JOINTAVBENCH: A BENCHMARK FOR JOINT AUDIO-VISUAL REASONING EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding videos inherently requires reasoning over both visual and auditory information. To properly evaluate Omni-Large Language Models (Omni-LLMs), which are capable of processing multi-modal information including vision and audio, an effective benchmark must comprehensively cover three key aspects: (1) multi-modal dependency (i.e., questions that cannot be answered using vision or audio alone), (2) diverse audio information types (e.g., speech, sound events), and (3) varying scene spans. However, existing datasets fall short in one or more of these dimensions, limiting strict and comprehensive evaluation. To address this gap, we introduce JointAVBench , a novel benchmark with strict audio-video correlation, spanning five cognitive dimensions, four audio information types (speech, sound events, music, vocal traits), and three scene spans (single-, cross-, and full-scene). Given the high cost of manual annotation, we propose an automated pipeline that leverages state-of-the-art vision-LLMs, audio-LLMs, and general-purpose LLMs to synthesize questions and answers that strictly require joint audio-visual understanding. We evaluate leading vision-only, audio-only, and Omni-LLMs on our dataset. Results show that even the best-performing Omni-LLM achieves only 56.2% average accuracy, outperforming uni-modal baselines but revealing substantial room for improvement, especially in cross-scene reasoning.
Project page: `https://jointavbench.github.io`

1 INTRODUCTION

Humans can understand videos and the real world by seamlessly perceiving and integrating both visual and auditory information across different scenes, where diverse audio signals (*e.g.* speech, sound, music, or even vocal traits) are used to complement the visual scene in analyses. As illustrated in Figure 1(a), for a multi-scene video understanding task, such as determining the order across a visual object in scene 3 with the dialogue in scene 1 and scene 23, requires complex joint audio-visual reasoning. This process involves recognizing visual and auditory cues, correlating them across distinct temporal and spatial contexts, and reasoning with the acquired relations. Toward the goal of artificial general intelligence, equipping multimodal large language models (MLLMs) with such joint audio-visual reasoning ability is paramount.

While newly developed Omni-LLMs (Team et al., 2023; Bai et al., 2025; Han et al., 2024; Wu et al., 2024a; Su et al., 2023) aim to process both audio and visual inputs jointly, progress is hindered by the lack of a comprehensive benchmark dedicated to evaluating this crucial capability. Existing benchmarks exhibit several limitations: some lack strict audio-visual correlation controls (Hong et al., 2025; Geng et al., 2024), others primarily focus on static images or simple videos (Li et al., 2024c; Gong et al., 2024), and mostly cover only a limited range of audio types (Yang et al., 2025). Furthermore, nearly all existing benchmarks neglect the complexities of multi-scene reasoning, which is a core component of human cognition.

To address this critical gap, we introduce JointAVBench, the first comprehensive benchmark for evaluating Omni-LLMs' joint audio-visual reasoning capabilities. Our benchmark features a systematic taxonomy covering five cognitive dimensions (e.g., temporal, plot, and long-form reasoning), four audio signal types (vocal traits, music, speech, and sound event), and three distinct scene spans (single-, cross-, and full-scene). These features enable us to construct 15 challenging tasks with strict audio-visual correlations, providing a unified and rigorous evaluation framework. For example,

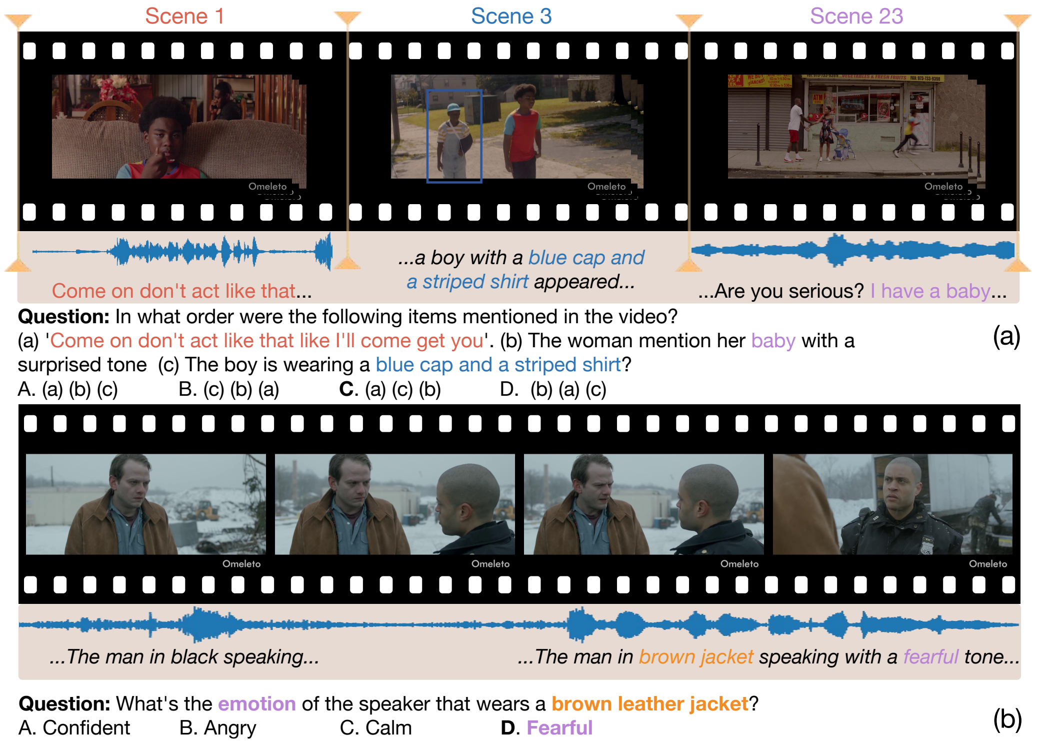

Scene 1
Scene 3
Scene 23
Omeleto
Come on don't act like that...
...a boy with a blue cap and a striped shirt appeared...
...Are you serious? I have a baby...
Question: In what order were the following items mentioned in the video?
(a) 'Come on don't act like that like I'll come get you'. (b) The woman mention her baby with a surprised tone (c) The boy is wearing a blue cap and a striped shirt?
A. (a) (b) (c) B. (c) (b) (a) C. (a) (c) (b) D. (b) (a) (c)
(a)
...The man in black speaking...
...The man in brown jacket speaking with a fearful tone...
Question: What's the emotion of the speaker that wears a brown leather jacket?
A. Confident B. Angry C. Calm D. Fearful
(b)

Figure 1: **Examples of JointAVBench.** (a) asks a cross-scene plot-related question that needs the visual information in Scene 3 and the speech information in Scene 1 and Scene 23 to reason the right answer. (b) asks a single-scene emotion-related question that needs the visual information of the speaker and his vocal traits to answer.

the task in Figure 1 (b) tests Speaker Emotion Recognition (SER), i.e., a single-scene, vocal traits involved, and emotion-related task.

To overcome the immense cost of manual annotation, we propose a semi-automated pipeline to generate high-quality question-answer (QA) pairs. This three-stage process first generates detailed multimodal captions, then synthesizes questions that strictly require joint audio-visual reasoning, and finally performs a rigorous quality assurance step to ensure data fidelity. We then use human labor to filter out unqualified data. This approach enables us to construct a high-quality multi-choice benchmark of 2,853 samples that are designed to probe complex reasoning abilities.

We conduct extensive experiments on JointAVBenchto evaluate three types of MLLMs: Omni-LLMs, Video-LLMs, and Audio-LLMs. Our results demonstrate that current Omni-LLMs, such as Qwen-omni and Gemini-2.5 flash, significantly outperform their single-modal counterparts. However, our analysis also reveals that these models exhibit uneven capabilities across different audio types and suffer from a substantial performance degradation with increasing scene complexity. Our comprehensive assessment highlights critical limitations in current models' audio-visual reasoning capacities, posing the potential for future improvement.

In summary, our contributions can be summarized as follows:

- We introduce JointAVBench, the first-ever comprehensive benchmark to evaluate joint audio-visual reasoning capability across five cognitive dimensions, four audio types, and three scene complexities.
- We propose a novel three-stage semi-automated pipeline for generating high-quality QA pairs with strict audio-visual correlations while reducing annotation difficulties and costs.
- We provide a comprehensive evaluation of current MLLMs on JointAVBench, demonstrating their limitations and highlighting the importance of developing truly integrated audio-visual reasoning Omni-LLMs.

Table 1: Comparison between our benchmark and previous ones. **Anno.:** the construction method, where A for automatic pipeline, A+M for pipeline involving manual inspection, and M for manual pipeline. **Modality:** the modality involved. V for video, I for image, and A for audio. **Aud. Type:** number of different audio signal types included in the dataset or benchmark. **AV Corr. Ratio:** the ratio of true audio-visual correlated questions, discussed in detail in Appendix C.1. **Div. Scenes:** whether the benchmark includes the evaluation across diverse scenes.

Benchmark/Dataset	**Avg. Duration**	**#QA**	**Anno. Method**	**Modality**	**#Audio Types**	**AV Corr. Ratio**	**Diverse Scenes**
Video Benchmarks/Datasets							
EgoSchema (Mangalam et al., 2023)	180s	5,063	A+M	V	-	0	✗
Video-MME (Fu et al., 2025a)	1,017.9s	2,700	M	V	-	0	✗
MVBench (Li et al., 2024b)	16.0s	4,000	A	V	-	0	✗
LVBench (Wang et al., 2024b)	4,101s	1,549	M	V	-	0	✗
MMBench-Video (Fang et al., 2024)	165.4s	1,998	M	V	-	0	✓
Audiovisual Benchmarks/Datasets							
Music-AVQA (Li et al., 2022a)	60s	45,867	M	V&A	1	56.7%	✗
ACVUBench (Yang et al., 2025)	67.8s	13,774	A+M	V&A	2	-	✗
OmniBench (Li et al., 2024c)	-	1,142	M	I&A	3	100%	✗
AV-Odyssey (Gong et al., 2024)	-	4,555	M	V/I&A	3	100%	✗
LongVALE (Geng et al., 2024)	235s	-	A+M	V&A	3	76.2%	✓
WorldSense (Hong et al., 2025)	141.1s	3,172	M	V&A	3	62.9%	✗
JointAVBench (ours)	97.2s	2,853	A+M	V&A	4	100%	✓

2 Related Works

2.1 Multimodal Large Language Models

The rise of Large Language Models (LLMs) has spurred interest in extending their capabilities beyond text to multimodal inputs (Bi et al., 2024; Achiam et al., 2023; Radford et al., 2018). Early efforts, such as (Radford et al., 2021; Hurst et al., 2024; Li et al., 2022b; 2023b), demonstrate effective fusion of visual and textual modalities for cross-modal understanding. Subsequent studies (Chu et al., 2023; Radford et al., 2023) expand this paradigm to incorporate audio-text integration and achieve significant improvements. Later advances in hardware and memory optimization enable video-text modeling in MLLMs (Team et al., 2023; Gao et al., 2017; Geng et al., 2022; Huang et al., 2024), spanning across various domains and achieving progress such as long video understanding (Wang et al., 2024a; Yuan et al., 2025; Chen et al., 2024a) and movie understanding (He et al., 2024; Song et al., 2024;?). Recent works (Chowdhury et al., 2025; Shu et al., 2023; Cheng et al., 2024; Tang et al., 2025; Fu et al., 2024; 2025b; Lu et al., 2022; 2024; Su et al., 2023; Wu et al., 2024a; Han et al., 2024; **?**) focus on achieving human-like audio-visual joint reasoning ability by interleaving inputting audio, video, and text. This requires datasets to contain QAs with strict audio-visual correlations. To facilitate the development of Omni-LLMs, we present JointAVBench to evaluate the models' audiovisual joint reasoning ability with questions fully audio-visual correlated.

2.2 Audio-Visual Benchmarks

With the development of MLLMs, various benchmarks have been constructed for the evaluation of MLLMs' comprehensive abilities (Wu et al., 2024b; Li et al., 2024a; 2023a; Yue et al., 2024; Zhang et al., 2024a; Sakshi et al., 2024; Liu et al., 2024; Li et al., 2024b). Early datasets or benchmarks, such as AVQA (Yang et al., 2022), Music-AVQA (Li et al., 2022a), and AVInstruct (Ye et al., 2024) only focus on certain types of audio signals and lack strict audio-visual correlation. Subsequent works such as Omni-bench (Li et al., 2024c) and AV-Odyssey (Gong et al., 2024) consist primarily of only image and audio, lacking the evaluation of videos. The recent WorldSense (Hong et al., 2025) has dived into the problem. However, it lacks strict audio-visual correlation and emphasizes the evaluation of visual tasks. These datasets cannot capture the complex and interleaved auditory and scene details in video (such as the details in Figure 1). In contrast, our proposed JointAVBench focuses on the evaluation of diverse audio signal types and multilevel scenes, aiming to conduct a comprehensive and systematic assessment of current MLLMs for joint audio-visual understanding.

Table 2: Task categories of our designed taxonomy. In audio signal type, we use SPE for speech, VOT for vocal traits, SEV for sound event, and MUS for music.

Scene Type	Cognitive Dimension	Audio Signal Type	Task Name	Task Code
Single	Temporal	SPE	Speech-based Timepoint Localization	STL
		SPE	Vision-Speech Sequence Recognition	VSSR
	Spatial	VOT	Speaker Spatial Localization	SPL
		SEV	Sounding Object Grounding	SOOG
		SEV	Sound Event Recognition	SOER
	Emotion	VOT	Speaker Emotion Recognition	SPER
		MUS	Musical Tone Inference	MPTI
Multiple	Long-form	SPE	Cross-scene Association	CSA
	Plot	SPE, VOT	Multi-plot Ordering	MPO
		SPE	Plot Development Prediction	PDP
		SEV, MUS	Audio Function Analysis	AFA
	Temporal	SPE	Plot Temporal Grounding	PTG
Full	Long-form	SPE, VOT, SEV, MUS	Audio-Visual Detail Memory	AVDM
	Emotion	MUS	Musical Emotion Shift Inference	MESI
	Plot	SPE	Character Relationship Inference	CRI

3 JOINTAVBENCH

We propose JointAVBench, a benchmark for evaluating Omni-LLMs' joint audio-visual reasoning ability. This section will first detail the benchmark's core requirements, then the carefully designed data generation pipeline, including (i) omni-caption generation, (ii) QA pair creation, and (iii) rigorous quality control, with statistics provided at the end of this section.

3.1 BENCHMARK REQUIREMENTS

The benchmark construction adheres to three fundamental requirements, ensuring comprehensive evaluation of joint audio-visual reasoning capabilities.

Strict Audio-Visual Correlation. We design a hierarchical taxonomy comprising 15 tasks (detailed in Table 2), ensuring that each task requires the integration of both visual and audio information to generate answers.

High-quality Video Source. Movie scenes are a natural source of extensive and diverse multimodal data. For our benchmark, we leverage the Short-Films 20K (SF20K) (Ghermi et al., 2024) dataset, which comprises 1,072 professionally produced movies rich in narrative and balanced audiovisual features. We then remove unavailable or grayscale videos, retaining 1,046 films that can be used to construct our benchmark.

Multi-dimensional Task Taxonomy. To ensure a comprehensive and fine-grained evaluation, we categorize our tasks along three key dimensions:

- **Cognitive Dimension:** Derived from a systematic analysis of previous studies (Fu et al., 2025a; Hong et al., 2025), this dimension assesses core cognitive abilities essential for video understanding. We define 5 types of cognitive dimensions: temporal, spatial, emotional, plot, and long-form.
- **Audio Types:** This dimension enables a comprehensive evaluation of audio understanding capabilities across all audio signal types. We divide audio into four types of signals: speech, vocal traits, sound event, and music.
- **Scene Complexity:** This dimension evaluates model performance across videos with varying temporal characteristics, using different scene types to quantify temporal information. We define three types of scene complexity: single-scene, multi-scene, and full-scene.

3.2 BENCHMARK CONSTRUCTION

Our dataset construction pipeline is illustrated in Figure 2, which adopts a semi-automated process capable of handling diverse modality characteristics. More details can be found in the appendix.

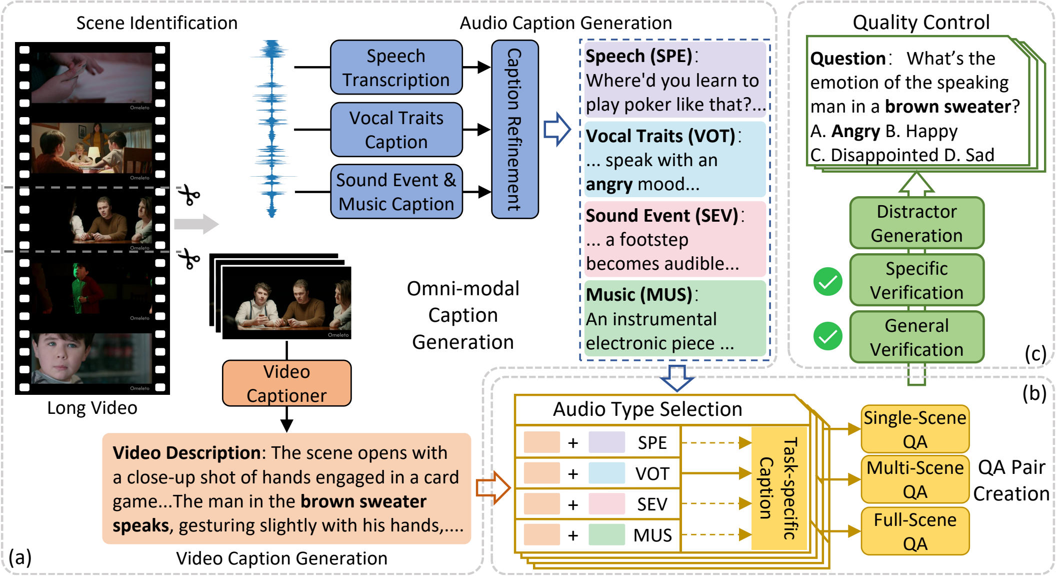

Scene Identification
Audio Caption Generation
Quality Control
Speech Transcription
Vocal Traits Caption
Sound Event & Music Caption
Caption Refinement
Speech (SPE): Where'd you learn to play poker like that?...
Vocal Traits (VOT): ... speak with an angry mood...
Sound Event (SEV): ... a footstep becomes audible...
Music (MUS): An instrumental electronic piece ...
Question: What's the emotion of the speaking man in a brown sweater?
A. Angry B. Happy
C. Disappointed D. Sad
Distractor Generation
Specific Verification
General Verification
(c)
Long Video
Omni-modal Caption Generation
Video Captioner
Video Description: The scene opens with a close-up shot of hands engaged in a card game...The man in the brown sweater speaks, gesturing slightly with his hands,....
(a)
Video Caption Generation
Audio Type Selection
+ SPE
+ VOT
+ SEV
+ MUS
Task-specific Caption
Single-Scene QA
Multi-Scene QA
Full-Scene QA
QA Pair Creation
(b)

Figure 2: **Pipeline for JointAVBench.** Our construction pipeline is three-fold: (a) Omni-modal caption generation, (b) QA pair creation, and (c) Quality control.

3.2.1 Stage 1: Omni-Modal Caption Generation

Scene Identification. We first split the video into scenes with semantic consistency. Specifically, we follow the procedure in Panda-70m (Chen et al., 2024b) to divide long videos into distinct scenes with PySceneDetect[1] and then merge scenes with high semantic similarity, ensuring in-scene consistency. These segmented scenes retain considerable length, enabling us to capture richer contextual information within each scene.

Video Caption Generation. After scene identification, we directly generate visual descriptions for all segmented scenes, ensuring that static features (*e.g.* in-scene objects and characters) and dynamic features (*e.g.* transitions between shots and movements of characters) are well captured.

Audio Caption Generation. To ensure the diversity of audio types, we follow the requirements to generate captions for each audio type as shown in Figure 2. Notably, we observe that existing audio models have limitations in distinguishing between sound event and music, and therefore generate their captions simultaneously. Subsequently, we refine the audio captions by addressing the hallucination in the caption and separating the sound event caption and music caption using different LLM judges.

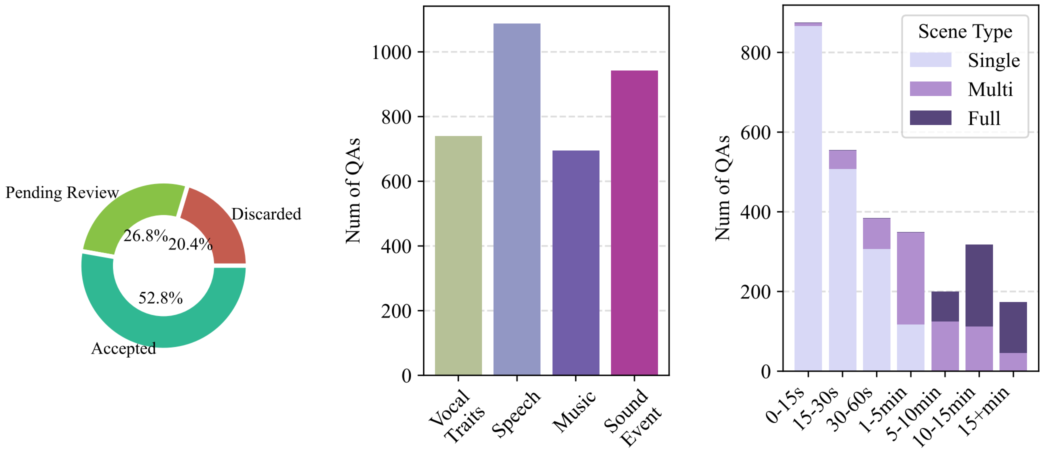
3.2.2 Stage 2: QA Pair Creation

To create QA pairs with strict audiovisual correlation, we design various question templates for tasks that LLM cannot easily understand (temporal, plot tasks that require complex audio-visual relation), while leaving other tasks to be curated by LLM to ensure question diversity (general tasks such as Character Relationship Inference). Additionally, when providing cross-modal descriptions, we strictly adhere to each task's modality and scene requirements by inputting only the required modality descriptions from the designated scenes. For example, when generating data for the task Speaker Spatial Localization, we provide video captions along with vocal traits descriptions from only one scene. This procedure can eliminate possible interference from extraneous modalities and scenes.

3.2.3 Stage 3: Quality Control

We implement a multi-stage quality control process to address issues identified in the collected 9,109 QA pairs, such as mismatched question-answer pairs and redundant information. This process

[1] https://www.scenedetect.com/



(a) Distribution of MCQ quality (b) Distribution of audio types (c) Distribution of scene durations

Figure 3: Statistics of JointAVBench.

employs a general-to-specific verification strategy, where we guide models to use a chain-of-thought approach for step-by-step data filtering.

General Verification. We validate all QA pairs to ensure they meet fundamental standards. This includes a **Modality Check** to confirm that each QA pair necessitates both audio and video information, and a **Logic Check** to verify that answers are directly derivable from the question's context. For instance, a question like *"What is the emotion of the adult male speaker?"* will be discarded if the audio contains only one male speaker, as the answer can be inferred from a single modality.

Specific Verification. This stage focuses on task-specific validations. We design the following three specific checks based on QA's task: 1) **Sequence Check** to ensure the correct element order for sequence-based tasks; 2) **Ambiguity Check** to filter out overly generic QA pairs for complex reasoning tasks (e.g., *"What makes the door closing sound?"* with *"door closing"* as the answer); 3) **Audio Signal Type Check** to confirm that the required auditory information cannot be deduced from visual information for sound event and music.

Distractor Generation. For each verified QA pair, we craft three plausible but incorrect distractors to create challenging multiple-choice questions. These distractors incorporate diverse misdirections, such as replacing the sound source or confusing details.

3.3 Human Verification

From the automated three-stage generation process, we obtain 3,974 MCQs. To ensure their quality and factual accuracy, we conducted a rigorous human verification process, and the results are illustrated in 3a. Specifically, a team of human annotators rated QAs based on four key criteria: (i) answer correctness, (ii) information correctness, (iii) audio-visual dependency, and (iv) question difficulty. Based on these ratings, we categorize these data into three subsets: (1) Accepted: QAs that pass the answer correctness check and score highly on all other criteria, which are directly retained in the final dataset; (2) Pending Review: QAs that pass the answer correctness check but receive lower ratings on one or more additional criteria, which are subject to further selection according to their ratings; and (3) Discarded: QAs that fail the answer correctness check and are removed from the dataset. In total, we retained 2,853 QAs, achieving a data retention rate of 71.8%, which demonstrates that our automatic pipeline is highly effective at generating data of sufficient quality.

3.4 Benchmark Statistics

JointAVBench consists of 2,853 high-quality, manually verified MCQs spanning all scene levels and audio types, with an average duration of 97.2s (Table 1). A detailed statistical analysis of the benchmark is presented in Figure 3. The number of QA pairs is balanced across diverse audio signal

Table 3: Evaluation results of three types of mainstream MLLMs. We evaluate the performance of Omni-LLMs, Video-LLMs and Audio-LLMs on JointAVBench to provide a comprehensive analysis.

Models	STL	SPL	SOOG	SOER	SPER	MPTI	VSSR	CSA	MPO	PTG	AFA	PDP	AVDM	MESI	CRI	**Avg**
								Omni-LLMs								
Qwen2.5-Omni	**71.3**	35.3	**59.8**	**72.3**	30.6	63.4	**77.6**	51.2	40.4	20.8	**69.9**	47.3	47.3	69.9	70.3	**56.2**
Gemini2.5-Flash	62.6	43.8	55.7	68.6	23.3	59.1	42.6	39.2	**46.2**	25.3	65.3	**48.4**	**64.8**	67.7	**81.3**	52.6
VideoLLaMA2	20.9	38.8	56.1	67.0	29.6	47.5	48.5	24.0	35.3	30.9	63.6	36.6	38.0	61.7	58.8	46.6
OneLLM	33.0	**44.2**	45.6	37.8	29.9	29.7	33.9	**55.4**	31.7	32.9	46.6	44.1	34.5	34.6	50.3	38.5
video-SALMONN-o1	32.2	30.0	35.1	43.5	14.0	44.7	32.0	25.6	20.0	**36.2**	55.4	30.1	35.5	66.9	58.8	37.3
video-SALMONN	52.2	25.1	37.8	52.2	19.0	33.3	33.5	30.5	31.7	26.1	48.9	26.9	24.2	37.9	50.3	35.8
AVicuna	31.9	29.3	35.1	38.8	16.6	31.2	21.4	25.0	21.2	30.9	43.7	30.1	27.6	29.5	44.3	30.6
								Video-LLMs								
InternVL-2.5	28.7	37.9	59.8	71.1	23.6	64.1	52.2	42.5	44.2	27.5	63.6	41.9	50.0	68.4	68.3	51.3
VideoLLaMA3	43.5	41.1	58.8	55.8	17.9	**69.2**	50.0	34.7	43.3	33.6	61.9	40.9	51.8	73.7	64.8	49.9
Qwen2.5-VL	33.9	38.8	55.3	59.3	22.9	57.2	47.2	31.7	40.4	32.2	62.5	39.8	40.7	62.9	61.6	47.1
LLaVA-Video	37.4	33.0	48.6	64.7	10.0	68.5	53.9	27.3	43.3	30.9	51.1	36.6	46.4	**76.7**	61.8	47.0
GPT-4o	30.4	34.8	55.7	69.7	11.6	53.6	24.8	40.5	13.5	14.1	51.7	47.3	50.9	56.4	70.9	43.3
								Audio-LLMs								
Kimi-Audio	56.5	21.9	48.6	61.7	32.9	53.3	34.3	38.0	33.0	26.2	65.3	38.7	40.2	56.1	69.5	45.9
Qwen2-Audio	54.1	24.3	39.5	54.3	**34.6**	40.0	34.3	33.0	32.7	27.6	55.0	29.8	32.9	46.6	58.1	40.0

types (Figure 3b), showcasing our benchmark's comprehensive coverage. Moreover, our dataset spans a wide range of video durations (Figure 3c), with single-scene, multi-scene, and full-scene tasks mainly comprising videos of less than 1 min, 1-10 min, and over 10 min, respectively.

4 Experiment

This section first demonstrates a comprehensive evaluation of mainstream MLLMs on our proposed benchmark, and then key factors that influence performance to provide valuable insights for future Omni-LLMs.

4.1 Experiment Setup

Models. Our experiments are conducted on a diverse set of mainstream MLLMs. To comprehensively evaluate their joint audio-visual reasoning capability across different modalities, we categorize them into three groups: (i) Omni-modal LLMs: Qwen2.5-Omni (Xu et al., 2025), VideoLLaMA2 (Cheng et al., 2024), video-SALMONN-o1 (Sun et al., 2025), video-SALMONN (Sun et al., 2024), OneLLM (Han et al., 2024), AVicuna (Tang et al., 2025), and Gemini2.5-Flash (Team et al., 2023); (ii) Video-LLMs: Qwen2.5-VL (Bai et al., 2025), LLaVA-Video (Zhang et al., 2024b), Video-LLaMA3 (Zhang et al., 2025), InternVL2.5 (Chen et al., 2024c), and GPT-4o (Hurst et al., 2024); and (iii) Audio-LLMs: Kimi-Audio (Ding et al., 2025) and Qwen2-Audio (Chu et al., 2024).

Metrics and Experiment Settings. To achieve evaluation consistency, we follow previous works (Hong et al., 2025; Fu et al., 2025a) and use accuracy as the evaluation metric. For a fair evaluation, we adopt the following protocols for all experiments. For open-source models, we use their official codebase with default configurations, while for closed-source models, we leverage their official APIs. To maintain comparability, we select open-source models with comparable parameter sizes and enforce a unified sampling of 32 frames across all models. We also ensure that the text input to all models is limited to the question text, without any additional contextual information.

4.2 Results and Findings

Overall Performance. Our results, summarized in Table 3, reveal that current mainstream MLLMs perform sub-optimally on our benchmark, with the best performing model having an average accuracy of only 56.2%. This suggests a significant gap in their ability to process omni-modal information. Importantly, Omni-LLMs consistently outperform Video-LLMs and Audio-LLMs, highlighting the critical role of native modality integration. For instance, Qwen2.5-Omni significantly improves upon Intern-VL across most tasks.

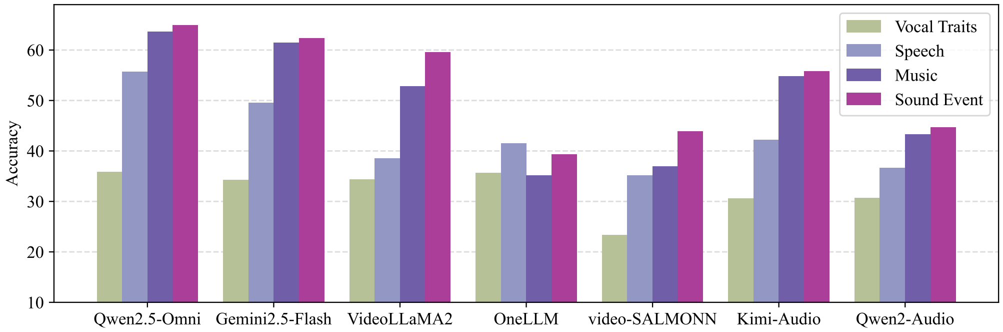


Figure 4: Results on JointAVBench across different audio types.

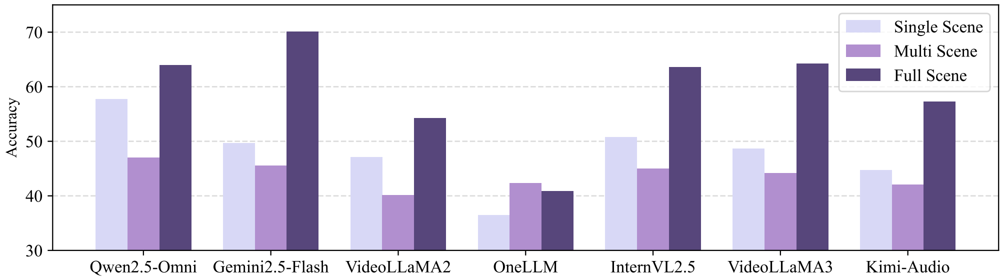


Figure 5: Results on JointAVBench across different scene types.

Breakdown Findings. To gain a deeper understanding, we analyze model performance across various task categories and have the following observations.

1) Models perform unevenly across different audio types, failing on tasks requiring vocal traits and speech. We find a significant performance gap among different audio types (Figure 4). Models excel at tasks involving sound events and music, likely due to their stronger visual correspondence (where audio often matches visible objects or atmosphere). However, they struggle with more abstract audio, such as speech and vocal traits. This is likely due to a lack of training data focused on vocal traits, as most audio-visual datasets (Li et al., 2022a; Yang et al., 2022) overlook information like emotion and gender, leading to tasks like SPL, SPER, and MPO being the worst-performing overall.

2) Multi-scene tasks usually yield worse results compared to single-scene tasks, while full-scene tasks often achieve better results. Figure 5 shows the pronounced impact of scene complexity. Models perform well on single-scene tasks, particularly those requiring speech-based reasoning like STL and VSSR. This is likely because single scenes offer stable visual contexts and limited speech content, simplifying cross-modal correspondence. Conversely, multi-scene tasks requiring speech, such as MPO and PTG, yield worse performance, as they demand more complex processing of diverse scenes and cross-scene connections. Interestingly, while most models struggle with multi-scene tasks, they perform better on full-scene tasks, which focus on global narratives rather than fine-grained details. This highlights that improving models' cross-scene reasoning capabilities is needed.

3) Omni-models perform worse on emotional and spatial tasks than single-modal models. As demonstrated in Figure 6, while Omni-LLMs generally perform best in 11 of our 15 tasks, they surprisingly fall behind single-modality models on emotion-based tasks. This suggests that in some cases, single-modality models can better focus on emotion cues without the distraction of additional modalities. Furthermore, Omni-LLMs perform poorly on spatial tasks like SOOG and SOER, even falling behind Video-LLMs. This is likely because models primarily rely on spatial information from video and fail to effectively integrate complementary audio cues. This finding highlights a critical limitation in current models' ability to perform true audio-visual spatial reasoning.

4) Increased scene number leads to models' performance degradation on multi-scene tasks. Our analysis reveals that increasing the number of scenes adversely affects multi-scene task performance

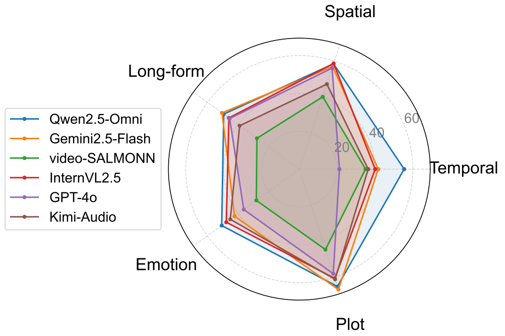


Figure 6: Results on JointAVBench across 5 cognitive dimensions.

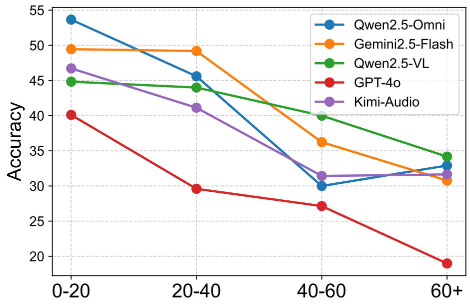


Figure 7: Results on JointAVBench with different number of scenes.

Table 4: Evaluation results of open-source omni-LLMs with different modality utilization. † For neatness, we use short names of models, where Qwen2.5 represents Qwen2.5-Omni, ViLLaMA2 represents VideoLLaMA2, and vid-SAL represents video-SALMONN

Model[†]	Modality	N_o	N_u	**Avg**	STL	SPL	SOOG	SOER	SPER	MPTI	VSSR	CSA	MPO	PTG	AFA	PDP	AVDM	MESI	CRI
Qwen2.5	A+V	8	1	56.2	71.3	35.3	59.8	72.3	30.6	63.4	77.6	51.2	40.4	20.8	69.9	47.3	47.3	69.9	70.3
	V			49.3	38.3	33.0	57.8	64.6	18.6	60.1	50.4	43.8	40.4	24.2	68.8	37.6	48.2	70.7	68.5
	A			52.3	71.3	23.9	64.4	66.5	34.2	56.2	57.0	41.3	45.2	18.1	69.3	41.9	49.1	64.7	68.3
VidLLaMA2	A+V	6	3	46.6	20.9	38.8	56.1	67.0	29.6	47.5	48.5	24.0	35.3	30.9	63.6	36.6	38.0	61.7	58.8
	V			46.6	27.8	37.9	56.8	67.4	27.9	49.3	43.9	28.9	46.0	26.2	62.5	38.7	38.0	62.4	52.7
	A			41.4	24.3	35.7	56.1	61.4	18.3	37.3	32.6	35.5	35.6	30.2	60.2	29.0	31.8	50.4	56.4
OneLLM	A+V	8	3	38.5	33.0	44.2	45.6	37.8	29.9	29.7	33.9	55.4	31.7	32.9	46.6	44.1	34.5	34.6	50.3
	V			32.7	27.8	35.7	37.8	28.6	15.3	28.6	31.7	31.4	33.7	32.2	38.6	50.5	34.5	35.3	53.3
	A			38.5	28.7	43.8	43.6	38.9	31.6	28.6	27.4	47.9	23.1	30.2	41.5	62.4	45.5	39.8	60.0
vid-SAL	A+V	5	4	35.8	52.2	25.1	37.8	52.2	19.0	33.3	33.5	30.5	31.7	26.1	48.9	26.9	24.2	37.9	50.3
	V			34.8	23.5	24.3	44.4	49.0	16.8	35.5	29.1	28.3	36.5	26.2	47.2	40.9	31.3	38.3	44.1
	A			35.7	53.9	22.6	39.0	50.6	21.4	34.9	32.6	36.4	28.7	25.7	44.3	31.2	23.8	39.7	44.2

(Figure 7), with accuracy dropping sharply by approximately 20% from 0-20 to over 60 scenes. This underscores the uneven performance of current MLLMs across diverse scenes, highlighting a critical area for improvement.

5) Omni-modal Models Demonstrate Effective Modality Fusion. We quantify the effectiveness of joint reasoning by defining N_o and N_u as the number of tasks where a model's audio-visual (A+V) performance is better and worse than its single-modality baseline, respectively, in Table 4. For all models, N_o significantly outweighs N_u, confirming that integrating audio and video fundamentally enhances overall reasoning capability. Furthermore, as a model's overall performance increases, its N_o count rises while its N_u count falls. This pattern indicates that more advanced models are better adept at modality fusion. For instance, while early Omni-LLMs (Cheng et al., 2024; Han et al., 2024; Sun et al., 2024) show only marginal gains over single-modality baselines, Qwen2.5-Omni's dual-modality performance significantly surpasses its single-modality results. This confirms that true joint reasoning is a hallmark of mature omni-modal architectures.

5 Conclusion

In this paper, we propose JointAVBench, a comprehensive benchmark for evaluating joint audio-visual reasoning, distinguished by a hierarchical taxonomy and a high-quality, automated generation pipeline. Each question in JointAVBenchis meticulously designed to necessitate the integrated understanding of both visual and a specific type of audio input. We further ensure benchmark quality through human verification. Our extensive experiments reveal that even the best-performing models achieve an accuracy of only 56.2%, underscoring the substantial need for more powerful omni-modal models with enhanced audio-visual fusion capabilities.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024a.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

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

Sanjoy Chowdhury, Hanan Gani, Nishit Anand, Sayan Nag, Ruohan Gao, Mohamed Elhoseiny, Salman Khan, and Dinesh Manocha. Aurelia: Test-time reasoning distillation in audio-visual llms. *arXiv preprint arXiv:2503.23219*, 2025.

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

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

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025a.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025b.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pp. 5267–5275, 2017.

Tiantian Geng, Feng Zheng, Xiaorong Hou, Ke Lu, Guo-Jun Qi, and Ling Shao. Spatial-temporal pyramid graph reasoning for action recognition. *IEEE Transactions on Image Processing*, 31: 5484–5497, 2022.

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. *arXiv preprint arXiv:2411.19772*, 2024.

Ridouane Ghermi, Xi Wang, Vicky Kalogeiton, and Ivan Laptev. Short film dataset (sfd): A benchmark for story-level video understanding. *arXiv preprint arXiv:2406.10221*, 2024.

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

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26584–26595, 2024.

Yichen He, Yuan Lin, Jianchao Wu, Hanchong Zhang, Yuchen Zhang, and Ruicheng Le. Storyteller: Improving long video description through global audio-visual character identification. *arXiv preprint arXiv:2411.07076*, 2024.

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

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14271–14280, 2024.

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

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

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

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19108–19118, 2022a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022b.

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 *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b.

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

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

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

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

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

Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding. *arXiv preprint arXiv:2312.06720*, 2023.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 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. *arXiv preprint arXiv:2406.15704*, 2024.

Guangzhi Sun, Yudong Yang, Jimin Zhuang, Changli Tang, Yixuan Li, Wei Li, Zejun Ma, and Chao Zhang. video-salmonn-o1: Reasoning-enhanced audio-visual large language model. *arXiv preprint arXiv:2502.11775*, 2025.

Yunlong Tang, Daiki Shimada, Jing Bi, Mingqian Feng, Hang Hua, and Chenliang Xu. Empowering llms with pseudo-untrimmed videos for audio-visual temporal understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7293–7301, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*, 2024a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024b.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. In *Forty-first International Conference on Machine Learning*, 2024a.

Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, et al. Scimmir: Benchmarking scientific multi-modal information retrieval. *arXiv preprint arXiv:2401.13478*, 2024b.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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*, pp. 3480–3491, 2022.

Yudong Yang, Jimin Zhuang, Guangzhi Sun, Changli Tang, Yixuan Li, Peihan Li, Yifan Jiang, Wei Li, Zejun Ma, and Chao Zhang. Acvubench: Audio-centric video understanding benchmark. *arXiv preprint arXiv:2503.19951*, 2025.

Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*, pp. 146–164. Springer, 2024.

Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *arXiv preprint arXiv:2501.07888*, 2025.

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

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

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

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

A More Details on JointAVBench

A.1 Task Definition

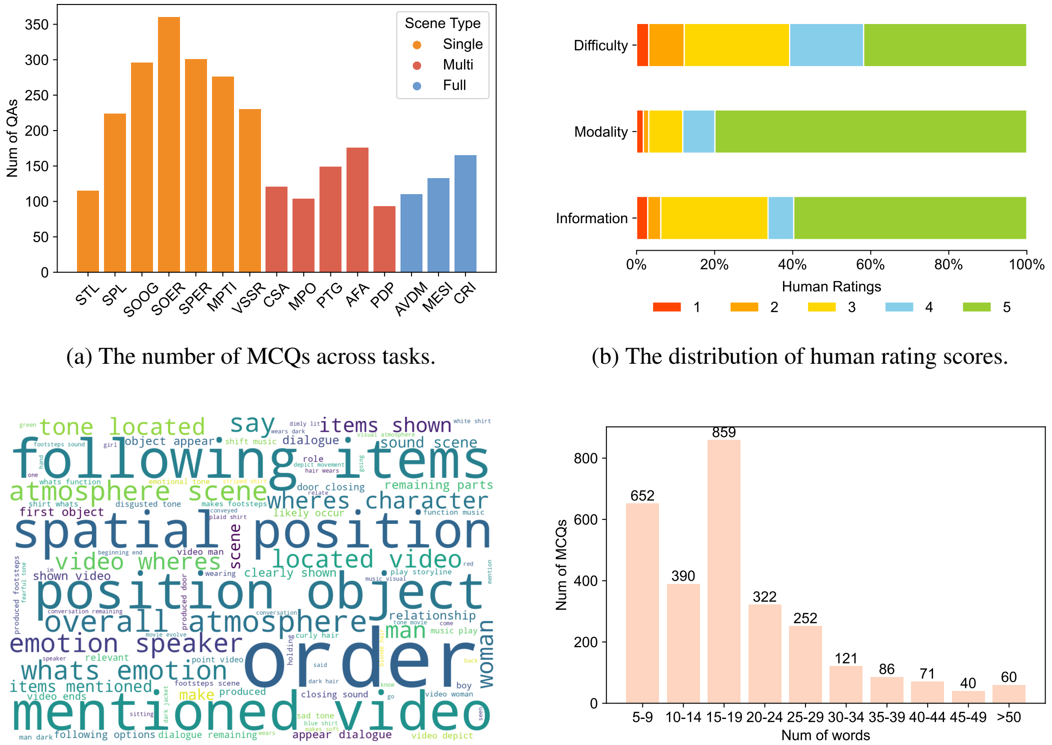
We construct a taxonomy of 15 tasks requiring audio-visual joint reasoning ability based on the benchmark's requirements. The detailed task descriptions with examples are presented in Table 5. Since single-scene tasks and multi-scene tasks require focus on movie details (e.g., visual cues, temporal relationships), question templates are designed to ensure the generated questions both capture critical information and comply with task specifications.

Table 5: Problem Categories and Definitions of JointAVBench's Taxonomy

Task Name	Code	Category Description	Question Example
Speech-based Timepoint Localization	STL	Locate the temporal position of an object in the dialogue	Which objects are mentioned only in the dialogue but not clearly shown in the video, and when does the first object appear in the dialogue?
Speaker Spatial Localization	SPL	Locate the spatial position of a character in the video	Where's the character that says Ï'm gonna give you a compliment nowẅith a contemptuous tone located in the video?
Sounding Object Grounding	SOOG	Locate the spatial position of a sound-emitting object in the movie	What is the spatial position of the object that produced the loud bang in the scene?
Sound Event Recognition	SOER	Infer what action occurred that caused the sound	What makes the high-pitched, bright sound?
Speaker Emotion Recognition	SPER	Identify the emotions of characters	What's the emotion of the speaker that wears a brown jacket over a blue shirt?
Musical Tone Inference	MPTI	Determine the overall tone of a scene	What is the overall atmosphere of the scene?
Vision-Speech Sequence Recognition	VSSR	Determine the chronological order of element appearances	In what order were the following items mentioned in the video? (a) 'Did you see this guy smoking pot?' (b) The police officer holding an object. (c) 'Excuse me, sir.'
Cross-scene Association	CSA	Identify associations between elements across different scenes	Which dialogue in the remaining parts is most relevant to what the man does in 18.56s-35.16s?
Multi-plot Ordering	MPO	Order different audio-visual details across different movie segments	In what order were the following items mentioned in the video? (a) The girl is seen adjusting an oxygen mask on a child (b) The woman is seen trimming flower stems. (c) The man says, "This isn't funny anymore."
Plot Temporal Grounding	PTG	Identify the approximate temporal position of plot segments	When did the woman in the green tulle outfit reveal her success in the music industry?
Audio Function Analysis	AFA	Analyze the purpose of sound effects and background music in movie segments	How does the video depict the movement of the man in the beige suit?
Plot Development Prediction	PDP	Predict future plot developments based on existing plot elements	Which of the following options is most likely to occur after this video ends?
Audio-Visual Detail Memory	AVDM	Test the model's long-term memory capability	What was the man doing when the police officer asked if they were smoking?
Musical Emotion Shift Inference	MESI	Identify emotional changes and trends throughout the movie	How does the emotional tone evolve from the beginning to the middle of the movie?
Character Relationship Inference	CRI	Infer complex relationships between characters based on the overall plot	What is the relationship between the man in the brown jacket and the police officer?

A.2 More statistics

To further analyze the descriptive characteristics of audiovisual events in our dataset, we performed lexical frequency analysis on both captions and QA pairs, visualized through the word cloud in Figure 8c. The results highlight frequent visual descriptors (e.g., "position", "object", "location") and auditory terms (e.g., "sound", "speaker", "speech"). Furthermore, we examined the distribution of question lengths (in words), as illustrated in Figure 8d, which confirms the conciseness of our formulated questions with minimal redundancy. Figure 8a presents the quantitative distribution of



(a) The number of MCQs across tasks.

(b) The distribution of human rating scores.

(c) The word cloud presentation of our benchmark.

(d) The distribution of word count.

Figure 8: More details of JointAVBench. The number 1-5 in Figure (b) are human ratings, 1 represents the lowest score and 5 represents the highest score.

QA pairs across different tasks, revealing that single-scene tasks exhibit particularly rich audiovisual correlations and simplicity in question designs.

A.3 Manual Check

To ensure the quality of JointAVBench, we engage a group of annotators to evaluate and correct the questions according to the criteria specified in Section 3.3. The annotation process consists of two main steps: First, annotators rate the generated MCQs; then, for incorrect MCQs where the correct answer appears in the distractors, they replace the designated correct answer with the appropriate distractor. Through this process, we maintain high-quality MCQs while discarding only those that fail to meet our quality standards. After human verification, we discard 17.4% of the data for incorrect answer. Other data are filtered based on the other 3 judging criteria, resulting in 2,853 MCQs.

Figure 8b presents the human evaluation scores of the automatically generated MCQs. The results demonstrate that our pipeline generates high-quality questions, with the majority of MCQs receiving the highest rating across all evaluation criteria. Specifically, we find that our QAs are relatively difficult, with the difficulty criteria having high ratings.

A.4 Experiment Details

To ensure comparability between MLLMs, we selected similar model parameters and video frames. For open-source models, we selected 7B as the parameter size (we selected InternVL2.5-8B (Chen et al., 2024c) since it does not provide a 7B model). We also selected 32 frames as the maximum number of frames per video to ensure that the frame number did not affect performance. For closed-source models, we adhered to the official model settings. Notably, for GPT-4o (Hurst et al., 2024), we only input video frames alongside the question text, and therefore the model is unable to access timestamps. During the experiment, we randomly shuffled the options to ensure that the distribution of correct answer prefixes was uniform and free from biases.

All experiments were conducted on NVIDIA H-100 GPUs and can be reproduced using a single H-100 (80G) GPU. During the automatic generation process, we used the official API (model name: 'qwen2.5-72b-instruct') from Qwen to ensure long-context capability and generation stability.

A.5 Cases

Figure 9 presents detailed examples from our benchmark. The questions and options in JointAVBench are designed to incorporate both audio and video modalities and to evaluate audio-visual joint reasoning capabilities. We present the task names, a few frames of the video, questions, options, and correct answers in the cases.

B Details on Generation Pipeline

B.1 Video Caption Generation

We generate visual captions using Qwen2.5-VL (Bai et al., 2025) at a rate of 1 frame per second (fps) for each identified scene, regardless of its quality level (low or high). To ensure the captions capture both static and dynamic scene information, we carefully design a video captioning prompt as shown in Figure 10.

B.2 Audio Caption Generation

To ensure the audio captions contain rich information about diverse audio signal types, we generate separate captions for each type. We then utilize an LLM with carefully designed Chain-of-Thought (CoT) prompts to reduce caption hallucination.

Vocal Traits, Sound Event, and Music Description. We find that directly using general-purpose audio-language models (ALMs) to generate overall audio captions often overlooks important details and may produce inaccurate results. Therefore, we employ Qwen2.5-Omni (Xu et al., 2025), an open-source multimodal model, to separately generate descriptions of vocal traits, sound events, and music components. Notably, current ALMs cannot reliably distinguish between sound events and music. We address this limitation by generating both sound event and music descriptions initially, then separating them during post-processing. The detailed prompt templates are provided in Figure 10.

Subtitle Transcription. For dialogue transcription, our primary objectives are accurate speech recognition and precise timestamp generation. Since general ALMs underperform in timestamp estimation, we utilize Whisper-v3 (Radford et al., 2023), an advanced and widely utilized automatic speech recognition (ASR) system, to ensure transcription quality.

Audio Caption Refinement. The initial audio descriptions contain hallucinations (*i.e.*, factually incorrect or repetitive outputs). We employ Qwen-2.5 (Yang et al., 2024) to perform three refinement steps: (1) distinguishing between background music and sound events, (2) aligning vocal characteristics with dialogue transcripts, and (3) removing redundant content. The detailed prompt engineering for this process is illustrated in Figure 11.

B.3 QA Pair Generation

We utilize Qwen2.5 (Yang et al., 2024) to generate QA pairs following predefined question templates. For single-scene and multi-scene tasks, we only provide descriptions for high-quality scenes. For full-scene tasks, we include all scene descriptions regardless of quality to ensure no details are omitted. After generating multi-scene QA pairs, we verify the interval between questions to ensure they require information from multiple scenes, using the prompt shown in Figure 16 to identify their information intervals. Additionally, we require the model to generate a brief justification for each answer while generating QA pairs.

Table 6: AV Correlation Scores for Different Datasets

Dataset	Automatic Score	Human Score
Music-AVQA	56.7%	54.5%
OmniBench	100%	94.0%
AV-Odyssey	100%	99.5%
LongVALE	76.2%	75.0%
WorldSense	62.9%	60.5%
JointAVBench (ours)	100%	94.5%

B.4 Quality Control

In the quality control stage, we employ extensive CoT techniques to ensure that Qwen2.5 (Yang et al., 2024) achieves optimal performance in identifying potential hallucinations.

During the general check, we utilize only the QA pair and its explanation to filter out unqualified QA pairs. This stage includes four checks: modality, format, content, and speculation checks. The details of each check are as follows: (i) modality check assesses whether the modality clues used in the QA pair are derived from dual modalities; (ii) format check evaluates whether the answer corresponds to the question in format (*e.g.*, the answer explains two items, but the question asks about only one item); (iii) content check verifies whether the answer can be logically inferred from the question based on the explanation; (iiii) speculation check examines whether the answer relies excessively on speculation rather than concrete evidence. The prompts used in this stage are shown in Figure 13.

For the specific check, we design task-specific prompts based on the definition of each task. The prompts for each specific check are presented in Figure 14 and Figure 15. Note that for better adaptation to different tasks, the prompt for the sequence check varies slightly between VSSR and MPO, and the ambiguity check varies slightly between SPL and SOER.

B.5 Distractor Generation

Since generating distractors requires additional information from the video, we generated them after filtering the QA pairs. In this process, we designed a generation prompt incorporating various error types to ensure option diversity and complexity, as illustrated in Figure 16. This includes common error categories such as incorrect details and temporal/spatial misplacement.

C More Experiments

C.1 Evaluation of AV Correlation for Previous Works

We evaluate the AV correlation ratio (the proportion of questions that truly require auditory and visual information to answer) for previous works in Table 1. However, since some benchmarks do not evaluate the ratio themselves, we utilize Qwen-2.5 (Yang et al., 2024) to judge the correlation. We utilize our modality check prompt and sample 1,000 QA pairs from each dataset to judge. We also recruit human volunteers to verify the AV correlation ratio by asking them to assess whether the randomly sampled 100 data items from each benchmark require both auditory and visual information to answer. The overall comparison is described in Table 6, indicating that our automatically calculated score has high correspondence to human scores. This score indicates that our benchmark can assess joint audio-visual reasoning ability.

C.2 Error Study

We'd like to provide some failure cases and deep analysis in this section. The failure cases are selected from the worst-performing tasks.

Based on the failure cases, we find that:

- **Models fail to understand vocal traits.** In the first example, the model fails to find the spatial information based on speech and vocal traits. In the second example, the model can't

Table 7: Examples of failure cases.

Task	Question	Groundtruth Answer	Model Answer
SPL	Where's the character that says 'Oh, my God' with a vibrant shocked tone located in the video? A. In the left B. In the center C. Standing behind the loveseat D. On the right	A	D. On the right.
SPER	What's the emotion of the speaker that wears a vibrant green tulle outfit with hair rollers and dramatic makeup? A. Surprised B. Amused C. Confused D. Excited	D	A. Surprised.
MPO	In what order were the following items mentioned in the video? (a) 'When do you need to have the van back'. (b) The man talked about cars with a joyful tone (c) The man is driving a car along a road lined with bare trees A. (c) (b) (a) B. (a) (b) (c) C. (b) (c) (a) D. (c) (a) (b)	D	The correct answer is A.
PTG	When did the woman in the green tulle outfit reveal her success in the music industry? A. 51.05s-80.79s B. 126.38s-174.51s C. 86.46s-126.38s D. 45.84s-51.05s	A	The correct answer is B.

find vocal traits based on visual information. These two examples indicate that the ability to align visual information with vocal traits remains low in current models and needs to be improved.

- **Models fail to understand temporal information.** In the third example, the question tests the model's ability to understand the storyline and arrange the detailed information. And the fourth example tests the model's temporal grounding ability from a long video. These two examples showcase that future works should focus on increasing the model's ability to understand temporal relationships in audio-visual scenarios.

D Limitations

We acknowledge that the JointAVBench has several limitations in its generation and experimental evaluation. First, the dataset is exclusively derived from SF20K, which may introduce biases in data distribution. Second, our designed taxonomy, while comprehensive, may not encompass all dimensions of audio-visual joint reasoning capabilities. Nevertheless, we have rigorously ensured that the included tasks cover critical aspects and effectively assess the target abilities. Third, due to computational constraints, our experiments were limited to selected representative MLLMs rather than an exhaustive evaluation. We intend to address these limitations in future work through dataset expansion and more extensive benchmarking.

E Broader Impacts

We constructed JointAVBench to facilitate research and development in omni-LLMs and video understanding. We anticipate that this dataset may yield both positive and negative societal impacts. JointAVBench offers several potential benefits, including: (1) enabling development of human-like agent systems, (2) advancing video understanding tools, and (3) creating assistive software for people with disabilities. However, the dataset also presents certain risks, such as privacy concerns and copyright issues. We believe a thorough discussion of these benefits and challenges will lead to a more comprehensive understanding of the dataset's societal implications.

F Declaration of LLM Usage

During our research, we use LLMs as a major dataset construction tool, including dataset generation, quality control, and experiments. During our paper writing, we use LLMs to polish our paper and correct the defects in our paper.

G SAFEGUARDS

To ensure our benchmark excludes unsafe content, we adopted two key measures. First, we used the publicly released SF20K dataset (Ghermi et al., 2024) as the foundation, which provides pre-filtered safe content. Second, we employed the official Qwen API during benchmark generation, whose built-in safety mechanisms automatically screen both input prompts and output responses for potentially unsafe video recommendations.

H LICENSE

The JointAVBench dataset is released under the CC BY-NC-SA 4.0 license. Subsequent research using this dataset must comply with the license terms.

Speech-based Timepoint Localization

Question:
Which objects are mentioned only in the dialogue but not clearly shown in the video, and when does the first object appear in the dialogue?
A. The frisbee, mentioned at around 329s
B. The pig in a blanket, mentioned at around 329s
C. The green box, mentioned at around 502s
D. The dead porcupine, mentioned at around 445s
Correct Answer: A

Sounding Object Grounding

Question:
Where's the sounding object located in the video?
A. Near the sink.
B. By the exit door.
C. In the hallway.
D. Near the stalls.
Correct Answer: D

Sound Event Recognition

Question:
What makes the footsteps sound?
A. The man getting up from the table
B. The door closing as the woman enters the cabin
C. The woman approaching the man
D. The woman pacing back and forth in the cabin
Correct Answer: C

Musical Tone Inference

Question:
What is the overall atmosphere of the scene?
A. Busy and chaotic
B. Calm and relaxed
C. Somber and focused
D. Bright and cheerful
Correct Answer: C

Plot Temporal Grounding

Question:
At what point in the video does the boy in the orange sweater explain to the man in the dark turtleneck that the money he brought is his own, not his father's?
A. 668.04s-710.88s
B. 778.08s-816.42s
C. 759.88s-772.00s
D. 744.29s-751.62s
Correct Answer: B

Cross-scene Association

Question:
Which dialogue in the remaining parts is most relevant to what the man does in 18.56s-35.16s?
A. "Excuse me."
B. "I'll take care of them."
C. "This is true."
D. "Stop."
Correct Answer: B

Charater Relationship Inference

Question:
What is the relationship between the man with the blood-stained hands and the woman in the red jacket?
A. The man and the woman are father and daughter.
B. The man and the woman are brother and sister.
C. The man and the woman are uncle and niece.
D. The man and the woman are husband and wife.
Correct Answer: A

Musical Emotion Shift Inference

Question:
What is the emotional shift depicted in the scenes leading up to and including the car ride after the "debs"?
A. From celebratory and joyful to tense and anxious.
B. From calm and serene to lively and energetic.
C. From celebratory and joyful to reflective and tender.
D. From reflective and tender to celebratory and joyful.
Correct Answer: C

Figure 9: Additional cases of JointAVBench. The first and second row represents single-scene tasks, the third row represents multi-scene tasks, and the last fourth row represents full-scene tasks.

Prompt for video captioning

Please generate a clear, concise, and detailed caption for the provided movie video clip. Ensure the description is entirely based on the visual content of the video, without any speculation, assumptions, or uncertain information. Focus on capturing the most essential visual elements while avoiding unnecessary repetition or overly verbose language. Describe the scene as follows:
1. Scene Setting: Briefly describe the environment, location, time of day, lighting, and any notable objects or elements in the background. Only include details that are clearly visible in the video.
2. Characters and Actions: Highlight the appearance, clothing, and key actions of any characters present. Focus on their most significant movements, gestures, and interactions. Do not infer emotions, intentions, or backstory unless explicitly shown through visual cues.
3. Scene Dynamics: Describe any important changes in the scene. Include the sequence of events and the pacing of the scene to convey how it unfolds over time. Only describe what is visually evident.
4. Emotional Tone: Convey the mood or atmosphere through the most impactful visual cues, such as facial expressions, body language, or environmental details. Only describe emotions that are clearly expressed through visible actions or expressions.
5. Key Events: Highlight any significant events or actions that occur within the scene, focusing on their narrative importance or impact on the characters. Only include events that are explicitly shown in the video.
Important Notes:
- Strictly factual: Ensure the description is entirely based on the visual content of the video. Do not include any information that is not explicitly shown or clearly visible.
- Avoid speculation: Do not infer or assume anything about characters' thoughts, feelings, or motivations unless they are directly expressed through visible actions or expressions.
- Acknowledge uncertainty: If certain details are unclear or ambiguous, do not include them in the description. Focus only on what is clearly visible.
- Exclude text: If there are subtitles or text displayed at the bottom of the video, do not include them in the caption. Focus only on the visual elements of the scene.
- Be concise: Avoid unnecessary details or repetition. Focus on the most critical visual elements that define the scene.
Use vivid but economical language to paint a clear and engaging picture of the scene for someone who cannot see the video.

Prompt for sound event & music captioning

Describe the audio by focusing ONLY on the following detectable elements:
1. Sound Events (if clearly present):
- Non-speech sounds: (e.g., crumpling, footsteps, door closing, glass breaking).
- Non-verbal vocalizations: (e.g., laughter, sighing, coughing, crying, humming, screaming).
- Characteristics (only if unambiguous):
- Pitch (high/low), timbre (bright/muffled), rhythm (steady/erratic), volume (loud/soft).
- Rules:
- Do NOT guess sound sources (e.g., no ”paper crumpling” → just ”a crumpling sound”).
- If uncertain, omit the detail entirely.
2. Background Music (if clearly present):
- Instruments (e.g., piano, strings, electric guitar).
- Mood(e.g., cheerful, tense, melancholic).
- Avoid technical terms (BPM, key, scales).
- Rules:
- Do NOT describe lyrics or vocal melodies.
- If the music is ambiguous (e.g., genre unclear), only state observable features.
Critical Constraints:
- Do NOT describe speech, spoken words, conversation content, or verbal interactions.
- Never use speculative language (e.g., ”likely,” ”probably,” ”seems like”).
- If a sound cannot be confidently identified, skip it.
- No timestamps, durations, or speaker demographics (e.g., ”a child laughing” → just ”laughter”).

Prompt for vocal traits captioning

Please analyze the provided speech audio and generate a strictly factual description following these requirements:
1. Output format for each utterance:
Speech Content: [Exact dialogue content]
Emotion: [Observed emotional tone] (eg. happy, sad, angry, fearful, surprised, disgusted, excited)
Speaker traits: [Directly discernible characteristics like age/gender if evident from voice]
2. Rules:
(1) Only describe emotions clearly conveyed through vocal tone
(2) Note speaker characteristics ONLY when immediately apparent from voice (e.g. ”child-like voice”)
(3) Never add interpretations beyond what the audio contains
(4) Process each utterance separately
(5) Non-speech audio (music or sound only): output ”[Non-speech audio: skip analysis]” and stop

Figure 10: Prompts for generating omni-modal caption.

Prompt for cleaning and separating sound event caption and music caption

You are an expert in audio data cleaning. Your task is to clean and organize audio captions by separating them into two distinct categories: [music] and [sound event].
Please analyze the provided audio caption and separate it into two parts:
`[music]`: Include descriptions related to music, such as background music, musical instruments, and any emotional or atmospheric content (e.g., tense, sad). If no music is described, output ”None”.
`[sound event]`: Include descriptions of Non-musical sounds (e.g., actions, engine noises, wind) and Non-verbal vocal sounds (e.g., laughter, sighing, coughing) (EXCEPT for any mention of ”Crumpling sound” - these should be completely omitted). If no sound events are described (after filtering out Crumpling sounds), output ”None”.
Do not include any content related to speech, dialogue, or verbal communication in either section. Remember not to generate any explanations or notations after the separated part output.
Output Format:
`[music]` ¡description of music and emotional characteristics if any¿ or None
`[sound event]` ¡description of sound events¿ or None
Example: ¡example¿

Prompt for cleaning vocal traits caption

You are an expert in speech and subtitle data analysis. You have access to the following:
Vocal traits: The text that describes the emotion and other characteristics of speech sentence by sentence, each utterance contains speech content, emotion and speaker traits.
Subtitle: The transcription of speech. Each line in the subtitle represents a single phrase.
Your task is to evaluate the provided vocal traits by comparing it with the corresponding subtitle. The goal is to determine whether the vocal traits contains usable information based on its alignment with the subtitle. Your analysis should be rigorous and follow a structured approach.
Please follow the guidelines below:
1. Comparison with Subtitle:
- Directly compare the Speech Content text with the provided subtitle.
- Discard any speech content marked with neutral emotion (e.g., ”neutral tone”, ”neutral mood”) immediately, regardless of subtitle alignment.
- For non-neutral speech content:
- If the speech content contains phrases that overlap or align with the subtitle (e.g., shared keywords or contextual similarity), proceed to the next stage.
- If no alignment is found, the vocal traits is unusable.
2. Output of Emotional Information:
For each utterance:
- Only if the vocal traits aligns with the subtitle and is non-neutral:
- Replace any speech content in the vocal traits (i.e., quoted or referenced dialogue) with the exact matching phrase from the subtitle.
- Preserve all other emotional/tonal descriptors (e.g., ”sad mood,” ”English accent”).
- Format the output as a coherent description combining the subtitle content and emotional features.
- If no alignment exists, output ‘[Unavailable]‘.
- Always prefix the final output with ‘[Output]‘ or ‘[Unavailable]‘.
Provide the final cleaned utterances in the following format:
`[Output]` ¡utterances with emotional/tonal information¿ if subtitle roughly matches or `[Unavailable]` if NO utterance is available
Example: ¡example¿
Input:
Vocal Traits: {vocal_traits}
Subtitle: {subtitle}
Please follow the guidelines to clean the vocal traits text, and provide the final output in the specified format. Remember to output ’`[Unavailable]`’ only when no utterance is available.

Figure 11: Prompt for audio caption refinement.

Prompts for generating single-scene QA pairs

Your task is to generate a question-answer pair based on the instructions. The question must utilize both visual and audio information (e.g. speech, sound event, or music). Generate a question-answer pair based on this analysis, ensuring the question does not contain any hints or details about the answer. The answer should be precise and concise, and include an explanation justifying the answer. If the material does not provide sufficient information to generate a valid question-answer pair, respond with '[Unavailable]'.
Format the output as follows:
`[Question]`......
`[Answer]`......
`[Explanation]`......
Please follow the instructions below:{instruction}
Example:{examples}
Here are the input material:{input_modality}
Please follow the instructions and refer to the examples provided to assist in your question design.

Prompts for generating multi-scene QA pairs

Your task is to generate five different question-answer pairs based on the instructions. Each question must integrate information from both the audio and video modalities, ensuring that neither modality alone can provide the answer. Importantly, the information used to formulate each question should be derived from a few consecutive segments of the material, rather than from a single segment or the entire content.
The generated question-answer pairs should be unique and avoid overlapping in content. The questions should be designed without revealing hints or details about the answers. The answers should be precise and concise, accompanied by a brief explanation that justifies the response based on the combined audio-visual information from the selected segments. Use the video description to enhance your understanding of the material.
Format the output as follows:
¡Output id¿
`[Question]`......
`[Answer]`......
`[Explanation]`......
Please follow the instructions below:
Generate five different question-answer pairs that {instruction}
Example:{examples}
Here is the input material:{segments_info}
Please follow these instructions and refer to the examples provided to guide your question design.

Prompts for generating full-scene QA pairs

Your task is to generate five different question-answer pairs based on the instructions. The questions must utilize both visual and audio information, and cannot be answered by information from only one modality. The generated five questions should be different from each other. The questions should also be derived based on the whole movie, not just a few segments. Generate the question-answer pairs based on this analysis, ensuring each question does not contain any hints or details about the answer. The answers should be precise and concise, and include an explanation justifying the answer for each question. You can use the video description to help you better understand the material.
Format the output as follows:
¡Output id¿
`[Question]`......
`[Answer]`......
`[Explanation]`......
Please follow the instructions below:
Generate five different question-answer pairs that {instruction}
Example:{examples}
Here are the input material:{segments_info}
Please follow the instructions and refer to the examples provided to assist in your question design.

Figure 12: Prompts for generating QA pairs

Prompt for modality check in general check

You are a multimodal understanding assistant. You have access to the following:
1. Question: A question related to the video clip.
2. Answer: An answer provided to the question.
3. Explanation: An explanation supporting the answer.
Your task is to evaluate whether the question can be answered using only one modality (either video or audio) or if it requires both modalities. Please strictly base your judgment on the information explicitly required to answer the question, as well as the content of the provided answer and explanation. Avoid making assumptions about the content of the modalities beyond what is explicitly stated in the question, answer, or explanation.
Please follow these steps to complete your evaluation:
1. Information Analysis:
Analyze the question to identify the specific visual and auditory details required to answer it. Does the question require visual details (e.g., objects, actions, or settings) or auditory details (e.g., speech, sound effects, or music)?
Extract the visual and the auditory information from the explanation to determine which modalities are used to support the answer.
2. Modality Assessment:
Based on the analysis of the question and explanation, determine if the required information can be obtained entirely from one modality (either video or audio) or if both audio and visual modalities are necessary.
- Determine if the question can be answered using only the extracted video text.
- Determine if the question can be answered using only the extracted audio text.

If the question requires information from both the video text and the audio text to be answered, then it is considered feasible to use both modalities.
3. Conclusion:
Provide your final determination: Output [YES] if the question explicitly requires information from both video and audio modalities to be answered correctly, or if the answer and explanation rely on information from both modalities. Otherwise, output [YES]. Here is the question: {question}
Here is the answer: {answer}
Here is the explanation: {explanation}
Please complete the Information Analysis, Modality Assessment, and Conclusion stages with the special answer token [YES] or [NO].

Prompt for checking format content and speculation in general check

You are a quality evaluation assistant. You have access to the following:
1. Question: A question related to a given context.
2. Answer: An answer provided to the question.
3. Explanation: An explanation supporting the answer.
Your task is to evaluate the quality of the question-answer pair by performing two checks: format check and content check. Please strictly base your judgment on the information explicitly provided in the question, answer, and explanation. Avoid making assumptions beyond what is stated. Please follow these steps to complete your evaluation:
1. Format Check:
- Analyze the question to determine how many distinct pieces of information it is asking for.
- Check if the answer addresses all the pieces of information requested in the question.
- If the question asks for only one piece of information and the answer fully addresses it, proceed to the content check.
- If the question asks for multiple pieces of information but the answer only addresses some of them, output ‘[NO]‘ and stop.

2. Content Check:
- Analyze the explanation to determine if it is reasonable and logically sound.
- Check if the answer can be derived from the explanation and if the answer is correct based on the context of the question.
- If the explanation is reasonable and the answer is correct and supported by the explanation, output ‘[YES]‘.
- If the explanation is unreasonable or the answer cannot be derived from the explanation, output ‘[NO]‘.

3. Speculation Check:
- Analyze the explanation to determine if the answer relies too heavily on speculation rather than concrete evidence or logical reasoning.
- If the explanation provides clear, evidence-based reasoning or logical steps to derive the answer, proceed to the final output.
- If the explanation relies on assumptions, guesses, or unsupported claims, output [NO] and stop.

Final Output:
- If both the format check and content check pass, output ‘[YES]‘.
- If either check fails, output ‘[NO]‘.

Here is the question: {question} Here is the answer: {answer} Here is the explanation: {explanation} Please complete the Format Check, Content Check, and Final Output stages with the special answer token ‘[YES]‘ or ‘[NO]‘.

Figure 13: Prompts for general checks

Prompt for sequence check

You are a quality control assistant for video-based question-answering pairs. Your task is to validate whether a given QA pair about element ordering in videos is correct and properly sourced. Follow this 4-stage process:
Stage 1: Element Extraction
- Extract all elements to be sorted from the question (format: (a)[element1], (b)[element2], etc.)
- Verify exactly 3 elements exist. If not, immediately output "[NO]"

Stage 2: Occurrence Localization
For each extracted element:
1. Search through all video segments to find its first occurrence
2. Record for each element:
- Segment ID of first appearance
- Modality type (video caption/subtitle/speech emotion)
3. If any element cannot be found → Output "[NO] (unverifiable element: [element_name])"

Stage 3: Element Validation
Verify the located elements meet these criteria:
1. Unique Segment Check:
- All elements must appear in different segments
- If any segment ID is shared → Output "[NO] (co-occurring elements: [element1] & [element2] in segment X)"
2. Modality Diversity Check:
- Elements must come from $\geq$ 2 different modalities
- If all same modality → Output "[NO] (single modality: [modality_type])"

Stage 4: Order Verification
1. Sort elements by their first appearance segment ID (ascending)
2. Compare against provided answer:
- If orders match → Output "[YES]"
- If orders differ → Output "[Corrected]" with proper order and explanation

Output Format:
`[Validating]` ¡4-stage analysis¿
`[Output]`: [YES/NO/Corrected]
`[Corrected:  (a) (b) (c)]` (if applicable)
`[Explanation]` (if Corrected):
- (a) [element1]: first appears in segment [X] ([modality])
- (b) [element2]: first appears in segment [Y] ([modality])
- (c) [element3]: first appears in segment [Z] ([modality])

Provided Information:
Video Text: {segments_info}
Question: {question}
Answer: {answer}
Explanation: {explanation}
Please perform the stages above carefully in [Validating] and provide the final output in the specified format.

Prompt for ambiguity check

You are a QA evaluation assistant tasked with filtering incorrect or low-quality question-answer pairs based on video and audio context. Follow this structured evaluation:
Phase 1: Specificity Check
- Check if the 'answer' is overly generic (e.g., fails to distinguish between objects/agents).
- Example: If the 'question' asks "Where is the sound source located?" and the 'answer' is "in a wooden house" (while the entire video occurs in a wooden house), mark as "[NO]" (lacks specificity).
- Output: Proceed only if "[PASS]"; else, output "[NO]".

Phase 2: Sound Source Ambiguity (Video Context)
- Using the 'Video Caption', verify if other objects in the scene could plausibly produce the sound mentioned in the 'question'.
- Example: If the 'question' asks "What caused the splashing sound?" and the 'Video Caption' only describes a "person by a pool," but no other water-related objects exist, mark as "[NO]".
- Output: Proceed only if "[PASS]"; else, output "[NO]".

Phase 3: Cross-Modality Dependency
- Determine if the 'answer' can be derived solely from 'Video Caption' + 'question' + commonsense (ignoring 'Sound Event').
- Example: If the 'question' asks "Where is the splashing sound coming from?" and the 'Video Caption' mentions "a beach with waves," commonsense suggests "ocean" → mark as "[NO]" (audio not needed).
- If 'Sound Event' is required (e.g., to distinguish between similar objects), mark as "[PASS]".
- Output: If "[PASS]" in all phases, output "[YES]"; else, "[NO]".

Final Output:
- Only "[YES]" or "[NO]" based on the above checks.

Provided Information:
{segments_info}
Here is the question-answer pair to be analyzed:
- Question: {question}
- Answer: {answer}
- Explanation: {explanation}

Perform the analysis step-by-step and output either "[YES]" or "[NO]" based on your evaluation.

Figure 14: Prompts for sequence check and ambiguity check.

Prompt for sound event check

You are a QA evaluation assistant tasked with filtering incorrect question-answer pairs based on video and sound event information.
Follow this phased approach:
Phase 1: Off-Screen Sound Check
- If the Answer describes an off-screen sound event (e.g., ”An off-screen object”), output [NO].
Phase 2: Sound Event Presence Validation
- Extract the sound event mentioned in the Question, Answer, or Explanation.
- Check if this sound event exists in the provided Audio Text. If not, output [NO].
Phase 3: Contextual Consistency with Video
- Using only the Video Caption (ignore Explanation), verify if the sound event logically fits the scene.
- Example: If the sound event is ”glass breaking” but the Video Caption lacks glass-related objects/actions, output [NO].
- Do not speculate; reject if the video lacks supporting evidence.
Phase 4: Final Judgment
- If all phases pass, output [YES]. Otherwise, output [NO].
Here is the provided information:{segments_info}
Question: {question}
Answer: {answer}
Explanation: {explanation}
Perform the four-phase analysis and output either ’[YES]’ or ’[NO]’.

Prompt for music check

You are a QA pair evaluation assistant. Your task is to determine whether a given question-answer pair is valid based on the provided video description and music content. Follow these steps:
1. Phase 1: Music Information Validation
- Extract the music-related information used in the ‘Answer‘ and ‘Explanation‘ of the QA pair.
- Check if this music information appears in the provided ‘Music Content‘.
- If the music information is not found in the ‘Music Content‘, output ‘[NO]‘ (invalid QA pair).
2. Phase 2: Visual Information Cross-Check
- If the music information is valid (from Phase 1), analyze whether the emotion/atmosphere described in the music can also be inferred from the ‘Video Caption‘ (e.g., character expressions, scene mood, or events).
- Example: If the music mentions a ”sad atmosphere” and the video shows ”a character crying,” the music info can be inferred visually → ‘[NO]‘; If the music describes a ”warm atmosphere” and the video caption mentions ”bright lighting,” the music info can be inferred visually → ‘[NO]‘; If the music’s emotional/atmospheric cues can be derived from the video alone, output ‘[NO]‘.
3. Phase 3: Final Judgment
- If the QA pair passes both Phase 1 and Phase 2 (i.e., music info is valid and cannot be inferred visually), output ‘[YES]‘.
- Otherwise, output ‘[NO]‘.
Here is the provided information:{segments_info}
- Question: {question}
- Answer: {answer}
- Explanation: {explanation}
Perform the three-phase analysis and output either ’[YES]’ or ’[NO]’.

Prompt for vocal traits check

You are an assistant tasked with validating question-answer (QA) pairs generated from video content, specifically focusing on the use of speech emotion data (Speech Content, Emotion, Speaker Traits). Your goal is to filter out incorrect or weakly supported QA pairs by following this phased approach:
Phase 1: Extract Utilized Speech Emotion Information
- From the Question and Explanation (if provided), identify:
- Speech Content: Exact phrases/words from the audio used to answer the question.
- Emotion: The emotion label (e.g., ”angry,” ”joyful”) tied to the speech content.
- Speaker Traits: Any speaker characteristics (e.g., ”deep voice,” ”child”) referenced.
- Output: List only the explicitly used components. If none are used, stop and output ‘[NO]‘.
Phase 2: Verify Grounding in Provided Speech Emotion Text
- Check if the extracted Speech Content, Emotion, and Speaker Traits from Phase 1 appear verbatim or unambiguously in the provided speech emotion text.
- Example: If the QA pair uses ”confused” emotion for ”What?”, but the speech emotion text lacks this pairing, it fails.
- Output: If any extracted component is missing, output ‘[NO]‘.
Phase 3: Assess Text-Based Emotion Inferrability
- For the Speech Content and Emotion pair used in the QA pair, determine if the emotion could be directly inferred from the text alone (e.g., ”I’m furious” → ”angry”).
- Disqualify: Obvious cases (e.g., sarcasm-free explicit statements, clear interrogatives).
- Output: If inferrable, output ‘[NO]‘.
Phase 4: Check Video-Based Emotion Redundancy
- Determine if the Emotion used in the QA pair could also be clearly deduced from the video caption (e.g., ”she frowns” → ”sad”).
- Note: Assume video captions describe visible emotions unless stated otherwise.
- Output: If deducible, output ‘[NO]‘.
Final Decision
- Only output ‘[YES]‘ if all phases are passed (i.e., the QA pair uses non-inferrable, video-independent speech emotion data with explicit grounding).
- For any phase failure, output ‘[NO]‘.
Provided Information:{segments_info}
- Question: {question}
- Answer: {answer}
- Explanation: {explanation}
Perform the analysis step-by-step and output either ”[YES]” or ”[NO]” based on your evaluation.

Figure 15: Prompts for audio check.

Prompt for interval identification

You are an expert in finding all the continuous segments needed to answer a question-answer pair. Your task is to identify the minimal continuous sequence of movie segments (neither the first nor last segment) that contains all necessary information to answer the given question.
You will be provided with:
1. A question-answer pair
2. An explanation of how the answer is derived
3. Complete information about all movie segments (timestamps, video descriptions, audio descriptions, and subtitles)

Instructions:
1. Carefully analyze the question and answer explanation to understand what information is required
2. Examine all movie segments sequentially to locate where the relevant information begins
3. Determine where the last necessary piece of information appears
4. Select the earliest segment where required information starts (start segment) and the latest segment where required information ends (end segment)

Ensure:
1. The selected segments form a continuous sequence
2. The sequence is not from the very first to the very last segment
3. All information needed to answer the question is contained within this sequence
4. The sequence is as compact as possible

Output Format:
Provide your response in this exact format:
`[Start]`: ¡segment number¿ `[End]`: ¡segment number¿ `[Rationale]`: ¡brief explanation of why these segments were chosen¿
Important Notes:
- If the answer requires information that only appears in disjoint segments, select the smallest continuous sequence that contains all relevant segments
- The start and end segments must be different (cannot be the same segment)
- Never choose segment 0 and the final segment at the same time

Input:
Here is the question: {question}
Here is the answer: {answer}
Here is the explanation: {explanation}
Here are the segments information:{segments_info}
Please follow the guidelines and use the input material to identify start segment and end segment for the question-answer pair. Make sure that the generated output follow output format.

Prompt for generating distractors

You are an expert in generating multi-choice question. Your task is to generate distractors based on the guidelines.
Given the following background information, question, correct answer, and answer rationale, generate three incorrect answer options (distractors) that closely mimic the correct answer in terms of length, format, and style. The distractors should appear reasonable to someone who doesn't fully grasp the concept but must contain subtle errors (factual, logical, or contextual).
Please follow these guidelines to generate distractors:
1. **Selective Modification**: Alter specific elements such as character actions, dialogue, objects, or settings to create plausible yet incorrect options.
2. **Maintain Plausibility**: Ensure each distractor could feasibly occur within the context of the video, making them appear credible based on the visual and audio cues.
3. **Incorporate Diverse Misdirections**:
- **Action Confusion**: Modify or swap character actions or events in ways that fit the context but are incorrect.
- **Dialogue Adjustments**: Propose believable alterations to dialogue or audio cues that didn't actually occur.
- **Object or Setting Misdirection**: Suggest plausible but incorrect details about objects, settings, or visual elements.
- **Speech Emotion Alteration**: Modify the described emotional tone of speech content while keeping the words themselves accurate.
- **Sound Event Manipulation**: Change specific sound effects or environmental audio cues to similar but incorrect versions that could plausibly exist in the context.
- **Musical Atmosphere Shift**: Adjust the described mood or emotional impact of background music to a different but related atmosphere.
4. **Incorporate Partial Truths**: Use true audio-visual details or partial truths within the distractors to add complexity, ensuring these elements do not directly answer the question but make the distractors more compelling.
5. **Avoid Obvious Falsities**: Shift the context or details significantly without creating options that are blatantly wrong or unrelated to the video.
6. **Ensure Distinct Incorrectness**: Craft distractors that will be clearly identifiable as incorrect by someone who has closely watched and listened to the video, challenging their attention to detail.

Requirements for Distractors:
1. Plausibility: Each distractor should seem correct at first glance, matching the tone and structure of the correct answer.
2. Variety: Errors should vary (e.g., minor inaccuracies, flipped terms, oversimplifications, or common misconceptions).
3. Consistency: Maintain the same verb tense, technicality, and formatting as the correct answer. Format the output as follows:

`[Distractor 1]` ¡Incorrect but plausible option¿
`[Distractor 2]` ¡Incorrect but plausible option¿
`[Distractor 3]` ¡Incorrect but plausible option¿

Provided Information:
Background infromation:{segments_info}
Question: {question}
Correct Answer: {answer}
Answer Rationale: {explanation}
Now generate three high-quality distractors for the given question and correct answer in the specified format. DO NOT provide option rationale.

Figure 16: Prompts for interval check and distractor generation.