limited capability to
 358 translate such implicit cues into effective reasoning, indicating that cross-modal alignment for
 359 emotional and atmospheric understanding remains an urgent challenge to be addressed.
- 360 • **Current MLLMs still have room for improvement in long videos.** Although some leading
 361 models like Gemini-2.5-Pro demonstrate relatively robust performance on long videos, most
 362 MLLMs (e.g., Gemini-2.0-Flash, Qwen3-Omni) still struggle in long videos, which highlights
 363 the widespread challenge in understanding long videos.

364 3.4 FURTHER ANALYSIS

365 **Performance of Models on Tasks across Different Types.** Figure 5 presents a fine-grained com-
 366 parison of model accuracy on the 13 reasoning categories in OmniVideoBench. Several consistent
 367 patterns emerge. (1). Closed-source MLLMs demonstrate superior performance across nearly all
 368 task types. Gemini-2.5-Pro achieves the highest accuracy on 11 out of 13 tasks, demonstrating par-
 369 ticularly strong performance in *Relationship Reasoning*, *Spatial Reasoning*, *Referential Reasoning*,
 370 and *Cause and Effect Reasoning*. These tasks require long-term sequence integration and multi-step
 371 cross-modal reasoning, highlighting Gemini’s strengths in long-context modeling and multimodal
 372 fusion. (2). MLLMs’ understanding of audio remains limited to relatively superficial surface-level
 373 information. Whether open-source or closed-source models, *Background and Music Under-
 374 standing* remains the most challenging task, with even Gemini-2.5-Pro achieving accuracy below 50%.
 375 This is probably because such tasks require linking low-semantic acoustic cues (e.g., musical style,
 376 tempo changes) with high-level reasoning, while current models struggle to master the capability.
 377 In contrast, *Relationship Reasoning* and *Summarization* are relatively easier. This may be because

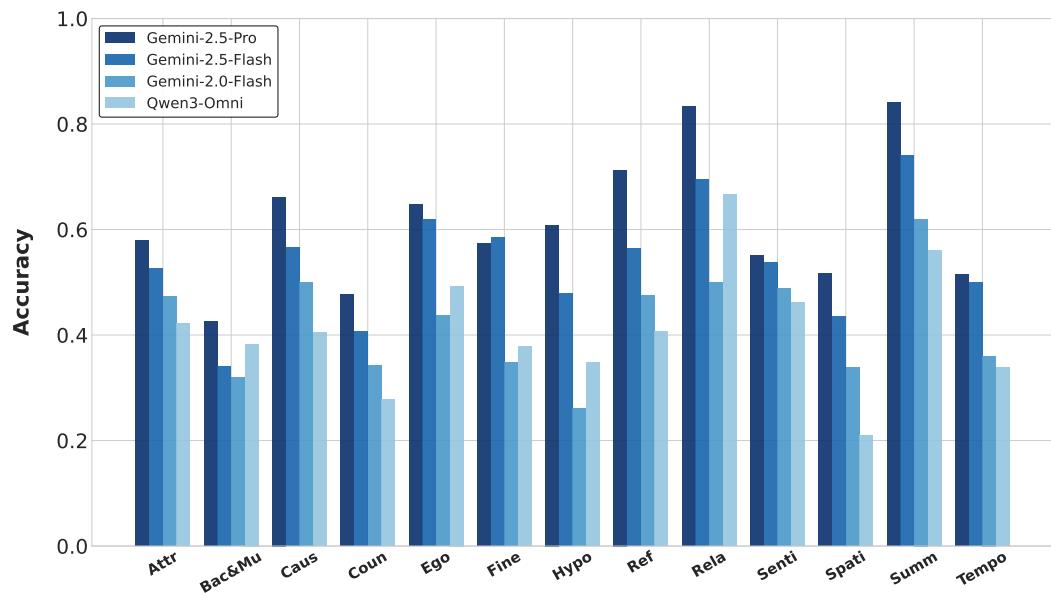
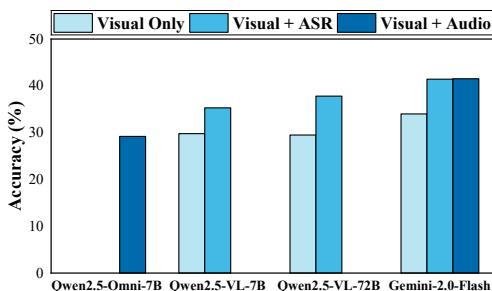


Figure 5: Performance Comparison of some Open-Source and Closed-Source Omni Models on 13 Tasks in OmniVideoBench. Here, “**Attr**”: Attribute Comparison, “**Bac&Mu**”: Background and Music Understanding, “**Caus**”: Cause and Effect Reasoning, “**Coun**”: Counting, “**Ego**”: Ego Reasoning, “**Fine**”: Fine-grained Perception, “**Hypo**”: Hypothetical Reasoning, “**Ref**”: Referential Reasoning, “**Rela**”: Relationship Reasoning, “**Senti**”: Sentiment Analysis, “**Spati**”: Spatial Reasoning, “**Summ**”: Summarization, “**Tempo**”: Temporal Sequencing Understanding.

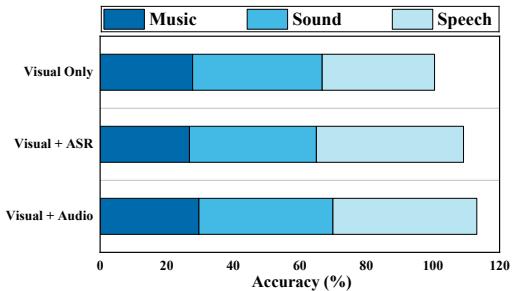
they rely more on recognizing language within audio and visual observation capabilities, and less on cross-modal abstraction abilities.

Effect of ASR Transcripts for Visual Only MLLMs. To further investigate the role of audio information in MLLMs’ reasoning performance, we evaluate several models using both the automatic speech recognition (ASR) transcripts generated by the Voxtral-Mini-3B model (Liu et al., 2025a) and silent video frames as inputs. The results are shown in Figure 6. The observations are as follows: (1). Open-source models demonstrate weaker integration capabilities for audio information compared to their understanding of textual information. In Figure 6a, all tested models demonstrate significantly improved accuracy after extracting ASR text information compared to receiving only visual inputs. However, the Qwen2.5-Omni-7B model, which processes both visual and audio inputs simultaneously, performed even worse than the Qwen2.5-VL-7B model with equivalent parameters. This highlights a common challenge faced by most open-source Omni-Modal Language Models: insufficient cross-modal reasoning capabilities for audio-visual information. (2). In cross-modal video reasoning, audio comprehension capabilities remain irreplaceable by ASR. In Figure 6b, although ASR can help MLLMs achieve decent performance on certain tasks requiring speech recognition capabilities, its effectiveness is extremely limited for tasks demanding deeper and more abstract audio comprehension such as the videos whose audio type is *Music* or *Sound*.

Effect of Different Numbers of Frames. We conduct experiments on Qwen2.5-Omni-7B and Qwen3-Omni-30B-A3B with total frame counts fixed at 32, 64, 128, and 256, respectively, and observe that both models benefit from more frequent time sampling. In Figure 7a, as the total frame counts increase, accuracy steadily improves, likely because richer temporal coverage provides more complete motion cues and reduces the risk of missing key events. As shown in Figure 7b, this improvement becomes more pronounced for longer videos. The consistent gains across different video durations further indicate that dense frame sampling not only captures fine-grained visual dynamics but also strengthens cross-modal alignment. This highlights the importance of dense temporal information and long-context processing for achieving robust audiovisual reasoning.

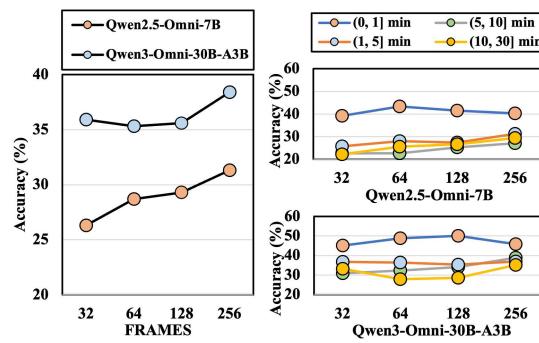


(a) Accuracy rates of selected MLLMs under different inputs.

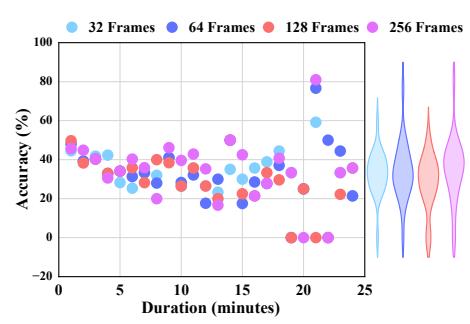


(b) Accuracy of Gemini-2.0-Flash on videos with different audio types.

Figure 6: Accuracy comparison of MLLMs with and without ASR transcripts on OmniVideoBench.



(a) Performance of Qwen2.5-Omni-7B and Qwen3-Omni-30B-A3B at different numbers of frames.



(b) Accuracy of Qwen3-Omni-30B-A3B on questions with videos of varying durations across different numbers of frames.

Figure 7: Performance of selected models when inputting videos with different numbers of frames.

Open-ended QA vs. MCQ. To investigate whether the multiple-choice question (MCQ) format overstates model performance, we additionally evaluated several representative models on open-ended question-answering (QA) tasks, where no predefined answer options are provided. In this setting, models must directly generate textual responses, eliminating both the possibility of random guessing and any lexical cues potentially present in candidate options. In Table 4, the accuracy of all models drops significantly compared to their performance on multiple-choice questions. For instance, the Gemini-2.5-Pro, which leads in MCQ benchmarks, experiences a relative accuracy decline exceeding 14 percent in open-ended scenarios, while open-source models exhibit even steeper drops.

4 RELATED WORKS

Omni-Understanding MLLMs. The development of MLLMs (Chen et al., 2022; Awadalla et al., 2023; Liu et al., 2023) began with a foundational focus on integrating the two primary modalities of vision and language. A recent paradigm shift aims to develop Omni-modal MLLMs capable of processing and generating information across an arbitrary combination of modalities (“Any-to-Any”). This approach positions the LLM as a central cognitive engine, unifying diverse data types like audio, video, and text within its semantic space. This has driven a move from integrating pre-trained unimodal components towards developing “natively multimodal” architectures trained from the ground up, as exemplified by models like GPT-4o (Hurst et al., 2024). This ambition is showcased by state-of-the-art models (Xu et al., 2025a; Zhao et al., 2025; Li et al., 2024b; 2025; Yao

Table 4: Comparison of performance on Open-ended Question Answering (QA) and Multiple- Choice Questions (MCQ) across various models.

Models	Open-ended QA	MCQ
Gemini-2.0-Flash	27.06	41.50
Qwen2.5-Omni-7B	17.25	29.30

486 et al., 2024; Sun et al., 2025; Liu et al., 2025b), which pioneer end-to-end streaming capabilities for
 487 simultaneously processing video and audio to generate text and speech. At the forefront of this
 488 paradigm, proprietary models like Gemini series (Team, 2024; Comanici et al., 2025) demonstrate
 489 pinnacle performance, powered by a natively multimodal design and a massive context window that
 490 together unlock superior understanding of complex, interwoven data streams.

491 **MLLM Benchmarks.** The landscape of MLLM evaluation has matured significantly, evolving from
 492 foundational perception benchmarks (Fu et al., 2023; Liu et al., 2024b; Li et al., 2024a; Yu et al.,
 493 2024a;b; Chen et al., 2024a; Jiang et al., 2025) to more sophisticated frameworks. Recent efforts
 494 probe deeper cognitive abilities, with MLLM-Bench (Ge et al., 2025) assessing a hierarchy of cog-
 495 nitive skills. MMMU (Yue et al., 2023) and MMMU-Pro (Yue et al., 2024b) challenging models
 496 with expert-level, multi-disciplinary reasoning under stricter protocols like vision-only inputs. Si-
 497 multaneously, evaluation has specialized into high-stakes domains such as finance (Gan et al., 2024)
 498 and medicine (Chen et al., 2024b). For video, some benchmarks (Wang et al., 2019; Li et al., 2021;
 499 2023; Fang et al., 2024; Wu et al., 2024; Fu et al., 2024) now focus on the critical challenge of
 500 long-context temporal understanding, revealing key limitations in current models.

501

502 5 CONCLUSION

503

504 We presented OmniVideoBench, a large-scale benchmark for evaluating audio–visual collaborative
 505 reasoning in MLLMs, with diverse videos, carefully verified QA pairs, and explicit reasoning an-
 506 notations. Experiments show that both open- and closed-source models still struggle with modality
 507 complementarity, long-form temporal reasoning, and music understanding, underscoring a large gap
 508 from human-level performance. We hope this benchmark will drive future research toward more ro-
 509 bust and generalizable multimodal reasoning systems.

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540
541
ETHICAL STATEMENTS

542 This work fully adheres to the ICLR Code of Ethics in all aspects of research conduct. The processes
 543 of data collection, usage, annotation, and benchmark construction strictly comply with ethical stan-
 544 dards regarding privacy, consent, and responsible AI practices. Videos in OmniVideoBench are
 545 strictly limited to academic research purposes. Any form of commercial use is prohibited. All video
 546 copyrights remain the property of their original owners. To the best of our knowledge, this study
 547 does not involve any data, methodologies, or applications that raise ethical concerns. The authors
 548 confirm that they have reviewed and followed the ICLR Code of Ethics throughout the entirety of
 549 this research.

550
551
REPRODUCIBILITY STATEMENT
552

553 To ensure the reproducibility of our work, we have made the following comprehensive efforts. We
 554 provide detailed descriptions of the video collection, filtering, annotation, and reasoning-chain con-
 555 struction pipeline in Sec. 2 and Appendix B. We also present complete statistics of the dataset
 556 and explicit definitions of task types, reasoning modalities, and evaluation metrics in Sec. 2.1 and
 557 Appendix A. Furthermore, the prompts used in the experiments are documented in detail in Ap-
 558 pendix C. Extensive replication experiments demonstrate only minor variations across runs, con-
 559 firming the stability and full reproducibility of the reported results. We promise to release the exper-
 560 imental code for evaluation in the future to facilitate verification and benchmarking.

561
562
REFERENCES

563 Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani S.
 564 Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith,
 565 Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo:
 566 An open-source framework for training large autoregressive vision-language models. *ArXiv*,
 567 abs/2308.01390, 2023.

568 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang,
 569 Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*,
 570 2025.

572 Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A
 573 survey on vision-language pre-training. *Machine Intelligence Research*, 20:38–56, 2022.

575 Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Xiangdi
 576 Meng, Tianyu Liu, and Baobao Chang. Pca-bench: Evaluating multimodal large language models
 577 in perception-cognition-action chain. *ArXiv*, abs/2402.15527, 2024a.

578 Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tian-Xin Li, Haodong
 579 Duan, Ziyan Huang, Yan-Cheng Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan
 580 Zhuang, Eric J. Seibel, Junjun He, and Yu Qiao. Gmai-mmbench: A comprehensive multimodal
 581 evaluation benchmark towards general medical ai. *ArXiv*, abs/2408.03361, 2024b.

583 Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M³ cot: A
 584 novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint*
 585 *arXiv:2405.16473*, 2024c.

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

588 Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi
 589 Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and
 590 audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

592 Sanjoy Chowdhury, Sayan Nag, Subhrajyoti Dasgupta, Yaoting Wang, Mohamed Elhoseiny, Ruohan
 593 Gao, and Dinesh Manocha. Avtrustbench: Assessing and enhancing reliability and robustness in
 594 audio-visual llms. *arXiv preprint arXiv:2501.02135*, 2025.

594 Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit
595 Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the
596 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-
597 bilities. *arXiv preprint arXiv:2507.06261*, 2025.

598 Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen.
599 Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *ArXiv*,
600 abs/2406.14515, 2024.

602 Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei
603 Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive
604 evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394, 2023.

606 Chaoyou Fu, Yuhua Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu
607 Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li,
608 Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever
609 comprehensive evaluation benchmark of multi-modal llms in video analysis. *2025 IEEE/CVF
610 Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2024.

611 Ziliang Gan, Yu Lu, Dong Zhang, Haohan Li, Che Liu, Jian Liu, Ji Liu, Haipang Wu, Chaoyou Fu,
612 Zenglin Xu, Rongjunchen Zhang, and Yong Dai. Mme-finance: A multimodal finance benchmark
613 for expert-level understanding and reasoning. *ArXiv*, abs/2411.03314, 2024.

615 Wentao Ge, Shunian Chen, Hardy Chen, Nuo Chen, Junying Chen, Zhihong Chen, Wenya Xie, Shuo
616 Yan, Chenghao Zhu, Ziyue Lin, Dingjie Song, Xidong Wang, Anningzhe Gao, Zhang Zhiyi, Jian-
617 quan Li, Xiang Wan, and Benyou Wang. MLLM-bench: Evaluating multimodal LLMs with
618 per-sample criteria. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025
619 Conference of the Nations of the Americas Chapter of the Association for Computational Linguis-
620 tics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4951–4974, Albuquerque,
621 New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-
622 6. doi: 10.18653/v1/2025.nacl-long.256. URL [https://aclanthology.org/2025.nacl-long.256/](https://aclanthology.org/2025.nacl-long.256).

623 Kaixiong Gong, Kaituo Feng, Bohao Li, Yibing Wang, Mofan Cheng, Shijia Yang, Jiaming Han,
624 Benyou Wang, Yutong Bai, Zhuoran Yang, and Xiangyu Yue. Av-odyssey bench: Can your
625 multimodal llms really understand audio-visual information? *ArXiv*, abs/2412.02611, 2024.

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

630 Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Os-
631 trow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint
632 arXiv:2410.21276*, 2024.

633 Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liupei Wang, Jianhan
634 Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. Mme-cot:
635 Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness,
636 and efficiency. *ArXiv*, abs/2502.09621, 2025.

638 Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan.
639 Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF
640 Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024a.

641 Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji rong Wen, and Di Hu. Learning to answer
642 questions in dynamic audio-visual scenarios. *2022 IEEE/CVF Conference on Computer Vision
643 and Pattern Recognition (CVPR)*, pp. 19086–19096, 2022.

645 Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen,
646 Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video under-
647 standing benchmark. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition
648 (CVPR)*, pp. 22195–22206, 2023.

648 Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohith Krishnan Pillai, Yu Cheng, Luowei
 649 Zhou, Xin Eric Wang, William Yang Wang, Tamara L. Berg, Mohit Bansal, Jingjing Liu, Lijuan
 650 Wang, and Zicheng Liu. Value: A multi-task benchmark for video-and-language understanding
 651 evaluation. *ArXiv*, abs/2106.04632, 2021.

652 Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei
 653 Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Dawei Pan, Shusen Zhang, Xin Wu, Zheng
 654 Liang, Jun Liu, Keer Lu, Yaqi Zhao, Yan-Bin Shen, Fan Yang, Kaicheng Yu, Tao Lin, Jianhua
 655 Xu, Zenan Zhou, and Weipeng Chen. Baichuan-omni technical report. *ArXiv*, abs/2410.08565,
 656 2024b.

657 Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Gu-
 658 osiheng Dong, Dawei Pan, Chong Li, Yuanbo Fang, Dong-Ling Kuang, Mingrui Wang, Chenglin
 659 Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, Yuran Wang, Bowen Ding, Wei Song, Xu Li,
 660 Yuqiu Huo, Zheng Liang, Shusen Zhang, Xin Wu, Shuai Zhao, Lin-Xiao Xiong, Yozhen Wu,
 661 Jia-Reng Ye, Wenhao Lu, Bowen Li, Yan Zhang, Yaqi Zhou, Xin Chen, Lei Su, Hongda Zhang,
 662 Fuzhong Chen, Xu Dong, Na Nie, Zhiying Wu, Bin Xiao, Ting Li, Shunya Dang, Ping Zhang, Yi-
 663 jia Sun, Jincheng Wu, Jinjie Yang, Xionghai Lin, Zhi-Xing Ma, Ke-Ye Wu, Jia Li, Ai-Min Yang,
 664 Hui Liu, Jianqiang Zhang, Xiaoxi Chen, Guangwei Ai, Wentao Zhang, Yicong Chen, Xiaoqin
 665 Huang, Kun Li, Wenjing Luo, Yi qun Duan, Lingling Zhu, Ran Xiao, Zhengquan Su, Jiani Pu,
 666 Dian Wang, Xu Jia, Tianyu Zhang, Mengyu Ai, Mang Wang, Yu Qiao, Lei Zhang, Yanjun Shen,
 667 Fan Yang, Miao Zhen, Yijie Zhou, Mingyang Chen, Fei Li, Chenzheng Zhu, Keer Lu, Yaqi Zhao,
 668 Hao Liang, Youquan Li, Yanzhao Qin, Lin-Lin Sun, Jianhua Xu, Haoze Sun, Mingan Lin, Zenan
 669 Zhou, and Weipeng Chen. Baichuan-omni-1.5 technical report. *ArXiv*, abs/2501.15368, 2025.

670 Yizhi Li, Ge Zhang, Yi Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Jian
 671 Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhen Yang, Xiangzhou Wang, Zhaoxi-
 672 ang Zhang, Zachary Liu, Emmanouil Benetos, Wenhao Huang, and Chenghua Lin. Omnipbench:
 673 Towards the future of universal omni-language models. *ArXiv*, abs/2409.15272, 2024c.

674 Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao,
 675 Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint*
 676 *arXiv:2412.19437*, 2024a.

677 Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lam-
 678 ple, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Mud-
 679 direddy, et al. Voxtral. *arXiv preprint arXiv:2507.13264*, 2025a.

680 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tun-
 681 ing. *ArXiv*, abs/2304.08485, 2023. URL <https://api.semanticscholar.org/CorpusID:258179774>.

682 Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi
 683 Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahu Lin. Mmbench: Is your multi-modal model
 684 an all-around player? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy,
 685 September 29–October 4, 2024, Proceedings, Part VI*, pp. 216–233, Berlin, Heidelberg, 2024b.
 Springer-Verlag. ISBN 978-3-031-72657-6. doi: 10.1007/978-3-031-72658-3_13.

686 Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola:
 687 Pushing the frontiers of omni-modal language model with progressive modality alignment. *ArXiv*,
 688 abs/2502.04328, 2025b.

689 Gorjan Radevski, Teodora Popordanoska, Matthew B Blaschko, and Tinne Tuytelaars. Dave: Diag-
 690 nostic benchmark for audio visual evaluation. *arXiv preprint arXiv:2503.09321*, 2025.

691 S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ra-
 692 mani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio
 693 understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

694 Shezheng Song, Xiaopeng Li, Shasha Li, Shan Zhao, Jie Yu, Jun Ma, Xiaoguang Mao, Weimin
 695 Zhang, and Meng Wang. How to bridge the gap between modalities: Survey on multimodal large
 696 language model. *IEEE Transactions on Knowledge and Data Engineering*, 37(9):5311–5329,
 697 2025. doi: 10.1109/TKDE.2025.3527978.

Wei Sun, Linhan Cao, Yu Shan Cao, Weixia Zhang, Wen Wen, Kaiwei Zhang, Zijian Chen, Fangfang Lu, Xiongkuo Min, and Guangtao Zhai. Engagement prediction of short videos with large multimodal models. *ArXiv*, abs/2508.02516, 2025.

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. *arXiv preprint arXiv:2410.18325*, 2024.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, abs/2403.05530, 2024.

Xin Eric Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4580–4590, 2019.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *ArXiv*, abs/2407.15754, 2024.

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

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfu Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Weiyue Xu, Jiahao Wang, Weiyun Wang, Zhe Chen, Wengang Zhou, Aijun Yang, Lewei Lu, Houqiang Li, Xiaohua Wang, Xizhou Zhu, et al. Visulogic: A benchmark for evaluating visual reasoning in multi-modal large language models. *arXiv preprint arXiv:2504.15279*, 2025c.

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*, 2022.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qi-An Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A gpt-4v level mllm on your phone. *ArXiv*, abs/2408.01800, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12): nwae403, November 2024. ISSN 2095-5138. doi: 10.1093/nsr/nwae403. URL <https://doi.org/10.1093/nsr/nwae403>. eprint: <https://academic.oup.com/nsr/article-pdf/11/12/nwae403/61201557/nwae403.pdf>.

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

Weihao Yu, Zhengyuan Yang, Linfeng Ren, Linjie Li, Jianfeng Wang, Kevin Qinghong Lin, Chung-Ching Lin, Zicheng Liu, Lijuan Wang, and Xinchao Wang. Mm-vet v2: A challenging benchmark to evaluate large multimodal models for integrated capabilities. *ArXiv*, abs/2408.00765, 2024b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, 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 Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2023.

756 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens,
757 Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multi-
758 modal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF*
759 *Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024a.

760 Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun,
761 Ming Yin, Bota Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. Mmmu-
762 pro: A more robust multi-discipline multimodal understanding benchmark. In *Annual Meeting of*
763 *the Association for Computational Linguistics*, 2024b.

764 Jiaxin Zhao, Qize Yang, Yi-Xing Peng, Detao Bai, Shimin Yao, Boyuan Sun, Xiang Chen, Shenghao
765 Fu, Weixuan Chen, Xihan Wei, and Liefeng Bo. Humanomni: A large vision-speech language
766 model for human-centric video understanding. *ArXiv*, abs/2501.15111, 2025.

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

771

772

773

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

799

800

801

802

803

804

805

806

807

808

809

810
811 A FULL VIDEO CATEGORY TAXONOMY
812813 Table 5 shows that videos in OmniVideoBench span 8 major categories and 68 subcategories.
814815 Table 5: Full taxonomy of the video dataset.
816

Main Category	Subcategories
Vlog	Cooking & Cuisine; Travel & Outdoor; Art; Animals; Daily Life at Home; DIY & Handcraft; Gardening; Fitness; Sports; Interviews; Party Games; Makeup & Beauty; Fashion & Styling; Hiking & Trekking
News	Politics; Economy; Society; Technology; Education; Healthcare; Military; Law & Justice; Sports; Culture; Entertainment; Weather; Disaster; Transportation
Cartoon	2D Animation; 3D Animation
Sports	Basketball; Football (Soccer); Volleyball; Badminton; Table Tennis; Swimming; Figure Skating; Skiing; Gymnastics; Wrestling & Judo; Track & Field; Esports; Others
Documentary	Nature & Wildlife; History & Archaeology; Society & Humanity; Politics & Military; Science & Engineering; Medicine & Health; Crime & Law; Art & Culture; Education & Growth; Economy; Environment & Climate; Food & Culinary Culture; Religion & Belief
TV	Short; Dramas & Web Series; Variety; Stage Plays; Dance; Mime; Movies
Others	Live; Advertisement; Course Replay; Short Video
Ego	First-person: People; First-person: Pets

839
840 B DETAILED PRINCIPLES OF VIDEO COLLECTION
841842 To ensure an objective and reliable evaluation of MLLMs, the videos included in the benchmark
843 must satisfy multiple requirements, ensuring diversity in both type and duration. The content should
844 provide rich information across audio and visual modalities, while maintaining complementarity
845 between the two. In other words, the benchmark avoids cases where the visual content can be fully
846 inferred from the audio alone, or where the audio is redundant given the visual stream. Furthermore,
847 since many existing video training datasets overlap with the sources of our benchmark—for ex-
848 ample, clips from Friends—evaluation may otherwise reduce to simple “answer memorization.” To
849 mitigate this unfairness, we additionally consider the publication year of videos when constructing
850 the dataset. The detailed principles for video collection are as follows:
851

- **Video publication date.** Given that most existing training datasets are constructed from YouTube videos, similar to ours, or contain overlapping content such as identical TV shows, we restrict our selection to videos published after June 2024. We use the most recent videos possible to mitigate unfairness and potential overestimation issues arising from the model having already been exposed to similar content during training.
- **Rich dynamic visual information.** The distinguishing feature of videos compared to images lies in their rich dynamic visual information. A prerequisite for evaluating a model’s ability to understand visual information in videos is that the videos themselves contain sufficient dynamic content to be captured and analyzed. Consequently, videos lacking diverse dynamic visual information are excluded, such as those consisting of only several static scenes or perspectives throughout, or those that remain largely static with minimal motion confined to a small corner of the frame.
- **Effective audio information.** In some videos, the audio is completely unrelated to the visual content, such as when only an independent background track is added. We consider

864 such audio to be invalid. To fairly evaluate the model’s capability in audio-visual collabora-
 865 tive reasoning, the audio—whether speech, environmental sound, or music—must align
 866 with the visual content.

867

- 868 • **Absence of subtitle.** We excluded videos with embedded subtitles, as such practices con-
 869 vey most of the audio information visually, enabling models to “cheat” through vision
 870 alone. Likewise, videos containing large text overlays were regarded as undesirable, since
 871 these overlays often directly reveal information about characters’ speech, mental states, or
 872 ongoing events, thereby undermining the assessment of the model’s genuine understanding
 873 and reasoning abilities.
- 874 • **Video resolution.** To ensure video quality, we require a minimum resolution of 480p, and
 875 the visual content must be free from issues such as distortion or blurriness that would hinder
 876 comprehension.

877

878 C PROMPTS USED IN THIS WORK

879 C.1 PROMPT FOR OVERALL EVALUATION

880

881

882 # **Instruction:** You are given a video. Based on the content of the video, answer the follow-
 883 ing question:

884 # **Question:** {Question}

885 # **Options:**

886 A: {Option A} B: {Option B} C: {Option C} D: {Option D}

887 # **Task:**

888 Answer with the option’s letter directly(e.g., A, B, C, or D).

889 If your access to the video content is limited, at least one option that is more likely than the
 890 others must be chosen.

891 Mustn’t give any other reason for can not choose!

892

893 C.2 PROMPT TO SELECT QUESTIONS THAT CAN BE ANSWERED WITHOUT RELYING ON 894 OPTIONS

895

896

897

898 # **Role:** You are an impartial judge.

899 # **Instruction:** Your task is **NOT** to answer the question, but to determine whether the ques-
 900 tion is inherently **DEPENDENT** on the multiple-choice options in order to be answered.

901 # **Task:**

902 We aim to convert this multiple-choice question into an open-ended question.

903 The video content is **NOT** provided here, but you should assume you have fully watched the
 904 video and know everything about it.

905 Your job is **ONLY** to decide whether the question itself ***requires*** the options to be an-
 906 swerable.

907

908 # **Guidelines:**

909 - If the question can still be reasonably answered ****without needing the options****
 910 (even if the exact wording might change slightly), return “**No**”.

911 - If the question cannot be answered at all without the options (e.g., it explicitly asks
 912 “Which of the following...”), return “**Yes**”.

913

914 # **Question:** {Question}

915 # **Answer:** {Answer}

916 **Respond ONLY with “Yes” or “No”.**

917

918 C.3 PROMPT FOR MULTIPLE-CHOICE QUESTIONS WITH STEP-BY-STEP REASONING
919

920
921 **# Instruction:** You are given a video. Based on the content of the video, answer the follow-
922 ing question:
923 **# Question:** {Question}
924 **# Options:**
925 A: {Option A} B: {Option B} C: {Option C} D: {Option D}
926 **# Task:**
927 Note that you should first reason step by step, and then you should give your final choice in
928 A, B, C, or D.
929 Your answer format should be as follows:
930 **Step X:** [Reasoning step X]
931 The final choice is:
932 **\bbox**{ {Answer with the option's letter directly(A, B, C, or D).}}.
933

935 D CASE STUDY
936938 Table 6: Error Distribution Across Models
939

941 Error Category	942 Gemini2.0-Flash	943 Qwen3.0-Omni	944 Qwen2.5-Omni
943 Audio Understanding	17.77%	25.6%	29.2%
944 Temporal Reasoning	6.15%	39.9%	28.2%
945 Multimodal Fusion	61.03%	23.0%	25.6%
946 Visual Understanding	12.03%	8.3%	9.9%
947 Counting	1.9%	2.0%	4.7%
948 Spatial Understanding	1.12%	1.2%	2.4%

950
951 Q: When the man and woman in the picture were discussing ice cubes, why did they notice Superman behind them?
952 Task: causal reasoning
953

```
954 "reasoning_steps": [  
955   {  
956     "modality": "vision",  
957     "evidence": "they notice Superman at  
958     0:37",  
959     "inference": "get the Superman."  
960   },  
961   {  
962     "modality": "vision",  
963     "evidence": "Superman just turned  
964     around and took a step.",  
965     "inference": "get the point."  
966   },  
967   {  
968     "modality": "audio",  
969     "evidence": "Superman made a sound  
970     when he stepped on the wooden floor.",  
971     "inference": "Because Superman made a  
972     sound when he stepped on the wooden  
973     floor."  
974   },  
975   {  
976     "modality": "vision",  
977     "evidence": "B. Because Superman made  
978     a sound when he stepped on the  
979     wooden floor."  
980   }]
```

981 Based on the audio and visual
982 cues in the video, we can
983 determine the reason why the man
984 and woman noticed Superman:
985 1. The man and woman are discussing
986 a glass with ice cubes. The man says,
987 "Yes, that's what happens when
988 water gets cold." This indicates they
989 are focused on the ice cubes.
990 2. Immediately after the woman says,
991 "No, what I'm saying is...", there is a
992 loud, crashing sound.
993 3. The camera cuts to the man in the
994 white robe, who is holding a
995 sandwich. He is startled by the noise,
996 drops his sandwich, and turns around
997 with a shocked expression. The
998 sound is clearly the noise of the
999 sandwich falling and hitting the floor.
1000 4. Therefore, the noise that made them
1001 notice Superman was the sound of
1002 him dropping his sandwich.
1003 Model's answer: "D. Because Superman
1004 made too much noise while eating."

Audio Understanding Error

- ✓ Locate the correct event.
- ✓ Locate the correct event.
- ✗ Misidentifying the content and source of the audio!
- ✗ Incorrect answer.
- ✗

971 Figure 8: A case about Audio Understanding Error.
972

972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994

Q: What is the fourth animal seen in the Safari Zone?
Task: temporal understanding

```
"reasoning_steps": [
  {
    "modality": "audio",
    "evidence": "At 4:50, a woman said 'We are going to the Safari Zone!'.",
    "inference": "This indicates that they would go to see the animals in the Safari Zone afterwards."
  },
  {
    "modality": "vision",
    "evidence": "At 5:08, a brown bear appear in the scene.", ...
  },
  {
    "modality": "vision",
    "evidence": "At 5:32, several siberian tigers appears in the scene.", ...
  },
  {
    "modality": "vision",
    "evidence": "At 5:53, two white lions appear in the scene.", ...
  },
  {
    "modality": "vision",
    "evidence": "At 6:04, a jackal appear in the scene.", ...
  },
]
"answer": "C. Jackal"
```

Based on the sequence of animals shown in the Safari Zone, we can identify the fourth one by observing the order of appearance.

1. The video shows a large bird's nest on the ground, which is not an animal.

2. The camera pans to show a group of elephants in an enclosure.

3. The view shifts to a lion and lioness resting on a grassy slope.

4. The camera then shows a group of cheetahs, with one lying down and two standing alert.

5. Therefore, the fourth animal seen in the Safari Zone is the cheetah.

Model's answer: "D. Cheetah"

Temporal Reasoning Error









Incorrect temporal and sequential understanding !

Incorrect answer.

Figure 9: A case about Temporal Reasoning Error.

995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018

Q: When a blogger presents the process of searching for a skier, which of the following is correct?
Task: summarization

```
"reasoning_steps": [
  {
    "modality": "audio",
    "evidence": "The blogger said he will search for a skier.",
    "inference": "locate this point."
  },
  {
    "modality": "vision",
    "evidence": "The blogger is showing the search for a skier.",
    "inference": "locate this point."
  },
  {
    "modality": "vision",
    "evidence": "There are 10 clips searched overall.",
    "inference": "locate this point."
  },
  {
    "modality": "vision",
    "evidence": "there are 6 clips highly relevant with the video.",
    "inference": "there are 4 videos found that were not highly relevant."
  },
]
"answer": "B. There were 4 clips found that were not highly relevant."
```

Let's analyze the video content to determine the correct answer.

1. The scene where a blogger presents the process of searching for a skier occurs from 2:47 to 2:58.

2. Analyze the options B: The video shows a search for "skier" and displays 10 clips.

3. The blogger does not mention that 4 of them were not relevant. This option is incorrect.

Model's answer: "C. The length of the second viewed clip is 8 seconds."

Multimodal Fusion Error







Locate the correct event.
Correct number of whole clips.
Couldn't reason with cross-modality information!

Incorrect answer.

Figure 10: A case about Multimodal Fusion Error.

1019
1020
1021
1022
1023
1024
1025

To better understand the underlying causes of model successes and failures on OmniVideoBench, we conducted a detailed analysis using three representative models: Qwen2.5-Omni, Qwen3-Omni-30A3B, and Gemini2.0-Flash. We aligned each model's incorrect reasoning chain with the OmniVideoBench ground-truth chains using DeepSeek-V3.1 and categorized the resulting discrepancies into 44 concise error types. These types were further grouped into **six high-level categories: temporal reasoning error, audio comprehension error, multimodal fusion error, visual comprehension error, counting error, and spatial reasoning error**. A human verification study covering 50 randomly sampled errors per category confirmed a high level of agreement with the automatic

1026 Q: While Peggy is complaining to her friends, where is the person with hair the color of the jacket she made for
1027 her husband located in relation to Peggy?

Task: spatial understanding

"reasoning_steps": [
{
 "modality": "vision",
 "evidence": "In 1:16, the jacket appears in the water. The main part of the jacket is kind of orange and the sleeves are red.",
 "inference": "The hair the person is either orange or red."},
 {
 "modality": "audio",
 "evidence": "In 3:16, Peggy is complaining to her friends.",
 "inference": "Her friends appear."},
 {
 "modality": "vision",
 "evidence": "In 3:46, the only person with the red hair is shown, sitting across from Peggy.",
 "inference": "The hair color of the woman across from Peggy matches the color of the jacket."}
],
 "answer": "A. Across from Peggy."}

Based on the scene where Peggy is complaining to her friends, we can determine the location of the person with hair the color of the jacket she made for her husband.

1. Peggy is the woman with brown hair and glasses, sitting in the middle of the three women.
2. The blonde-haired woman is sitting on the right side of the table from Peggy's perspective.
3. Therefore, the person with hair the color of the jacket is located to the right of Peggy.

Model's answer: "C. To the right of Peggy."

Visual Understanding Error

Locate the correct event.
incorrect recognition of jacket and hair colors.

Figure 11: A case about Visual Understanding Error.

annotations. The quantitative comparisons appear in Table 6 and some cases are shown as figs. 8 to 11.

As shown in Figure 12, the first three categories constitute the majority of observed failures for open-source models, indicating that the core bottlenecks of current MLLMs lie in the intersection of *temporal modeling*, *non-speech audio interpretation*, and *cross-modal integration*. These capabilities align closely with the design focus of *OmniVideoBench*, emphasizing long-range temporal dependencies, general acoustic semantics, and robust multimodal alignment.

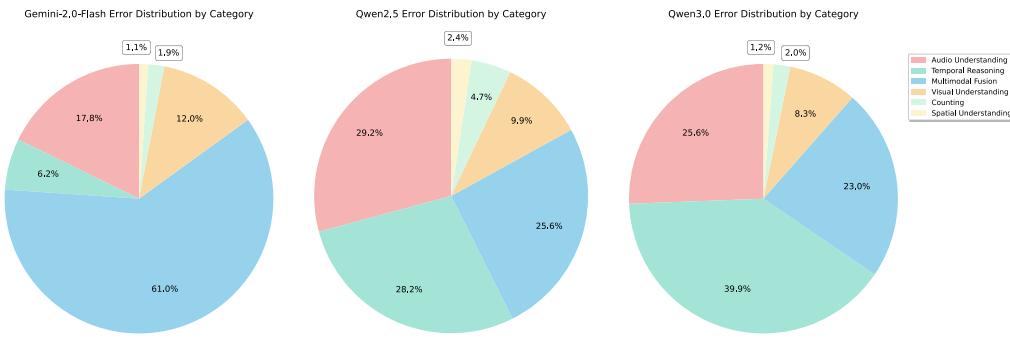


Figure 12: Error distribution of Gemini2.0-Flash, Qwen2.5-Omni and Qwen3-Omni.

Temporal Reasoning Error. By analyzing the distribution of error types across different tasks, as shown in Table 7, we find that models frequently struggle with capturing long-range dependencies and sequential relationships, especially when relevant evidence is distributed across multiple clips. Difficulties are further amplified when visual and auditory streams are not perfectly synchronized, requiring the model to integrate misaligned cues over extended durations. These trends suggest that current MLLMs lack sufficiently expressive mechanisms for hierarchical temporal modeling and consistent aggregation of temporal evidence.

Audio Understanding Error. By analyzing the distribution of error types across different audio types, as shown in Table 8, we find that non-speech audio—such as music and environmental sounds—remains a major challenge because it lacks stable symbolic anchors comparable to those

1080
1081
1082
1083 Table 7: Error distribution by video duration (percentage) for Gemini2.0-Flash, Qwen2.5-Omni, and
1084 Gemini2.0-Flash, Qwen2.5-Omni, and Qwen3-Omni.
1085
1086
1087
1088
1089

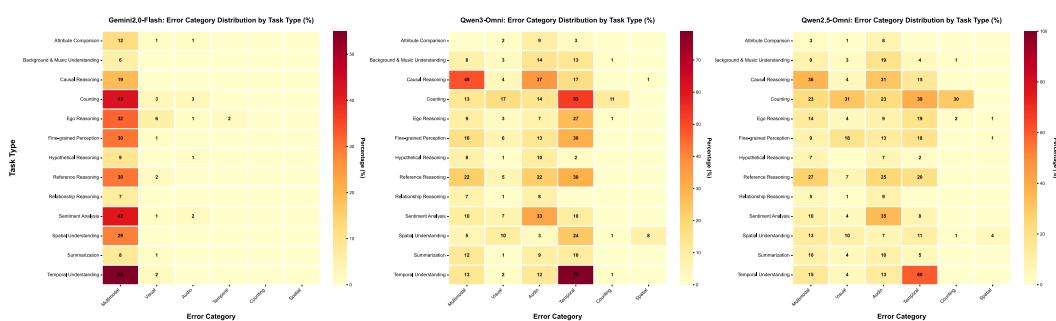
Model	Video Duration	Audio Understanding (%)	Temporal Reasoning (%)	Multimodal Fusion (%)	Visual Understanding (%)	Counting (%)	Spatial Understanding (%)
Gemini2.0-Flash	(0,1] min	16.3	10.8	39.8	9.6	19.3	4.2
	(1,5] min	14.5	10.4	39.7	9.6	18.6	7.2
	(5,10] min	22.5	16.7	31.3	9.3	13.2	7.0
	(10,30] min	27.1	17.2	22.9	16.0	11.5	5.3
Qwen2.5-Omni	(0,1] min	27.7	31.9	27.7	6.7	4.2	1.7
	(1,5] min	23.3	36.8	24.5	10.3	3.2	2.0
	(5,10] min	22.9	45.3	21.2	10.0	0.0	0.6
	(10,30] min	29.3	43.9	20.0	5.4	1.0	0.5
Qwen3-Omni	(0,1] min	25.2	31.1	30.1	7.8	5.8	0.0
	(1,5] min	28.0	26.8	25.9	12.1	5.9	1.3
	(5,10] min	29.7	33.5	24.1	8.2	4.4	0.0
	(10,30] min	33.8	25.6	25.6	10.3	3.1	1.5

1090
1091 in spoken language. Models often fail to map continuous acoustic features (e.g., rhythm, timbre,
1092 intensity) to higher-level semantic interpretations involving events, emotions, or actions. Attribution
1093 mistakes are also common, including misidentifying background music as speech or incorrectly as-
1094 signing off-screen sounds to on-screen entities. This reflects the speech-centric nature of existing
1095 audio encoders and insufficient grounding in general acoustic semantics.

1096
1097 Table 8: Error distribution by audio type (percentage) for Gemini2.0-Flash, Qwen2.5-Omni, and
1098 Qwen3-Omni.
1099

Model	Audio Type	Audio Understanding (%)	Temporal Reasoning (%)	Multimodal Fusion (%)	Visual Understanding (%)	Counting (%)	Spatial Understanding (%)
Gemini2.0-Flash	Music	50.5	9.9	14.3	2.2	17.6	5.5
	Sound	19.0	12.9	23.1	15.6	23.1	6.1
	Speech	16.4	14.3	37.7	11.4	13.9	6.3
Qwen2.5-Omni	Music	15.9	47.8	24.6	8.7	1.4	1.4
	Sound	18.7	52.3	18.7	5.6	4.7	0.0
	Speech	28.0	36.6	23.6	8.8	1.6	1.4
Qwen3-Omni	Music	23.4	28.1	31.2	9.4	7.8	0.0
	Sound	17.8	42.1	21.5	12.1	6.5	0.0
	Speech	32.8	26.0	26.3	9.7	4.0	1.1

1100
1101
1102
1103
1104
1105
1106
1107
1108 **Multimodal Fusion Error.** Multimodal fusion failures manifest primarily in two ways: (1) im-
1109 perfect alignment between visual and auditory cues and (2) *modality neglect*, where the model over-
1110 relies on a single modality while disregarding complementary or corrective information present in
1111 another. For example, a model may rely solely on visual cues to infer an answer while failing to
1112 incorporate crucial auditory evidence. Such issues highlight weaknesses in cross-modal attention
1113 robustness and balanced multimodal reasoning.

1124
1125 Figure 13: Proportion of error types across different tasks for Gemini2.0-Flash, Qwen2.5-Omni and
1126 Qwen3-Omni.

1127
1128 **Counting and Spatial Understanding Error.** Although models demonstrate comparatively weak
1129 performance on *Counting* and *Spatial Understanding* tasks, only a limited portion of errors are
1130 explicitly attributed to these categories. Further inspection, as shown in Figure 13 reveals that
1131 most failures stem from earlier deficits in *fine-grained perception* and *precise temporal localiza-
1132 tion*. These tasks require detecting multiple small entities, identifying subtle spatial orientations,
1133 or localizing objects at specific moments in long video sequences. Errors in early perception or
temporal alignment therefore cascade into incorrect counting or spatial reasoning outcomes.

E USE OF LLMs

Large language models (LLMs) were utilized in this work solely as research tools to assist with data quality control and ancillary writing support. Specifically, we employed advanced multimodal LLMs (e.g., Gemini-2.0-Flash and DeepSeek-V3) to help filter out questions that could be answered using only a single modality and to identify potential textual biases during dataset refinement, as described in Sec. 2.4. In addition, LLMs were used to perform minor language polishing of the manuscript after the main content was written by the authors. All experimental design, dataset construction, analysis, and conclusions were conceived and executed by the authors without automated decision-making. No confidential, private, or sensitive data were provided to any external LLM services.

1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187