CAN VISION-LANGUAGE MODELS ANSWER FACE TO FACE QUESTIONS IN THE REAL-WORLD?

Anonymous authors
Paper under double-blind review

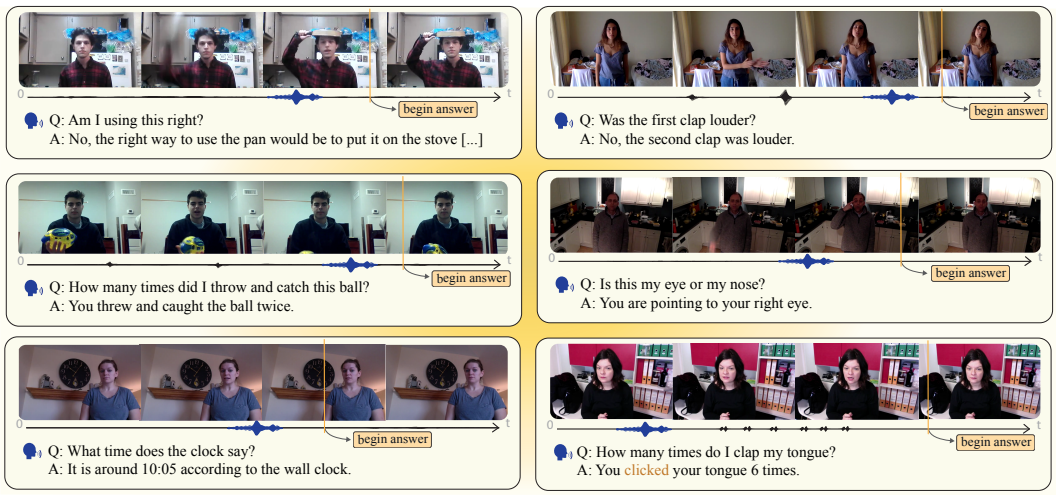

Figure 1: We present Interactive Video Dataset (IVD), a dataset collected in an online question-answering setup, where users pose open-ended questions using their camera and microphone. IVD offers videos with raw audio, annotated textual transcriptions of the spoken questions, and text answers with annotated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

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 (Hurst et al., 2024) 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 (Figure 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 Interactive Video Dataset (IVD), a new dataset and benchmark designed for end-to-end trained systems aimed at real-time user interaction. 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 (Hurst et al., 2024), even if they are otherwise good at performing visual reasoning. Our experiments indicate that the failure modes of existing systems can be attributed to their: 1. difficulty integrating visual and auditory information in real-time to disambiguate questions, 2. inability to determine the appropriate time at which to answer, and 3. inability to answer questions whose answers require situational common sense. 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:

1. We introduce IVD, a novel multi-modal dataset designed to evaluate online situated audio-visual reasoning and real-time conversational skills.
2. We benchmark existing LMMs and identify critical weaknesses in their ability to handle real-life conversations.
3. We demonstrate that these limitations can be effectively mitigated by fine-tuning models on appropriate audio-visual conversational data.
4. 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 (Li et al., 2024b; Liu et al., 2024b; Xu et al., 2023; Patraucean et al., 2023a; Ning et al., 2023; Hu et al., 2025), hand movements (Goyal et al., 2017; Materzynska et al., 2019), articulated motion (Dagli et al., 2024), full human body motion (Panchal et al., 2024), robotics (Haresh et al., 2024; Yu et al., 2024; Bao et al., 2023a; Gu et al., 2023; Brohan

Table 1: Comparison of various benchmarks encompassing several key aspects.

Benchmark	#Videos	#QA-Pairs	Annotation	Audio	Subtitle	Interactive	Face-to-Face
AVSD (DSTC7) (Alamri et al., 2018)	11156	∼111560	Manual	✓	✗	✓	✗
KnowIT VQA (Garcia et al., 2020)	207	24282	Manual	✓	✓	✗	✗
LifeQA (Castro et al., 2020)	275	2326	Manual	✓	✓	✗	✗
How2QA (Li et al., 2020)	9035	44007	Manual	✓	✓	✓	✗
MedVidQA (Gupta et al., 2023)	899	3010	Manual	✓	✓	✓	✗
Social-IQ (Zadeh et al., 2019)	1250	7500	Manual	✓	✗	✗	✓
Video-MME (Fu et al., 2024)	900	2700	Manual	✓	✓	✗	✗
CodeVidQA (Raja et al., 2025)	2104	2104	Automatic	✓	✓	✓	✗
Ego4D Social Interactions (et. al., 2022)	667	task-specific *labels*	Manual	✓	✗	✓	✓
TVQA (Lei et al., 2018)	21793	152545	Manual	✓	✓	✗	✗
NExT-GQA (Xiao et al., 2024)	1557	10531	Manual	✓	✓	✓	✗
STAR (Wu et al., 2024)	22000	60000	Automatic	✓	✗	✓	✗
IVD	2900	2900	Manual	✓	✓	✓	✓

et al., 2023; Jiang et al., 2022), and embodied reasoning (Yang et al., 2025b). 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 (Wang et al., 2024b), and MoVQA (Zhang et al., 2023b), which extend the context window but fail to simulate real-time constraints. In contrast, our IVD dataset and benchmark focuses on real-world questions answering.

Situated Video Evaluation Benchmarks: Situated question answering has also been studied by (Das et al., 2018; Ma et al., 2023) and follow-up works (*e.g.* (Wang et al., 2023; 2025; Wu et al., 2023)). A separate line of work has studied "common sense" situational understanding for AI models, albeit not a VQA format. This includes the work by (Goyal et al., 2017; et. al., 2022; Patraucean et al., 2023b) and recent work on situated live dialogue (Bao et al., 2023b; Panchal et al., 2024). 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 (Chen et al., 2024) and FlashVStream (Zhang et al., 2024), 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 (Panchal et al., 2024) 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 IVD

The purpose of the IVD 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. A side-by-side comparison of the features offered by IVD and other related datasets is presented in Table 1.

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 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 categories listed in Appendix B.2.

Table 2: 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.

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

Table 3: 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.

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

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 IVD dataset consists of 2900 video clips and thus 2900 unique question-answer pairs. Table 2 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 3.

Temporal Characteristics: A distinctive feature of the 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. We analyze the temporal characteristics by category in Table 3, which shows the distribution of optimal answer timestamps relative to video duration for each category. As we observe from Table 3, 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 IVD include:

Table 4: **ASR performance comparison.** Evaluation of Automatic Speech Recognition (ASR) systems on the IVD dataset using standard text similarity metrics. The value Δt represents the mean absolute error in the optimal time to answer.

Model	METEOR ↑	BLEU ↑	ROUGE-L ↑	$\Delta t \downarrow$	$\Delta t\,(-) \downarrow$	$\Delta t\,(+) \downarrow$
Whisper (Radford et al., 2022)	90.01	80.95	90.32	-	-	-
Whisper-Streaming (Machácek et al., 2023)	92.34	74.57	91.82	0.83	-0.94	0.61

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 (Machácek et al., 2023) to identify "when to answer". The Streaming-Whisper model (Machácek et al., 2023) uses the LocalAgreement algorithm (Liu et al., 2020) 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 IVD.

5.1 Experiments Setup

Configurations: The experiments are performed within four 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. Impact of audio: Among existing LMMs, the VideoLLaMA family of models (Zhang et al., 2023a) 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

Table 5: Evaluation of baseline LMMs on the IVD dataset using (a) questions and estimated when-to-answer timestamps by Whisper (Radford et al., 2022) and (b) ground-truth questions and timestamps. Corr. represents correctness by LLM judge.

	ASR Questions and Timestamps					Human Questions and Timestamps				
Model	Corr. ↑	BERT ↑	METEOR ↑	BLEU ↑	ROUGE-L ↑	Corr. ↑	BERT ↑	METEOR ↑	BLEU ↑	ROUGE-L ↑
Chat-UniVi (Jin et al., 2024)	34.66	89.94	37.47	6.08	28.45	40.79	90.50	40.02	7.24	31.22
InstructBLIP (Dai et al., 2023)	35.03	82.19	4.35	0.02	10.00	39.14	82.03	4.54	0.07	10.72
LLaMA-VID (Li et al., 2024c)	39.41	90.51	37.19	5.84	29.80	43.0	90.78	37.55	5.42	29.82
LLaVA-NeXT (Liu et al., 2024a)	19.45	85.29	22.85	1.38	11.64	22.66	85.78	24.50	1.67	13.22
Video-ChatGPT (Maaz et al., 2024)	32.45	90.53	38.13	7.58	31.08	36.59	91.01	40.59	9.07	33.58
VideoChat (Li et al., 2024a)	3.69	85.05	23.48	1.08	12.22	3.52	85.20	24.39	1.03	12.54
VideoChat2 (Li et al., 2024b)	44.66	91.13	45.49	11.35	41.38	50.35	91.52	47.93	12.43	43.87
Video-LLaVA (Zhu et al., 2023; Lin et al., 2023)	20.28	87.77	27.15	1.98	19.31	15.0	83.38	2.90	0.00	15.66
VideoLLaMA (Zhang et al., 2023a)	30.76	89.50	39.06	7.62	30.84	35.93	90.45	43.88	9.86	34.93
VideoLLaMA2-7B (Cheng et al., 2024)	43.34	91.18	47.20	13.93	40.63	50.07	91.71	51.08	16.41	43.97
VideoLLaMA2-72B (Cheng et al., 2024)	46.52	91.42	46.58	14.03	41.70	50.83	92.29	51.13	16.12	45.76
VideoLLaMA3-7B (Zhang et al., 2025)	50.59	90.92	45.20	11.21	40.54	56.38	91.63	48.56	12.72	43.84
VideoLLM-online (Chen et al., 2024)	–	–	–	–	–	23.76	88.67	33.73	4.16	26.27
Qwen2.5-VL-7B (Wang et al., 2024a)	44.90	87.17	34.95	3.88	26.52	50.62	87.58	37.37	4.66	29.44
Qwen2.5-Omni-7B (Xu et al., 2025)	43.97	86.65	33.45	2.77	20.57	45.90	86.73	33.98	2.87	20.98
GPT-4o (Hurst et al., 2024)	–	–	–	–	–	58.76	89.36	51.18	15.72	42.55
Human (subset)	–	–	–	–	–	87.33	93.01	53.21	17.40	49.76

models by examining the impact of additional audio on the accuracy of their question-answering capabilities.

4. Impact of when-to-answer: This experiment investigates how the timing of the when-to-answer moment affects model performance. We utilize the Qwen2.5-Omni model (Xu et al., 2025), the only publicly available model capable of concurrently processing both audio and video modalities while also transcribing speech. The model is provided with both the ground-truth and ASR-derived when-to-answer timestamps. It then transcribes the question and generates an answer, allowing us to compare the outputs and assess the influence of timing on response quality.

Baseline Models: We experiment with various open-source and closed-source LMMs.

The open-source models we evaluate include InstructBLIP (7B) (Dai et al., 2023), Video-ChatGPT (7B) (Maaz et al., 2024), VideoChat (7B) (Li et al., 2024a), VideoChat2 (7B) (Li et al., 2024b), LLaVA-NeXT (7B) (Liu et al., 2024a), LLaMA-VID (13B) (Li et al., 2024c), Video-LLaMA (13B) (Zhang et al., 2023a), VideoLLaMA2 (7B/72B) (Cheng et al., 2024), VideoLLaMA2.1 (7B) (Cheng et al., 2024), VideoLLaMA3 (7B) (Zhang et al., 2025), Video-LLaVA (7B) (Zhu et al., 2023; Lin et al., 2023), Chat-UniVi (13B) (Jin et al., 2024), Qwen2.5-VL (7B) (Wang et al., 2024a), and Qwen2.5-Omni (7B) (Xu et al., 2025) The model sizes range from 7B to 13B parameters for the language backbone, with the exception of VideoLLaMA2-72B (Cheng et al., 2024). 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 (Dai et al., 2023), an image model, we sample 4 frames from each video, process these frames with the image encoder and a Q-Former (Zhang et al., 2023c) 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 (Hurst et al., 2024), in a zero-shot fashion. Videos are preprocessed by uniformly selecting 4 frames from each video and down-scaling the resolution to half. The query used to prompt GPT-4o is provided in the appendix.

Evaluation Metrics: Since the answers in 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 Qwen-32B model (Yang et al., 2024) as the LLM judge (see Appendix C.5 for comparisions with other LLM judges). The prompts that were used are provided in the appendix. In addition, we report Bert (Zhang* et al., 2020), METEOR (Lavie & Agarwal, 2007), BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004) scores between the ground-truth answers and the predicted answers.

5.2 Results

We now present the results obtained from the three settings described in Section 5.1.

Streaming setup: Table 4 presents the transcription results obtained from Whisper-Streaming (Machácek et al., 2023), where the transcription quality is quantified using BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Lavie & Agarwal, 2007) scores, by comparing

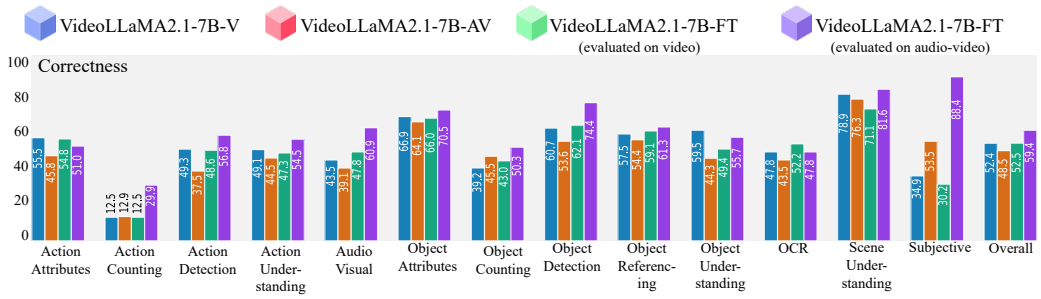

Figure 2: Evaluations of the public and finetuned VideoLLaMA2.1-7B-AV (Cheng et al., 2024) in vision + audio and vision-only settings.

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. In addition to the overall MAE, we also report the mean values of both negative and positive Δt instances. Notably, the average negative Δt is larger in magnitude, suggesting that models which initiate responses immediately after detecting the end of a question tend to answer prematurely—often before sufficient contextual information has been received. We also report the results obtained from the standard Whisper model (Radford et al., 2022) as an additional baseline. It is worth noting that this model does not return any timestamps alongside the transcriptions.

The baseline LMMs are evaluated using a video trimmed at the when-to-answer timestamp and the question, both extracted via Whisper-Streaming (Machácek et al., 2023). Table 5 summarizes the baseline results.

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.

Impact of audio: The only publicly available checkpoint from the VideoLLaMA (Cheng et al., 2024) family that supports concurrent audio and video processing is VideoLLaMA2.1-7B-AV (Cheng et al., 2024). 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 2 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.

We fine-tune this model on IVD using both audio and video modalities. Due to the dataset's small size, we apply 5-fold cross-validation. The vision encoder remains frozen, while the LLM backbone and audio pathway are fine-tuned for two epochs per fold. We repeat the initial experiments with the fine-tuned model. Results in Figure 2 show setting (1) in purple and setting (2) in green. The fine-tuned model performs best when both modalities are available but underperforms the pretrained model in some video-only cases, likely due to its adaptation to multimodal inputs. Since IVD relies heavily on audio cues, missing audio during inference significantly impairs performance.

Impact of when-to-answer: In this experiment, we evaluate the Qwen2.5-Omni (7B) (Xu et al., 2025) model to assess the impact of accurately identifying the when-to-answer moment. The model is tested using timestamps derived from both ground-truth (GT) annotations and ASR-Stream predictions. As illustrated in Figure 3, the results demonstrate that precise estimation of the when-to-answer moment can lead to substantial improvements in model performance.

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 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

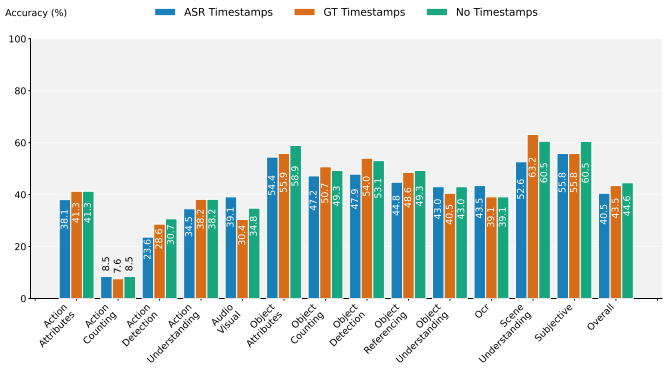

Figure 3: Evaluations of Qwen2.5-Omni (7B) with when-to-answer moment provided from ground-truth and ASR.

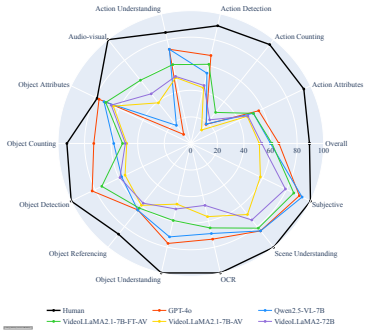

Figure 4: Correctness of selected baseline LMMs across individual categories of IVD.

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, (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 2, 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 sophisticated architectural inductive biases.

As shown in Figure 2, 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 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 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

Huda Alamri, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Jue Wang, Irfan Essa, Dhruv Batra, Devi Parikh, Anoop Cherian, Tim K Marks, et al. Audio visual scene-aware dialog (avsd) challenge at dstc7. *arXiv preprint arXiv:1806.00525*, 2018.

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

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?, 2023b.

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.

Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada Mihalcea. Lifeqa: A real-life dataset for video question answering. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4352–4358, 2020.

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.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers, 2022. URL https://arxiv.org/abs/2212.09058.

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. URL https://arxiv.org/abs/2406.07476.

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.

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.

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

Grattafiori et. al. The llama 3 herd of models, 2024.

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

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2024.

Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 10826–10834, 2020.

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.

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.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158, 2023.

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

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.

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.

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.

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.

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.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018.

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

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*, 2024b.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

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

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.

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

Danni Liu, Gerasimos Spanakis, and Jan Niehues. Low-latency sequence-to-sequence speech recognition and translation by partial hypothesis selection. In Helen Meng, Bo Xu, and Thomas Fang Zheng (eds.), *INTERSPEECH*, 2020.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

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*, August 2024b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

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.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *ACL*, August 2024.

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

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

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.

NVIDIA. Cosmos-Tokenizer: A suite of image and video neural tokenizers. https://github.com/NVIDIA/Cosmos-Tokenizer, 2025. Archived on 2025-02-11; accessed 2025-05-15.

NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL https://arxiv.org/abs/2501.03575.

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.

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

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*, 2023a.

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*, 2023b.

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

Sana Javaid Raja, Adeel Zafar, and Aqsa Shoaib. A dataset for programming-based instructional video classification and question answering. In *Proceedings of the First Workshop of Evaluation of Multi-Modal Generation*, pp. 1–9, 2025.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020. URL https://arxiv.org/abs/1910.02054.

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*, 2024a.

Ruijie Wang, Zhiruo Zhang, Luca Rossetto, Florian Ruosch, and Abraham Bernstein. Nlqxform: A language model-based question to sparql transformer, 2023. URL https://arxiv.org/abs/2311.07588.

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

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. URL https://arxiv.org/abs/2311.17331.

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. URL https://arxiv.org/abs/2310.18983.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos, 2024. URL https://arxiv.org/abs/2405.09711.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214, 2024. doi: 10.1109/CVPR52733.2024.01254.

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.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025. URL https://arxiv.org/abs/2503.20215.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin

Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

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, 2025b.

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.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

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.

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

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.

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*, 2023b.

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

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

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.

Appendix

A Social Impacts of IVD

A.1 Data Release

Our data set and code will be publicly released upon acceptance.

A.2 Limitations of IVD

While our experiments with IVD indicate that it presents a significant challenge for current multimodal models—and despite the dataset being human-validated for annotation accuracy—there are several potential limitations and sources of bias to consider:

1. Relatively small class sizes, which may limit the diversity of questions and answers; 2. Recordings conducted in controlled environments, potentially reducing variability in lighting, background, and camera angles, which may affect model generalization; 3. Possible demographic biases in terms of gender, age, and ethnicity, which could impact model performance across diverse user groups.

A.3 Privacy and Ethics in IVD

The data was collected under direct agreements with crowd workers, permitting both research and commercial use ensuring compliance with applicable privacy regulations, including GDPR-equivalent standards. All videos were manually reviewed to identify and exclude any content containing issues such as the presence of individuals in the background. Personally identifiable information, including metadata, was removed to the extent possible to ensure participant privacy. Additionally, all contributors received fair and appropriate compensation in accordance with the standards of their respective regions. All contributors signed a consent form that explicitly permits both research and commercial use of their video and audio data, including use in training AI models. We will provide a contact email on the dataset release page, allowing participants to request data removal at any time.

A.4 Broader Impact of IVD

In addition to the aforementioned sources of bias, language models trained on IVD may generate harmful or biased content, propagate misinformation, or offer inappropriate advice. These risks must be carefully considered when interacting with, deploying, or building upon such models.

B Additional Dataset Details

B.1 Additional Examples

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

B.2 Semantic Taxonomy

A detailed definition of the categories used in IVD is provided here.

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.

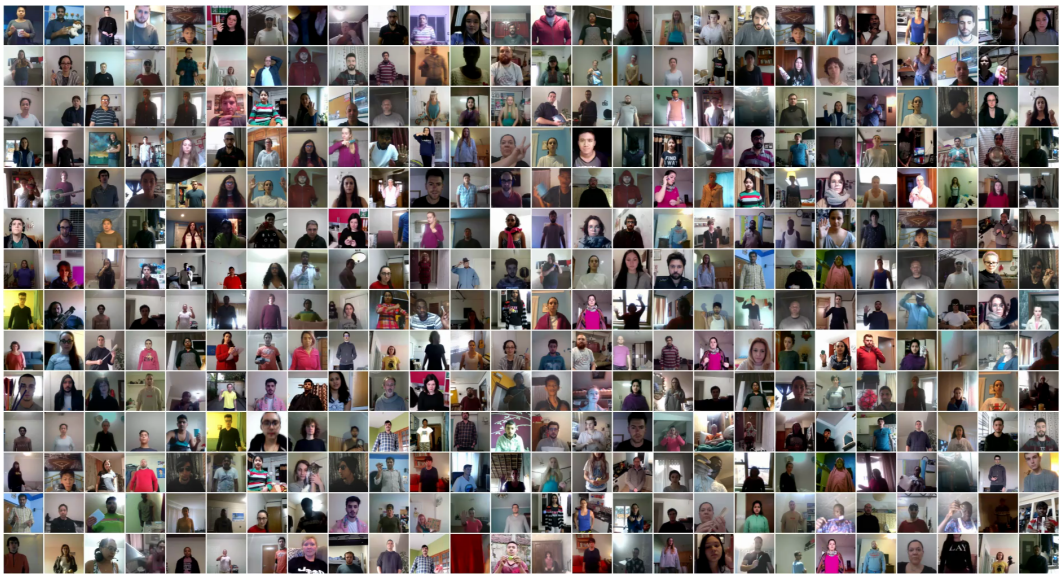

Figure B.1: Each image showcases a different video from our collection, demonstrating the substantial variation in visual scenarios captured within the dataset. These examples highlight the diversity of environments (indoor and outdoor settings), participants, objects, actions, lighting conditions, camera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

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.

B.3 Comparing IVD With Other Datasets

We examine whether IVD overlaps visually or semantically with prior video–QA corpora by embedding every clip with the *Cosmos-CV8×8×8* tokenizer (NVIDIA, 2025; NVIDIA et al., 2025) and measuring distances to clips from the closest public datasets: AVSD (DSTC7) (Alamri et al., 2018), and Social-IQ (Zadeh et al., 2019).For each video, we first normalise the frame-rate to 8 FPS

Table B.1: Nearest–neighbour L2 distances between IVD clips and each dataset.

Comparison	Mean	Median	Min	Max	5th Percentile
QIVD vs. QIVD	0.0157	0.0148	0.0031	0.1124	0.0062
QIVD vs. AVSD	0.0386	0.0369	0.0125	0.1238	0.0173
QIVD vs. Social-IQ	0.0894	0.0871	0.0257	0.2156	0.0458

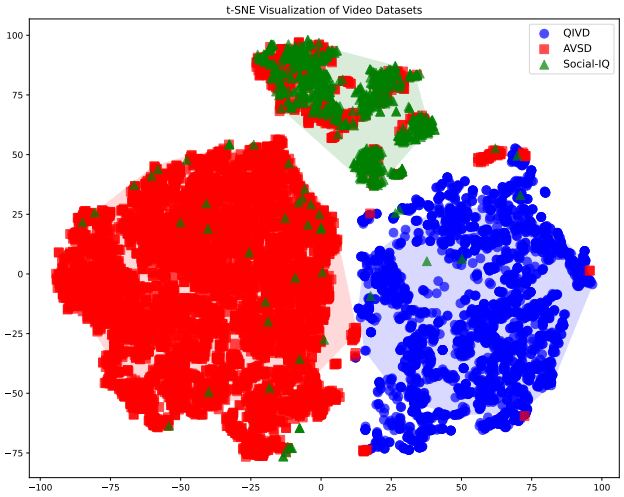

Figure B.2: Two-dimensional t-SNE projection of the 1024-dimensional embeddings for IVD (blue), AVSD (red), and Social-IQ (green). IVD clips form a tight, coherent cluster that is clearly separated from those of AVSD and Social-IQ, illustrating their distinct distributions in latent space.

and truncate (or zero-pad) to a maximum of 64 frames (8 seconds). The input frames are resized to 224×224 pixels. The resulting tensor, after batching and permuting to a $\mathcal{V} \in \mathbb{R}^{1\times3\times64\times224\times224}$ (Batch $\times$ Channels $\times$ Time $\times$ Height $\times$ Width) format, is passed through the tokeniser's encode method. This yields a continuous latent representation. We average these latent features across the temporal and spatial dimensions (dimensions 2, 3, and 4), obtaining a single embedding vector $\mathbf{e}$ per clip.

Table B.1 reports, for each clip in IVD, the L2 distance to its nearest neighbour within IVD itself as well as to AVSD (Alamri et al., 2018) and Social-IQ (Zadeh et al., 2019). The intra-IVD distances (mean 0.0157, 5th percentile 0.0062) are smaller than the inter-dataset distances: AVSD (mean 0.0386, 5th percentile 0.0173) and Social-IQ (mean 0.0894, 5th percentile 0.0458). Only a small fraction of IVD clips find their closest counterpart outside the test split indicating minimal overlap with prior benchmarks and underscoring that IVD brings substantially novel visual–semantic content.

Figure B.2 further illustrates this separation in a two-dimensional t-SNE projection of the 1024-dimensional embeddings: IVD points form a tight cluster on the right, clearly distinct from AVSD (red) and Social-IQ (green), which occupy disjoint regions. We demonstrate that IVD is significiantly different than other closet datasets.

B.4 When-to-Answer Statistics

Figure B.3 plots, for every clip, the temporal offset between the moment an answer becomes valid and the end of the video. Most questions are answerable within the last quarter of the clip, yet the long, asymmetric tail indicates that a non-trivial fraction require substantially earlier or later responses, confirming the need for models to reason over the full temporal span rather than assume a fixed "answer now" point.

We also quantify how often the correct answer becomes valid only after the question has finished. Because ground-truth end-of-question timestamps are unavailable, we use the end-of-question de-

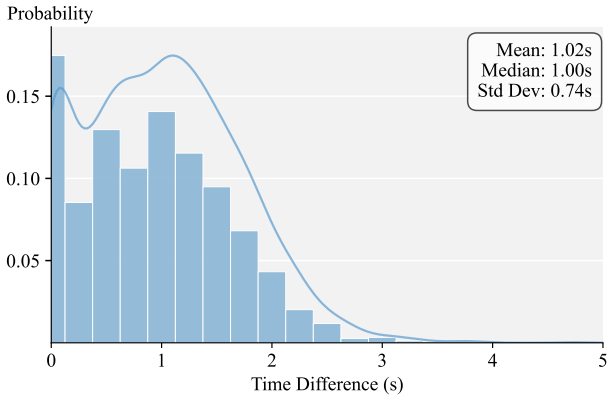

Figure B.3: Distribution of *optimal answer time* relative to the end of each clip. A value of x on the horizontal axis means the ground-truth "answer now" moment occurs x seconds before the video finishes.

Table B.2: Time difference statistics between ASR end-of-question and ground-truth when-to-answer.

Time Threshold	Count	Percentage
$\geq$ 0.0 s after	2054	100.0%
$\geq$ 0.5 s after	1351	65.8%
$\geq$ 1.0 s after	745	36.3%
$\geq$ 1.5 s after	375	18.3%
$\geq$ 2.0 s after	186	9.1%
$\geq$ 2.5 s after	90	4.4%
$\geq$ 3.0 s after	34	1.7%
$\geq$ 4.0 s after	5	0.2%

tected by Streaming-Whisper as a proxy. Table B.2 reports the count and fraction of clips whose ground-truth when-to-answer time occurs at least a given threshold after the ASR end-of-question time.

Figure B.4 shows the empirical distribution of Δt, the error of our streaming-ASR estimate relative to the human "when-to-answer" annotation. The skew toward negative values reveals a systematic tendency of the ASR system to answer questions prematurey often before sufficient visual context is available. Together, the figures highlight both the variability of optimal answer timing in real-world interactions and the practical challenge of detecting that moment reliably in a streaming setting.

C Additional Experimental Details

C.1 Development Environment

All experiments were conducted in PyTorch. Every open-source LMM checkpoint was loaded in half-precision (FP16), except for the 72B parameter VideoLLaMA2 (Zhang et al., 2023a) model, which was run with post-training INT8 quantization to satisfy memory limits. Inference code for each baseline was taken unmodified from the authors' public repositories and executed with the best-performing hyper-parameter settings provided by the authors. All of our experiments were run on a single A100-80 GB GPU.

C.2 Finetuning Details

We initialize from the publicly released VideoLLaMA 2.1-7B-AV (Zhang et al., 2023a) checkpoint and reuse the authors' training recipe with minimal modifications. The 2900 clips in IVD are partitioned into 5 non-overlapping folds via a deterministic hash of the video filename.Each fold in

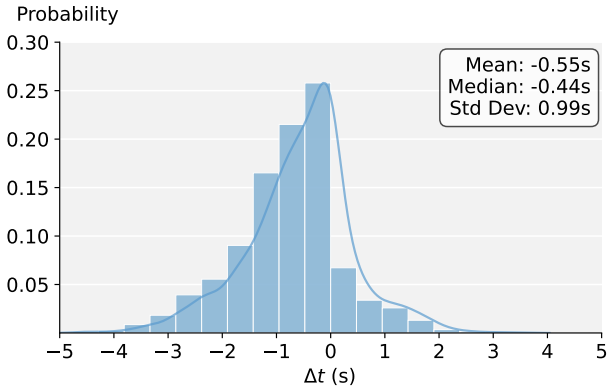

Figure B.4: Histogram of the signed error $\Delta t = t_{\mathrm{ASR}} - t_{\mathrm{GT}}$ between the streaming-ASR trigger and the human "answer now" annotation.

Table C.1: Trainable versus frozen modules during fine-tuning.

Module	Status
Vision encoder (SigLIP-SO400M/16F (Zhai et al., 2023))	❄ Frozen
Audio tower (BEATs iter-3 (Chen et al., 2022) + AS-2M)	🔥 Trainable
Multimodal projector (A)	🔥 Trainable
Multimodal projector (V)	❄ Frozen
LLM backbone (Qwen2-7B-Instruct (Yang et al., 2024))	🔥 Trainable

turn serves as validation, while the remaining four constitute the training split (∼2.32K clips). The architecture components updated during fine-tuning are summarized in Table C.1. We present all the hyperparameters in Table C.2.

C.3 LMM Evaluation

The prompts supplied to the LLM-based judge are reproduced in Table C.3 and Table C.4. Since *subjective* questions require a qualitatively different notion of correctness, we evaluate those cases with a dedicated prompt that deems responses acceptable provided they are contextually appropriate, friendly, and affirmatively phrased.

Table C.2: Hyper-parameters and optimisation settings for each cross-validation fold.

Hyper-parameter	Value
Training precision	bf16
Global batch size	8 videos (1×8)
Frames per clip	8
Epochs	2
Optimizer	AdamW (Loshchilov & Hutter, 2019)
Adam $(\beta_1, \beta_2, \varepsilon)$	$(0.9,\ 0.999,\ 10^{-8})$
Weight decay	0
Learning rate schedule	$2 \times 10^{-5} \rightarrow 0$ (cosine), 3% warm-up
Gradient accumulation steps	8
Gradient clip-norm	1.0
Distributed strategy	Deepspeed ZeRO-2 (parameter off-load) (Rajbhandari et al., 2020)

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)}`.

Table C.3: We use these prompts to evaluate the correctness of LMM-generated answers.

C.4 Using Different Video Sampling strategies

We study how restricting visual evidence to a short temporal window around the moment the question is spoken affects performance. For a given clip, we consider the segment spanning ± 0.5, ± 1.0, ± 2.0, ± 3.0, ± 4.0, or ± 5.0 seconds around the question timestamp and uniformly sample frames from that segment. We evaluate Qwen2.5-VL-7B under these settings as well as the full-video baseline. Results are reported in Table C.5. Consistent with the intuition that many questions require context before and after the utterance, very short windows harm performance. Wider windows recover accuracy, with ± 3–± 4 seconds yielding the best results; beyond ± 5 seconds, inputs effectively cover the full clip and performance plateaus.

C.5 LLM Judge Accuracy

To evaluate the accuracy of the LLM judge, we randomly select a subset of 300 samples from IVDand collect human ratings of GPT-4o's (Hurst et al., 2024) answers. In addition to the human evaluation, we use three automatic judges: LLaMA3-8B (et. al., 2024) and two recent Qwen3 (Yang et al.,

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 C.4: We use these prompts to evaluate the correctness of LMM-generated answers.

Table C.5: Effect of sampling frames within temporal windows around the question timestamp for Qwen2.5-VL-7B (Wang et al., 2024a). Values are proportions or similarity scores (higher is better).

Model	Corr. ↑	BERT ↑	METEOR ↑	BLEU ↑	ROUGE-L ↑
Qwen2.5-VL-7B (full video)	60.00	87.58	37.37	4.66	29.44
Qwen2.5-VL-7B (sampled around question) ± 0.5 s	38.28	78.90	25.40	2.60	20.30
Qwen2.5-VL-7B (sampled around question) ± 1.0 s	42.00	80.10	28.90	3.10	22.80
Qwen2.5-VL-7B (sampled around question) ± 2.0 s	60.41	87.00	42.02	4.60	29.42
Qwen2.5-VL-7B (sampled around question) ± 3.0 s	61.20	87.90	38.10	4.72	29.80
Qwen2.5-VL-7B (sampled around question) ± 4.0 s	60.75	88.01	37.80	4.68	29.55
Qwen2.5-VL-7B (sampled around question) ± 5.0 s	59.95	87.72	37.55	4.64	29.48

2025a) models (32B and 8B). The fraction of answers deemed correct by each evaluator is reported in Table C.6. Based on Table C.6, we use Qwen3-8B (Yang et al., 2025a) as the main judge throughout our experiments. We provide the results with LLaMA3-8B (et. al., 2024) as the judge in Table C.7.

C.6 GPT-4o PROMPT

To process 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.8, and this query is used to prompt GPT-4o.

Table C.6: Correctness of GPT-4o answers on a 300-sample subset under different evaluators. Values are the proportion marked correct.

Evaluator	Correctness
Human Evaluation	0.64
LLaMA3-8B (et. al., 2024)	0.68
Qwen3-32B (Yang et al., 2025a)	0.57
Qwen3-8B (Yang et al., 2025a)	0.59

Table C.7: Evaluation of baseline LMMs on the IVD dataset using (a) questions and estimated when-to-answer timestamps by Whisper (Radford et al., 2022) and (b) ground-truth questions and timestamps. Corr. represents correctness by LLM judge with LLaMA3-8B (et. al., 2024) as the judge.

	ASR Questions and Timestamps					Human Questions and Timestamps				
Model	Corr. ↑	BERT ↑	METEOR ↑	BLEU ↑	ROUGE-L ↑	Corr. ↑	BERT ↑	METEOR ↑	BLEU ↑	ROUGE-L ↑
Chat-UniVi (Jin et al., 2024)	39.69	89.94	37.47	6.08	28.45	45.10	90.50	40.02	7.24	31.22
InstructBLIP (Dai et al., 2023)	37.17	82.19	4.35	0.02	10.00	41.14	82.03	4.54	0.07	10.72
LLaMA-VID (Li et al., 2024c)	43.48	90.51	37.19	5.84	29.80	48.48	90.78	37.55	5.42	29.82
LLaVA-NeXT (Liu et al., 2024a)	24.97	85.29	22.85	1.38	11.64	28.90	85.78	24.50	1.67	13.22
Video-ChatGPT (Maaz et al., 2024)	35.38	90.53	38.13	7.58	31.08	40.76	91.01	40.59	9.07	33.58
VideoChat (Li et al., 2024a)	8.00	85.05	23.48	1.08	12.22	8.31	85.20	24.39	1.03	12.54
VideoChat2 (Li et al., 2024b)	46.07	91.13	45.49	11.35	41.38	53.07	91.52	47.93	12.43	43.87
Video-LLaVA (Zhu et al., 2023; Lin et al., 2023)	23.52	87.77	27.15	1.98	19.31	18.62	83.38	2.90	0.00	15.66
VideoLLaMA (Zhang et al., 2023a)	33.52	89.50	39.06	7.62	30.84	39.21	90.45	43.88	9.86	34.93
VideoLLaMA2-7B (Cheng et al., 2024)	44.31	91.18	47.20	13.93	40.63	52.69	91.71	51.08	16.41	43.97
VideoLLaMA2-72B (Cheng et al., 2024)	47.69	91.42	46.58	14.03	41.70	53.41	92.29	51.13	16.12	45.76
VideoLLaMA3-7B (Zhang et al., 2025)	52.31	90.92	45.20	11.21	40.54	59.62	91.63	48.56	12.72	43.84
VideoLLM-online (Chen et al., 2024)	–	–	–	–	–	25.48	88.67	33.73	4.16	26.27
Qwen2.5-VL-7B (Wang et al., 2024a)	53.55	87.17	34.95	3.88	26.52	60.00	87.58	37.37	4.66	29.44
Qwen2.5-Omni-7B (Xu et al., 2025)	44.76	86.65	33.45	2.77	20.57	46.97	86.73	33.98	2.87	20.98
GPT-4o (Hurst et al., 2024)	–	–	–	–	–	66.38	89.36	51.18	15.72	42.55
Human (subset)	–	–	–	–	–	89.00	93.01	53.21	17.40	49.76

C.7 GPT-4o REFUSAL CASES

: GPT-4o declines to answer 76 questions in IVD due to `ResponsibleAIPolicyViolation`. Given that the samples in 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 C.1.

C.8 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 C.2). Notably, these examples highlight the shortcomings of several advanced models, including the robust GPT-4o, the large-scale VideoLLaMMA2-72B (Zhang et al., 2023a), and even the fine-tuned VideoLLaMA2.1-7B-AV (Zhang et al., 2023a).

C.9 STATISTICAL SIGNIFICANCE

We report the standard deviation values corresponding to table 4 and table 5 in table C.9, table C.10, and table C.11.

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"{encoded_frame_1}",
          "detail": "high"
        }
      },
      {
        "type": "image_url",
        "image_url":
        {
          "url": f"{encoded_frame_2}",
          "detail": "high"
        }
      },
      {
        "type": "image_url",
        "image_url":
        {
          "url": f"{encoded_frame_3}",
          "detail": "high"
        }
      },
      {
        "type": "image_url",
        "image_url":
        {
          "url": f"{encoded_frame_4}",
          "detail": "high"
        }
      }
    ]
  }
]
```

Table C.8: The prompt used to run inference with GPT-4o.

Table C.9: ASR performance comparison.

Model	METEOR ↑	BLEU ↑	ROUGE-L ↑	$\Delta t \downarrow$
Whisper (Radford et al., 2022)	90.01 ± 23.11	80.95 ± 35.13	90.32 ± 22.66	-
Whisper-Streaming (Machácek et al., 2023)	92.34 ± 15.31	74.57 ± 33.52	91.82 ± 15.72	0.83 ± 0.77

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

Table C.10: Evaluation of baseline LMMs on the IVD dataset using questions and when-to-answer timestamps extracted by Whisper-Streaming (Radford et al., 2022). Corr. represents the correctness score calculated by the LLM judge.

Model	Corr. ↑	BERT ↑	METEOR ↑	BLEU ↑	ROUGE-L ↑
Chat-UniVi (Jin et al., 2024) (Jin et al., 2024)	34.66 ± 47.58	89.94 ± 3.56	37.47 ± 23.53	6.08 ± 16.44	28.45 ± 22.41
InstructBLIP (Dai et al., 2023)	35.03 ± 47.71	82.19 ± 3.0	4.35 ± 6.53	0.02 ± 0.73	9.99 ± 14.4
LLaMA-VID (Li et al., 2024c)	39.41 ± 48.88	90.51 ± 3.56	37.18 ± 23.25	5.84 ± 16.39	29.8 ± 22.03
LLaVA-NeXT (Liu et al., 2024a)	19.45 ± 39.53	85.29 ± 3.24	22.85 ± 15.72	1.38 ± 8.68	11.64 ± 15.21
Video-ChatGPT (Maaz et al., 2024)	32.45 ± 46.83	90.53 ± 3.78	38.14 ± 24.78	7.58 ± 19.46	31.09 ± 24.45
VideoChat (Li et al., 2024a)	3.69 ± 18.82	85.05 ± 2.77	23.48 ± 15.29	1.08 ± 6.47	12.22 ± 12.29
VideoChat2 (Li et al., 2024b)	44.66 ± 49.72	91.13 ± 3.88	45.49 ± 26.63	11.35 ± 23.38	41.38 ± 26.04
Video-LLaVA (Zhu et al., 2023; Lin et al., 2023)	20.28 ± 40.20	87.77 ± 3.37	27.15 ± 18.88	1.98 ± 9.73	19.31 ± 17.63
VideoLLaMA (Zhang et al., 2023a)	30.76 ± 46.12	89.5 ± 4.56	39.05 ± 26.06	7.62 ± 18.87	30.84 ± 24.83
VideoLLaMA2-7B (Cheng et al., 2024)	43.34 ± 49.56	91.18 ± 4.18	47.2 ± 27.92	13.93 ± 26.57	40.63 ± 27.22
VideoLLaMA2-72B (Cheng et al., 2024)	46.52 ± 49.88	91.42 ± 5.68	46.6 ± 28.88	14.04 ± 27.41	41.71 ± 28.5
VideoLLaMA3-7B (Zhang et al., 2025)	50.59 ± 50.01	90.92 ± 5.34	45.2 ± 27.14	11.21 ± 23.54	40.55 ± 26.55
Qwen2.5-Omni-7B (Xu et al., 2025)	43.97 ± 49.63	86.65 ± 1.95	33.45 ± 17.12	2.77 ± 5.94	20.57 ± 12.71
Qwen2.5-VL-7B (Wang et al., 2024a)	44.90 ± 49.75	87.17 ± 2.71	34.95 ± 20.21	3.89 ± 10.62	26.52 ± 23.25

Table C.11: Evaluation of baseline LMMs on the IVD dataset using ground-truth questions and timestamps. Corr. represents correctness by LLM judge.

Model	Corr. ↑	BERT ↑	METEOR ↑	BLEU ↑	ROUGE-L ↑
Chat-UniVi (Jin et al., 2024)	40.79 ± 49.14	90.5 ± 3.49	40.02 ± 23.64	7.24 ± 18.29	31.22 ± 22.7
InstructBLIP (Dai et al., 2023)	39.14 ± 48.80	82.03 ± 3.13	4.54 ± 6.81	0.07 ± 1.7	10.72 ± 14.56
LLaMA-VID (Li et al., 2024c)	43.00 ± 49.51	90.78 ± 3.32	37.55 ± 22.42	5.42 ± 15.59	29.82 ± 21.12
LLaVA-NeXT (Liu et al., 2024a)	22.66 ± 41.81	85.78 ± 3.4	24.5 ± 16.66	1.67 ± 9.53	13.22 ± 16.54
Video-ChatGPT (Maaz et al., 2024)	36.59 ± 48.16	91.01 ± 3.78	40.59 ± 25.2	9.07 ± 21.51	33.58 ± 25.11
VideoChat (Li et al., 2024a)	3.52 ± 18.56	85.2 ± 2.72	24.39 ± 15.51	1.03 ± 5.52	12.54 ± 12.11
VideoChat2 (Li et al., 2024b)	50.35 ± 50.01	91.52 ± 3.81	47.93 ± 26.62	12.43 ± 24.04	43.87 ± 25.97
Video-LLaVA (Zhu et al., 2023; Lin et al., 2023)	15.00 ± 35.72	83.38 ± 1.85	2.9 ± 5.27	0.0 ± 0.0	15.66 ± 16.0
VideoLLaMA (Zhang et al., 2023a)	35.93 ± 47.97	90.45 ± 4.15	43.88 ± 25.81	9.86 ± 21.99	34.93 ± 25.09
VideoLLaMA2-7B (Cheng et al., 2024)	50.07 ± 50.01	91.71 ± 4.15	51.08 ± 27.91	16.41 ± 28.98	43.97 ± 27.56
VideoLLaMA2-72B (Cheng et al., 2024)	50.83 ± 50.00	92.29 ± 4.35	51.13 ± 27.95	16.12 ± 28.86	45.76 ± 28.06
VideoLLaMA3-7B (Zhang et al., 2025)	56.38 ± 49.60	91.63 ± 4.24	48.56 ± 26.81	12.72 ± 24.92	43.84 ± 26.11
Qwen2.5-Omni-7B (Xu et al., 2025)	45.90 ± 49.84	86.73 ± 1.93	33.98 ± 17.22	2.87 ± 5.96	20.98 ± 12.71
Qwen2.5-VL-7B (Wang et al., 2024a)	50.62 ± 50.01	87.58 ± 2.63	37.37 ± 20.46	4.66 ± 11.67	29.44 ± 24.18
GPT-4o (Hurst et al., 2024)	58.76 ± 49.25	89.36 ± 15.25	51.18 ± 27.32	15.72 ± 28.27	42.55 ± 28.17
Human (subset)	89.00 ± 33.32	93.01 ± 3.89	53.21 ± 25.22	17.4 ± 30.9	49.76 ± 25.18

Figure C.2: Simple daily face-to-face questions that strong baseline LMMs such as GPT-4o, VideoLLaMMA2-72B, and VideoLLaMA2.1-7B-AV fail to answer.