JOINTAVBENCH: A Benchmark for Joint Audio-Visual Reasoning Evaluation

Jianghan Chao, Jianzhang Gao, Wenhui Tan, Yuchong Sun, Ruihua Song, Liyun Ru

Code and data are available on: **jointavbench.github.io**

百川智能 BAICHUAN AI
中国人民大学 RENMIN UNIVERSITY OF CHINA

Introduction of JointAVBench

- Existing AV benchmarks usually lack at least one of three essentials: **strict AV dependency**, **diverse audio signals**, or **multi-scene reasoning**.
- JointAVBench targets the real Omni-LLM failure mode: **the answer should break if either audio or video is removed**.

Examples

Case 1: Cross-scene Plot Reasoning

- **Multi-scene speech + visual grounding**
- Hard for current models

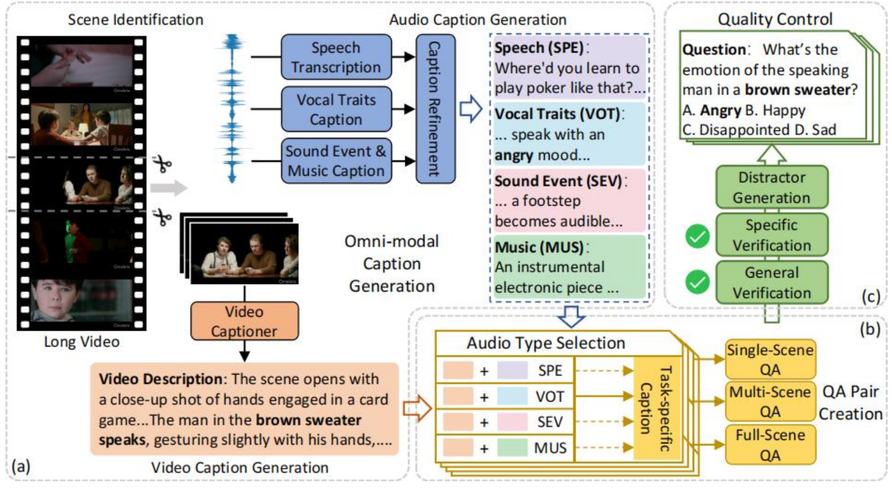
~20% drop on multi-scene tasks

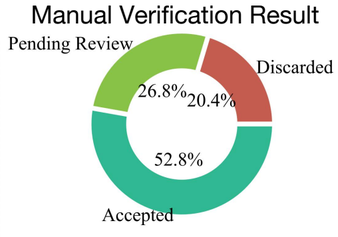
Case 2: Speaker Emotion & Vocal Traits

- Requires **audio + visual fusion**
- Weakness in **abstract audio understanding**

Major Weakness in emotion & speech understanding

Benchmark Construction





Stage 1. Generate scene captions plus audio-specific captions.
Stage 2. Build QA pairs with only the required modality and scene span.
Stage 3. Verify, add distractors, and keep only approved samples.

Taxonomy

Benchmark/Dataset	Avg. Duration	#QA	Anno. Method	Modality	#Tasks	#Audio Types	AV Corr. Ratio
Video Benchmarks/Datasets							
EgoSchema (Mangalam et al., 2023)	180s	5,063	A+M	V	1	-	0
Video-MME (Fu et al., 2025a)	1,017.9s	2,700	M	V	12	-	0
MVBench (Li et al., 2024b)	16.0s	4,000	A	V	20	-	0
LVBench (Wang et al., 2024b)	4,101s	1,549	M	V	6	-	0
MMBench-Video (Fang et al., 2024)	165.4s	1,998	M	V	26	-	0
Audiovisual Benchmarks/Datasets							
Music-AVQA (Li et al., 2022a)	60s	45,867	M	V&A	9	1	56.7%
OmniBench (Li et al., 2024c)	-	1,142	M	I&A	8	3	100%
AV-Odyssey (Gong et al., 2024)	-	4,555	M	V/I&A	26	3	100%
LongVALE (Geng et al., 2024)	235s	-	A+M	V&A	3	3	76.2%
AVUT (Yang et al., 2025)	67.8s	13,774	A+M	V&A	8	2	77.8%†
WorldSense (Hong et al., 2025)	141.1s	3,172	M	V&A	26	3	62.9%
JointAVBench (ours)	97.2s	2,853	A+M	V&A	15	4	100%

Key Design Dimensions:
5 cognitive dimensions
4 audio types
3 scene spans
Finalizing 15 tasks with **2,853 manually verified MCQs** of strict AV correlation.

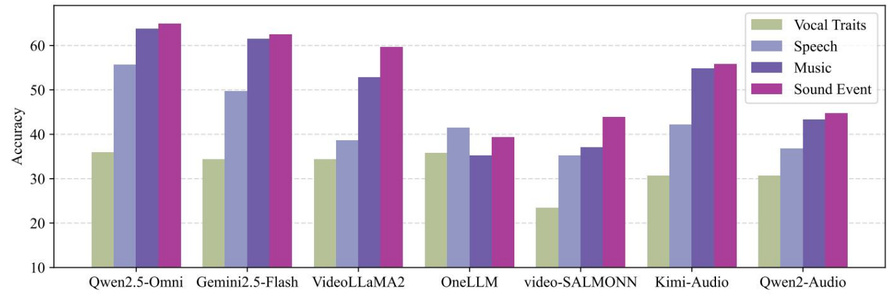
Evaluation & Main Findings

Overall Performance

Model Type	Model	Avg Accuracy (%)
Omni-LLMs	Gemini-2.5-Pro	**62.6**
	Qwen3-Omni (30B)	62.1
Video-LLMs	InternVL-2.5 (8B)	51.3
	VideoLLaMA3 (7B)	49.9
Audio-LLMs	Kimi-Audio (7B)	45.9
	Qwen2-Audio (7B)	40.0

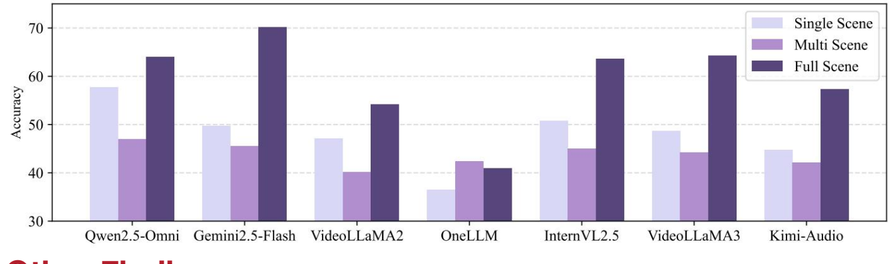
- Best Omni-LLM only achieves **62.6%** accuracy, indicating **large room for improvement**
- Omni-LLMs outperform Video-LLMs and Audio-LLMs, indicating importance of modality fusion

Audio Type Breakdown



Models struggle with abstract audio (emotion, speech)

Scene Complexity Analysis



Cross-scene reasoning is a major bottleneck

Other Finding

Weakness in **emotion & spatial reasoning**