AURELIA: Test-time Reasoning Distillation in Audio-Visual LLMs

Sanjoy Chowdhury*, Hanan Gani*, Nishit Anand, Sayan Nag, Ruohan Gao,
Mohamed Elhoseiny, Salman Khan, Dinesh Manocha

Introduction and Motivation

ICCV OCT 19-23, 2025 HONOLULU HAWAII

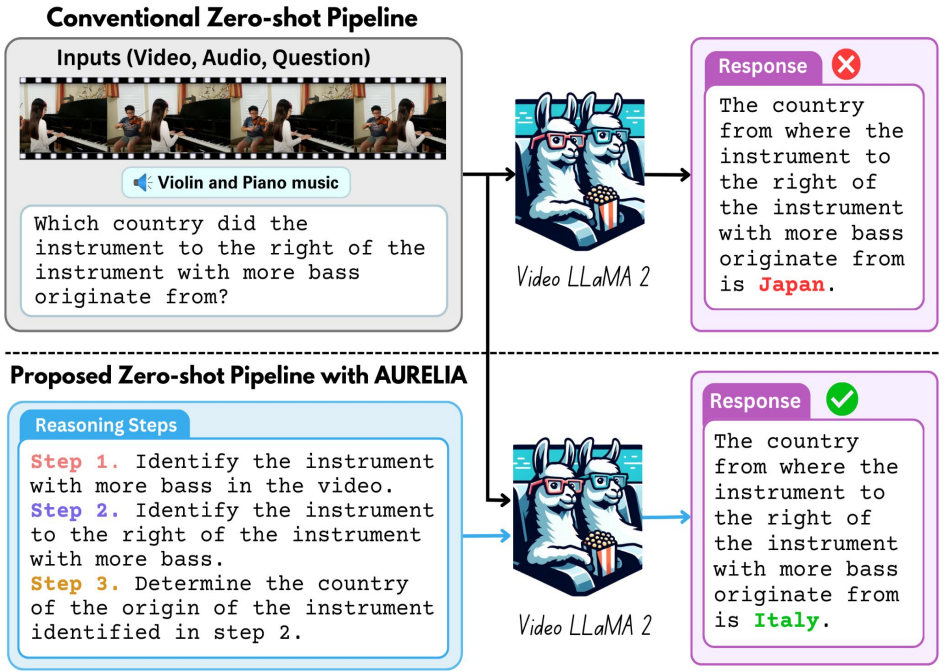
- Audio-Visual (AV) reasoning is crucial for capturing abstract nuances that text or images alone cannot convey.
- Current AVLLMs are prone to cultural, contextual, and perceptual biases in training data, relying on dominant visual or auditory cues over true reasoning.
- Recent advances in reasoning remain largely unexplored for AV models.

Contributions

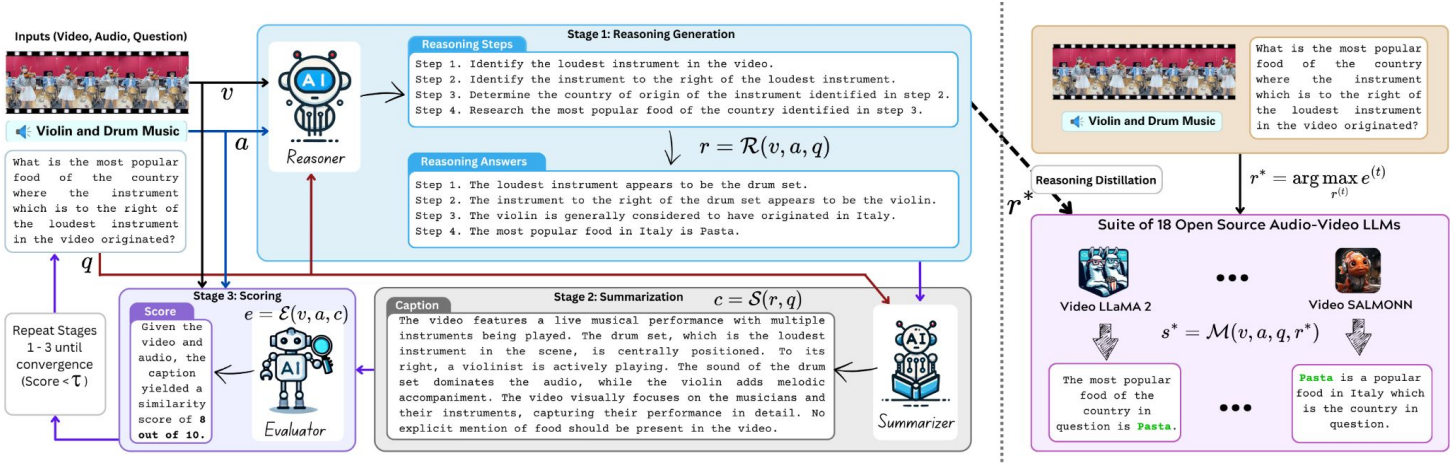
- We introduce **AURELIA**, a test-time multi-agent reasoning distillation framework for addressing challenges in audio-visual cross-modal comprehension by mitigating visual and auditory biases without the need for additional training
- Using our pipeline, we build **AVReasonBench** with 4,500 samples across six tasks, plus new *AV-GeoIQ*, *AV-Compositional,* and *AV-Meme* subsets.

Overview of AURELIA

- AURELIA systematically breaks down the problem into intermediate reasoning steps, guiding the model toward more accurate and interpretable answer.



AURELIA Framework



- **Multi-Agent Framework:** AURELIA employs an interactive Actor-Critique setup where a **Reasoning Generator** proposes structured, step-by-step reasoning based on input audio, video, and query. A **Summarizer** Agent condenses these steps into a coherent caption, and an **Evaluator** Agent scores how well this reasoning aligns with the given AV input. **Feedback** refines the reasoning iteratively until it surpasses a quality threshold or exhausts iterations.

AURELIA Framework: Properties

- **Zero-Shot Reasoning Distillation:** Rather than requiring new training or fine-tuning, AURELIA distills the refined reasoning steps into AVLLMs at inference time, enabling improved reasoning without modifying model weights.
- **Bias Mitigation via Structured Reasoning:** By breaking the problem into logical steps, AURELIA helps AVLLMs move away from relying on superficial visual or auditory cues, addressing cultural and perceptual biases.

AVReasonBench

- Leveraging AURELIA, we build **AVReasonBench**, a benchmark with 4,500 audio-visual question–answer samples, each paired with detailed reasoning. It includes six task categories such as commonsense, music understanding, humor, and a new **AV-GeoIQ** for geo-cultural reasoning.

Task ID	Question Category	Task Name	Class	Number
1	Country Recognition	AV-GeoIQ	17	21
2	Famous Landmark	AV-GeoIQ	18	23
3	Popular Dish/Food	AV-GeoIQ	16	19
4	Currency	AV-GeoIQ	12	13
5	Continent	AV-GeoIQ	5	17
6	Flag Specifics	AV-GeoIQ	10	15
7	Popular Dance Form	AV-GeoIQ	N/A	20
8	Geographical	AV-GeoIQ	N/A	31
9	Language	AV-GeoIQ	11	13
10	Commonsense Reasoning	AV-GeoIQ, AV-Meme, AV-Dance Match	N/A	165
11	Musical Performances	Music-AVQA, AV-GeoIQ	N/A	1014
12	Dynamic Scene	AVSD	N/A	931
13	Meme and Humor	AV-Meme	N/A	50
14	Dance Performances	AV-Dance Match	N/A	100
15	Indoor/Kitchen Scenarios	VALOR	N/A	945
16	Compositional	AV-Comp	N/A	968
17	Miscellaneous	AV-GeoIQ, AVSD, VALOR	N/A	159

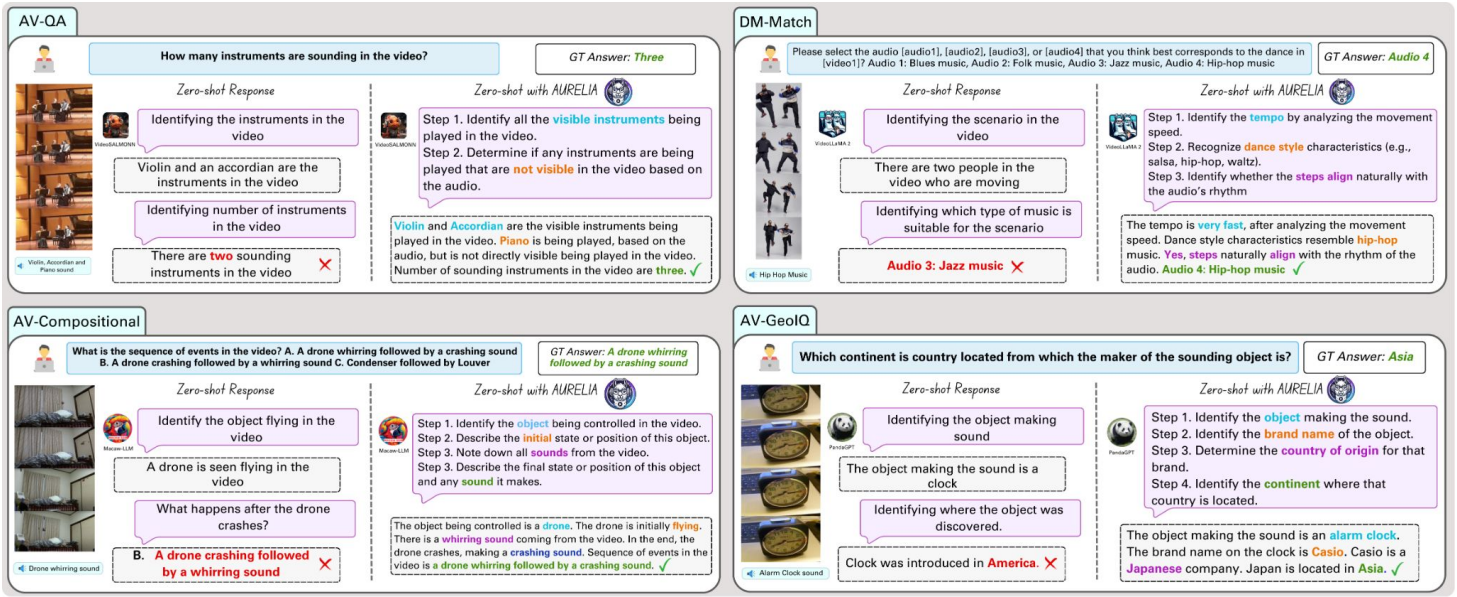
Experimental Results: Quantitative

- By leveraging the zeroshot reasoning distillation through AURELIA, we observe consistent boost in the performance of all the AVLLMs

Models	AV-QA		AV-Captioning	AV-Compositional	AV-GeoIQ	AV-Meme	DM-Match
	Music-AVQA	AVSD					
Closed-Source Models							
Gemini 1.5 Pro	70.6 / 68.9	74.7 / 72.5	84.9 / 82.7	38.9 / 36.8	71.2 / 68.0	52.0 / 49.0	43.4 / 41.5
Reka Core	67.9 / 64.3	74.5 / 69.5	83.2 / 80.4	38.6 / 35.3	45.7 / 42.5	24.0 / 19.0	35.8 / 32.5
Open-Source Models in ZS							
PandaGPT (13B)	35.8 / 33.7	29.1 / 26.1	67.8 / 64.7	28.8 / 24.1	17.2 / 12.5	25.0 / 21.0	30.2 / 27.0
Macaw-LLM (7B)	34.7 / 31.8	38.4 / 34.3	67.7 / 65.9	26.1 / 24.3	17.2 / 14.0	18.0 / 14.0	24.5 / 20.0
VideoLLaMA (7B)	39.1 /36.6	40.0 / 36.7	68.4 / 66.2	28.8 / 25.8	19.3 / 16.5	18.0 / 16.0	26.6 / 23.0
ImageBind-LLM	44.2 / 43.9	42.7 / 39.2	69.0 / 66.9	28.8 / 25.4	18.0 / 13.0	17.7 / 15.0	26.2 / 22.5
X-InstructBLIP (13B)	47.8 / 44.5	43.9 / 40.1	69.5 / 66.1	27.5 / 25.9	27.6 / 14.5	18.7 / 15.0	27.3 / 24.5
AV-LLM (13B)	48.2 / 45.2	55.4 / 52.6	70.1 / 67.6	29.6 / 26.1	18.0 / 14.5	24.4 / 20.0	29.4 / 27.0
OneLLM (7B)	49.9 / 47.6	52.3 / 49.8	71.6 / 68.1	29.7 / 26.3	20.9 / 17.0	24.5 / 18.0	28.8 / 26.5
AVicuna (7B)	51.6 / 49.6	56.2 / 53.1	71.2 / 67.9	29.6 / 26.6	19.7 / 16.5	28.4 / 23.0	29.6 / 27.0
CREMA (4B)	56.8 / 52.6	62.3 / 58.6	73.8 / 68.4	31.6 / 27.0	23.8 / 19.0	29.0 / 26.0	31.5 / 28.5
VideoLLaMA2 (7B)	-	-	70.4 / 68.3	29.7 / 26.8	25.7 / 22.0	27.5 / 23.0	28.4 / 25.5
AnyGPT (7B)	53.7 / 50.7	59.2 / 56.9	72.5 / 68.1	28.8 / 26.2	25.7 / 22.5	24.0 / 19.0	28.9 / 25.5
NExT-GPT (7B)	53.5 / 50.9	58.4 / 56.3	68.7 / 67.9	28.0 / 26.4	23.8 / 22.0	19.5 / 16.0	32.3 / 28.0
Unified-IO-2 L (6.8B)	58.3 / 55.1	60.0 / 57.9	73.8 / 70.1	31.8 / 27.2	25.6 / 21.5	26.5 / 22.0	29.3 / 27.5
Unified-IO-2 XL	61.3 / 57.2	59.7 / 58.6	73.7 / 71.8	30.0 / 28.5	24.7 / 22.5	29.0 / 26.0	29.6 / 27.0
Bay-CAT (7B)	55.6 / 53.8	58.3 / 56.5	71.9 / 69.5	31.9 / 28.2	24.4 / 20.5	22.0 / 18.0	29.8 / 27.5
Video-SALMONN (7B)	56.8 / 54.9	58.7 / 57.2	71.1 / 70.2	29.8 / 27.5	24.7 / 22.0	21.0 / 17.0	27.5 / 26.5
VITA (7B)	59.0 / 58.6	61.2 / 60.1	73.8 / 72.9	30.1 / 29.2	26.7 / 25.5	44.0 / 41.0	29.2 / 27.5
Open-Source Models with AURELIA							
PandaGPT (13B)	41.9$^{+24.33\%}$	32.7$^{+25.28\%}$	72.9$^{+12.67\%}$	28.6$^{+18.67\%}$	25.0$^{+100\%}$	25.0$^{+19.04\%}$	31.0$^{+14.81\%}$
Macaw-LLM (7B)	41.6$^{+30.81\%}$	38.1$^{+11.07\%}$	73.5$^{+11.53\%}$	29.3$^{+20.57\%}$	25.5$^{+82.14\%}$	24.0$^{+71.42\%}$	28.5$^{+42.5\%}$
VideoLLaMA (7B)	45.8$^{+25.13\%}$	41.5$^{+13.07\%}$	74.2$^{+12.08\%}$	29.6$^{+14.72\%}$	28.5$^{+72.72\%}$	28.0$^{+75.0\%}$	29.0$^{+26.08\%}$
ImageBind-LLM	49.7$^{+13.21\%}$	44.2$^{+12.75\%}$	72.8$^{+8.81\%}$	30.1$^{+18.50\%}$	28.0$^{+100\%}$	23.0$^{+53.33\%}$	31.0$^{+37.77\%}$
X-InstructBLIP (13B)	52.3$^{+17.52\%}$	46.9$^{+16.95\%}$	72.6$^{+9.83\%}$	29.8$^{+15.05\%}$	29.0$^{+100\%}$	27.0$^{+80.0\%}$	30.0$^{+22.45\%}$
AV-LLM (13B)	52.7$^{+16.59\%}$	57.9$^{+10.07\%}$	73.4$^{+8.57\%}$	31.1$^{+19.15\%}$	28.5$^{+83.87\%}$	29.0$^{+45.0\%}$	34.0$^{+25.92\%}$
OneLLM (7B)	54.1$^{+13.65\%}$	55.3$^{+11.04\%}$	73.9$^{+8.51\%}$	30.7$^{+16.73\%}$	29.0$^{+70.58\%}$	29.0$^{+61.11\%}$	33.5$^{+26.41\%}$
AVicuna (7B)	55.3$^{+11.49\%}$	57.8$^{+8.85\%}$	73.1$^{+7.65\%}$	30.4$^{+14.28\%}$	29.5$^{+79.09\%}$	34.0$^{+47.80\%}$	34.5$^{+27.78\%}$
CREMA (4B)	59.8$^{+13.68\%}$	67.2$^{+14.67\%}$	74.2$^{+8.47\%}$	31.9$^{+18.14\%}$	32.5$^{+71.05\%}$	40.0$^{+53.84\%}$	34.0$^{+19.29\%}$
VideoLLaMA2 (7B)	-	-	74.7$^{+9.37\%}$	31.6$^{+17.91\%}$	38.0$^{+72.72\%}$	35.0$^{+40.0\%}$	34.5$^{+35.29\%}$
AnyGPT (7B)	56.2$^{+10.84\%}$	62.5$^{+9.84\%}$	73.3$^{+7.63\%}$	31.4$^{+19.84\%}$	35.5$^{+57.77\%}$	33.0$^{+73.68\%}$	33.0$^{+29.41\%}$
NExT-GPT (7B)	57.8$^{+13.55\%}$	60.8$^{+7.99\%}$	73.5$^{+8.25\%}$	31.8$^{+20.45\%}$	36.0$^{+63.63\%}$	32.0$^{+100\%}$	33.5$^{+19.64\%}$
Unified-IO-2 L (6.8B)	61.9$^{+12.34\%}$	62.0$^{+7.08\%}$	74.6$^{+6.41\%}$	32.4$^{+19.11\%}$	36.5$^{+69.76\%}$	35.0$^{+59.09\%}$	33.5$^{+21.81\%}$
Unified-IO-2 XL (6.8B)	62.3$^{+8.91\%}$	62.8$^{+7.16\%}$	75.6$^{+5.29\%}$	33.6$^{+17.89\%}$	38.5$^{+71.11\%}$	40.0$^{+53.84\%}$	34.0$^{+25.92\%}$
Bay-CAT (7B)	58.5$^{+8.73\%}$	61.1$^{+8.14\%}$	75.0$^{+7.91\%}$	32.7$^{+15.95\%}$	34.0$^{+65.85\%}$	35.0$^{+94.40\%}$	32.5$^{+18.18\%}$
Video-SALMONN (7B)	59.8$^{+8.92\%}$	61.7$^{+7.86\%}$	75.2$^{+7.12\%}$	32.5$^{+18.18\%}$	37.5$^{+70.45\%}$	32.0$^{+88.23\%}$	33.0$^{+24.52\%}$
VITA (7B)	62.6$^{+6.82\%}$	66.5$^{+10.64\%}$	78.8$^{+8.09\%}$	33.8$^{+15.75\%}$	39.0$^{+52.94\%}$	50.0$^{+21.95\%}$	35.0$^{+27.27\%}$

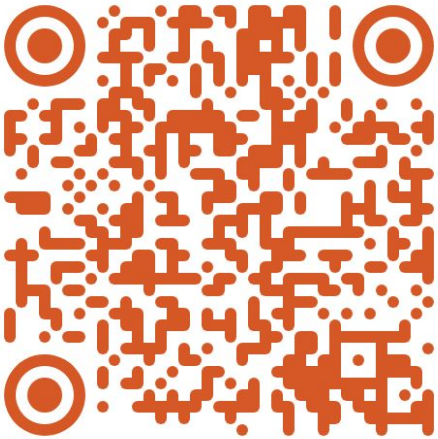
Experimental Results: Qualitative

- Compared to vanilla zero-shot inference, AURELIA augments the target model with reasoning capabilities, leading to the improved answers.



ICCV OCT 19-23, 2025 HONOLULU HAWAII

Thank You!



Project Page