Audio-centric Video Understanding Benchmark without Text Shortcut

Yudong Yang[1,3*], **Jimin Zhuang**[1,3*], **Guangzhi Sun**[2,3], **Changli Tang**[1,3], **Yixuan Li**[1,3], **Peihan Li**[3], **Yifan Jiang**[3], **Wei Li**[3], **Zejun Ma**[3], **Chao Zhang**[1✉]

[1]Tsinghua University [2]University of Cambridge [3]ByteDance

{yang-yd21, zhuangjm21}@mails.tsinghua.edu.cn, cz277@tsinghua.edu.cn

Abstract

Audio often serves as an auxiliary modality in video understanding tasks of audio-visual large language models (LLMs), merely assisting in the comprehension of visual information. However, a thorough understanding of videos significantly depends on auditory information, as audio offers critical context, emotional cues, and semantic meaning that visual data alone often lacks. This paper proposes an audio-centric video understanding benchmark (**AVUT**) to evaluate the video comprehension capabilities of multimodal LLMs with a particular focus on auditory information. AVUT introduces a suite of carefully designed audio-centric tasks, holistically testing the understanding of both audio content and audio-visual interactions in videos. Moreover, this work points out the text shortcut problem that largely exists in other benchmarks where the correct answer can be found from question text alone without needing videos. **AVUT** addresses this problem by proposing a answer permutation-based filtering mechanism. A thorough evaluation across a diverse range of open-source and proprietary multimodal LLMs is performed, followed by the analyses of deficiencies in audio-visual LLMs. Demos and data are available at `https://github.com/lark-png/AVUT`.

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) (Hurst et al., 2024; Team et al., 2024; Wang et al., 2024; Zhang et al., 2024; Chen et al., 2024) have led to significant progress in video understanding. Evaluation of these models has been focusing predominantly on their visual abilities, with a large number of video benchmarks proposed in the past few years (Xiao et al., 2021; Maaz et al., 2024; Patraucean et al., 2023; Li et al., 2024a; Mangalam et al., 2023). While several benchmarks (Yang et al., 2022; Geng et al., 2023; Chen et al., 2023b,a; Fu et al., 2024a; Geng et al., 2024; Sung-Bin et al., 2024) incorporate audio in their evaluations, they are still largely visual-centric where the audio modality only acts as a secondary source of information.

However, a comprehensive understanding of videos relies heavily on auditory information, as it provides essential context, emotional cues, and semantic meaning often absent from visual data alone. Existing benchmarks either overlooks the importance of audio in video understanding, or only focus on the speech content and ignoring other information. Consequently, there is an urgent need for dedicated efforts to evaluate model capabilities to understand and process speech and audio in videos. Moreover, most existing video understanding benchmarks suffer from the text shortcut problem, where the model can deduce the correct answer by simply focusing on the text alone. To fill the gaps, we propose **AVUT**, the text-shortcut-free benchmark specifically designed for audio-centric video understanding. The key differences of **AVUT** compared with examples of existing benchmarks are shown in Table 1.

AVUT introduces a suite of carefully designed audio-centric video understanding tasks, which cover not only content understanding of the video but also the interaction between audio and visual information, such as time synchronization or content matching. Besides, **AVUT** adopts a dual-dataset design which combines high-quality human annotation with scalable, semi-automatic annotation to obtain a holistic evaluation benchmark. Specifically, a novel Gemini-powered in-context learning data creation approach is proposed to generate high-quality question-answer pairs that are further validated by human annotators. **AVUT** incorporates **2,662** videos meticulously selected from YouTube, covering 18 audio-centric domains. This collection is split into two partitions: (1) AV-Human, comprising 698 videos with 1,734 expertly

[*] Equal contribution; ✉: Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6569–6587
November 4-9, 2025 ©2025 Association for Computational Linguistics

Benchmarks	Num. of Videos	Avg.Duration(s)	QA Pairs	Tasks	Annotation	w. Audio	Ori. Videos
Video Benchmark							
NExT-QA (2021)	1,000	39.5	8,564	3	A	✗	✗
Video-Bench (2023)	5,917	56.0	17,036	10	SA	✗	✓
EgoSchema (2023)	5,063	180.0	5,063	1	SA	✗	✗
MVBench (2024a)	3,641	16.0	3,641	20	A	✗	✗
MMBench-Video (2024)	609	165.4	1,998	26	M	✗	✗
Audio-Visual Benchmark							
LongVALE (2024)	8,400	235.0	-	3	SA	✓	✗
AVHBench (2024)	2,327	-	5,816	4	SA	✓	✗
Video-MME (2024a)	900	1017.9	2,700	12	M	✓	✓
AVUT	2,662	67.8	11,609	8	M+SA	✓	✓

Table 1: Comparison of AVUT with other representative video and audio-visual benchmarks. “Annotation” is the annotation type (A: Automatic, SA: Semi-Automatic, M: Manual, M+SA: Mixed); “w. Audio” indicates the presence of audio; and “Ori. Videos” denotes whether the videos are newly collected (✓) or repurposed (✗).

human-annotated question-answer pairs; and (2) AV-Gemini, containing the remaining 1,964 videos with **9,875** question-answer pairs obtained from the proposed in-context learning data creation method. In addition, to address the text shortcut problem, a novel answer permutation-based filtering mechanism is proposed and applied in **AVUT**, ensuring the necessity of multimodal inputs. The main contributions are as follows:

- We introduce **AVUT**, the **A**udio-centric **V**ideo **U**nderstanding benchmark without **T**ext shortcut, featuring diverse tasks encompassing both audio content understanding and audio-visual alignment.
- **AVUT** first systematically investigate the text shortcut problem in existing audio and video benchmarks, and proposes a novel filtering mechanism to minimize the text shortcut and ensure sufficient necessity of the multimodal inputs.
- We evaluate a diverse range of audio-visual, visual-only, and audio-only models on **AVUT** to demonstrate their capabilities on audio-centric video understanding.

2 Related Work

MLLMs. Early research in multimodal alignment explored the creation of shared representations across images, audio, and text (Aytar et al., 2017), demonstrating the potential of transferable classifiers. Further research on shared representations across modalities leads to encoders that can align with LLMs like Whisper (Radford et al., 2023) or CLIP (Radford et al., 2021), making MLLMs technically possible. Recent progress in audio-visual learning has spurred the development of large language models (LLMs) capable of processing both audio and visual information. Early video understanding works like VideoChat (Li et al., 2023a) uses external subtitles for speech. Subsequent models have explored approaches to fuse audio and visual modalities. PandaGPT (Su et al., 2023) integrates ImageBind (Girdhar et al., 2023) and Vicuna (Chiang et al., 2023), leveraging image-text pairings for enhanced cross-modal understanding, while VAST (Chen et al., 2023b), trained on the large-scale VAST-27M (Chen et al., 2023b) dataset, further advances multimodal capabilities. Models like CAT (Ye et al., 2025) employ specialized architectures, such as clue aggregators, to effectively process audio-visual events. Video-LLaMA (Zhang et al., 2023), VideoLLaMA2 (Cheng et al., 2024) and Baichuan Omni (Li et al., 2024b) have been fine-tuned specifically on audio data to improve multimodal comprehension. Architectures like the multi-resolution causal Q-former in video-SALMONN (Sun et al., 2024) aim to bridge pre-trained audio-visual encoders with LLMs efficiently. Furthermore, models like NExTGPT (Wu et al., 2024), OneLLM (Han et al., 2024), and VITA (Fu et al., 2024b) use “omni-encoders” to uniformly process various data modalities, including audio and video.

Video Benchmark. Numerous efforts have focused on developing video benchmarks (Xiao et al., 2021; Li et al., 2021, 2024a; Ning et al., 2023; Li et al., 2023b; Liu et al., 2024; Mangalam et al., 2023; Fu et al., 2024a; Fang et al., 2024), predominantly targeting the visual aspects of video content. However, the crucial role of audio in video understanding has received comparatively less attention. Several early datasets explore the use of audio in video understanding. AVSD (Alamri et al., 2019) focuses on scene-aware dialogue, Pano-AVQA (Yun et al., 2021) on panoramic videos, MUSIC-AVQA (Li et al., 2022) specifically on musical contexts, and VGGSound (Chen et al., 2020) on short videos of audio events. Thus, these datasets target specific scenarios. While AVQA (Yang et al., 2022) explores the synergy between audio and visual information but is limited by shorter video lengths and simplistic task design. More recent datasets like UnAV-100 (Geng et al., 2023), VALOR (Chen et al., 2023a), and VAST-27M (Chen et al., 2023b) offer large-scale data but prioritize specific tasks, such as audio event detection, retrieval, and captioning. Similarly, Long-VALE (Geng et al., 2024) focuses on audio-visual-language events, and AVHBench (Sung-Bin et al., 2024) on cross-modal hallucination. These benchmarks, often lacking human annotation, are less suited for evaluating the nuanced and complex aspects of audio-centric understanding, particularly alignment capabilities, targeted by AVUT.

3 Our Benchmark

3.1 Data Curation Principles

Focusing on Audio-Centric Understanding in Video. Unlike existing video benchmarks that predominantly target visual understanding, AVUT focuses specifically on the crucial role of audio in video comprehension. To this end, the tasks in AVUT are designed to focus on the following two aspects: (1) **Audio Content Understanding**, requiring models to perceive and understand audio information; (2) **Audio Visual Alignment**, requiring the models to align audio with the corresponding visual information.

Covering Diverse Audio-Centric Video Domains and Comprehensive Annotation. To thoroughly evaluate audio-centric understanding capabilities in MLLMs, AVUT incorporates a diverse selection of 18 video domains where audio plays a central role. For robust evaluation, AVUT offers both **AV-Human**, a meticulously human-annotated dataset, and **AV-Gemini**, a significantly larger dataset generated via a novel Gemini-augmented approach. This combination of diverse content and rich annotation enables AVUT to effectively support the research in audio-centric video understanding.

3.2 Audio-Centric Task Definition

The tasks are designed to probe specific aspects of audio-centric video understanding, presenting unique challenges for MLLMs. Examples of each task are shown in Table 2.

3.2.1 Audio Content Understanding

The ability to perceive and process audio is fundamental to audio-centric video understanding. Inspired by classic audio processing challenges, we designed tasks to probe a model's capability to extract meaningful information, quantify audio events, and locate them in a video. Therefore, we define three tasks as follows:

- **Audio Information Extraction (AIE)** is to extract specific information from spoken language, such as sentiment, key facts, or main points.
- **Audio Content Counting (ACC)** is to test the accuracy of counting the number of occurrences for a specific audio event.
- **Audio Event Location (AEL)** is to pinpoint the precise start and end time of a specific audio event in the video.

3.2.2 Audio Visual Alignment

Existing benchmarks often lack a dedicated focus on this cross-modal alignment, which is fundamental to a comprehensive understanding of video content. To bridge this gap, we define five matching-based tasks as follows.

- **Audio-Visual Character Matching (AVCM)** tests whether the model links spoken content to the corresponding character's visual appearance.
- **Audio-Visual Object Matching (AVOM)** is to draw the connection between utterances and depicted objects or background scenes.
- **Audio-Visual Text Matching (AVTM)** requires the model to associate spoken information with on-screen text, e.g. subtitles.
- **Audio-Visual Segment Matching (AVSM)** tests whether the model aligns audio clips with their corresponding video segments.
- **Audio-Visual Speaker Diarization (AVDiar)** requires the model to determine "who spoke when" by integrating audio and visual cues.

3.3 Dataset Construction

AVUT contains **2,662** videos with **11,609** questions in total with each question either created or validated by human annotators. This section describes the data construction in detail.

Theme	Task Type	Format	Example
Audio Content Understanding	Audio Information Extraction	Multiple Choice	*The person in the audio said when they would have a trip from Central Africa to the farthest edge of Norway?* (A) The spring of 2025. (B) The summer of 2025. (C) The spring of 2026. (D) The summer of 2026.
	Audio Content Counting	Multiple Choice	*How many times does the rumbling sound appear in the video?* (A) 2. (B) 3. (C) 4. (D) 5.
	Audio Event Localization	Multiple Choice	*In what seconds does the person begin to sing In the video?* (A) 8 to 9. (B) 10 to 11. (C) 13 to 14. (D) 15 to 16.
Audio Visual Alignment	Audio Visual Character Matching	Multiple Choice	*Who says "there's no way of knowing who did it" in the video?* (A) The woman on the bottom right of the screen. (B) The man on the top right of the screen. (C) The man on the top left of the screen. (D) The woman on the bottom left of the screen.
	Audio Visual Object Matching	Multiple Choice	*What happens when the singer in the video finishes singing "talk to me"?* (A) A picture appears. (B) A man appears. (C) A woman appears. (D) Nothing.
	Audio Visual Text Matching	Multiple Choice	*When the number 48KG appears on the screen, what is being explained at this time?* (A) J2 Weight requirements for competitors. (B) Classification of judo. (C) Judo techniques. (D) The duration of a judo competition.
	Audio Visual Segment Matching	Open Ended	*The model is prompted to sort the audio pieces of a video whose audio was clipped into 4 pieces and shuffled.* Success rate of sorting the entire audio and any two clips is calculated.
	Audio Visual Speaker Dirization	Open Ended	*The model is instructed to transcribe from only the speaker with a certain visual characteristic.* WER calculation is then applied to the model's transcription.

Table 2: Task examples of AVUT. "Audio Visual Segment Matching" is to sort four randomly shuffled audio segments in a video into their original order. "Audio Visual Speaker Diarization" is to transcribe the speech of a visually specified speaker with the Word Error Rate (WER) as the metric.

3.3.1 Video Collection

We define 18 audio-centric video domains (see Appendix A) to cover a diverse range of scenarios in which audio plays a crucial role in comprehension. Our annotators searched YouTube for videos within these audio-centric domains to manually collect a total of 2,662 raw videos, as depicted in Figure 1. The collected videos were required to be in English, rich in audio content, and recently published (mostly after 2023) to minimize overlap with existing training sets. Videos were intentionally limited to two minutes as driven by the following considerations: First, the nature of our audio-centric tasks, which focus on analyzing specific audio events or aligning audio cues with corresponding visual information within localized segments, lends itself well to evaluation within this shorter time frame. Second, most popular open-source MLLMs struggle to process longer videos, and hence short videos allow us to perform a wider range of comparisons. Finally, extending the length of the video would lead to an exponential increase in the cost and potential inaccuracies of manual annotation, without significantly enhancing the discriminatory power of our benchmark. These videos further went through a validation process, where independent annotators confirmed the designated domains and the presence of sufficient and relevant audio information. The detailed video collection process is provided in Appendix D.

3.3.2 Human Annotation Partition

The video collection and annotation for AV-Human was conducted by a team of trained annotators with expertise in data labeling and rigorous quality control procedures (see Appendix C for details).

Annotation Procedure. We selected 698 videos for full annotation where annotators were given detailed task definitions and annotation examples. Then, for each video, annotators selected the three most suitable task types and created corresponding multiple-choice questions and answer options, resulting in a total of **1,734** question-answer pairs. For AVDiar, annotators selected 260 videos from the full annotation set that contain a higher number of speakers (3.59 speakers per video on average). Detailed annotations for these videos included speaker appearance, start and end times of utterances, and speech transcriptions. The Audio-Visual Matching task did not require manual annotation, as audio-video segment pairings were generated automatically using a script.

Quality Assurance. For each task in AVUT, we selected and trained annotators with relevant expertise. A rigorous two-stage quality control process, involving specialist review and cross-validation, ensured high-quality annotations. During annotation, questions were carefully designed to require genuine multimodal understanding, avoiding commonsense reasoning and text-based answers.

3.3.3 Gemini-Augmented Partition

Human annotations are both costly and often have a bias towards visual-grounded questions (see Appendix J). To mitigate these limitations and to expand to a larger scale, we employed a semi-automatic annotation strategy using Gemini 1.5 Pro and created the AV-Gemini subset. Specifically, we propose an **in-context learning** approach which includes 4 stages as shown in Figure 1:

Stage 1: Demonstration Selection. We carefully selected four representative examples from the human-annotated data for each of the six multiple-choice question task types. These exam-

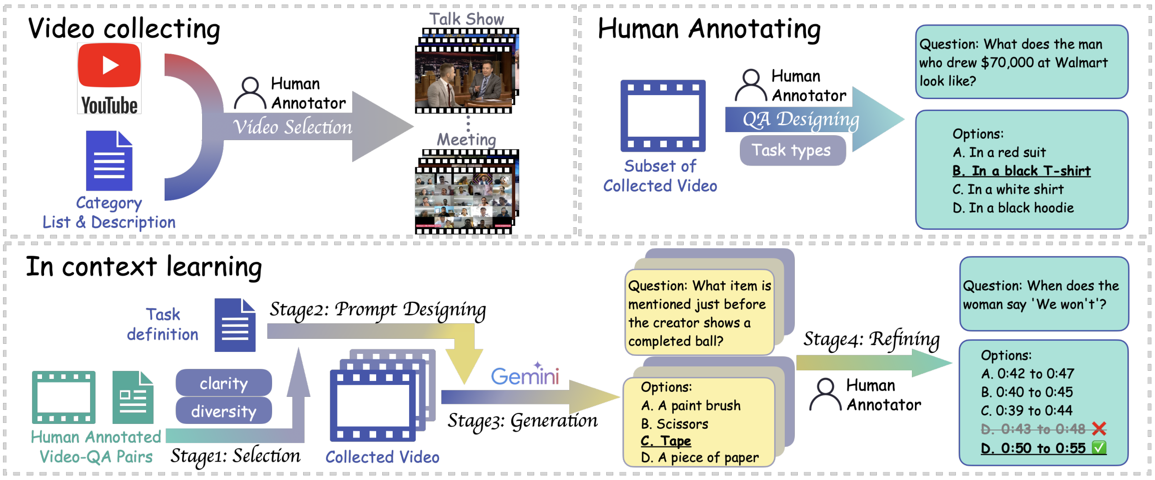


Figure 1: The Pipeline of AVUT. The pipeline consists of three main parts: video collection, human annotation and in-context learning creation. Annotators select videos from YouTube and these videos are then used for human annotation, where question-answer pairs are designed based on various task types. The remaining videos are used for in-context learning with Gemini 1.5 Pro. The in-context learning process involves four stages: (1) Demonstration Selection; (2) Prompt Designing; (3) Generation; (4) Refining.

ples, chosen for their clarity and diversity, served as high-quality demonstrations for Gemini.

Stage 2: Prompt Construction. We constructed prompts for Gemini by combining the task definition with the selected demonstrations. Each prompt provided clear instructions and illustrative examples to guide Gemini's generation process.

Stage 3: Automated Annotation Generation. We provided Gemini with the remaining 1,964 raw videos from our collected set of 2,662. For each video and each task type, Gemini generates a question, four options and the correct option, resulting in a total of **9,875** generated QA pairs.

Stage 4: Human Review and Refinement. All generated QA pairs underwent meticulous human review, focusing on accuracy, relevance, adherence to task requirements, and answer correctness. Annotators manually corrected any discrepancies. (See Appendix F for correction examples.)

4 Text Shortcut Problem

Text shortcut occurs when models give correct answers by only looking at the textual content of the question without requiring the corresponding video, which results in unreliable evaluation of the video understanding capabilities. This section analyzes the causes and consequences of text shortcut and details the methodology employed in AVUT to mitigate this issue.

Text shortcut arises from several factors. Bias in question formulation can inadvertently provide textual cues or biases that hint at the correct answer, even without viewing the video. For instance, a question about a "fast-moving animal" might bias the model towards choosing "cheetah" even if the video depicts a different animal. Furthermore, models may have encountered similar question-answer pairings during their training, effectively memorizing the associations rather than learning to reason about video content. Finally, the powerful text processing capabilities of advanced language models like GPT-4o allow them to exploit subtle textual cues that humans annotators might overlook.

Text shortcut severely compromises evaluation validity. Models may appear to perform better than they genuinely do, masking their true video understanding capabilities and leading to inflated performance metrics. This can, in turn, misdirect research efforts towards optimizing for textual cues rather than genuine video comprehension, hindering true progress in multimodal learning.

AVUT adopts a filtering mechanism based on answer permutation. For each multiple-choice question, we created four cyclic permutations of the answer options (ABCD, BCDA, CDAB, DABC). The permutations help exclude the influence of potential positional bias in the LLM. We then present these variations of the same question to GPT-4o and check the correctness under each permutation. Without video inputs, if the model can still find the correct answer under all four permutations, the question is likely to have a text shortcut. These questions, deemed solvable through textual cues alone, were subsequently removed from AVUT.

This filtering mechanism ensures that the video content is essential to finding the correct answer in the remaining questions.

5 Experiment

5.1 Experimental Setup

We evaluated a range of prominent and representative models on AVUT, encompassing audio-visual, visual-only, and audio-only architectures, to comprehensively assess performance across different modalities. **Audio-Visual MLLMs:** We evaluated Gemini 1.5 Pro (Team et al., 2024), Video-SALMONN(13B) (Sun et al., 2024), VideoLLaMA2(7B) (Cheng et al., 2024), and PandaGPT(13B) (Su et al., 2023). These models received both video and audio inputs. **Visual MLLMs:** We evaluated GPT-4o (Hurst et al., 2024), Qwen-2-VL(7B) (Wang et al., 2024), LLaVA-video(7B) (Zhang et al., 2024), InternVL2(76B and 8B) (OpenGVLab, 2024), VILA-1.5(8B) (Chen et al., 2024), and VideoLLaVA(7B) (Lin et al., 2023). These models received video input with the audio track removed. **Audio MLLM:** We evaluated SALMONN(13B) (Tang et al., 2024) using only the audio track extracted from the videos.

For the closed-source models, GPT-4 and Gemini 1.5 Pro, we used the API versions "gpt-4o-2024-05-13" and "gemini-1.5-pro-preview", respectively, with a sampling rate of 1 frame per second (fps) and a temperature of 1.0. For open-source models, we adhered to their official configurations, including default parameter settings, decoding strategies, temperature settings, and frame sampling rates[2]. During testing, audio-visual models received videos with audio, visual-only models received videos with the audio removed, and audio-only models received only the audio track extracted from the videos. All experiments were conducted on A100 GPUs. Experiments with closed-source models, utilizing their respective APIs, took an average of 6 hours to complete. Inference with open-source models required an average of 4 hours.

5.2 Results

Table 4 presents the overall performance of all evaluated models on the AV-Human, AV-Gemini, and combined Overall (AV-Human + AV-Gemini) datasets. Performance is measured by accuracy on the multiple-choice question tasks. To provide a more detailed analysis, Table 3 presents the per-task accuracy of all models on the AV-Human, AV-Gemini. This result leads to further discussion on how the annotation method can influence the annotation result, and how different abilities of a model can be revealed through different task types.

AV-Human compared to AV-Gemini. Most models show a performance drop when switching from AV-Human to AV-Gemini, particularly in audio content understanding tasks. However, they perform better on AV-Gemini in audiovisual alignment tasks due to differences in annotation methods. Gemini, leveraging in-context learning, captures finer details for QA pair creation, while human annotators focus on a video's main theme, often missing subtleties and favoring visually dominant content. This makes human annotations less challenging for visual-only models. In contrast, human annotations excel in audiovisual matching tasks due to access to full-frame-rate video and synchronized audio, enabling the creation of fine-grained therefore more challenging questions, which are more difficult for models to handle.

The difficulty of different tasks varies. Most open-source models perform well in audio-visual character-matching, as human figures are visually prominent. However, their performance drops in audio-essential tasks like content counting and event localization, highlighting reliance on audio, showcased in case studies in Appendix B. Nevertheless, visual models perform well on audio information extraction tasks due to the high correlation between audio and visual content. Such correlation enables models to leverage visual cues to answer questions, even in tasks designed to depend primarily on audio. This explains why visual-only models also perform well on the matching tasks, since the audio content can be inferred from visual to help solve the question. Besides, the audio-only model, SALMONN, has a more consistent performance among tasks, indicating a stable reliance on audio information across different tasks.

AVUT is designed to be a multi-modal benchmark where audio plays a crucial role. To sup-

[1] We evaluated GPT-4 on a subset of 2100 questions randomly sampled from the full question set with balanced task types, sufficiently representing the original task distribution.

[2] Number of frames or frame rates are: Video-SALMONN, 2 fps; VideoLLaMA2, 16 frames total; PandaGPT, 10 frames total; Qwen2-VL, 1 fps; LLaVA-Video, 32 frames total; InternVL2, 32 frames total; VILA-1.5, 6 frames total; VideoLLaVA, 8 frames total; and SALMONN, 16kHz audio.

Models	AIE (%)	ACC (%)	AEL (%)	AVCM (%)	AVOM (%)	AVTM (%)
Audio Visual MLLMs						
Gemini 1.5 Pro (2024)	**90.02 / 94.70**	**50.85 / 39.99**	**83.43 / 73.32**	**74.05 / 83.13**	**74.79 / 76.62**	**81.22 / 83.68**
VideoLLaMA2 (7B) (2024)	42.30 / 35.65	32.77 / 32.88	32.02 / 25.75	46.80 / 57.41	49.60 / 46.11	48.92 / 40.90
video-SALMONN (13B) (2024)	52.45 / 32.86	33.33 / 30.17	36.05 / 32.48	34.47 / 41.42	30.25 / 31.42	39.95 / 36.21
PandaGPT (13B) (2023)	24.94 / 30.82	25.00 / 29.13	26.16 / 30.61	24.31 / 27.54	26.28 / 28.44	25.77 / 26.36
Visual MLLMs						
GPT-4o (2024)	**60.37 / 61.43**	35.00 / 31.14	33.14 / 34.57	51.81 / 55.71	59.34 / 55.43	**70.69 / 61.71**
Qwen2-VL (7B) (2024)	59.21 / 53.48	**41.67 / 33.18**	**34.88 / 39.44**	**62.90 / 65.08**	55.85 / 57.01	69.74 / 58.37
LLaVA-Video (7B) (2024)	54.19 / 55.77	39.05 / 30.62	24.83 / 38.17	62.83 / 62.53	**61.87 / 58.21**	62.94 / 60.04
InternVL2 (76B) (2024)	53.15 / 45.49	25.83 / 26.83	30.23 / 26.89	58.42 / 57.79	53.18 / 50.13	61.70 / 53.25
InternVL2 (8B) (2024)	44.52 / 37.70	30.00 / 25.24	29.07 / 26.58	49.25 / 53.54	47.43 / 43.38	53.19 / 43.69
VILA-1.5 (8B) (2024)	41.72 / 38.28	28.33 / 26.78	28.49 / 36.61	53.30 / 62.38	46.41 / 47.77	46.34 / 46.64
VideoLLaVA (7B) (2023)	34.73 / 27.97	15.00 / 19.10	22.67 / 17.30	36.67 / 42.36	34.09 / 35.86	35.93 / 29.69
Audio MLLM						
SALMONN (13B) (2024)	45.92 / 45.18	33.33 / 31.68	29.65 / 31.56	32.62 / 32.70	36.76 / 33.11	34.52 / 38.39

Table 3: Task Results (Multiple-Choice Questions). This table presents the performance of each model on the multiple-choice question tasks. AIE: Audio Information Extraction, ACC: Audio Content Counting, AEL: Audio Event Localization, AVCM: Audio-Visual Character Matching, AVOM: Audio-Visual Object Matching, AVTM: Audio-Visual Text Matching. Results are presented as accuracy scores (%). Each cell shows the performance on AV-Human (left) and AV-Gemini (right) datasets, separated by a forward slash (/).

Models	AV-Human	AV-Gemini	Overall
Audio Visual MLLMs			
Gemini 1.5 Pro (2024)	78.34	75.21	75.67
VideoLLaMA2 (7B) (2024)	44.90	39.75	40.56
video-SALMONN (13B) (2024)	38.33	34.08	34.74
PandaGPT (13B) (2023)	25.38	28.83	28.31
Visual MLLMs			
GPT-4o (2024)	56.62	50.00[1]	53.31
Qwen2-VL (7B) (2024)	58.38	51.05	52.26
LLaVA-Video (7B) (2024)	56.52	50.62	51.48
InternVL2 (76B) (2024)	52.62	43.29	44.72
InternVL2 (8B) (2024)	45.90	38.31	39.47
VILA-1.5 (8B) (2024)	44.48	43.05	43.27
VideoLLaVA (7B) (2023)	33.14	28.70	29.37
Audio MLLM			
SALMONN (13B) (2024)	36.48	35.43	35.59

Table 4: Overall Performance on AVUT. This table presents the overall performance of each model on the AV-Human, AV-Gemini, and Overall (combined AV-Human and AV-Gemini) datasets. Performance is measured by accuracy (%) on MCQs.

Modalities	Audio-Visual	Transcription	Visual-Only	Audio-Only
Accuracy (%)	78.34	69.23	62.78	54.10

Table 5: Gemini 1.5 Pro's performance when provided different modalities.

port this, we choose the most competitive audio-visual model, Gemini 1.5 Pro, to perform ablation studies as shown in Table 5. We compared the model's performance across several input configurations: full audio-visual, video-only, audio-only, and a detailed transcription setting. The "detailed transcription" includes speaker identification, corresponding ASR transcripts, and start/end timestamps for each utterance. This result demonstrated that our benchmark relies heavily on audio modality, as its performance drops significantly when missing any of the two modalities. While providing detailed transcriptions with speaker diarization and timestamps improves performance compared to video-only, it still falls short of the performance achieved with full audio. This suggests that beyond the transcribed words themselves, the rich information encoded in audio, such as intonation, prosody, and overlapping speech, is crucial for understanding the video content and is not fully captured even by detailed textual transcriptions. Transcription annotation examples are shown in Appendix K.

Most audio-visual models underperform visual-only models even on audio-visual alignment tasks. For weaker models, having audio-visual inputs may instead introduce more hallucinations, causing them to confuse visual-only and audio-only distractors, whereas video-only models only deal with visual distractors.

Open-ended QA tasks need stronger models. As shown in Table 6, although segmentation and diarisation are popular speech processing tasks, only Gemini 1.5 Pro achieved sensible results while other models failed to perform those tasks. This is partly due to the poor instruction-following abilities of audio-LLMs, which highlights an aspect that audio-visual LLMs need improvements on.

Models	AVSM		AVDiar
	Pair (%) ↑	Full (%) ↑	DWER (%) ↓
Gemini 1.5 Pro (2024)	72.84	26.29	66.31
VideoLLaMA2 (7B) (2024)	38.05	0.00	116.92
PandaGPT (13B) (2023)	39.87	1.22	155.94

Table 6: Open-Ended Task Results on Audio-Visual Segment Matching (AVSM) and Audio-Visual Speaker Diarization (AVDir). "Pair ↑" denotes the percentage of correctly sequenced segment pairs (not necessarily adjacent), "Full ↑" denotes the percentage of fully matched videos, "DWER ↓" represents Word Error Rate (WER) between ground truth and diarized transcriptions.

Models	GPT rank	Rouge-L	MCQ Acc
Audio Visual MLLMs			
Gemini 1.5 Pro (2024)	3.331	0.108	78.34
VideoLLaMA2 (7B) (2024)	1.839	0.062	44.90
video-SALMONN (13B) (2024)	2.032	0.080	38.33
PandaGPT (13B) (2023)	1.384	0.035	25.38
Visual MLLMs			
Qwen2-VL (7B) (2024)	2.233	0.074	58.38
LLaVA-Video (7B) (2024)	2.217	0.099	56.52
InternVL2 (76B) (2024)	2.022	0.053	52.62
InternVL2 (8B) (2024)	1.959	0.052	45.90
VILA-1.5 (8B) (2024)	2.079	0.074	44.48
VideoLLaVA (7B) (2023)	1.024	0.022	33.14
Audio MLLM			
SALMONN (13B) (2024)	1.758	0.053	36.48

Table 7: Open-ended QA test on AV-Human. "GPT rank" denotes the average ranking of the model's answers ranked by GPT-4o, "Rouge-L" denotes the average Rouge-L between model's answer and reference answer, and MCQ Acc stands for the model's accuracy(%) on MCQs.

MCQs are capable enough for revealing models' ability. We turned MCQs in AV-Human into open-ended questions, where the original correct option is used as the reference answer. The generated answer was ranked using GPT-4o and Rouge-L. Performance comparison between open-ended questions and MCQs in Table 7 showed that models performed similarly with different question formats. Despite differences from real-world scenarios, the MCQs still effectively reflect models' performance on audio-centric video understanding in real life.

5.3 Evaluating Text Shortcut on Benchmarks

To demonstrate the prevalence of text shortcut and the effectiveness of our mitigation strategy, we evaluated several video understanding benchmarks alongside AVUT. We used GPT-4o as our test model, assessing its performance in two conditions: (1) using the Text-Only Accuracy (only the question text is given) and (2) using the Cycled Accuracy metric (requiring correct answers across all four permutations).

Benchmarks	Text-Only Accuracy (%) ↓		Cycled Accuracy (%) ↓	
	GPT-4o	Gemini 1.5 Pro	GPT-4o	Gemini 1.5 Pro
VideoVista (2024c)	68.00	62.63	41.25	38.75
TemporalBench (2024)	65.42	59.37	73.57	52.98
NExT-QA (2021)	47.57	48.95	23.22	26.37
MLVU (2024)	42.96	40.98	19.38	12.81
MotionBench (2025)	41.24	36.37	16.17	11.86
Video-MME (2024a)	41.22	45.22	16.03	13.05
MVBench (2024a)	32.24	39.03	10.84	5.70
Perception Test (2023)	29.20	32.67	4.68	6.02
AVUT	**25.32**	**29.61**	**1.89**	**4.50**

Table 8: Benchmark Text Shortcut Result. "Text-Only Accuracy" denotes accuracy with text input only. "Cycled Accuracy" denotes accuracy requiring consistent correct answers across four option permutations (ABCD, BCDA, CDAB, DABC). Text shortcut is tested with both GPT-4o and Gemini 1.5 Pro to avoid model bias.

As shown in Table 8, existing benchmarks suffer from text shortcut, with both GPT-4o and Gemini-1.5-pro achieving surprisingly high accuracy **without accessing any video**. In particular, VideoVista had a text-only accuracy of **68.00%** with GPT-4o, and the audio-visual benchmark, Video-MME had **41.22%**, which are all far above the 25% accuracy of random guessing. This highlights the ubiquity of the problem and the potential for misleading performance evaluations. In contrast, AVUT **approaching the expected random guessing accuracy** on four-option multiple-choice questions with text-only input on both GPT-4o and Gemini-1.5-pro. Furthermore, AVUT also achieves a cycled accuracy of **1.89%** with GPT-4o that is significantly lower than others by large margins, confirming the effectiveness of our filtering mechanism in mitigating text shortcut. As a result, AVUT provides a more reliable and rigorous assessment of the actual capabilities of video understanding.

6 Conclusion

In this paper, we present **AVUT**, a text-shortcut-free benchmark designed to advance audio-centric video understanding. AVUT provides a diverse suite of tasks focused on audio understanding and audio-visual alignment, supported by a meticulously curated dataset including both human-annotated and Gemini-augmented subsets. An extensive evaluation of models highlights the importance of audio information in complete video comprehension. We believe AVUT is valuable for the research community, fostering innovation and progress in audio-centric video understanding and paving the way for more robust and comprehensive multimodal learning systems.

Limitations

One major limitation is that our conclusion is only applicable to the current set of open-source audio-only and audio-visual LLMs being slightly outdated, which yields generally worse results than visual-only models due to their deficiency in incorporating audio. We are aware of contemporaneous audio-visual LLMs with implementations yet to be released. We will include their results in our benchmark as soon as they are available.

Another limitation stems from our focus on short videos. While this choice allowed us to effectively evaluate fine-grained audio-centric understanding with existing models, it limits the assessment of long-range temporal reasoning and durational video comprehension abilities. Future work will explore extending our benchmark to longer videos as models improve in these areas.

While we endeavor to ensure a wide range of annotators, it is inevitable that certain annotator bias still exists in data annotation, such as preference towards visual information and bias due to the first language. Moreover, same as many other MCQ-based benchmarks, the MCQ parts of AVUT only explored 4 option cases, leaving other numbers of choices unexplored.

References

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, and Peter Anderson. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2017. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing*.

Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. 2023a. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*.

Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. 2023b. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *Advances in Neural Information Processing Systems*, 36:72842–72866.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, Ethan He, Hongxu Yin, Pavlo Molchanov, Jan Kautz, Linxi Fan, Yuke Zhu, Yao Lu, and Song Han. 2024. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*.

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

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

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

Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. 2023. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. 2024. Longvale: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. *arXiv preprint arXiv:2411.19772*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. 2025. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. *arXiv preprint arXiv:2501.02955*.

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

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

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023a. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

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

Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. 2021. Value: A multi-task benchmark for video-and-language understanding evaluation. *arXiv preprint arXiv:2106.04632*.

Shicheng Li, Lei Li, Shuhuai Ren, Yuanxin Liu, Yi Liu, Rundong Gao, Xu Sun, and Lu Hou. 2023b. Vitatecs: A diagnostic dataset for temporal concept understanding of video-language models. *arXiv preprint arXiv:2311.17404*.

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, Song Chen, Xu Li, Da Pan, Shusen Zhang, Xin Wu, Zheng Liang, Jun Liu, Tao Zhang, Keer Lu, Yaqi Zhao, Yanjun Shen, Fan Yang, Kaicheng Yu, Tao Lin, Jianhua Xu, Zenan Zhou, and Weipeng Chen. 2024b. Baichuan-omni technical report. *arXiv preprint arXiv:2410.08565*.

Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. 2024c. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pretraining for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

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

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

Team OpenGVLab. 2024. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan, Banarse, Mateusz Malinowski, Yezhou Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine, Miech, Skanda Koppula, Alexander Fréchette, Hanna Klimczak, R. Koster, Junlin Zhang, Stephanie, Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. 2023. Perception test : A diagnostic benchmark for multimodal models. In *Conference on Neural Information Processing Systems*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

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

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. 2024. Video-SALMONN: Speech-enhanced audio-visual large language models. In *Proceedings of the International Conference on Machine Learning*.

Kim Sung-Bin, Oh Hyun-Bin, JungMok Lee, Arda Senocak, Joon Son Chung, and Tae-Hyun Oh. 2024. Avhbench: A cross-modal hallucination benchmark for audio-visual large language models. *arXiv preprint arXiv:2410.18325*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations*.

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

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

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. NExT-GPT: Any-to-any multimodal LLM. In *Proceedings of the International Conference on Machine Learning*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, and Wenwu Zhu. 2022. Avqa: A dataset for audio-visual question answering on videos. In *Proceedings of the 30th ACM international conference on multimedia*.

Qilang Ye, Zitong Yu, Rui Shao, Xinyu Xie, Philip Torr, and Xiaochun Cao. 2025. Cat: Enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios. In *European Conference on Computer Vision*.

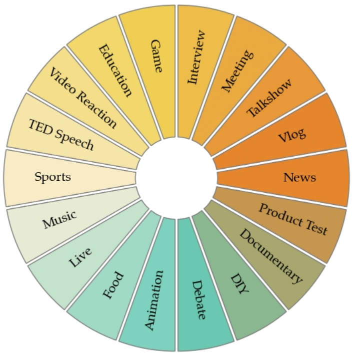


Figure 2: The 18 video domains of AVUT from which the videos are collected.

Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. 2021. Pano-avqa: Grounded audio-visual question answering on 360deg videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

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

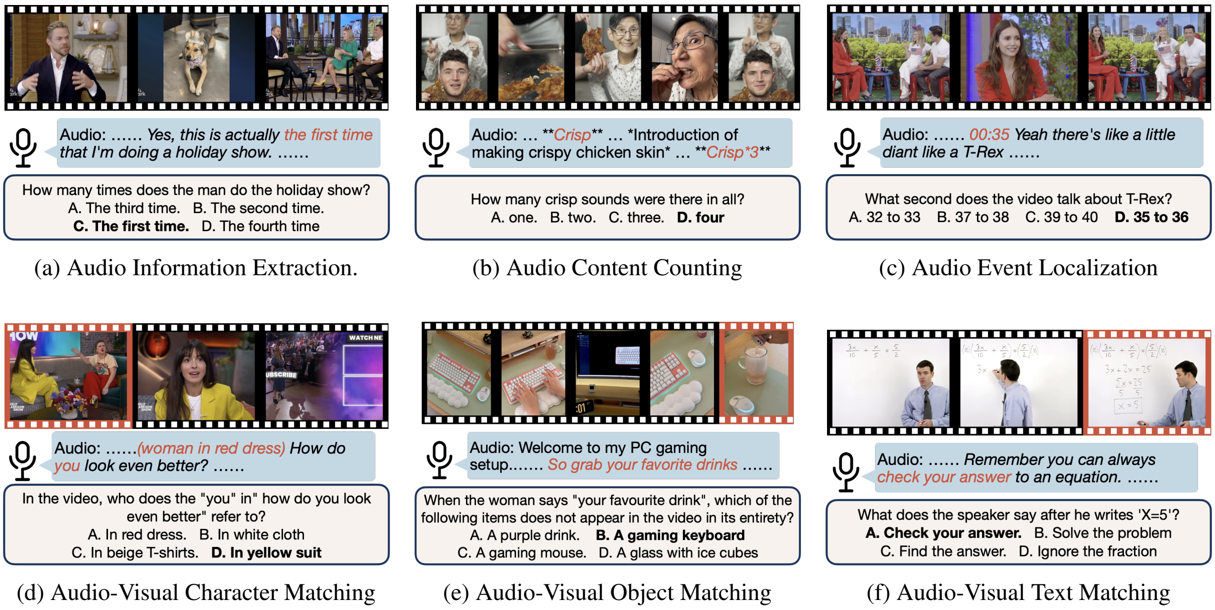
Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.

A Example Videos and Tasks

This section provides illustrative examples of the six multiple-choice question tasks in AVUT, the video data and corresponding task is in Figure 3 . Each figure showcases a video frame and the corresponding annotated question with answer options. And 18 audio-centric video domains are shown in Figure 2.

B Case Study on Audio Content Counting and Audio Event Localization Tasks

This section provides an illustrative and detailed case study for the task mentioned difficult for models in the result display section. Two cases on the Audio Content Counting task and one case on the Audio Event Localization task show how visual information is not helpful or even misleading in



Audio: ...... *Yes, this is actually the first time that I'm doing a holiday show.* ......

How many times does the man do the holiday show?
A. The third time. B. The second time.
C. The first time. D. The fourth time

(a) Audio Information Extraction.

Audio: ... ***Crisp*** ... *Introduction of making crispy chicken skin* ... ***Crisp*3***

How many crisp sounds were there in all?
A. one. B. two. C. three. **D. four**

(b) Audio Content Counting

Audio: ...... *00:35 Yeah there's like a little diant like a T-Rex* ......

What second does the video talk about T-Rex?
A. 32 to 33 B. 37 to 38 C. 39 to 40 **D. 35 to 36**

(c) Audio Event Localization

Audio: ......*(woman in red dress) How do you look even better?* ......

In the video, who does the "you" in" how do you look even better" refer to?
A. In red dress. B. In white cloth
C. In beige T-shirts. **D. In yellow suit**

(d) Audio-Visual Character Matching

Audio: Welcome to my PC gaming setup....... *So grab your favorite drinks* ......

When the woman says "your favourite drink", which of the following items does not appear in the video in its entirety?
A. A purple drink. **B. A gaming keyboard**
C. A gaming mouse. D. A glass with ice cubes

(e) Audio-Visual Object Matching

Audio: ...... *Remember you can always check your answer to an equation.* ......

What does the speaker say after he writes 'X=5'?
A. Check your answer. B. Solve the problem
C. Find the answer. D. Ignore the fraction

(f) Audio-Visual Text Matching

Figure 3: Six MCQ format task examples in AVUT.

these tasks, making them harder for models. Answers from outperforming visual-only models and audio-visual models are displayed and analysed for possible difficulties.

For the case shown in Figure 4a, the visual elements provide three relatively separate cues pointing to the word "AI," as shown in the first three example frames in the figure. The last instance includes a fast scene transition, where the visual cue pointing to AI appeared twice in sync with the audio mentions of "AI" and "it." Additionally, when "AI" is mentioned for the last time in the audio, there is no corresponding visual cue, as shown in the final example frame.

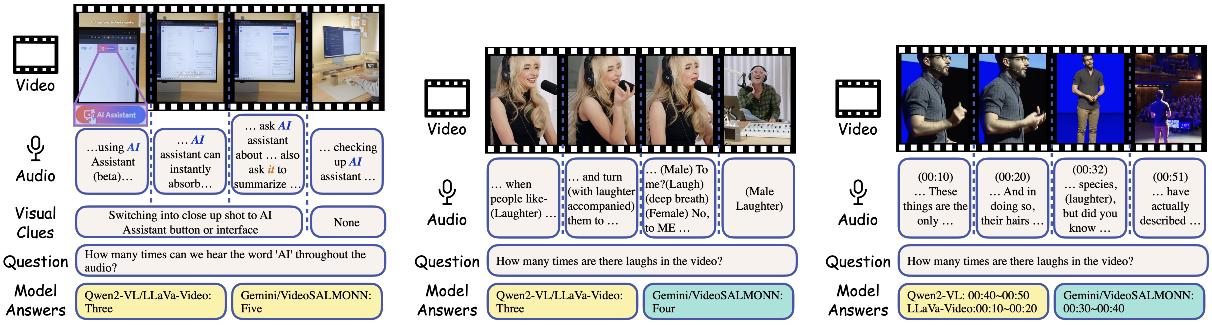
As a result, it can be observed that the best-performing visual models provided answers "three" corresponding to the visual cues. In contrast, the audio-visual models produced the answer "five", possibly because the strong visual cues and semantic associations in the earlier layers linked the meaning of "AI" to the representation of "it."

For the case shown in Figure 4b, the camera only captured three laughters among all four laughters heard in the audio, as the third laughter from a male interviewer was not captured by the camera, which focused on the female interviewee at the time, presented at the third displayed frame. As a result, models that captures only the visual clues would report three laughters, missing the ones that don't have the camera focused on the laughing characters. Models that can hear, like Gemini 1.5 Pro or video-SALMONN, would be more likely to answer this question correctly.

From the examples above, it can be seen that although the counted audio content generally aligns with the main theme of the video and sometimes corresponds to the visual content, such correspondence provides little help to the audio content counting task. Due to the nature of the counting task, the partial matching does not provide "partial assistance", but instead leads to incorrect answers. This makes the models benefit less from the natural, imperfect audio-visual consistency, thereby placing essential demands on the model's audio understanding capabilities.

For Figure 4c, the camera does not focus on the subject of "Laugh" (which is the audience) at the corresponding moment, as shown in the third frame. As a result, the visual model can only rely on guessing strategies, such as attempting to align with the closest appearance of the audience (like Qwen2-VL), capturing hallucinations, or making random guesses. Only models that really heard the relevant audio are able to locate the event correctly.

The example above illustrates that in the audio event localization task, the audio events involved may be relatively insignificant to the video's theme. As a result, video producers often do not ensure their alignment with visual elements, creating more cases where visual clues alone is insufficient. This further raises the demands on the model's audio understanding capabilities.


Video
Audio
Visual Clues
Question
Model Answers
AI Assistant
…using AI Assistant (beta)…
… AI assistant can instantly absorb…
… ask AI assistant about … also ask it to summarize …
… checking up AI assistant …
Switching into close up shot to AI Assistant button or interface
None
How many times can we hear the word 'AI' throughout the audio?
Qwen2-VL/LLaVa-Video: Three
Gemini/VideoSALMONN: Five
… when people like- (Laughter) …
… and turn (with laughter accompanied) them to …
… (Male) To me?(Laugh) (deep breath) (Female) No, to ME …
(Male Laughter)
How many times are there laughs in the video?
Qwen2-VL/LLaVa-Video: Three
Gemini/VideoSALMONN: Four
(00:10) … These things are the only …
(00:20) … And in doing so, their hairs …
(00:32) … species, (laughter), but did you know …
(00:51) … have actually described …
How many times are there laughs in the video?
Qwen2-VL: 00:40~00:50 LLaVa-Video:00:10~00:20
Gemini/VideoSALMONN: 00:30~00:40

(a) When visual elements misleads models in Audio Content Counting.

(b) When visual information doesn't help in Audio Content Counting.

(c) When visual information doesn't help in Audio Event Localization.

Figure 4: Three cases on models' under-performed tasks.

C Annotation Information

Our labelers are provided by a professional outsourcing labeling company, and all of our labelers have bachelor's degree or above and certificates of related professions. For the payment of the labeled person, the labeled person's income is higher than the local minimum income.

C.1 Annotator Skill Acquisition

Annotator selection and training were carefully managed to ensure high-quality annotations. Initial screening considered task-specific requirements, matching annotators' skills and experience to appropriate tasks. For example, annotators with sports knowledge were assigned to sports commentary videos, and those with strong listening skills were assigned to audio transcription tasks. This initial optimization aimed to maximize data quality by aligning expertise with task requirements. Following selection, dedicated managers conducted online training sessions, incorporating insights from pilot annotations to address potential challenges and difficulties. Training included detailed explanations of task guidelines and illustrative examples to facilitate annotator comprehension. During the formal annotation phase, managers continued to collaborate with annotators, providing feedback and corrections on initial annotations to further refine their understanding and skills. Additionally, annotators were encouraged to obtain relevant certifications to further enhance their expertise. This multi-faceted approach ensured annotators possessed the necessary skills and knowledge for producing high-quality annotations.

C.2 Task Allocation

Task allocation was carefully managed to ensure efficient and balanced workload distribution. The total workload was initially assessed and divided into smaller batches based on the project deadline and the number of available annotators. Tasks were then assigned based on data complexity and annotator skill level. For demanding tasks, such as turn-taking analysis requiring high listening comprehension and repeated video review, batch sizes were adjusted to maintain fairness and prevent annotator overload. The allocation process also included dynamic adjustments to incentivize participation and address workload fluctuations. Incentives, such as increased per-item rates and bonuses for additional batches or overtime work, were implemented to motivate annotators and ensure timely completion. A "task-grabbing" mechanism was employed, allowing annotators to claim new batches upon successful completion of previous ones. This prevented individual annotators from taking on excessive workloads and ensured consistent progress. Annotator feedback was actively solicited and incorporated into the allocation strategy. For example, batches identified as unusually difficult were further subdivided and distributed among more annotators. This flexible and responsive approach ensured timely project delivery and consistent annotation quality.

C.3 Quality Control Procedures

A rigorous, multi-stage quality control process, involving annotators, quality control specialists, and project managers, was implemented to ensure high-quality annotations. Annotators performed an initial annotation, actively communicating challenges and ambiguities to project managers for clarification and guidance. Unresolved issues were escalated to the researchers for further clarification. Upon completion of the initial annotation phase, dedicated quality control specialists conducted a thorough review, employing methods such as full review, sampling, or cross-validation, as appropriate for the specific task. Annotators then revised their work based on the specialists' feedback. A second round of quality control followed, with any remaining discrepancies adjudicated by a third-party reviewer to ensure final accuracy and consistency. Throughout this process, quality control adhered strictly to the provided guidelines, with specific criteria tailored to each task type (e.g., video retrieval, speech recognition, text annotation). Quality control specialists were selected based on their expertise and skills relevant to the specific task, further ensuring high-quality and consistent results. This meticulous approach guaranteed the accuracy, reliability, and overall quality of the final annotations.

D Video Collection Process

Annotators were tasked with collecting 2700 YouTube videos for each of 18 specified video domains. The following criteria were applied during video collection:

- **Audio Content.** Videos need to contain rich audio elements, such as diverse sounds (human speech, environmental sounds, artificial sounds) or substantial spoken dialogue.
- **Language.** Videos were required to be in English.
- **Duration.** Videos were limited to a maximum of two minutes. For longer videos, annotators provided start and end points for a relevant segment within the two-minute limit.
- **Publication Date.** Videos were required to be published in 2023 or later, with preference given to more recent videos (ideally from 2024) to minimize overlap with model training data.
- **Subtitles.** Videos could not contain manually added subtitles or captions.
- **Video Domains.** The 18 target video domains were: Game, Interview, Meeting, Talkshow, Vlog, News, Product Test, Documentary, DIY, Debate, Animation, Food, Live, Music, Sports, TED Speech, Video Reaction, and Education. Specific definitions and examples were provided for each video domain to ensure consistency.

E Annotation Guidelines and Examples

This section details the annotation process for creating question-answer pairs for the AV-Human. Annotators were provided with 700 videos and tasked with creating three question-answer pairs per video, totaling 1,734 annotations. The annotation process followed these steps:

(1) Video Comprehension: Annotators thoroughly watched each video to understand its content.

(2) Task Selection: Annotators reviewed the six multiple-choice format task types (listed below) and selected the three most relevant tasks for each video. Relevance was determined by whether the video's content could be used to create a meaningful question within the chosen task type.

(3) Question-Answer Design: Annotators designed a multiple-choice question and four answer options (A, B, C, and D) for each selected task type, based on the video content and the specific task guidelines. Only one answer option could be correct.

(4) Rationality Check: Annotators performed a rationality check on each question-answer pair to ensure it met the following criteria:

- **Audio Dependence.** The question could not be answered correctly by watching the video without audio.
- **Clear Distinctions.** The correct answer needed to be clearly distinguishable from the incorrect options, avoiding ambiguity.
- **No External Knowledge.** Questions and options could not rely on common sense knowledge or information beyond the video's content.
- **Data Formatting.** Annotations were recorded in an Excel spreadsheet with the following columns: video id, URL, task type, question, option A, option B, option C, option D, and answer.

To aid annotators in understanding and applying each task type, we provided two annotated examples with accompanying videos for each task. These examples served as templates and illustrated best practices for question-answer design. The task definitions provided for annotators are as follows:

- **Audio Information Extraction.** This task focuses on extracting specific details conveyed through spoken language in the video. Questions should be direct, clear, and objective, targeting factual information from the audio. Answer options should be relevant, concise, and clearly distinguishable, with distractor options designed to assess the depth of audio comprehension.

- **Audio Content Counting.** This task requires counting the occurrences of specific audio events within the video. Questions should clearly define the target event and the scope of counting (e.g., the entire video or a specific segment). Answer options should provide a reasonable range of numerical choices with appropriate intervals. Distractor options should be numerically close to the correct answer to test counting precision.
- **Audio Event Localization.** This task involves pinpointing the precise time a specific audio event occurs. Questions should clearly identify the target event and request its temporal location within the video. Answer options should offer distinct time ranges with reasonable granularity (seconds precision), using the format "time1 to time2". Distractor options should test the model's ability to precisely locate the event.
- **Audio-Visual Character Matching.** This task requires matching spoken information from a character to their visual appearance. Questions should link a character's utterance to their visual characteristics. Answer options should describe different characters or visual features present in the video. Distractor options should include other characters or similar visual features to test the model's matching accuracy.
- **Audio-Visual Object Matching.** This task focuses on matching spoken information to objects or scenes within the video. Questions should link spoken descriptions or references to the corresponding visual elements. Answer options should describe different objects or scenes present in the video. Distractor options should include other objects or similar scenes to evaluate the model's ability to connect audio and visual elements.
- **Audio-Visual Text Matching.** This task involves matching spoken information to on-screen text (OCR). Questions should connect spoken words or phrases to specific text appearing in the video, often specifying the timing and location of the text. Answer options should offer a variety of OCR text strings or combinations. Distractor options should include similar-looking text or text from different locations to test the model's ability to precisely match audio and text.

F Gemini-Generated Annotation Review Process

This section details the human review process for the 9,875 Gemini-generated question-answer pairs. Annotators were tasked with ensuring the generated content met specific quality criteria, focusing on question relevance, option validity, answer accuracy, and adherence to video content without reliance on external knowledge or common sense reasoning. The review process for each question-answer pair involved the following steps:

(1) Video Review: Annotators watched the entire video or the specified segment (for longer videos with provided start and end points) to fully understand the content.

(2) Task Type Verification: Annotators verified that the assigned task type aligned with the video content. The definitions of the six task types (Audio Character Matching, Audio Object Matching, Audio OCR Matching, Audio Event Location, Audio Content Counting, and Audio Information Extraction) were provided to the annotators (see Section E for definitions).

(3) Question Evaluation: Annotators assessed the question's relevance to both the task type and the video content, ensuring it could not be answered solely through text reading or common sense reasoning.

(4) Option Evaluation: Annotators checked the validity of all four answer options.

Our human review stage was critical for mitigating biases and ensuring annotation quality. Approximately 40% of the 9,875 generated QA pairs required modification, ranging from minor rephrasing to complete rewriting. Common categories of issues encountered during the review process are detailed below, along with illustrative examples:

- **Task Type Mismatch (20%):** Gemini generated questions that did not align with the intended task type. For example:
  - Intended Task Type: Audio Visual Character Matching (requiring matching characters based on both audio and visual information)
  - Gemini-Generated Question: "`What is the narrator wearing?`", `Options: A) Red shirt, B) Black suit, C) Red suit, D) Yellow suit`
  - Gemini-Generated Answer: `D`

- Explanation of Error: The “Audio Visual Character Matching” task requires identifying a character based on both their audio and visual presence. However, in this video, the narrator is only heard and never seen on screen. Thus, the generated question about the narrator’s visual appearance is inherently flawed and incompatible with both the task’s requirements and the video content itself.

- **Irrelevant to Video Content (15%):** Questions were generated that did not pertain to the content of the video. For example:
  - Intended Task Type: Audio Visual Object Matching
  - Gemini-Generated Question: “`What is the man talking about when he says, ‘This car is 10 years old and you would not know that by looking at it’?`”, `Options: A) A dryer, B) Clothes, C) A car, D) Plastic wrap`
  - Gemini-Generated Answer: `C`
  - Explanation of Error: The generated question can be answered solely by interpreting the provided quote. The video content is entirely superfluous.

- **Multiple Correct Answers (15%):** Questions were formulated in a way that allowed multiple valid answers. For example:
  - Intended Task Type: Audio Event Localization
  - Gemini-Generated Question: “`When does the speaker say the word ‘pumpkin’?`”, `Options: A) 0:00-0:05, B) 0:05-0:10, C) 0:10-0:15, D) 0:15-0:20`
  - Gemini-Generated Answer: `D`
  - Explanation of Error: The speaker says "pumpkin" in both time ranges C (0:10-0:15) and D (0:15-0:20). Thus, the question as initially formulated has two correct answers.

- **Incorrect Correct Answer (40%):** The answer provided by Gemini as "correct" was factually incorrect. For example:
  - Intended Task Type: Audio Content Counting
  - Gemini-Generated Question: “`How many times does the phone vibrate in the video?`”, `Options: A) 1, B) 2, C) 3, D) 4`
  - Gemini-Generated Answer: `A`
  - Explanation of Error: Gemini’s provided answer (“A: 1”) is factually incorrect. The phone does not vibrate in the video.

- **Other Issues (10%):** This category encompasses issues such as overly simplistic questions, lack of distractor quality, generated gibberish, etc.

G Data Availability and Copyright

Our benchmark AVUT is constructed using publicly available videos from YouTube. We adhere to YouTube’s copyright policies and terms of service. AVUT contains only links to these public videos; we do not host or distribute copies of the video content itself. This ensures compliance with copyright regulations and allows for easy access to the data while respecting intellectual property rights. Users of AVUT are responsible for complying with YouTube’s terms of service when accessing the linked videos.

H Use of Gemini 1.5 Pro and GPT-4o

This section clarifies the use of Gemini 1.5 Pro and GPT-4o within our research and our adherence to their respective terms of service.[3]Our use of these models is strictly for non-commercial, research purposes and does not involve any commercial applications or profit generation.

Gemini 1.5 Pro: Gemini 1.5 Pro was employed solely for data augmentation purposes, specifically to generate a larger set of question-answer pairs for our benchmark. In accordance with Google’s terms of service, we acknowledge that Google retains all rights to generate similar content for other users. We do not claim ownership over the Gemini-generated content. We also acknowledge our responsibility for the use of this generated content, both within our research and by anyone with whom we share the benchmark data. While we have used discretion in relying on the generated content, users of AVUT should be aware of its automated origin. As required by the API Terms, we will comply

[3] Gemini Terms of Service: `https://ai.google.dev/gemini-api/terms#use-generated`. GPT-4o Terms of Service: `https://openai.com/policies/business-terms/`.

with applicable laws regarding the use of generated content, including providing attribution when necessary.

GPT-4o: GPT-4o was utilized exclusively for evaluation purposes, specifically to assess the performance of various models on our benchmark. Our use of GPT-4o for evaluation is in compliance with its terms of service.

I Model Prompting Details

This section provides the specific prompt template used in the experiments.
Prompt 1 is used for evaluating models on the multiple-choice question tasks in AVUT. This prompt was provided to each model.
Prompt 2 is used for evaluating models on the open-ended question tasks in AV-Human. This prompt was provided to each model.
Prompt 3 is used for prompting GPT-4o to provide a ranking to the answer that models generated for questions in AV-Human. This prompt was provided to GPT-4o to create a ranking for each answer.

J Annotator Biases

While we implemented rigorous quality control measures, it's important to acknowledge potential biases in human-generated annotations. Analysis of the AV-Human revealed instances where annotations exhibited a stronger reliance on visual information than intended. This section presents a few examples to illustrate this bias and highlight the challenges of achieving truly audio-centric annotation. These examples underscore the importance of careful annotation guidelines, thorough training, and robust quality control processes to mitigate such biases. Furthermore, they motivate the use of semi-automatic annotation methods, such as our Gemini-based approach, to further diversify the data and reduce the impact of human biases.

```
{
    "video_id": 1,
    "url": "https://www.youtube.com/
        shorts",
    "video_type": "vlog",
    "task_type": "Audio Event
        Localization",
    "question": "During which time
        period, the video does not
        record the front face of the
        woman in the plaid shirt?",
    "option_A": "0s-15s.",
    "option_B": "16s-30s.",
    "option_C": "31s-45s.",
    "option_D": "46s-60s.",
    "answer": "C"
},
{
    "video_id": 2,
    "url": "https://www.youtube.com/
        watch",
    "video_type": "animation",
    "task_type": "Audio-Visual Object
        Matching",
    "question": "What is on the boy's
        cloth?",
    "option_A": "A boy.",
    "option_B": "An animal.",
    "option_C": "A girl.",
    "option_D": "A creature like fox.",
    "answer": "D"
},
{
    "video_id": 3,
    "url": "https://www.youtube.com/
        watch",
    "video_type": "education",
    "task_type": "Audio-Visual Object
        Matching",
    "question": "What scene does the
        video show at the beginning?",
    "option_A": "Home.",
    "option_B": "Shop.",
    "option_C": "Supermarket.",
    "option_D": "Theatre.",
    "answer": "C"
},
{
    "video_id": 4,
    "url": "https://www.youtube.com/
        shorts",
    "video_type": "music",
    "task_type": "Audio-Visual Object
        Matching",
    "question": "What color bass does
        the guy who plays the bass at
        the end of the video have?",
    "option_A": "Yellow and pink.",
    "option_B": "Black and white.",
    "option_C": "Grey and pink.",
    "option_D": "Pink and green.",
    "answer": "B"
}
```

K Transcription Generation with Gemini

This section provides examples of the detailed transcriptions generated using Gemini 1.5 Pro for AVUT. These transcriptions include speaker identification, corresponding ASR transcripts, and start/end timestamps for each utterance. The inclusion of detailed transcriptions allows us to investigate the extent to which textual representations of audio can substitute for the original audio in video understanding tasks.

```
[
    {
        "start_time": 0.0,
        "end_time": 2.0,
```

Algorithm 1 Unified prompt for testing.

```
Select the best answer to the following multiple-choice question based on the video.
data["question"]

A. data["option_A"]
B. data["option_B"]
C. data["option_C"]
D. data["option_D"]

Respond with only the letter (A, B, C, or D) of the correct option. The best answer is:
```

Algorithm 2 Unified prompt for open-ended ablation.

```
Answer the question based on the video: data["question"]
```

```
        "speaker": "Speaker 1",
        "characteristic": "female with
            long brown hair, wearing a
            blue dress with flowers",
        "transcript": "I heard that you
            brought your parents to the
            Oscars was that fun?"
    },
    {
        "start_time": 2.0,
        "end_time": 6.0,
        "speaker": "Speaker 2",
        "characteristic": "female with
            blonde hair, wearing a blue
            jacket with flowers",
        "transcript": "I did I did my
            dad actually called me and
            sort of invited himself."
    }
]
```

Listing 1: Example Transcription

Algorithm 3 Unified prompt for GPT-4o to rank the open-ended answer.

```
You are a helpful assistant. Given the question, predicted answer and the reference answer, rank the
     predicted answer from 1 to 5 by how similar it is compared with the reference answer. 1 means
     the predicted answer has nothing to do with the reference answer or conflicts with the reference
      answer, and 5 means the predicted answer and the reference answer expresses the same thing.
     Provide only the number.
    The question is: question
    The predicted answer is: model_answer
    The reference answer is: reference_answer
    Your ranking is:
```