ICLR

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

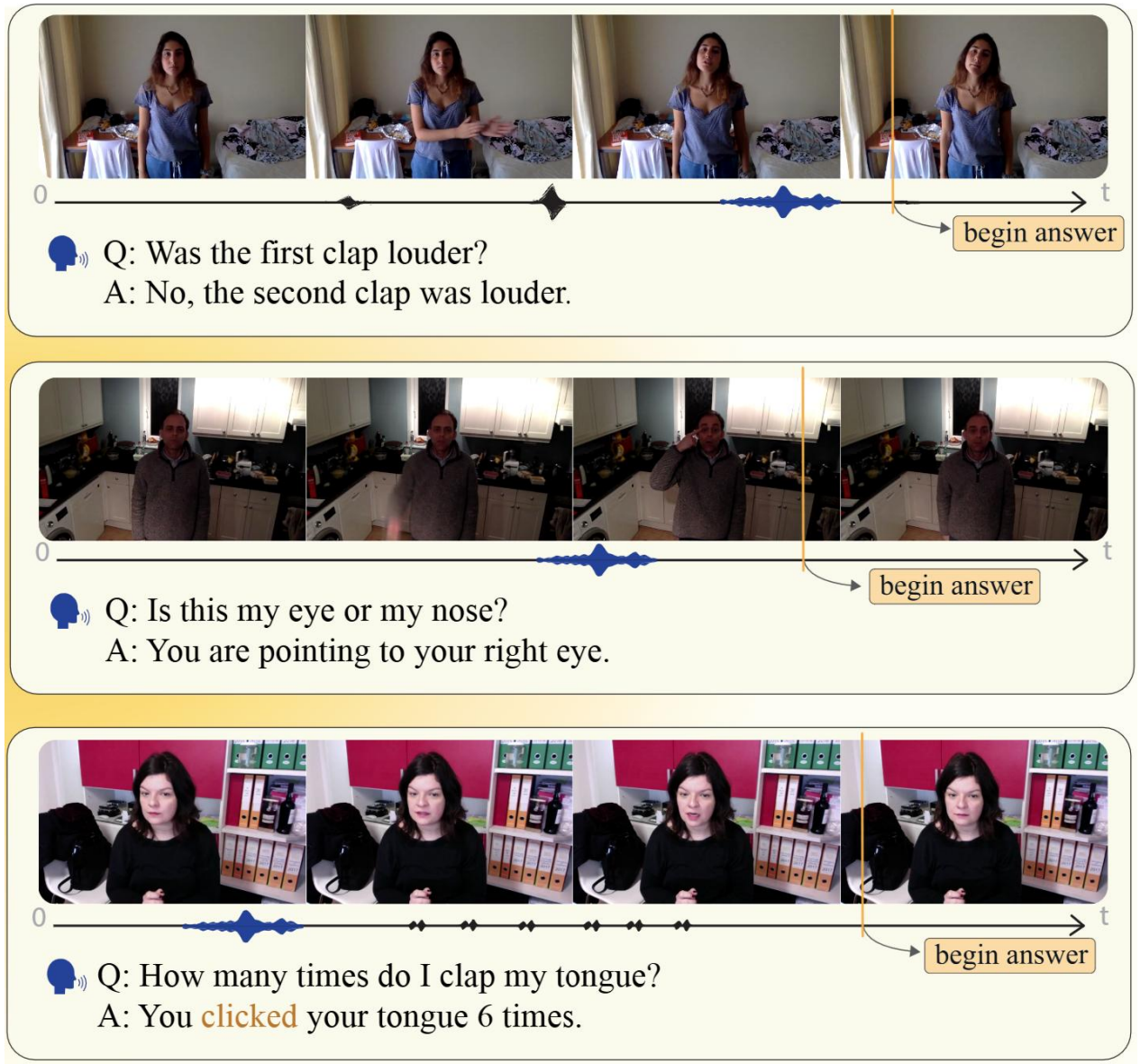
Reza Pourreza*, Rishit Dagli*, Apratim Bhattacharyya,
Sunny Panchal, Guillaume Berger, Roland Memisevic

Qualcomm AI Research

University of Toronto

Qualcomm Interactive Video Dataset (QIVD)

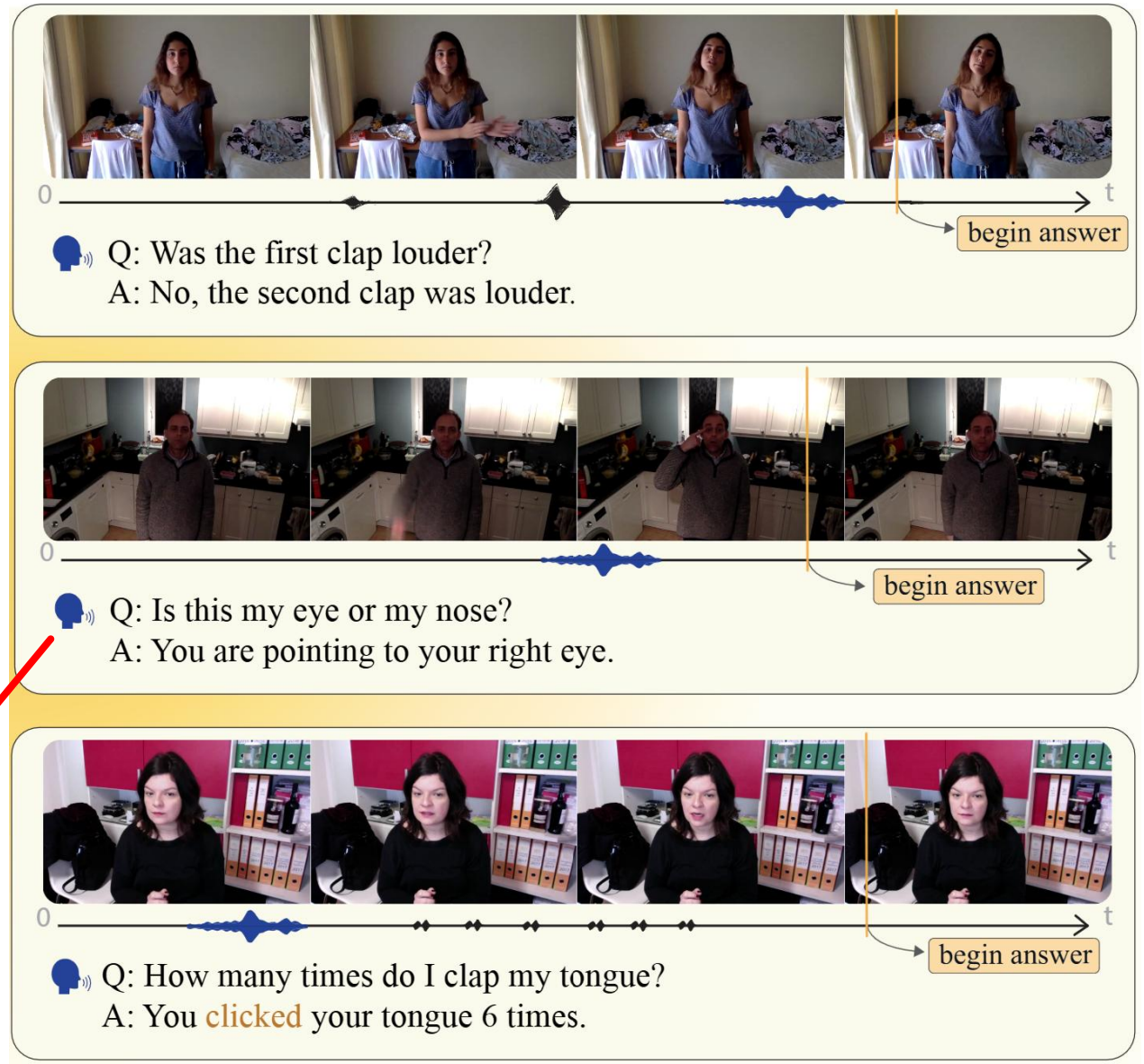
- We ask: Can MLLMs converse with users in real-time about scenes and events that are unfolding live in front of the camera?
- Users speak a question about the audio-visual content
- Models get access to a streaming camera and audio input
- System must see, listen, and respond appropriately in real-time



Qualcomm Interactive Video Dataset (QIVD)

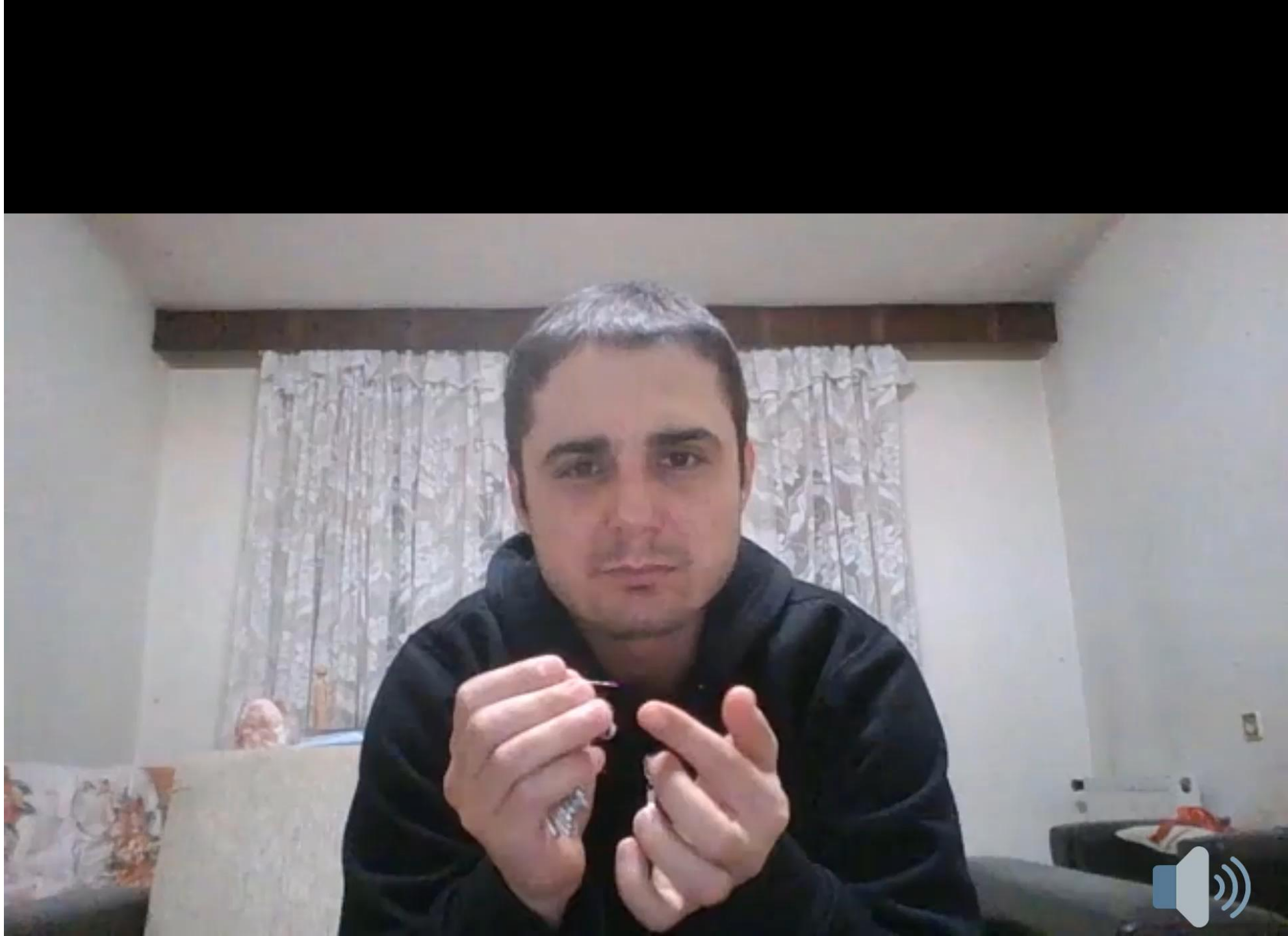
- We ask: Can MLLMs converse with users in real-time about scenes and events that are unfolding live in front of the camera?
- Users speak a question about the audio-visual content
- Models get access to a streaming camera and audio input
- System must see, listen, and respond appropriately in real-time

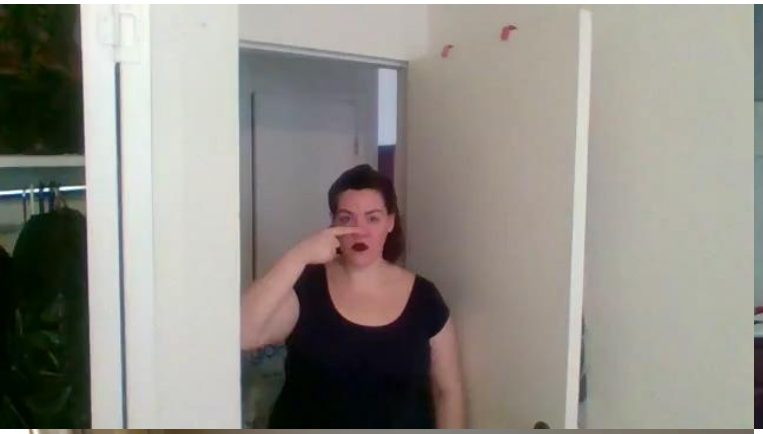
Audio-visual information maybe present before or during the process of asking the question



The information needed to answer the question was present when the question was asked

Thus, answer the question right after it has been spoken






Deictic References

Statistic	Value
Avg. Answer Timestamp	81.47% (±13.89)
Avg. FPS	30 (±0.00)
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

Most Relevant Information May Lie After the Question

Relevant audio-visual information may appear after asking the question

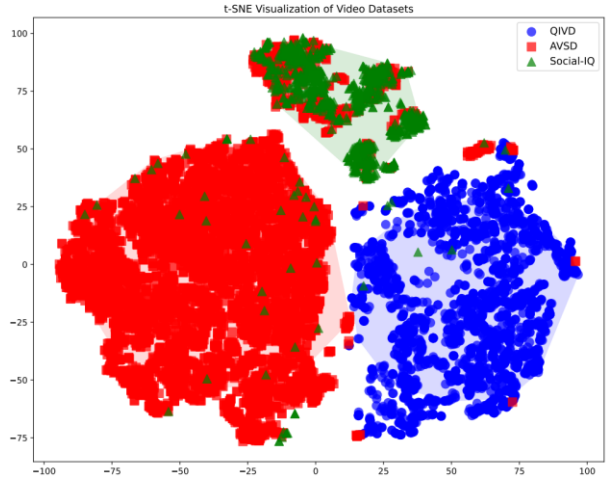
We are confident we have the needed information now


QIVD includes diverse data
with varying actions, lighting, views, backgrounds, etc.

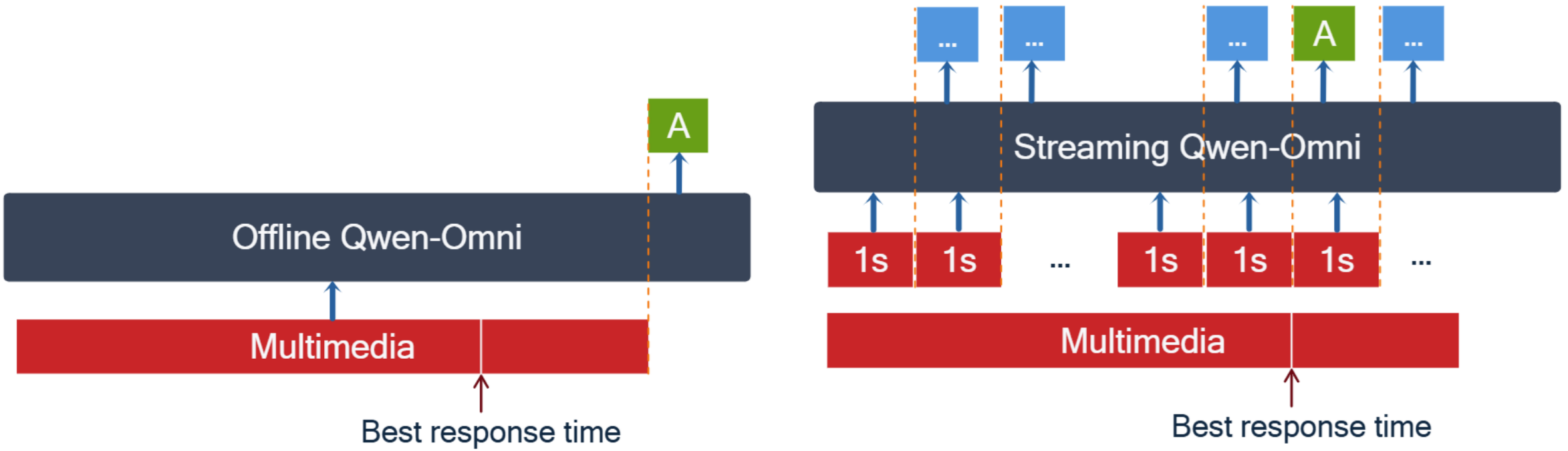
QIVD vs Other Datasets

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	✓	✗	✓	✗
VStream-QA (Zhang et al., 2024)	32	3500	Automatic	✓	✗	✓	✗
QIVD	2900	2900	Manual	✓	✓	✓	✓

Two-dimensional t-SNE projection of the 1024-dimensional embeddings for representative datasets AVSD, and Social-IQ versus QIVD,



Modifying MLLMs for Our Setting



"time_to_answer": "3.8s",
"answer": "You are holding a Rubik's cube in your left hand."

→

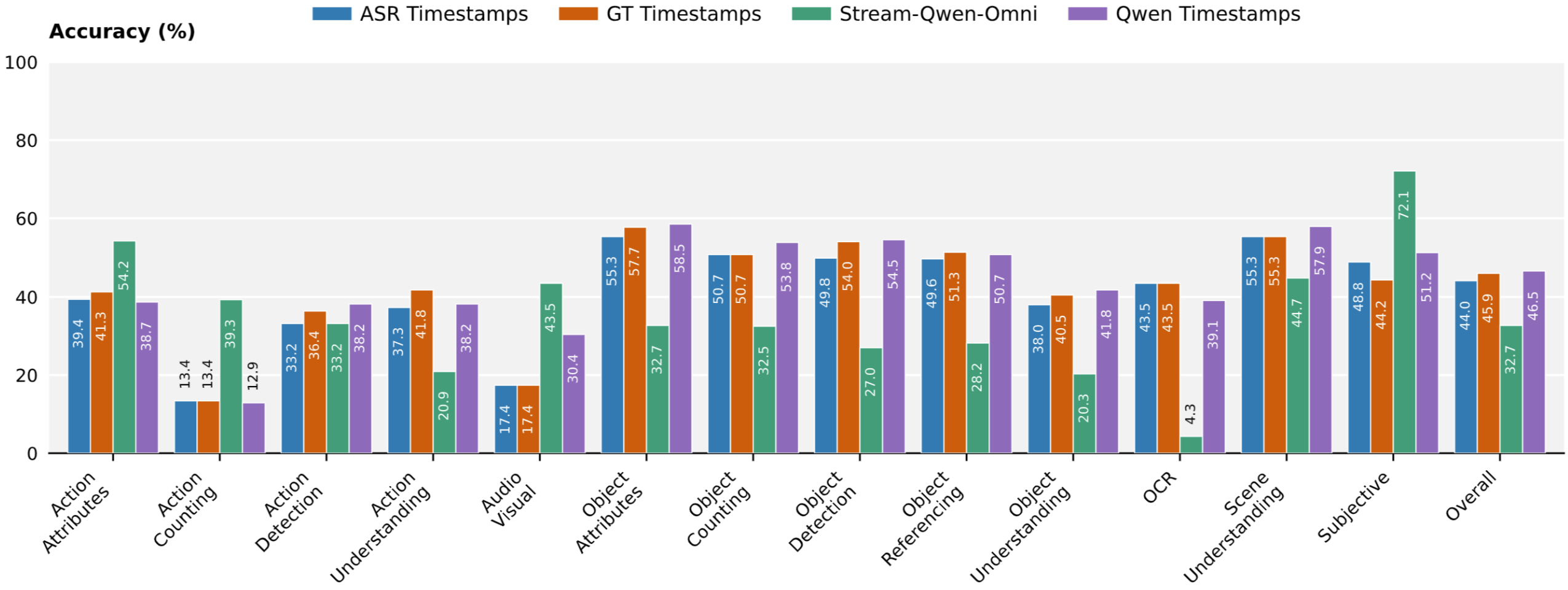
[0.0s, 1.0s, "..."],
[1.0s, 2.0s, "..."],
[2.0s, 3.0s, "..."],
[3.0s, 4.0s, "You are holding a Rubik's cube in your left hand."],
[4.0s, 5.0s, "..."],
[5.0s, 6.0s, "..."]

Comparing Models

Model	ASR Questions and Timestamps					Human Questions and Timestamps				
	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)	17.97	76.60	27.36	2.81	20.39	23.62	88.45	33.08	3.99	25.26
Flash-VStream (Zhang et al., 2024)	44.28	89.85	28.95	4.17	27.05	49.59	90.48	30.79	5.05	29.90
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
Qwen3-VL-8B (Yang et al., 2025a)	53.72	87.08	33.9	5.29	31.53	60.07	87.58	36.72	6.64	35.89
Gemini-2.5-Flash (Comanici et al., 2025)	–	–	–	–	–	58.07	90.43	43.07	8.33	36.05
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

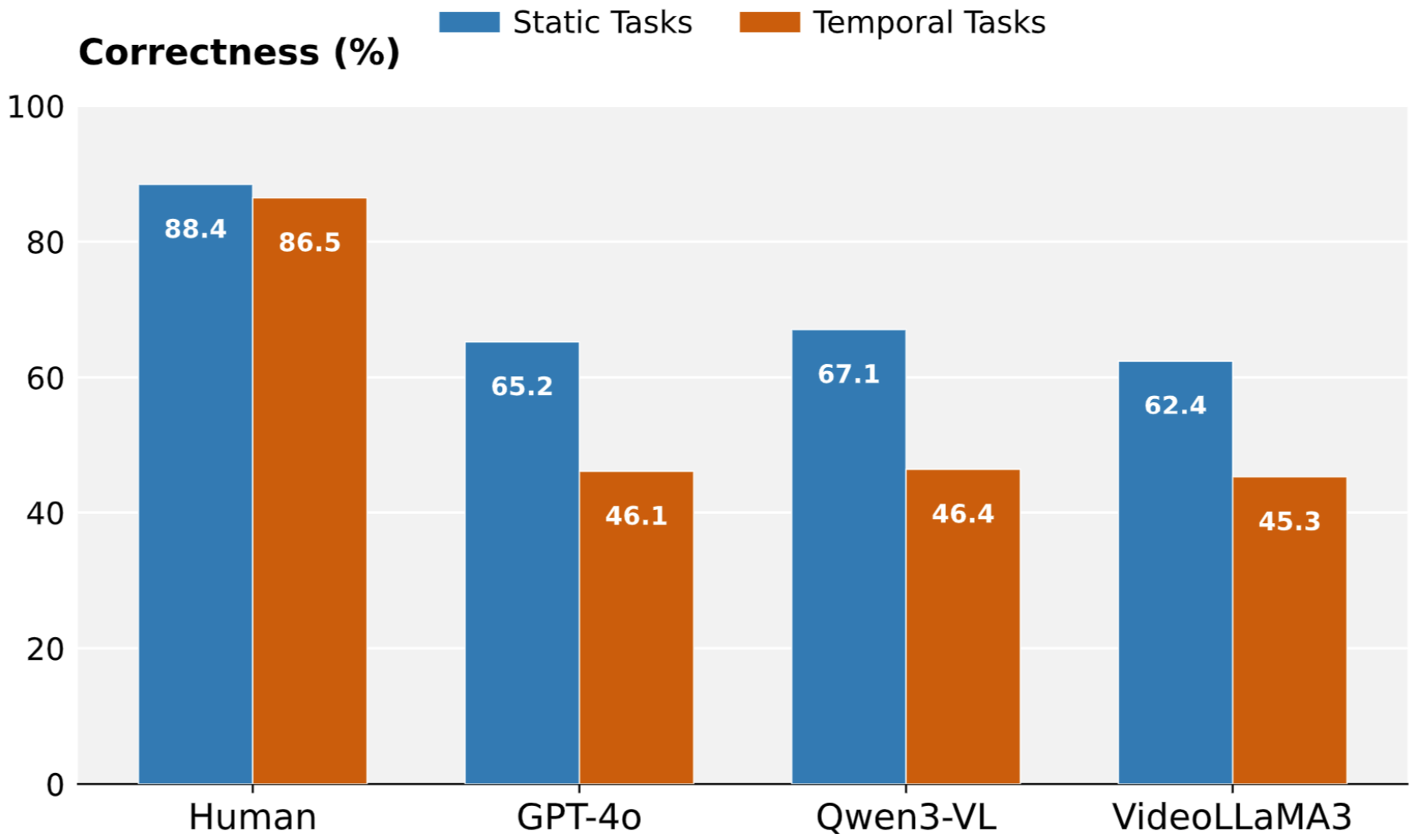
❖ Corr. denotes correctness measured by an LLM judge (Qwen3-32B)

Performance of Stream-Qwen-Omni by Categories

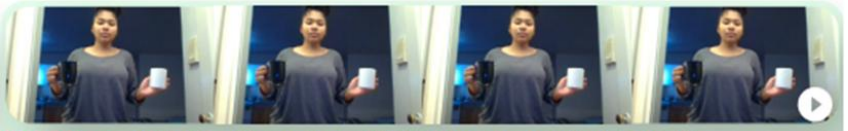


Performance of SoTA MLLMs Compared with a non-expert Human

Performance of SoTA MLLMs



Gap in Capabilities of SoTA MLLMs



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

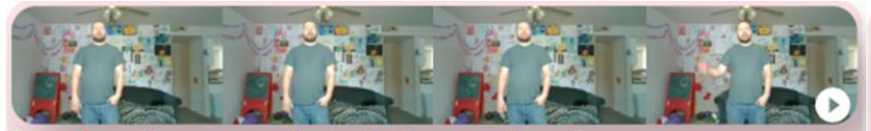
GPT-4o: The color of the mug in your left hand is black.

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 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: 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

Summary

- Introduce QIVD, a benchmark and dataset for training and evaluating LMMs (video, audio, language) in real-time human interaction tasks.
- Propose Stream-Qwen-Omni allowing MLLMs to adapt to online streaming settings
- Tests situated audio-visual understanding without requiring multi-hop conversation or domain-specific knowledge.
- Dataset uses a simple Q&A paradigm, avoiding complex reasoning but still poses significant challenges for current models.
- Extensive experiments highlight key limitations in existing multimodal models.

Limitations and Future Directions

- Small class sizes, limiting diversity of questions and answers.
- Controlled recording environments, reducing variability in lighting, background, and camera angles.
- Demographic biases (gender, age, ethnicity) could impact performance across diverse user groups.