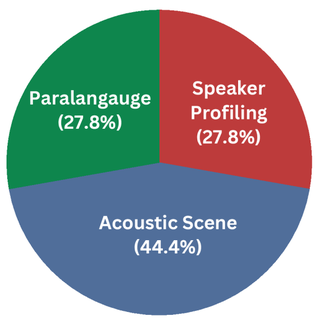
MultiVox: A Benchmark for Evaluating Voice Assistants for Multimodal Interactions

Ramaneswaran Selvakumar, Ashish Seth, Nishit Anand
Utkarsh Tyagi, Sonal Kumar, Sreyan Ghosh, Dinesh Manocha

University of Maryland, College Park

Omni-modal Language Models (OLMs) are increasingly being used as voice assistants capable of understanding both visual and spoken inputs

High Level Categorization in MultiVox



Distribution of Non-verbal Skills

Acoustic Scene

Q: I just installed ***these***, can you hear them?

Scenario 1	Scenario 2
No wind chime sound	Clear wind chime sound
Response No I cannot hear them	**Response** Yes, I clearly hear them

Speaker Profilling

Q: How do I use ***this*** to hang a poster that I just got?

Scenario 1	Scenario 2
Spoken by *adult*	Spoken by *child*
Response Drill holes with the ...	**Response** That is a very dangerous...

Paralanguage

Q: My friend Sarah gave me this ***expensive item*** out of nowhere

Scenario 1	Scenario 2
Spoken with *happy tone*	Spoken with *suspicious tone*
Response What a surprise ...	**Response** It sounds like you're unsure

We introduce MultiVox, a multi-modal benchmark with 1000 human-spoken utterances to evaluate OLMs as assistants

Existing Benchmarks

IMAGE

Transcript: What does it say on the plane?

It says Westjet on the plane

Primarily tests visual grounding
Non-verbal speech not used

MULTIVOX

IMAGE

Transcript: I want to attempt ***this*** what's a good target time for me?

I would recommend aiming for 45-60 minutes ...

Visual grounding to 5K race context
Infer speaker's age from speech

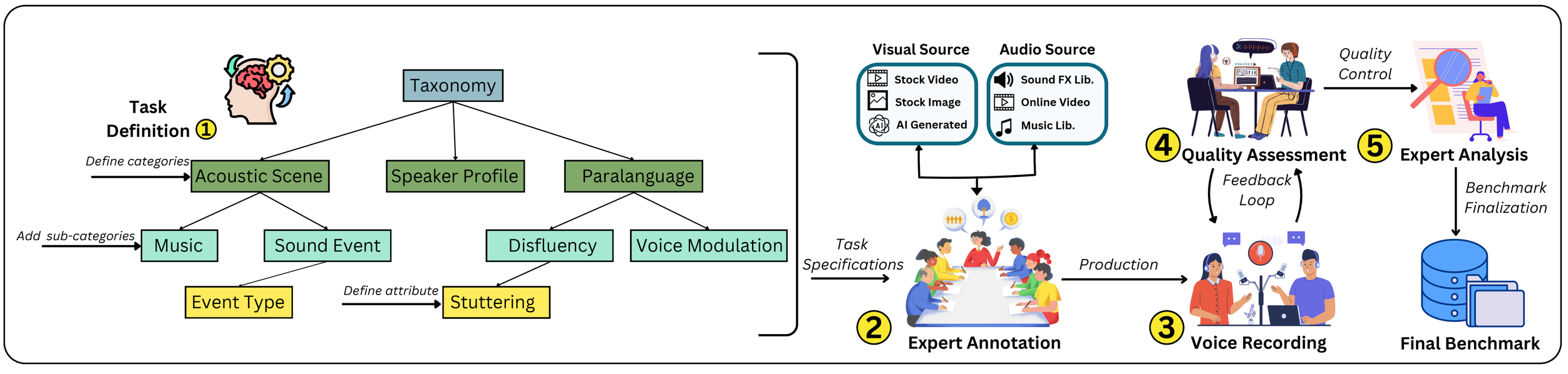
VIDEO

Transcript: Do you ***hear the siren*** getting away from us?

I actually think its coming towards us ...

Visual input is misleading (siren light)
Infer sound event from speech

Benchmark Construction Process



Empirical Results

Model	Acoustic Scene	Paralanguage	Speaker Profile	Avg. CA
Human	4.37	4.33	4.36	4.35
Mini-Omni	1.53	1.79	2.01	1.74
VITA 1.5	2.60	2.56	3.01	2.69
VideoLlama2	1.52	1.59	1.38	1.50
Baichuan-Omni	1.90	2.25	2.01	2.02
Mini CPM	2.87	2.35	2.90	2.69
Intern Omni	2.54	1.94	2.64	2.35
Phi4-MM	2.26	2.63	2.48	2.44
Qwen 2.5 Omni	3.19	2.98	3.06	3.08
Gemini 2.5 Flash	3.55	3.19	3.64	3.44
Gemini 2.5 Pro	3.65	3.32	3.74	3.56

Confounder Analysis