Can Vision-Language Models Answer Face to Face Questions in the Real-World?

Reza Pourreza*[1], Rishit Dagli*†[2], Apratim Bhattacharyya[1], Sunny Panchal[1],
Guillaume Berger[1], Roland Memisevic[1]
[1] Qualcomm AI Research‡ , [2] University of Toronto
{pourreza,aprabhat,sunnpanc,guilberg,rmemisev}@qti.qualcomm.com, rishit@cs.toronto.edu

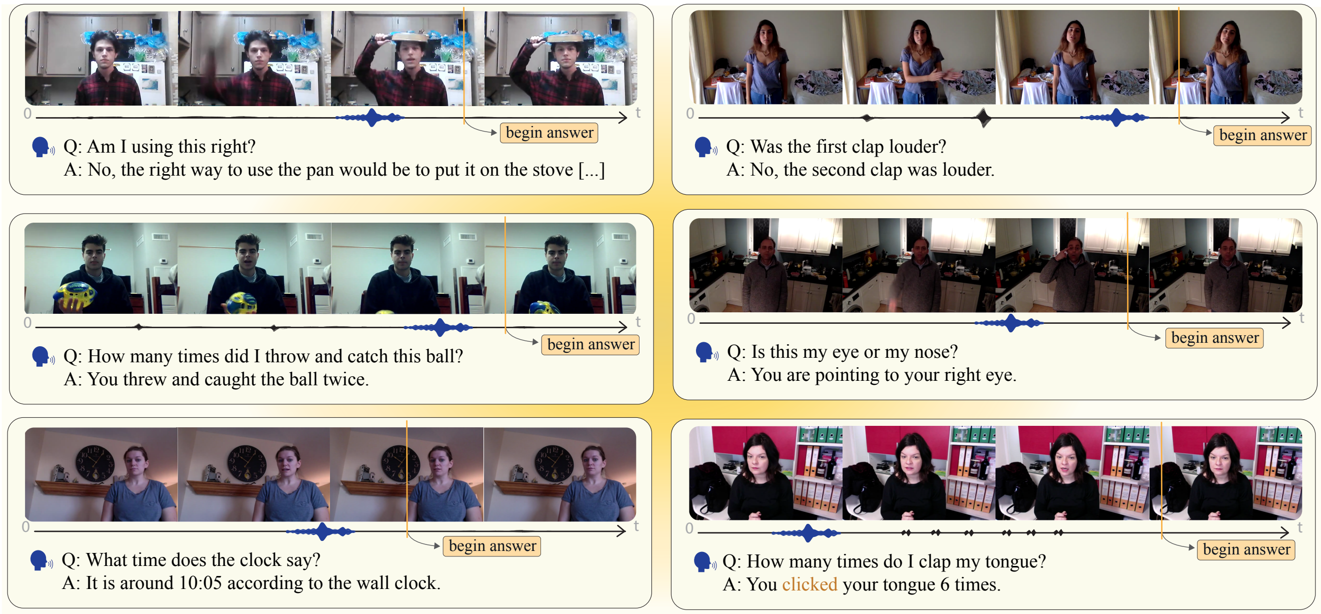

0
t
begin answer
Q: Am I using this right?
A: No, the right way to use the pan would be to put it on the stove [...]
Q: Was the first clap louder?
A: No, the second clap was louder.
Q: How many times did I throw and catch this ball?
A: You threw and caught the ball twice.
Q: Is this my eye or my nose?
A: You are pointing to your right eye.
Q: What time does the clock say?
A: It is around 10:05 according to the wall clock.
Q: How many times do I clap my tongue?
A: You clicked your tongue 6 times.

Figure 1. We present Qualcomm Interactive Video Dataset (IVD), a dataset collected in an online question-answering setup, where users pose open-ended questions using their camera and microphone. Qualcomm IVD offers videos with raw audio, annotated textual transcriptions of the spoken questions, and text answers with annotated timestamps. These timestamps indicate when a question can be sensibly answered given the video context. Qualcomm IVD serves as a realistic and challenging dataset for situated visual reasoning in Large Multi-modal Models.

Abstract

AI models have made significant strides in recent years in their ability to describe and answer questions about real-world images. They have also made progress in the ability to converse with users in real-time using audio input. This raises the question: have we reached the point where AI models, connected to a camera and microphone, can converse with users in real-time about scenes and events that are unfolding live in front of the camera? This has been a long-standing goal in AI and is a prerequisite for real-world AI assistants and humanoid robots to interact with humans in everyday situations. In this work, we introduce a new dataset and benchmark, the Qualcomm Interactive Video Dataset (IVD), which allows us to assess the extent to which existing models can support these abilities, and to what degree these capabilities can be instilled through fine-tuning. The dataset is based on a simple question-answering setup, where users ask questions that the system has to answer, in real-time, based on the camera and audio input. We show that existing models fall far behind human performance on this task, and we identify the main sources for the performance gap. However, we also show that for many of the required perceptual skills, fine-tuning on this form of data can significantly reduce this gap.

---

[*] Equal contribution.

[†] Work completed during internship at Qualcomm AI Research.

[‡] Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

[‡] CVPR 2025 competition version of Qualcomm IVD is available here. Additional labels will be released at a later date.

1. Introduction

Recent advancements in Large Multimodal Models (LMM) have significantly enhanced the ability of AI systems to interact naturally and fluently with users in real-time. Existing AI agents can process audio, speech, and visual inputs to engage in conversations about images or videos. However, the conversational capabilities of state-of-the-art LMMs such as GPT-4o [15] are limited to question answering on visual understanding and reasoning tasks, such as describing images or answering questions that require inferring object positions and relations in the visual input. These systems often fail to provide truly situated, live, conversational experiences (Fig. 1) that we may expect from humanoid robots or real-time video-call chatbots in the future.

We hypothesize that this limitation stems from the fact that current vision-language datasets and benchmarks are biased toward offline reasoning about images and videos. That is, the models receive the entire visual input and the entire question at once before being required to provide an answer. This is because the training data for such tasks can be easily sourced on the internet or easily generated through automated pipelines. There is a distinct lack of benchmarks and datasets that test genuine, real-time, "face-to-face" conversational skills. A separate but related problem is that models are not trained to respond at the appropriate time in a conversation – knowing "when to say" is crucial for conducting real-world conversations, yet this timing skill remains underdeveloped and understudied in current benchmarks.

To address these challenges and assess the limitations of existing models, we introduce the Qualcomm Interactive Video Dataset (IVD), a new dataset and benchmark designed for end-to-end trained systems aimed at real-time user interaction. Qualcomm IVD is structured as an online question-answering setup, where users pose open-ended questions using their camera and microphone, and the system must respond appropriately. Our work differs fundamentally from other related datasets and benchmarks by introducing an entirely online question-answering paradigm where both questions and answers evolve in real-time as the video unfolds, requiring models to maintain contextual awareness while handling inherent ambiguities in human references to visual elements. We show how this simple type of interaction allows us to capture a rich set of visual concepts that fall under the umbrella of situated visual understanding, including deictic (referring) expressions, pointing gestures, object ambiguities, behavior and action understanding, and counting, as well as audio-visual concepts. An overview of our dataset is shown in Figure 1. Due to the in-the-wild nature of the recordings, the videos exhibit considerable variation in lighting conditions, background settings, the range and nature of questions posed, actions performed by subjects, and other audio-visual characteristics.

To showcase the unique challenges our dataset presents, we conduct a series of experiments where we evaluate multiple open and closed-source state-of-the-art models, and fine-tuned models on our dataset. Our experiments reveal that the seemingly simple interaction of answering questions live, in real-time, is highly challenging for existing AI systems [15], even if they are otherwise good at performing visual reasoning. Our experiments indicate that the failure modes of existing systems can be attributed to: (1) their inability to answer questions whose answers require situational common sense, (2) their difficulty integrating visual and auditory information in real-time to disambiguate questions, and (3) their inability to determine the appropriate time at which to answer. Our dataset supports research on online LMMs capable of situated audio-visual reasoning, and can be leveraged to build conversational agents that interact with users in real-time.

Our contributions are summarized as follows:

- We introduce Qualcomm IVD, a novel multi-modal dataset designed to evaluate online situated audio-visual reasoning and real-time conversational skills.
- We benchmark existing LMMs and identify critical weaknesses in their ability to handle real-life conversations.
- We demonstrate that these limitations can be effectively mitigated by fine-tuning models on appropriate audio-visual conversational data.
- We develop a simple yet effective baseline to process streaming audio-visual inputs, departing from traditional offline paradigms.

2. Related Work

Offline Video Evaluation Benchmarks: Prior work on video understanding benchmarks has primarily focused on offline evaluation paradigms. There have been multiple temporal video understanding benchmarks for open-domain understanding [14, 20, 26, 31, 34, 42], hand movements [11, 30], articulated motion [6], full human body motion [32], robotics [1, 3, 12, 13, 16, 44], and embodied reasoning [43]. These benchmarks evaluate models' ability to comprehend temporal relationships but operate in a fully offline manner. Long-form video understanding has been addressed by datasets such as LVBench [39], and MoVQA [47], which extend the context window but fail to simulate real-time constraints. In contrast, our Qualcomm IVD dataset and benchmark focuses on real-world questions answering.

Situated Video Evaluation Benchmarks: Situated question answering has also been studied by [8, 27] and follow-up works (*e.g.*, [38, 40, 41]). A separate line of work has studied "common sense" situational understanding for AI models, albeit not a VQA format. This includes the work by

[9, 11, 35] and recent work on situated live dialogue [2, 32]. Our work is similar in that it involves real-world interaction. In contrast to the existing work, questions in our dataset are free-form and open-ended rather than task-specific and oriented towards a specific goal.

In contrast to existing question-answering tasks, the task introduced in our work involves real-world interaction with a user, and as such the input is not confined to only visual information. Moreover, we place the task into a truly situated context, where correct answers require a true understanding of the scene unfolding in the real world. In contrast to that line of work, in this paper, we study situated questions answering in a real-world not synthetic environment, by interacting "live" with a human subject, and by using audio and video input.

Online Models: Recent work on online video processing includes VideoLLM-online [4] and FlashVStream [48], which attempt to address real-time processing constraints but remain limited in their ability to handle deictic references and situated understanding and also do not include audio. The StreamVLM [32] supports situated understanding but is limited to the fitness domain and also lacks audio. Furthermore, existing benchmarks typically evaluate general visual understanding rather than modeling the situated, interactive nature of real-world human-AI conversations about visual content.

3. Qualcomm IVD

The purpose of the Qualcomm IVD dataset is to train and evaluate AI models on situated visual understanding. Each data instance comprises a video sequence annotated with temporally synchronized question-answer pairs. Furthermore, the dataset also includes the ground-truth answer to the question, making it possible to probe a model's understanding of the situation depicted in a given clip. Structuring the data as a simple question-answering task allows us to separate situated understanding from multi-hop conversational capabilities. The latter is a similarly difficult, but largely orthogonal, challenge for existing models.

3.1. Data Collection

Recording: crowd workers were instructed to record short videos using the camera and microphone of their mobile phone or laptop. They were free to choose the content of their videos but were shown examples featuring various gestures, actions, and objects to help them understand the dataset's purpose. The participants received written instructions explaining that these videos would be used to train and evaluate AI systems in understanding visual scenes. The instructions clarified that the AI system's purpose would be to correctly answer a single question rather than engage in a multi-step conversation. While recording their videos, crowd workers posed a question related to what was being shown. They were encouraged to be creative with their questions while ensuring they referenced the action or scene being recorded. After collection, all videos were inspected for audio and video quality, and their suitability for inclusion in the dataset.

Annotation Methodology: Each video in the Qualcomm IVD dataset has three annotations. First, it includes a human-generated transcript of the question asked during the recording. Second, we provide a human-generated answer to that question. Third, we added a timestamp that marks the specific moment when it would be appropriate to answer the question. The timestamp does not always coincide with the end of the spoken question–in many cases, additional video context is required after the question was asked. For example, if a participant asked "What action is this?" *before* performing the action, the appropriate moment to answer would be *after* the action was visible in the video. This approach ensures that annotations also reflect when sufficient information becomes available to answer the question correctly, if required, rather than simply when the question ends. Finally, all of our submissions were reviewed by humans to verify their accuracy.

Unlike datasets constructed from pre-recorded videos with post-hoc annotations, our contemporaneous question-asking approach places a strong demand on situational context understanding. Our videos capture authentic uncertainty about future events in the video, including questions that genuinely test temporal reasoning, and require situational awareness to answer at the appropriate time. The annotations for answer timing are particularly valuable as they acknowledge that certain queries require monitoring the audio or visual stream over time to aggregate relevant information, and ascertaining when to respond. Through our collection approach, we provide a robust benchmark for evaluating a model's proficiency in understanding and responding to situated audio-visual stimuli. We show a few examples from our dataset in Figure 1 using four frames per video.

3.2. Post-Processing Workflow

Following the initial data collection phase, we perform comprehensive post-processing to enhance dataset utility by adding structured metadata and further ensure dataset quality. This section details our approach to quality assurance and taxonomic categorization of the dataset.

Quality Checks: To ensure data quality and ethical standards, we used a multi-stage quality control process. Each video underwent automated evaluation followed by manual inspection by trained evaluators who assessed the content according to predefined exclusion criteria. Specifically, we examined all videos for the presence of 3rd persons, private data, and protected intellectual property; for the presence of inappropriate content, such as hate speech, and other potentially harmful elements; for linguistic compliance (clearly intelligible, English audio content); and for technical quality (absence of severe motion blur, compression artifacts, etc.).

After inspection, 2900 videos were deemed suitable and included in the dataset.

Semantic Categorization: To facilitate fine-grained analysis of model performance across different visual reasoning tasks, we developed a taxonomy of question types. The taxonomic structure allows for systematic evaluation of model performance across diverse visual reasoning tasks, allowing us to identify specific strengths and weaknesses in situated understanding capabilities. Each video-question pair was assigned to one or more of 13 predefined semantic categories representing distinct visual reasoning capabilities. The categorization process uses a semi-automated approach: first, a large language model (LLM) is used to perform preliminary classification based on question content and transcribed answers; next, human annotators verify and refine the categories. Our semantic taxonomy encompasses the following categories:

Action Attribute: Inquiries regarding the manner in which an action was performed, such as *Which hand did I use to wave?* or *How fast did I jump?*—tests ability to recognize fine-grained characteristics of dynamic events.

Action Counting: Questions about the frequency of an action's repetition, such as *How many times did I clap?*—evaluates temporal reasoning and event segmentation capabilities.

Action Detection: Identifying the specific action that was performed, such as *What am I doing right now?*—assesses basic activity recognition in dynamic scenes.

Action Understanding: Questions about the purpose or outcome of an action, such as *What does this gesture mean?* or *Why am I moving the chair?*—tests higher-level action interpretation and intention recognition.

Object Attributes: Inquiries about the characteristics of an object, such as *What color is this book?* or *Is this cup empty or full?*—evaluates fine-grained visual perception of static properties.

Object Counting: Determining the number of objects present, such as *How many pens are on the table?*—tests quantitative reasoning and object individuation.

Object Detection: Identifying an object within the scene, such as *Is there a lamp in this room?*—assesses basic object recognition capabilities.

Object Referencing: Indirectly pointing to an object within the scene, such as *What am I pointing at?* or *What is behind me?*—evaluates spatial reasoning and deictic reference resolution.

Object Understanding: Questions about the nature or function of an object, such as *What is this tool used for?*—tests semantic knowledge about objects beyond mere recognition.

Scene Understanding: Inquiries about the environment, such as *What room am I in?* or *Is it daytime or nighttime?*—evaluates holistic scene interpretation.

Audio-Visual: Questions that require audio information for a complete answer, such as *What sound am I making?* or *Am I speaking loudly or softly?*—tests cross-modal integration capabilities.

OCR: Extracting text from an object, such as *What does this sign say?*—evaluates the capability to recognize text in the real world and within the context of the conversation.

Subjective: Soliciting general opinions about an object or scene, such as *Does this outfit look good?*—tests a model's ability to respond sensibly to subjective questions.

Answer Normalization: To facilitate better quantitative evaluation and reduce ambiguity in model assessment, we implemented an answer normalization process. For each original free-form response, we generated a condensed "short-answer" version that retained only the essential information required to correctly address the question. We follow a similar semi-automated method as semantic categorization for generating short answers. During evaluation, we use both the short answer and the original ground truth to evaluate models.

3.3. Dataset Statistics

Dataset Composition: The Qualcomm IVD dataset consists of 2900 video clips and thus 2900 unique question-answer pairs. Table 1 summarizes the statistics of the dataset. The majority of clips have a length between 4 and 8 seconds. This range captures the natural timeframe in which a situated question about the visual scene can be posed and answered. We show the breakdown by the semantic taxonomy (Section 3.2) of the question-answer pairs in Table 2.

Temporal Characteristics: A distinctive feature of the Qualcomm IVD dataset is the temporal relationship between the point in time when a question is posed and the point in time when sufficient information is available to answer it. Figure 2 shows the distribution of answer timestamps relative to video end times. It is notable that there is significant variation in the optimal time to answer questions.

We further analyze the temporal characteristics by category in Table 2, which shows the distribution of optimal answer timestamps relative to video duration for each cat-

Statistic	Value
Total Videos	2900
Vocabulary Size (words)	3624
(tokens)	3072
Total Frames	443350
Average Video Length (s)	5.10 (±0.44)
Average Question Length (words)	6.09 (±1.94)
(tokens)	7.60 (±2.28)
Average Answer Length (words)	7.23 (±4.31)
(tokens)	9.73 (±5.61)
Average Short Answer Length (words)	1.38 (±0.82)
(tokens)	1.98 (±1.27)
Average Answer Timestamp (%)	81.47% (±13.89)
Average FPS	30 (±0.00)
Average Resolution (width)	640.00 (± 0.00)
(height)	382.29 (± 46.01)
Question Types (Total)	
Questions with "where"	47
Questions with "how"	512
Questions with "what"	1102
Deictic References (Total)	
Questions with "here"	32
Questions with "these"	39
Questions with "that"	45
Questions with "there"	105
Questions with "this"	568

Table 1. Dataset size metrics (total videos, vocabulary size), video characteristics (total frames, average length, frame rate, resolution), and linguistic properties of questions and answers. Average answer timestamps are represented by the average time in the video when the question should optimally be answered as a percentage of the video duration. The token statistics are calculated with the Llama-3 tokenizer. Standard deviations are shown in parentheses.

Category	Answer Timestamp	Samples
Action Attributes	84.31% (±13.56)	155 (5.34%)
Action Counting	92.22% (±8.73)	225 (7.76%)
Action Detection	85.46% (±13.22)	440 (15.17%)
Action Understanding	81.47% (±15.07)	110 (3.79%)
Object Attributes	79.52% (±13.41)	562 (19.38%)
Object Counting	78.41% (±12.75)	286 (9.86%)
Object Detection	76.95% (±15.65)	211 (7.28%)
Object Referencing	79.18% (±13.61)	706 (24.34%)
Object Understanding	80.63% (±14.07)	79 (2.72%)
Scene Understanding	79.91% (±13.58)	38 (1.31%)
Audio-Visual	90.09% (±11.49)	22 (0.76%)
OCR	83.04% (±13.08)	23 (0.79%)
Subjective	77.39% (±15.15)	43 (1.48%)
Total	81.47% (±13.89)	2900 (100%)

Table 2. Distribution of samples across the 13 semantic categories in our dataset, with the answer timestamp as a percentage of video duration for each category. Percentages in the Samples column show the relative distribution of categories within the dataset.

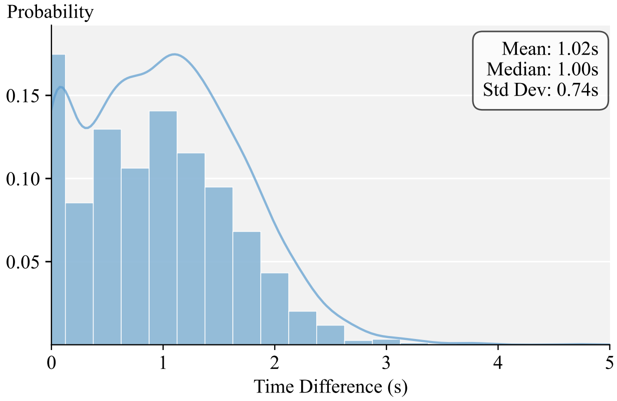


Figure 2. Temporal relationship between the end of the video and optimal answer timing. The horizontal axis represents seconds from the optimal time to answer to the the end of the video.

egory. As we can observe from Table 2, action-related categories generally require observing a larger portion of the video before answering, with Action Counting showing the highest optimal time (92.25% of video duration). This reflects the natural temporal dependency in action-related questions, where the answer often depends on observing the completion of an action sequence. In contrast, Object Detection (76.97% of video duration) and Subjective questions (77.39% of video duration) can typically be answered earlier in the video, often right after the question is asked.

4. Baseline Streaming Approach

Two critical features of Qualcomm IVD include:

Self-Contained Videos: The videos are self-contained, with the question embedded in the audio channel. An optimal model should be capable of answering these questions directly from the videos without the need for transcription.

When-to-Answer Desiderata: The videos are sufficiently long to include a scenario, a question, and any additional frames. An effective streaming model should identify the ideal moment to start answering the question, which is when both the question and any information necessary to answer it are present.

Current state-of-the-art LMMs do not integrate streaming and concurrent processing of audio and video information for situational interaction. To address this gap, we propose a novel streaming approach that combines a streaming automatic speech recognition (ASR) system to transcribe questions and detect answer moments, paired with a Video-LMM to analyze video content and provide answers.

In detail, our streaming approach relies on the Streaming-Whisper model [29] to identify "when to answer". The Streaming-Whisper model [29] uses the LocalAgreement algorithm [24] to transcribe text in a streaming setup. The LocalAgreement algorithm transcribes the input audio in chunks and a subset of previous chunks are used to condition the transcription of the next chunk. In practice, we found that a chunk size of 0.25 seconds is sufficient for accurate streaming transcription for this data. Processing the input audio in chunks allows us to detect the end of the question asked by the participant in the video. It is important to note that, as mentioned above, the end of a question does not necessarily capture the optimal moment for an answer, as some necessary information may arise later in the video. Thus, we consider this approach as a reasonable compromise given the current limitations of ASR solutions and LMMs. After the end of the question is detected, the input video and audio up to that timestamp, along with the transcribed question, are provided as input to the LMM backbone. The LMM backbone can then process the multi-modal video and audio inputs along with the transcribed question to provide an answer. We explore different LMM backbones as outlined in Section 5.1.

5. Experiments

We conduct comprehensive experiments to evaluate various open- and closed-source models on Qualcomm IVD.

5.1. Experiments Setup

Configurations: The experiments are performed within three distinct setups:

1. Streaming setup: Under this setup, we evaluate the baseline streaming approach introduced in Section 4.
2. Offline setup: In the baseline streaming approach, evaluating LMMs can be challenging due to potential inaccuracies in the questions extracted by the streaming ASR system, leading to accumulated errors. Therefore, in the offline setting, we use ground-truth questions to evaluate the models. This approach ensures that the evaluation is based on perfectly transcribed questions, allowing for an effective assessment of merely a model's answering performance. As a result, the resulting performance is an optimistic estimate of overall real-world performance.
3. Audio-visual models: Among existing LMMs, the VideoLLaMA family of models [46] are state-of-the-art models capable of simultaneously processing both audio and video content. Although these models cannot transcribe speech, they can utilize audio content as a complementary source of information, thereby potentially enhancing accuracy. We evaluate these models by examining the impact of additional audio on the accuracy of their question-answering capabilities.

Baseline Models: We experiment with various open-source and closed-source LMM backbone models.

The open-source models we evaluate include InstructBLIP (7B) [7], Video-ChatGPT (7B) [28], VideoChat (7B) [19], VideoChat2 (7B) [20], LLaVA-NeXT (7B) [25], LLaMA-VID (13B) [21], Video-LLaMA (13B) [46], VideoLLaMA2 (7B/72B) [5], VideoLLaMA2.1 (7B) [5], VideoLLaMA3 (7B) [45], Video-LLaVA (7B) [22, 51], Chat-UniVi (13B) [17], and Qwen2.5-VL (7B) [37]. The model sizes range from 7B to 13B parameters for the language backbone, with the exception of VideoLLaMA2-72B [5]. All models are evaluated in a zero-shot setting. We utilize the vision and audio heads provided with the checkpoints to process the input. For InstructBLIP [7], an image model, we sample 4 frames from each video, process these frames with the image encoder and a Q-Former [49] as individual images, and then treat all features as a long sequence of image tokens for the language model.

Additionally, we evaluate a closed-source model, GPT-4o [15], in a zero-shot fashion. Videos are preprocessed by uniformly selecting 4 frames from each video and downscaling the resolution to half. The query used to prompt GPT-4o is provided in the supplementary material.

Evaluation Metrics: Since the answers in Qualcomm IVD are in free-form, we determine the correctness of an answer using an LLM judge that receives a question, the ground-truth answer, and the predicted answer, alongside the short answer and the category of the question, and determines if the predicted answer is correct. We used a pre-trained Llama3-8B model [10] as the LLM judge. The prompts that were used are provided in the supplementary material. In addition, we report Bert [50], METEOR [18], BLEU [33], and ROUGE [23] scores between the ground-truth answers and the predicted answers.

5.2. Results

Here, we present the results obtained from the three settings described in Section 5.1.

Streaming setup: Table 3 presents the transcription results obtained from Whisper-Streaming [29], where the transcription quality is quantified using BLEU [33], ROUGE [23], and METEOR [18] scores, by comparing the transcribed questions to the ground-truth questions. The when-to-answer metric, denoted by Δt in the table, is measured as the Mean Absolute Error between the time-to-answer extracted by Whisper-Streaming and the ground-truth value. We also report the results obtained from the standard Whisper model [36] as an additional baseline. It is worth noting that this model does not return any timestamps alongside the transcriptions.

Model	METEOR ↑	BLEU ↑	ROUGE-L ↑	Δt ↓
Whisper [36]	90.01	80.95	90.32	-
Whisper-Streaming [29]	92.34	74.57	91.82	1.14

Table 3. **ASR performance comparison.** Evaluation of Automatic Speech Recognition (ASR) systems on the Qualcomm IVD dataset using standard text similarity metrics. The value Δt represents the mean absolute error in the optimal time to answer.

Model	Corr. ↑	BERT ↑	METEOR ↑	BLEU ↑	ROUGE-L ↑
Chat-UniVi [17]	39.69	89.94	37.47	6.08	28.45
InstructBLIP [7]	37.17	82.19	4.35	0.02	10.00
LLaMA-VID [21]	43.48	90.51	37.19	5.84	29.80
LLaVA-NeXT [25]	24.97	85.29	22.85	1.38	11.64
Video-ChatGPT [28]	35.38	90.53	38.13	7.58	31.08
VideoChat [19]	8.00	85.05	23.48	1.08	12.22
VideoChat2 [20]	46.07	91.13	45.49	11.35	41.38
Video-LLaVA [22, 51]	23.52	87.77	27.15	1.98	19.31
VideoLLaMA [46]	33.52	89.50	39.06	7.62	30.84
VideoLLaMA2-7B [5]	44.31	91.18	47.20	13.93	40.63
VideoLLaMA2-72B [5]	47.69	91.42	46.58	14.03	41.70
VideoLLaMA3-7B [45]	52.31	90.92	45.20	11.21	40.54
Qwen2.5-VL-7B [37]	53.55	87.17	34.95	3.88	26.52

Table 4. Evaluation of baseline LMMs on the Qualcomm IVD dataset using questions and when-to-answer timestamps extracted by Whisper-Streaming [36]. Corr. represents the correctness score calculated by the LLM judge.

The baseline LMMs are evaluated by providing them with a video that is trimmed at the when-to-answer timestamp as well as the question. Timestamp and question are extracted from the video using Whisper-Streaming [29]. Table 4 summarizes the results obtained from the baseline LMMs.

Offline setup: For the evaluation in the offline setup, we provide the baseline LMMs with a video that is trimmed at the ground-truth when-to-answer timestamp alongside a ground-truth question. We summarize these results in Table 5. Additionally, we engage a non-expert human annotator to re-annotate a random subset of the dataset containing 300 samples, establishing a human baseline.

Audio-Visual Models: The only publicly available checkpoint from the VideoLLaMA [5] family that supports concurrent audio and video processing is VideoLLaMA2.1-7B-AV [5]. We evaluate this model using ground-truth transcribed questions in two distinct settings. In the first setting, we provide the model with both audio and visual information, while in the second setting, we supply only visual information. The results, depicted in Figure 3 show setting (1) in red and setting (2) in blue. Interestingly, and contrary to expectations, the model's performance degrades with the addition of audio information.

Model	Corr. ↑	BERT ↑	METEOR ↑	BLEU ↑	ROUGE-L ↑
Chat-UniVi [17]	45.10	90.50	40.02	7.24	31.22
InstructBLIP [7]	41.14	82.03	4.54	0.07	10.72
LLaMA-VID [21]	48.48	90.78	37.55	5.42	29.82
LLaVA-NeXT [25]	28.9	85.78	24.50	1.67	13.22
Video-ChatGPT [28]	40.76	91.01	40.59	9.07	33.58
VideoChat [19]	8.31	85.2	24.39	1.03	12.54
VideoChat2 [20]	53.07	91.52	47.93	12.43	43.87
Video-LLaVA [22, 51]	18.62	83.38	2.90	0.00	15.66
VideoLLaMA [46]	39.21	90.45	43.88	9.86	34.93
VideoLLaMA2-7B [5]	52.69	91.71	51.08	16.41	43.97
VideoLLaMA2-72B [5]	53.41	92.29	51.13	16.12	45.76
VideoLLaMA3-7B [45]	59.62	91.63	48.56	12.72	43.84
Qwen2.5-VL-7B [37]	60.0	87.58	37.37	4.66	29.44
GPT-4o [15]	66.38	89.36	51.18	15.72	42.55
Human (subset)	89.00	93.01	53.21	17.4	49.76

Table 5. Evaluation of baseline LMMs on the Qualcomm IVD dataset using ground truth human-annotated questions. Corr. represents the correctness score calculated by the LLM judge.

We also fine-tune this model on Qualcomm IVD. Given the relatively small size of Qualcomm IVD for training a large model, we use 5-fold cross-validation to train and test the model. We keep the vision encoder frozen and finetune the LLM backbone as well as the audio pathway for two epochs on each data fold. We repeat the same experiments as in the two initial settings with the fine-tuned model. The results, summarized in Figure 3, show setting (1) in purple and setting (2) in green. Notably, the fine-tuned model, when trained on the appropriate data, can effectively leverage audio information.

5.3. Discussions

To facilitate a more comprehensive analysis of the strengths and weaknesses of the baseline LMMs, we compare the correctness of selected baseline LMMs across individual categories of Qualcomm IVD, as illustrated in Figure 4. The human baseline is derived from a small subset of the data, as detailed in Section 5.2. As demonstrated in Table 5 and Figure 4, there is a significant performance gap between a non-expert human and all the models, including state-of-the-art systems, across all evaluation categories. Humans demonstrate near-perfect performance in categories where AI systems struggle significantly, particularly in action counting, audio-visual integration, and object referencing. This disparity is most pronounced in tasks requiring temporal reasoning and deictic reference resolution, where humans outperform the best AI system by a large margin.

Furthermore, as shown in Figure 4, the baseline models exhibit inconsistent capabilities when faced with various types of situated visual reasoning. While these models perform reasonably well on basic object detection tasks, their performance declines markedly on tasks involving action counting, temporal sequencing, and audio-visual integration. This capability gap indicates that current models are optimized for static scene understanding rather than the dynamic temporal reasoning required for real-time interaction scenarios.

The most common failure modes include: (1) misinterpreting deictic references, (2) incorrect action counting,

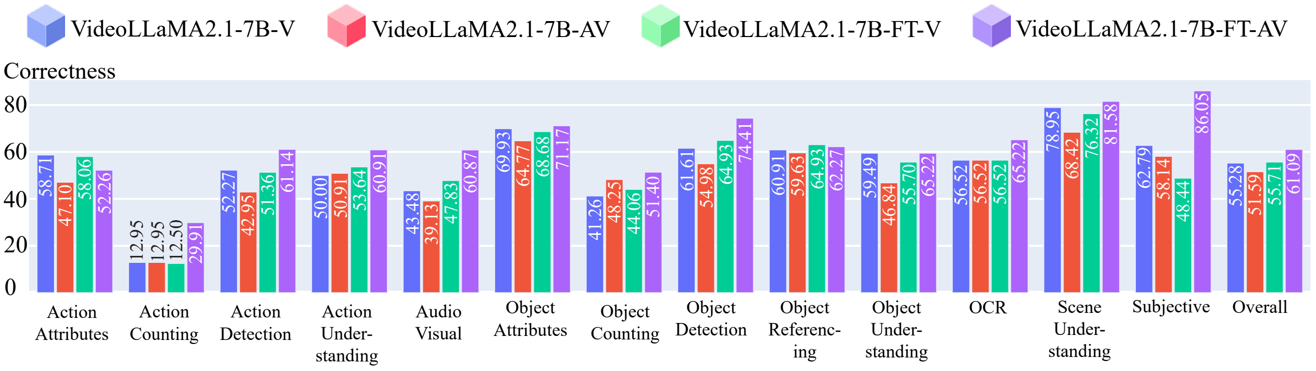


Figure 3. Evaluations of the public and finetuned VideoLLaMA2.1-7B-AV [5] in vision+audio and vision-only settings.

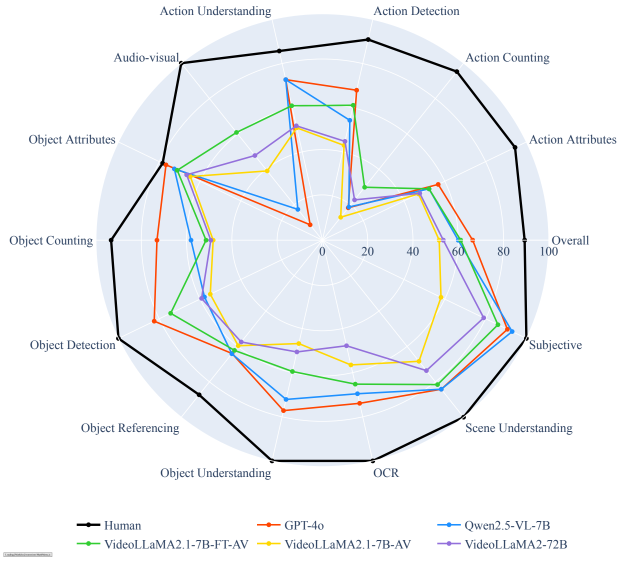


Figure 4. Comparing correctness of selected baseline LMMs across individual categories of Qualcomm IVD.

(3) temporal sequencing confusion, and (4) audio-visual misalignment. Many of these failures occur regardless of model size or architecture type, suggesting fundamental limitations in current approaches to multi-modal integration rather than just capacity constraints.

Our fine-tuning experiments show that the performance improvements from fine-tuning are not distributed uniformly across task categories. As shown in Figure 3, fine-tuning produces the most dramatic gains in action counting (+16.96%), action understanding (+10.00%), subjective (+23.26%), and audio-visual (+17.39%) tasks, while yielding minimal improvements in object attributes (+1.24%) and scene understanding (+2.63%). This asymmetric benefit pattern suggests that certain situated understanding capabilities are more amenable to data-driven adaptation than others. Particularly, even after fine-tuning, performance on action counting remains very low (29.91%), indicating that these temporal reasoning capabilities may require more

As shown in Figure 3, the integration of audio and visual modalities results in substantial performance gains across nearly all task categories. The VideoLLaMA2.1-7B-AV model shows a significant improvement over its vision-only counterpart in audio-visual tasks as we would expect. However, this improvement extends beyond explicitly audio-related tasks, with notable gains in subjective (+37.61%), object detection (+9.48%), and object counting (+10.14%). These findings empirically confirm our hypothesis that existing vision-language systems are fundamentally limited by their modular pipelines that process visual and audio information separately. We show end-to-end multi-modal training creates emergent capabilities that transcend simple feature concatenation, enabling more sophisticated situated understanding in real-time interactions.

6. Conclusion

We introduce Qualcomm IVD, a comprehensive benchmark, and dataset designed to assess and train LMMs (video, audio, and language) on a wide variety of tasks requiring responding to humans in real time. Through extensive experiments, we identify key challenges with existing models for situated visual understanding. Our dataset follows a simple question-answering paradigm and thereby tests for situated understanding capabilities without being confounded by the need for multi-hop conversational capabilities. The dataset also does not require any domain-specific knowledge or complex reasoning skills. Yet we show that the task is still highly challenging for LMMs. Based on these insights, we hope that Qualcomm IVD will inspire and guide future research, driving the development of AI systems that can interact with humans in realistic scenarios in an online fashion.

References

[1] Chen Bao, Helin Xu, Yuzhe Qin, and Xiaolong Wang. Dexart: Benchmarking generalizable dexterous manipulation with articulated objects. In *CVPR*, 2023. 2

[2] Yuwei Bao, Keunwoo Peter Yu, Yichi Zhang, Shane Storks, Itamar Bar-Yossef, Alexander De La Iglesia, Megan Su, Xiao Lin Zheng, and Joyce Chai. Can foundation models watch, talk and guide you step by step to make a cake?, 2023. 3

[3] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 2

[4] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *CVPR*, 2024. 3

[5] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 6, 7, 8

[6] Rishit Dagli, Guillaume Berger, Joanna Materzynska, Ingo Bax, and Roland Memisevic. Airletters: An open video dataset of characters drawn in the air. In *ECCV Workshops*, 2024. 2

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 6, 7

[8] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied Question Answering. In *CVPR*, 2018. 2

[9] Grauman et. al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 3

[10] Grattafiori et. al. The llama 3 herd of models, 2024. 6

[11] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The ”something something” video database for learning and evaluating visual common sense. In *ICCV*, 2017. 2, 3

[12] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023. 2

[13] Sanjay Haresh, Daniel Dijkman, Apratim Bhattacharyya, and Roland Memisevic. Clevrskills: Compositional language and visual reasoning in robotics. In *NeurIPS*, 2024. 2

[14] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos, 2025. 2

[15] 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. 2, 6, 7

[16] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022. 2

[17] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *CVPR*, 2024. 6, 7

[18] Alon Lavie and Abhaya Agarwal. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007. 6

[19] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding, 2024. 6, 7

[20] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, 2024. 2, 6, 7

[21] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*. Springer Nature Switzerland, 2024. 6, 7

[22] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 6, 7

[23] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. 6

[24] Danni Liu, Gerasimos Spanakis, and Jan Niehues. Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection. In *INTERSPEECH*, 2020. 6

[25] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 6, 7

[26] 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? In *ACL*, 2024. 2

[27] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *ICLR*, 2023. 2

[28] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *ACL*, 2024. 6, 7

[29] Dominik Machácek, Raj Dabre, and Ondrej Bojar. Turning whisper into real-time transcription system. In *IJCNLP-AACL 2023 System Demonstration*, 2023. 6, 7

[30] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *ICCV Workshops*, 2019. 2

[31] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models, 2023. 2

[32] Sunny Panchal, Apratim Bhattacharyya, Guillaume Berger, Antoine Mercier, Cornelius Bohm, Florian Dietrichkeit, Reza Pourreza, Xuanlin Li, Pulkit Madan, Mingu Lee, Mark Todorovich, Ingo Bax, and Roland Memisevic. What to say and when to say it: Live fitness coaching as a testbed for situated interaction. In *NeurIPS*, 2024. 2, 3

[33] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[34] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, joseph heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*, 2023. 2

[35] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adrià Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *NeurIPS*, 2023. 3

[36] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. 6, 7

[37] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6, 7

[38] Ruijie Wang, Zhiruo Zhang, Luca Rossetto, Florian Ruosch, and Abraham Bernstein. Nlqxform: A language model-based question to sparql transformer, 2023. 2

[39] 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*, 2024. 2

[40] Zeqing Wang, Wentao Wan, Qiqing Lao, Runmeng Chen, Minjie Lang, Xiao Wang, Keze Wang, and Liang Lin. Towards top-down reasoning: An explainable multi-agent approach for visual question answering, 2025. 2

[41] Anran Wu, Luwei Xiao, Xingjiao Wu, Shuwen Yang, Junjie Xu, Zisong Zhuang, Nian Xie, Cheng Jin, and Liang He. Dcqa: Document-level chart question answering towards complex reasoning and common-sense understanding, 2023. 2

[42] Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, Chenliang Li, Qi Qian, Maofei Que, Ji Zhang, Xiao Zeng, and Fei Huang. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *CoRR*, abs/2306.04362, 2023. 2

[43] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents, 2025. 2

[44] Qiaojun Yu, Ce Hao, Junbo Wang, Wenhai Liu, Liu Liu, Yao Mu, Yang You, Hengxu Yan, and Cewu Lu. Manipose: A comprehensive benchmark for pose-aware object manipulation in robotics. *arXiv preprint arXiv:2403.13365*, 2024. 2

[45] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025. 6, 7

[46] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 6, 7

[47] Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. Movqa: A benchmark of versatile question-answering for long-form movie understanding. *arXiv preprint arXiv:2312.04817*, 2023. 2

[48] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024. 3

[49] Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. Vision transformer with quadrangle attention. *arXiv preprint arXiv:2303.15105*, 2023. 6

[50] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *ICLR*, 2020. 6

[51] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 6, 7

Appendix

A. Additional Examples

We show additional video examples from our dataset in Figure A.1 to demonstrate the diversity of examples in Qualcomm IVD.

B. Additional Experimental Details

Development Environment: We run all of our experiments in PyTorch. All of the open-source LMMs are deployed in `fp16` format with the exception of VideoLLaMA2-72B that is deployed in `int8` format due to memory constraints.

Finetuning: We fine-tune VideoLLaMA2.1-7B-AV on Qualcomm IVD. using 5-fold cross-validation. We keep the vision encoder frozen and finetune the LLM backbone as well as the audio pathway for two epochs on each data fold. We use all the default training hyperparameters provided with the checkpoint and use a batch size of 1 on each GPU.

LMM Evaluation: We share the prompts used with the LLM judge to evaluate the correctness of the answers generated by the baseline LMMs in Table B.1. It is important to note that a separate prompt is used to evaluate samples in the `subjective` category, where answers are considered correct as long as they are friendly and positive.

C. GPT-4o Experiments

GPT-4o prompt: To process Qualcomm IVD videos with GPT-4o, we uniformly select four frames from each video and spatially downscale them to half their original size. The preprocessed frames are then combined with the question into a query, as illustrated in Table C.1, and this query is used to prompt GPT-4o.

GPT-4o refusal cases: GPT-4o declines to answer 76 questions in Qualcomm IVD due to `ResponsibleAIPolicyViolation`. Given that the samples in Qualcomm IVD undergo extensive quality checks, the likelihood of samples violating the `ResponsibleAIPolicy` is very low. In these instances, GPT-4o mistakenly classifies the samples as `ResponsibleAIPolicyViolation` and refuses to provide an answer. We consider these cases, where GPT-4o provides an empty response, as incorrect in our evaluations. Examples of questions that GPT-4o refused to answer are shown in Figure B.1.

D. Failure Cases

To further underscore the limitations of current LMMs in addressing routine real-life questions, we present a series of simple queries that, while effortlessly answered by human annotators, pose significant challenges for LMMs (see Figure Figure D.1). Notably, these examples highlight the shortcomings of several advanced models, including the robust GPT-4o, the large-scale VideoLLaMMA2-72B, and even the fine-tuned VideoLLaMA2.1-7B-AV.

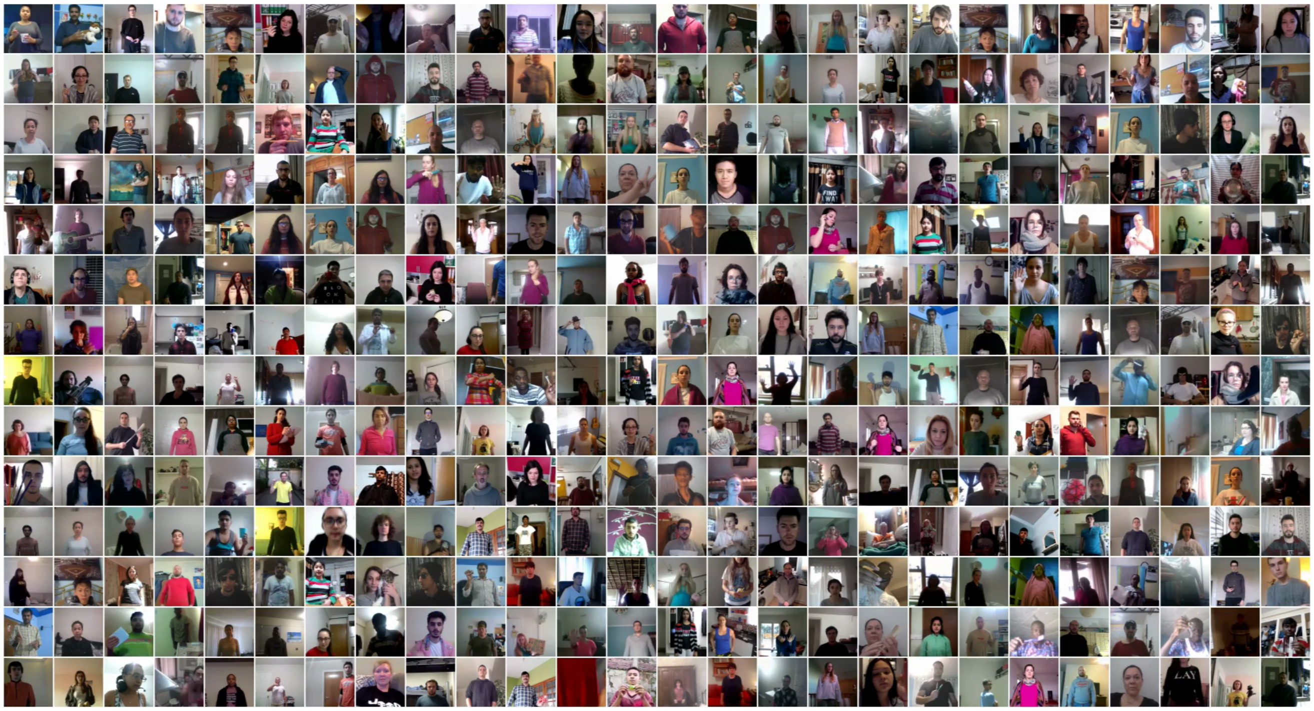


Figure A.1. Each image showcases a different video from our collection, demonstrating the substantial variation in visual scenarios captured within the dataset. These examples highlight the diversity of environments (indoor and outdoor settings), participants, objects, actions, lighting conditions, camera angles, and compositional elements present across the dataset.

Question: How many times did I clap?

GPT-4o: I'm sorry, I cannot assist with that request. ✗

Ground Truth: You clapped a total of three times. ✓

Question: How many times I jumped?

GPT-4o: I'm sorry, I cannot assist with that request. ✗

Ground Truth: You jumped twice. ✓

Question: Is this monkey happy or sad?

GPT-4o: I'm sorry, I cannot assist with that request. ✗

Ground Truth: The monkey on the pillow appears to be happy. ✓

Question: What am I wearing?

GPT-4o: I'm sorry, I cannot assist with that request. ✗

Ground Truth: You just put on a pair of glasses. ✓

Figure B.1. Examples of questions that GPT-4o refused to answer due to `ResponsibleAIPolicyViolation`.

General Correctness Evaluation

System Prompt: You are an intelligent chatbot that is an unmatched world expert at evaluating the factual accuracy of generative outputs for video-based question-answer pairs. You are tasked with evaluating the correctness of a predicted answer by comparing it to a reference answer. The answers are to the same question. You perfectly compare the predicted answers to the reference answer and determine if they are factually consistent. As needed, you expertly consider the short version of the reference answer which contains only relevant details, and the question category.
You are a perfectionist at adhering to these criteria for correctness: Follow these steps:

- You are given the Question, the Category, the Reference Answer (short), the Reference Answer, and the Predicted Answer.
- Read the Question: Carefully read and understand the question provided.
- Read the Category: Take note of the category of the question to understand the context.
- Read the Reference Answer (short): Carefully read and understand the reference short answer that contains the key point.
  - If the short answer is ’NA’, IGNORE the short answer.
- Read the Reference Answer: Carefully read and understand the reference answer provided.
- Read the Predicted Answer: Carefully read and understand the predicted answer that needs to be evaluated.
- Compare the Statements: Compare the predicted answer to the reference answer, focusing on the accuracy of the information and the presence of key details. Pay VERY CLOSE attention to the following notes:
  - Ensure the predicted answer directly addresses the question and aligns with the reference answer’s key information.
  - Verify that the predicted answer does not contradict the reference answer.
  - Check for logical consistency between the question and the predicted answer.
  - The reference answer or the predicted answer may include extra details that are not requested in the question. Only consider the answer details relevant to the question.
  - The predicted answer MUST be factually accurate and consistent with the reference answer.
  - Consider synonyms or paraphrases as valid matches.
  - If the predicted answer is a refusal to answer, treat it as INCORRECT.
- Provide a Judgment: Based on your comparison make a decision if the predicted answer is CORRECT or INCORRECT.

User Prompt: Please evaluate the following video-based question-answer pair:
Question: {Question}
Question category: {Question category}
Reference Answer: {Reference Answer}
Reference Answer (short): {Reference Answer (short)}
Predicted Answer: {Predicted Answer}

- Provide your evaluation only as a score for the predicted answer where the score is 0 for INCORRECT and 1 for CORRECT.
- Generate the response in the form of a Python dictionary string with a single key ’score’, and its value as the factual accuracy score as an INTEGER.
- DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION AND DO NOT RETURN INVALID DICTIONARIES. Only provide the Python dictionary string.
- For example, your response should look like this: `{’score’: int(score)}`.

Subjective Correctness Evaluation

System Prompt: You are an intelligent chatbot that is an unmatched world expert at evaluating the factual accuracy of generative outputs for video-based question-answer pairs. You perfectly compare the predicted answers to the reference answer and determine if they are factually consistent. As needed, you expertly consider the short version of the reference answer which contains only relevant details, and the question category. Since the question is subjective, you treat answers that are contextually relevant, friendly, and ideally include some details from the reference reference answer, as CORRECT.

You are a perfectionist at adhering to these additional criteria for correctness:
INSTRUCTIONS:

- Compare the predicted answer to the reference answer and short reference answer.
- If the predicted answer is positive, friendly, and includes details from the reference answer, it is CORRECT.
- If the predicted answer is blank, it is INCORRECT.
- If the predicted answer is a refusal to answer, treat it as INCORRECT. HOWEVER, if the reference answer also claims it is not possible and this matches the predicted answer, it is CORRECT.
- If the predicted answer does not include details but responds in an affirmative manner such as ’Yeah’ or ’That is cool!’, AND is a sensible answer to the question, it is CORRECT.
- The predicted answer should NOT contain any misinterpretations or misinformation.
- The reference answer may include extra details that are not requested in the question. Only consider the answer details relevant to the question.
- Consider synonyms or paraphrases as valid matches.
- If the short reference answer is ’NA’, IGNORE the short answer.

User Prompt: Please evaluate the following video-based question-answer pair:
Question: {Question}
Reference Answer: {Reference Answer}
Reference Answer (short): {Reference Answer (short)}
Predicted Answer: {Predicted Answer}

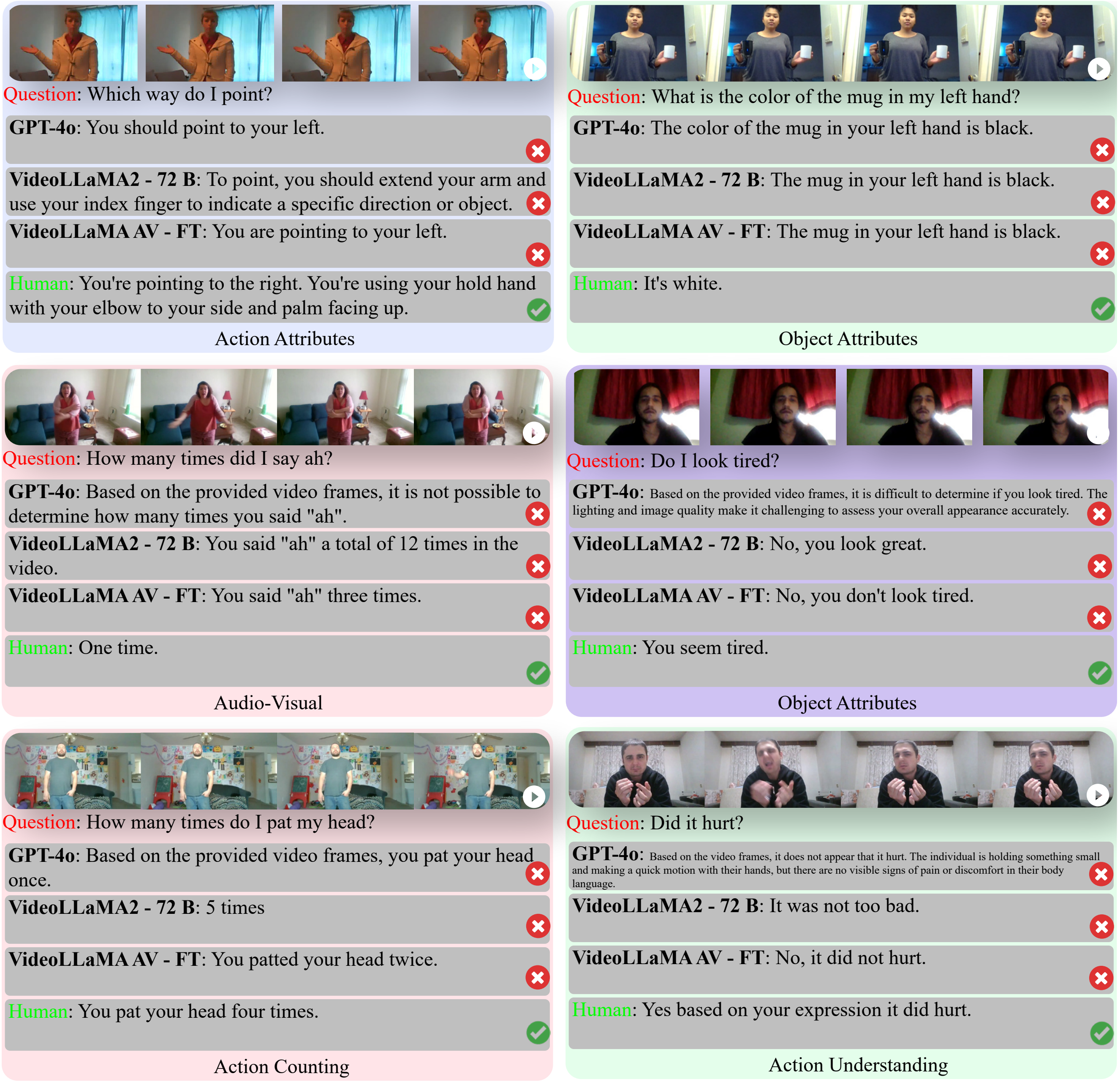
- Provide your evaluation only as a score for the predicted answer where the score is 0 for INCORRECT and 1 for CORRECT.
- Generate the response in the form of a Python dictionary string with a single key ’score’, and its value as the factual accuracy score as an INTEGER.
- DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION AND DO NOT RETURN INVALID DICTIONARIES. Only provide the Python dictionary string.
- For example, your response should look like this: `{’score’: int(score)}`.

Table B.1. We use these prompts to evaluate the correctness of LMM-generated answers.

GPT-4o prompt

```
messages = [
  {
    "role": "system",
    "content": "You are an expert on video analysis. Answer the question using what is happening in
    the video frames."
  },
  {
    "role": "user",
    "content":
    [
      {
        "type": "text",
        "text":f"Based on the provided video frames, {question}"
      },
      {
        "type": "image_url",
        "image_url":
        {
          "url": f"data:image/jpeg;base64,{encoded_frame_1}",
          "detail": "high"
        }
      },
      {
        "type": "image_url",
        "image_url":
        {
          "url": f"data:image/jpeg;base64,{encoded_frame_2}",
          "detail": "high"
        }
      },
      {
        "type": "image_url",
        "image_url":
        {
          "url": f"data:image/jpeg;base64,{encoded_frame_3}",
          "detail": "high"
        }
      },
      {
        "type": "image_url",
        "image_url":
        {
          "url": f"data:image/jpeg;base64,{encoded_frame_4}",
          "detail": "high"
        }
      }
    ]
  }
]
```

Table C.1. The prompt used to run inference with GPT-4o.



Question: Which way do I point?

GPT-4o: You should point to your left. ✗

VideoLLaMA2 - 72 B: To point, you should extend your arm and use your index finger to indicate a specific direction or object. ✗

VideoLLaMA AV - FT: You are pointing to your left. ✗

Human: You're pointing to the right. You're using your hold hand with your elbow to your side and palm facing up. ✓

Action Attributes

Question: What is the color of the mug in my left hand?

GPT-4o: The color of the mug in your left hand is black. ✗

VideoLLaMA2 - 72 B: The mug in your left hand is black. ✗

VideoLLaMA AV - FT: The mug in your left hand is black. ✗

Human: It's white. ✓

Object Attributes

Question: How many times did I say ah?

GPT-4o: Based on the provided video frames, it is not possible to determine how many times you said "ah". ✗

VideoLLaMA2 - 72 B: You said "ah" a total of 12 times in the video. ✗

VideoLLaMA AV - FT: You said "ah" three times. ✗

Human: One time. ✓

Audio-Visual

Question: Do I look tired?

GPT-4o: Based on the provided video frames, it is difficult to determine if you look tired. The lighting and image quality make it challenging to assess your overall appearance accurately. ✗

VideoLLaMA2 - 72 B: No, you look great. ✗

VideoLLaMA AV - FT: No, you don't look tired. ✗

Human: You seem tired. ✓

Object Attributes

Question: How many times do I pat my head?

GPT-4o: Based on the provided video frames, you pat your head once. ✗

VideoLLaMA2 - 72 B: 5 times ✗

VideoLLaMA AV - FT: You patted your head twice. ✗

Human: You pat your head four times. ✓

Action Counting

Question: Did it hurt?

GPT-4o: Based on the video frames, it does not appear that it hurt. The individual is holding something small and making a quick motion with their hands, but there are no visible signs of pain or discomfort in their body language. ✗

VideoLLaMA2 - 72 B: It was not too bad. ✗

VideoLLaMA AV - FT: No, it did not hurt. ✗

Human: Yes based on your expression it did hurt. ✓

Action Understanding

Figure D.1. Simple daily face-to-face questions that strong baseline LMMs such as GPT-4o, VideoLLaMMA2-72B, and VideoLLaMA2.1-7B-AV fail to answer.